COMPARISON BETWEEN FREQUENTIST AND BAYESIAN IMPLEMENTATION OF MIXED LINEAR MODEL FOR ANALYSIS OF MICROARRAY DATA

by

XIAOTING QIN

(Under the Direction of Romdhane Rekaya & Gauri Datta)

ABSTRACT

The objective of this study was to evaluate the performances of a mixed linear model under a frequentist and a Bayesian implementation for analysis of microarray data. A simulation was conducted following the structure of an existing Affymetrix chip data. PROC MIXED of SAS was used for the frequentist implementation. T-test, p-values, and the estimated difference between the two treatment levels were used to detect differentially expressed genes, as well as false positive and false negative cases. In the Bayesian implementation, the probabilities of a gene being in each of five pre-defined significance level classes were used for performances testing. The results indicate that both methods performed exceptionally well in identifying highly differentially expressed genes with a success rate of 0.96 and 0.98, respectively. However, the Bayesian approach was superior in clustering the most important genes. Both procedures performed similarity in detecting false positive and negative cases.

INDEX WORDS: Gene expression, Mixed linear model, Bayesian analysis,
Affymetrix, Simulation

COMPARISON BETWEEN FREQUENTIST AND BAYESIAN IMPLEMENTATION OF MIXED LINEAR MODEL FOR ANALYSIS OF MICROARRAY DATA

by

XIAOTING QIN

B.S., Beijing Agricultural University, P. R. of China, 1989

M.S., Beijing Agricultural University, P. R. of China, 1992

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2004

COMPARISON BETWEEN FREQUENTIST AND BAYESIAN IMPLEMENTATION OF MIXED LINEAR MODEL FOR ANALYSIS OF MICROARRAY DATA

by

QIN XIAOTING

Major Professor:	Romdhane Rekaya Gauri Datta
Committee:	Jaxk Reeves Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2004

AKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr. Romdhane Rekaya for his friendship, guidance and assistance. It has been a pleasure and honor to work with him for the duration of my studies. I would also like to thank Dr. Gauri Datta for his assistance, and for always having an interesting observation that kept things in perspective. Many thanks to Dr. Jaxk Reeves and Dr. Paul Schliekelman for their friendship, assistance and contribution to this work.

Special thanks must go to Dr. Steve Stice and Dr. Scott Martin, without their support, this thesis could not be finished.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In the last decade, scientists have marked a significant development in the study of biology. The completion of a "working draft" of the human genome and of some domestic animals, bacteria and viruses signals the beginning of a new stage of the modern biology in which most of the biological and biomedical research will be based on the sequences of the genome. This new approach promises not only a fast advance in the understanding of the basic biological processes but also in the prevention, diagnostic and cure of diseases of genetic origin. Although the optimal use of the genome information will take a long period of time (tens or perhaps hundreds of years), the implications of this new achievement for the basic forms of biological research, such as those in medicine or life sciences are very promising (Brazma, 2001).

Traditionally, medical and agricultural scientists have concentrated, for example, the study of the relationship between a single gene (or a very small assembly of genes) and animal or plant diseases. Although much progress has been observed, the understanding of the underlying mechanisms at the genetic level has been often incomplete. This can be due to not paying accounting for potential interactions between genes or the genes under consideration and the rest of the genome.

The advances at the genomic level have produced a massive amount of information and have created the need to develop quantitative methods directed toward the optimal use of this information, with the objective being to obtain a better understanding of the biological processes that take place. The knowledge of the coding sequences of virtually every gene in the genome invites the development of methodology which allows the identification of the function of these genes and their potential interactions. A possibility to exploring the function of an individual gene is the determination of its pattern of expression. At the moment, several techniques are available to quantify this level of expression. Northern blots, differential display, representational analysis and serial analysis of gene expression are just a few of these. cDNA microarrays are distinguished from the other methods by their potential to measure the level of expression on hundreds or even thousands of genes in a single experiment. This capacity allows measurement the expression of the complete genome during different stages of development, in different tissues or organisms, or as a the response to a specific drug. Thus, microarray technology has raised much interest both in the academic and commercial sector, especially in the field of human medicine, for obvious reasons.

At the moment, a detailed examination of the multidimensional genetic system is possible thanks to these new hybridization techniques. A single hybridization experiment using a DNA chip allows simultaneously the examination of profiles of expression for thousands of genes. This can change dramatically the strategies to study relationships between genes and agriculturally important traits, or diseases. Nevertheless, a new quantitative genomic framework must emerge for a complete and optimal use of the available information (Waston et al., 1998).

A typical microarray research project is a multi-step process (Schena, 1999). It starts with experimental design and array fabrication, and proceeds with array reading (scanning) and image processing. Subsequently, the information contained in the images must be reduced somehow; this is known as gene expression statistical analysis. Typically, a massive amount of data is generated, containing a variety of information that ranges from molecular sequences for genes or clones, to expression values (quantitative) for each gene under different experimental conditions

(Zhu et al., 2000). As mentioned by Douglas et al. (2001), a challenge posed by microarray technology resides in how to deal with the massive amounts of data, such that it can be explored and interpreted in the context of available biological knowledge. Arguably, the low number of primary research papers in the microarray field, relative to the number of review papers on gene expression, may not be due to a limited amount of primary data, but to difficult in sensibly analyzing the data which has been colleted.(Douglas et al., 2001).

In the early years of microarray technology, most statistical research done with gene expression data focused on the development of visualization tools, and standard statistical methods such as cluster analysis and principal components were applied (Carr et al., 2003). These techniques have been useful to summarize information, to identify clusters or groups of genes based on similarity or dissimilarity, and to predict biochemical and physiological pathways for some uncharacterized genes. Recently, more sophisticated statistical tools are being used to analyze expression data. Parametric and non-parametric methods have being developed to overcome the shortcomings of earlier procedures. The Mixed linear model (Wolfinger et al., 2002) is becoming a standard tool for several research groups in the analysis of microarray data; because of its flexibility in accommodating different experimental designs and its clear statistical proprieties.

In this study, performances of mixed linear model under frequentist and Bayesian implementations are evaluated based on simulated data. Quantities such as the correct identification of differentially expressed genes, false positive and false negative rates and computational cost will be used in performance assessment.

CHAPTER 2

REVIEW OF LITERATURE

1. DNA Microarrays

DNA microarrays constitute an extension of the hybridization methods that have been used for more then 20 years for the identification and quantification of nucleic acids in a sample. They consist of a great number of DNA molecules spotted in a systematic way on a solid substrate which can be crystal slides or a nylon membrane. Depending on the size of wells on the array where the DNA is spotted, this can be classified as microarray (the diameter of each well is inferior to 250 microns) or macroarray (if the diameter is superior to 300 microns). Two main methods exist to make DNA microarrays or DNA Chips. The first method consists of oligo nucleotides sequences having a size between 20-30 base pairs synthesized directly on a solid surface using the combination of photolithography technique and light directed chemical synthesis. This method allows the making of DNA microarrays with very high density (about 250000 oligo spots per cm^2). Although this approach has numerous advantages, its major disadvantage is the high cost of the necessary equipment for its manufacture and reading, in addition to the lack of design flexibility (Dieckgraefe et al., 2000). The second method, referred to as cDNA microarray, involves the analysis of a small amount of DNA which is spotted on a solid surface. The cDNA microarrays consist of long DNA sequences (500 -2000 base pairs) deposited with a high-speed robot. The density of the chip depends on the capacity of the dispensing device. Its greater advantage with respect to the method is the possibility of its routine manufacture in a regular laboratory and the design flexibility leading to an important cost reduction (Schena, 1999).

The underlying principle of DNA microarray technology is the spontaneous biochemical base-pairing process of complementary base pairs, called hybridization, which provides high sensitivity and specificity of detection as a consequence of exquisite, mutual selectively between complementary strands of nucleic acids (Southern et al., 1999). An array is an orderly arrangement of known cDNA sequences or oligonucleotides. It provides a medium for matching known DNA (probe sequences) and unknown, fluorescently labeled DNA or RNA samples (target sequences). The labeled target sequences allow a quantitative measurement of their abundance in a sample, i.e. tissue, cells, organ etc., being investigated. On the chip, target sequences are organized in so called spots. The sample spots size in microarrays are typically less than 200 microns in diameter and these arrays usually contain millions of spots (Duggan et al., 1999). Microarrays require specialized high-speed robotics for manufacturing and imaging equipment (scanner) for measuring the raw intensity data for each spot. Either fluorescence intensity, or extent of radio labeling at each spot, is proportional to the amount of target hybridized to each probe. Since the concentration of the probe is large relative to that of the target, hybridization occurs at a rate, which is proportional to the concentration of the target and to the incubation time (Duggan et al, 1999). Specific digital image processing procedures take advantage of the highly regular arrangement of the gene spots on the array to extract the intensity value of each spot (Cheng et al., 1999; Lipshutz et al., 1999).

1.1 Spotted cDNA Microarrays

The process of producing and using spotted cDNA microarray in comparative experiments is summarized in Figure 2.1. It consists of two major components: a) array production and b) the different stages of a comparative study**.**

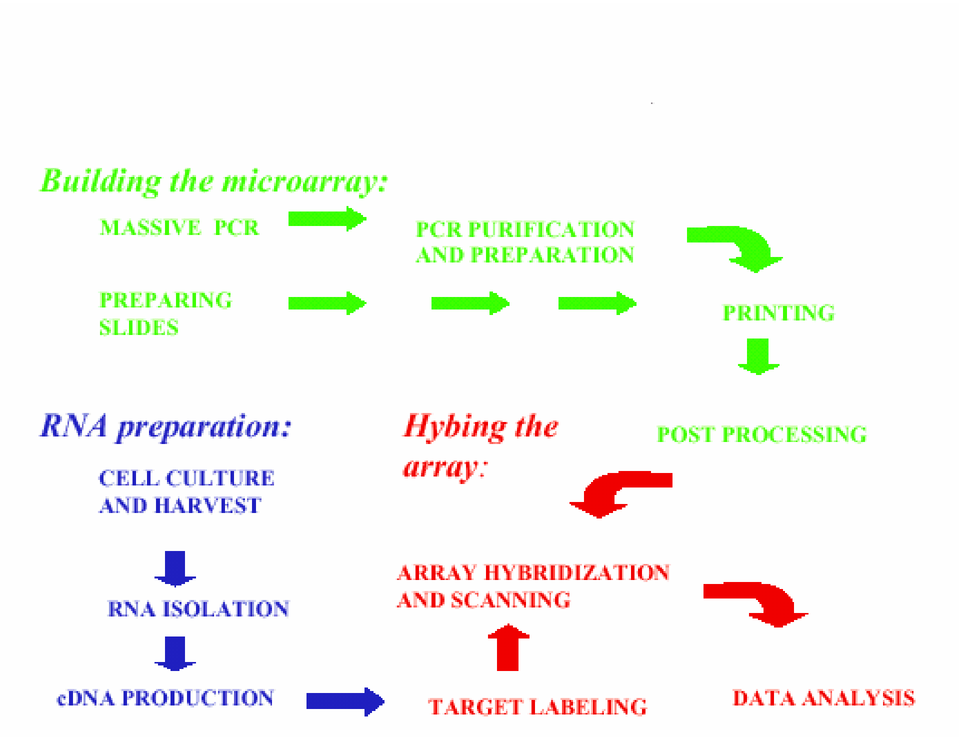


Figure 2.1: Different steps of a comparative microarray experiment (Dudoit et al., 2002)

1.1.1 Production of the array

The process of a microarray experiment begins with the hypotheses of the biologist and the selection of genes (EST) of interest that will be printed to the Array. Selected genes are amplified using a Polymerase Chain reaction (PCR). After purification, the products of PCR will be printed into the pretreated microscope slides using a robotic arrayer. Two methods of mechanical microspotting and ink jetting are used (Cheng et al., 1999).

Once the DNA Array has been made, it can be used in one of the two possible applications: genomic studies or gene expression studies. The genomic studies consist basically of the identification and genotyping of mutations and polymorphisms. Oligonucleotides microarrays have been used in the identification of single-nucleotide polymorphisms (SNPs), variations that happen frequently in the genome (each 100-300 bases). However, the majority of DNA microarray applications have focused on the study of changes in gene expressions (Bolstad et al., 2003)

1.1.2 Stages of a comparative study using cDNA microarray

The design of a comparative gene expression experiment is a multi-stage task (Figure 2.1) which needs the intervention of many specialists as well as the use of sophisticated and expensive equipment.

a) Cell lines or populations: This stage depends only on the hypotheses and objectives of the biologist or geneticist. It consists of choosing the cellular lines of which the genetic material will be used in the comparative study. As of now, many interesting applications have been carried out using the genes of the following cellular populations (Waston et al., 1998):

- *genes of specific tissues*: cells from different tissues on the same organism (i.e. nervous system and heart muscle) were used. This type of comparative study allows the identification of genes that are preferentially expressed in a specific tissue.
- *genes of the same tissue under different environmental conditions*: the objective of this type of experiments is to understand the adaptation of a cell line to environmental changes such as temperature, PH, radiation, pesticide or the response to a drug. This type of application is very frequent in comparative studies, especially in the search of treatments for many diseases. In its simplest form, one could study the alteration of

the expression of genes after subjecting them to an excitation or treatment and comparing them to a control population.

- *genes of the same tissue at different stages of development*: this type of experiment allows the study of genetic changes at the cellular level in the same tissue. Different stages of breast cancer for example, suggest the intervention of different genetic mechanism, and consequently a change in the gene expression patterns. In dairy cattle, one can investigate changes of expression in the mammary gland at different stages of lactation.
- *genes on the same tissue for diseases with genetic heterogeneity*: These types of comparative experiments are frequent in the studies of diseases such as cancer. A group of individuals with cancer in the same tissue can have different groups of missing or damaged genes. In this case, comparative methods can be considered as techniques for diagnosis but also for classification, and obviously they play a very important role in the design of the appropriate and effective treatment for each patient given his/her individual pattern of gene expression.

b) Extraction of mRNA and reverse transcription: Once the cellular lines for the comparative study have been chosen, the second stage consists on the extraction of the messenger RNA (mRNA) and the reverse transcription. Before that, it seems of interest to remind the reader of the main mechanisms that regulate the relationship between DNA and proteins as shown in Figure 2.2.

The transcription mechanism begins with the recognition of promoter; small DNA sequence that indicates the beginning of a gene. In Eukaryotes, the coding regions called exons are separated by non-coding zones called introns. Coding regions is DNA sequence which transcript as function protein. The final product of transcription is the

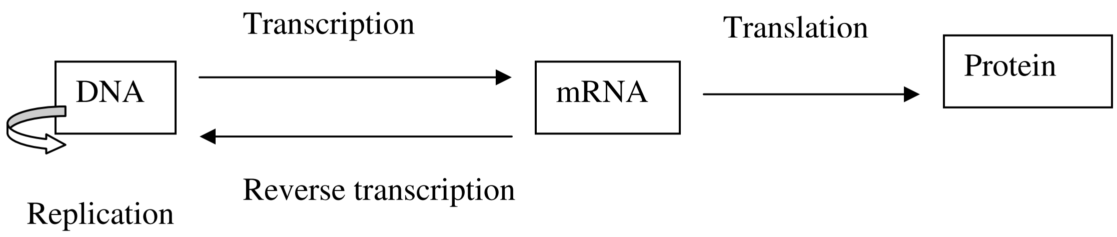


Figure 2.2: Mechanisms regulating the relation between DNA and protein (Lockhart et al., 1996)

messenger RNA (mRNA) obtained by copying the coding regions of the gene and eliminating the introns. Once obtained, mRNA must be purified from other cellular contents. Producing sufficient quantity of mRNA for a microarray experiment (1-2 micrograms) is not an easy task since the latter represent only 3% of the total RNA in a cell. Further, it is difficult to work with mRNA since it is not stable and it is susceptible to detection by environmental conditions. In order to avoid these problems, mRNA obtained from the two cell lines will be transformed to a more stable DNA form by means of reverse transcription. The products of this transformation are the complementary DNA or cDNA whose sequences are complementary to the original sequences of mRNA (Schena, 1999).

The major problem associated with the production of the cDNA comes from the fact that the reverse transcription is not equally efficient for all mRNA, leading to a well know phenomena called "reverse transcription bias". Although this does not prevent the comparison of the same mRNA between two cell lines, it prohibits the quantitative comparison between different mRNA in the same array (Yang et al., 2002; Schena et al., 1999).

c) cDNA labeling: In order to measure the relative abundance of the DNA sequences spotted in a specific position of the array, the two cDNA samples or targets are labeled with reporter molecules able to identify their presence. Quite often, two different fluorescent dyes (red-fluorescent: Cy5, green-fluorescent: Cy3) are used for labeling due to the clear difference in their wavelength of excitation.

d) Hybridization: Both labeled cDNA targets are mixed and then hybridized to the DNA sequences immobilized to the surface of the array. If a target contains cDNA whose sequence is complementary to the one immobilized in a specific well of the array, it will hybridize to it. The relative abundance of specific sequences of DNA in the two target samples will be reflected by the ratio of fluorescence intensities at each point of the array. Usually, there will be sufficient DNA in each well so that both targets can hybridize to it without interferences (Schena et al., 1999).

e) Reading of the array: After hybridization, the array is scanned to determine the magnitude of hybridization of each target with each well in the array. Since targets are labeled with reporter molecules that emit light after being stimulated with a laser, a detector able to measure its intensity will capture this light. As a result of the difference in their excitation waves, the light emitted by the fluorescents dyes can be filtered therefore allowing the separation of both intensities. In the opposite case, the intensities will be contaminated as a result of "cross talk" between the channels of the two fluorescent dyes.

In spite of precautions, the measured intensities during the scanning of the array are not clean of noise. Such noise typically originates from light emitted from the hybridization of some molecules to an incorrect well or the crystal of the array. This additional light represents the background of the image of the scanned array. Nevertheless, this background is relatively small using microarray technology compared with other hybridization techniques (Yang et al., 2001).

f) Image analysis: The final product of a comparative hybridization experiment is a colored image (two dimensional array), where the color at each point is a combination of the intensities of red and green fluorescent dyes (Figure 2.3). Spots in the array having DNA present in a high level in the red (green) labeled target are predominant red (green). A yellow spot indicates an equal amount of DNA bounded to each one of the two cell lines, since the yellow color is a mixture of an equal amount of red and green colors (Waston, 1998).

The next step in a comparative gene expression experiment is the extraction of the information in the scanned image. This is an image analysis task aiming to measure intensities at different points of the array, where such intensities reflect the level of expression. Under ideal conditions (all the wells are of the same size, constant distance between wells, all the wells are circular and of the same diameter, etc.), this step is reasonably simple. However, in true applications the idealized conditions are violated to diverse degrees, as a result of irregularities of dimension, size and well position, in addition to contamination. These and other factors have motivated scientists to develop software and algorithms for the process and detection of signal that are specific to microarray technology (Zhu et al., 2000). There is an extensive literature on techniques used in the gene expression field to extract the needed information from the scanned images. Several commercial and public software (i.e. BioDiscovery, ScanAlyse) are available. However, there is no universal solution to the problem and human intervention is still necessary in many cases.

When signal and background pixels in each well have been identified and their respective intensities have been measured during the image-processing step, some quantitative methods are need to allowance to deduce the level of expression of each gene in the array and in both channels. In most applications, the ratios of total or average intensities within a well have been used to evaluate the relative level of the expression in a comparative experiment. There is a certain tendency for the use of the ratio of averages of intensities between both channels is the best method of the measurement, although other ratios based on the midpoint, or the volume of values of the intensities of pixels have been used in some cases (Cheng et al., 1999).

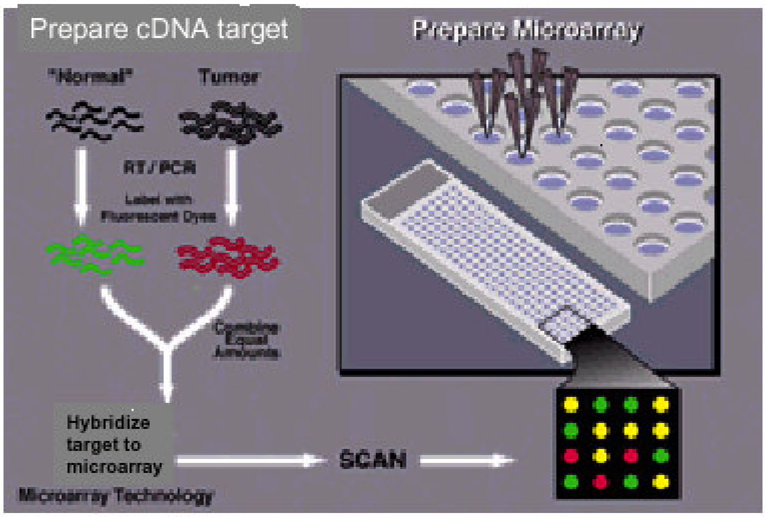


Figure 2.3: A typical image obtained in an comparative gene expression experiment (Dudoit et al., 2002).

In order to calculate the ratio of intensities between both channels, one must first needed to determine the background and signal intensities (Yang, et al., 2002; Brazma, 2001).

- Background intensity: in general, background intensity in a microarray image is not uniform, such that a local background adjustment is needed. There are many methods to determine the background intensity. The simple and most used method consists of estimating the background as the average of pixel intensities in the zone near the box for each well.

- Signal detection: one of the most complicated stages of the image processing is the determination of the region of interest (or signal) for each gene spotted in the array. The method of fixed threshold is widely used and consists of classifying a pixel as signal if its intensity is greater then a fixed quantity T (threshold).

$$\underline{T= m+c*s}$$

Where *m* is the average of the background intensity, *s* is the standard deviation and *c* is a constant subjectively determined (e.g., *c*=3). Other more sophisticated methods using variable thresholds or predicting the shape of the signal area were used in several applications.

Once the average background and signal intensities for every gene in both channels (red and green) have been determined, a corrected average is computed by subtracting the background from the signal intensity. The relative expression for each gene will be calculated as the ratio between the corrected averages in both channels. These two corrected signals will be symbolized in the remaining of this chapter as **R** and **G** for the red and green channels, respectively. At the logarithmic scale, the relative expression is given by (Cheng et al., 1999):

$$\mathrm{M}=\log_2(\mathbf{R}/\mathbf{G})$$

g) Normalization: Before using the ratio of gene expression in a statistical analysis, this ratio has to be calibrated (or normalized). There are many sources of systematic variation in microarray experiments that affect the measured levels of expression. Normalization is the term used to describe the process of removing such variation. For example, consider differences in labeling effectiveness between fluorescent dyes. In such cases, a constant adjustment is used to force the distribution of the ratios of expression to have an average (or medium) equal to zero. In statistical terms, this process consists of satisfying the null hypothesis in which signals in both channels (red and green) are assumed to be probabilistically the same.

In its simple form, normalization can be carried out by subtracting a constant from the expression ratio. Generally, this constant is calculated as the average or median of the intensities ratio of a particular assembly of genes frequently called "housekeeping genes". These genes are chosen based on biological reasons and their experimental behavior (ratio of expression near 1). After normalization, the ratio of expression will be equal to:

$$M=\log_2(\mathbf{R}/\mathbf{G})-k$$

where ‘k’ is the mean or median of the log ratio of expression level of the “housekeeping genes”.

Recently, more sophisticated statistical methods are being used to analyze microarray data. Background adjusted data is being used rather than the normalized log ratio. By doing so, the “heuristic” normalization step is replaced by more rigorous statistical modeling technique that accounts in a systematic way for all sources of variation in the data. In section 2, a detailed discussion of statistical methods used for microarray data analysis is presented.

1.2. Oligonucleotide Microarrays

Oligonucleotide microarrays or GeneChips have fundamental differences compared to the spotted cDNA microarrays both in their fabrication and the process of their use. In GeneChips, each gene is represented by 12 to 20 pairs of 25 base length oligonucleotide probes. One component of each pair is referred to as a perfect match (PM) probe and it is designed to be specific to the transcript from the intended gene. The second component of the pair is called mismatch (MM) probe, and it is designed the same way as the first component, except that middle base (base in the 13th position of 25) has been changed (Figure 2.4). The mismatch probe is used to account for the optical background and non-specific hybridization noises. Therefore, the observed intensities must be adjusted to yield accurate measurements of specific hybridization. The default adjustment approach, provided as part of the Affymetrix system, is based on the difference between perfect match and mismatch probe intensities (PM-MM).

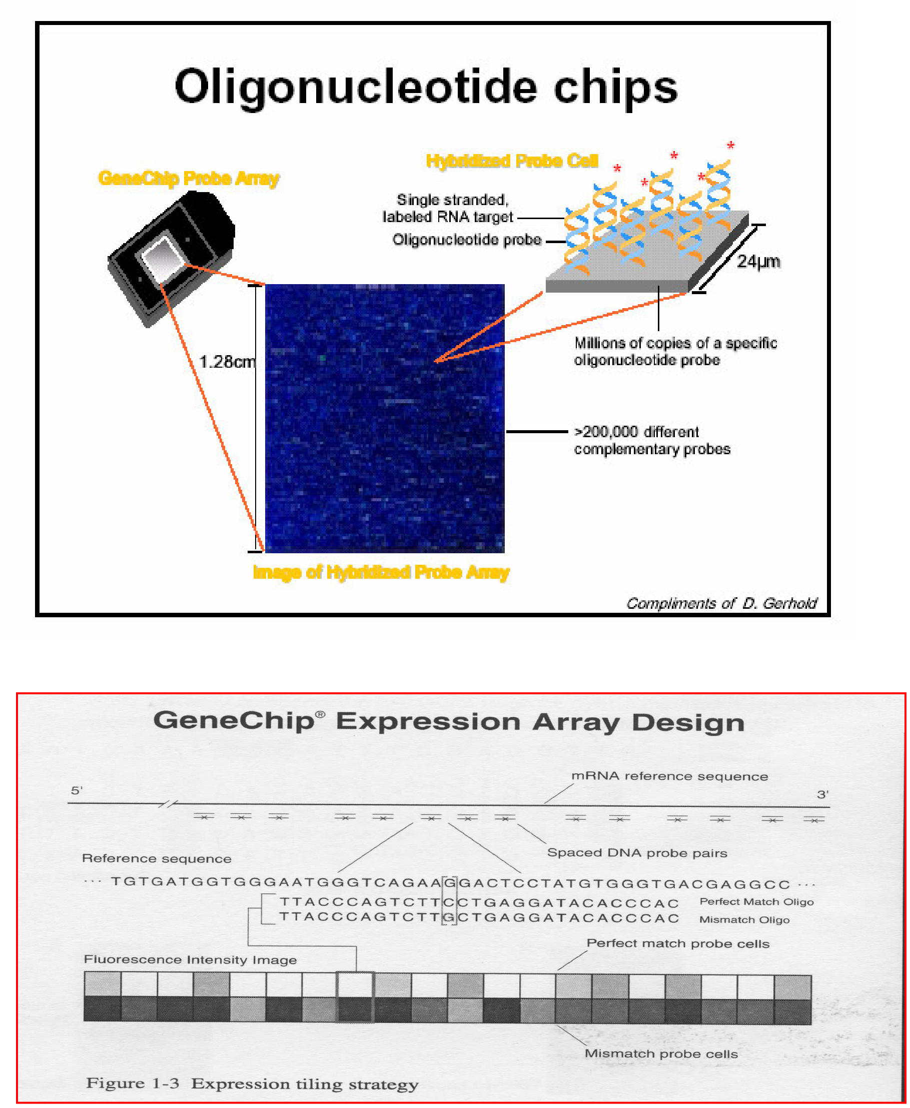

Oligonucleotide chips
GeneChip Probe Array
Hybridized Probe Cell
Single stranded, labeled RNA target
Oligonucleotide probe
24µm
1.28cm
Millions of copies of a specific oligonucleotide probe
>200,000 different complementary probes
Image of Hybridized Probe Array
Compliments of D. Gerhold
GeneChip® Expression Array Design
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
TTACCCAGTCTTCCTGAGGATACACCCAC
Perfect Match Oligo
TTACCCAGTCTTGCTGAGGATACACCCAC
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells
Figure 1-3 Expression tiling strategy

Figure 2.4: The process of oligonucleotide Chip design, fabrication and use (Dudoit, 2002)

After hybridization and scanning, the quantitative fluorescence image, along with the known identity of the probes, is used to assess the ’presence’ or ’absence’ of a particular molecule (such as a transcript), and its relative abundance in one or more samples. Because the oligonucleotides at each physical location (or address) is well described, and the recognition rules that govern hybridization are well characterized, the signal intensity at each position of the microarray gives a quantitative measurement of one single target sequence with known identity (http://www.affymetrix.com/index.affx).

Contrary to cDNA arrays, the GeneChip are not hybridized to samples (targets) from two populations (cell lines, treatments, etc.) to be compared at the same time. Instead, the hybridization is performed for each population separately. For other tasks, such as scanning and image analysis, both methods share many commonalities.

2. Statistical Analysis of Gene Expression Data

Once the image analysis process is completed, the results of the hybridization can be represented in a matrix form. Each element of the matrix represents either the level of expression or the ratio of expression of a given gene. In the majority of statistical analyses, the goal is to extract information about the underlying biological possess (Brazma et al., 2001).

Depending on the objective of the experiment and the availability of external information, the study of the expression information can be carried out in a supervised or unsupervised manner. Clustering methods are an example of unsupervised techniques used to cluster genes and/or samples based on some measure of similarity such as the distance. Several methods have been developed to implement cluster analysis (k-mean clustering, hierarchical clustering, self organizing maps). Classification techniques such as discriminant analysis, principal components or support vector machines are supervised methods and require additional information such as the phenotype (sick/healthy) or the functional class of the gene. For these methods, there is an extensive literature (Getz et al., 2000).

Although there are several differences in the experimental design and the quantitative output between the two platforms actually used for the study of gene expression profiling, they share a lot of similarities in their statistical analysis. Thereafter, we will focus more on GeneChips data analysis given that the simulation study carried out in the following chapter of this thesis is based on the Affymetrix chip platform.

A typical application of GeneChip technology is finding genes that are differentially expressed in different tissues or under different environmental conditions. Due to the large amount of information and the intrinsic variation in the data obtained in a microarray experiment, statistical methods have been applied to systematically extract biological information and to assess the in associated uncertainty. Successful analysis will detect all and only genes that are differently expressed due to biological variations. Here we review some widely used methods for detecting differentially expressed genes.

2.1 Fold Change Method

Fold change is the simplest method for identifying differentially expressed genes (Cui and Churchill, 2003). It is based on the observed ratio (or ratio of averages) between two treatment levels. An arbitrary cut-off value (for example, 2 folds) is often used to identify differentially expressed genes. This is not a statistical test and there is no associated level of confidence. The fold change method is subject to bias if the data are not properly normalized and may also be sensitive to variance heterogeneity across genes. For example, an excess of low intensity genes may be mis-identified as being differentially expressed due to an excess of variation relative to high intensity genes (Rocke et al., 2001; Cui and Churchill, 2003).

The most commonly used fold change estimate is AvDiff, the Affymetrix default. For each probe set on each array *i*, AvDiff is defined as:

$$AvDiff = \frac{\sum_{j=1}^{A}(PM_j - MM_j)}{\# A}$$

where *A* is the subset of probes for which $d_j = PM_j - MM_j$ is within ± 3 standard deviations (SD) away from the average, #*A* represents the cardinality of *A*. Many of the other expression measures commonly used are modifications of AvDiff which result from accounting somehow for outliers or for dealing with low expression values.

Realizing the inadequacy of the linear method, Affymetrix has presented a new algorithm MAS 5.0, where the log of the difference between perfect match and mismatch is being used. Specifically, the MAS 5.0 signal (measure) is defined as:

Signal = Tukey's Biweight {log(*PMj* - *CTj*)}

with *CTj* a quantity derived from the *MM*s that is never larger than its *PM* pair (Hubbell, 2001).

2.2 Robust Multi-array Analysis (RMA)

Given the exaggerated variance of the gene expression estimates using the log transformation of the difference between the perfect match (PM) and mismatch (MM) probe intensities, Irizarry et al. (2003) proposed the robust multi-array analysis (RMA) method. It consists of a global background adjustment step that ignores the MM intensities followed by quantile normalization. This intended to make the distribution of probe intensities similar across all arrays involved in the experiment. Consequently, the I- dimensional quantile –quantile plot of the normalization probe level data from all arrays ($i = 1,2,...,I$) will follow an I–dimensional identity line.

Finally, for each probe set of size *n*, the background adjusted, normalized and log transformed PM intensities, denoted by $\mathbf{y}$, will be modeled following an additive linear model.

$$y_{ijm} = \mu_i + \alpha_j + e_{ijm}; \qquad i = 1,2,...,I;\ j = 1,2,...,J \text{ and } m=1,2,...,M$$

where μ_i represents the log scale expression level for array *i*, α_j is a probe specific affinity effect, and e_{ijm} is an independent identically distributed error term with mean 0. To make all parameter identifiable, it is assumed that $\sum_j \alpha_j = 0$ for all probe sets. The estimate of μ_i represents the expression measures for probe set *n* in array *i*, which was referred to by Irizarry et al. (2003) as robust multi-array average (RMA). RMA has been implemented in the software developed in the Bioconductor project (http://www.bioconductor.org) and it has become a popular alternative to the default algorithm provided by Affymetrix.

2.3. Mixed Linear Model

Wolfinger et al. (2001) developed a statistically rigorous approach to analyze probe-level Affymetrix GeneChips data. It provides a general and powerful framework to fully utilize the available information in microarray experiments with multiple factors and/or a hierarchy of sources of variation. The method simultaneously considers the data across all chips in an experiment. It accommodates complex experiments involving many types of treatments and can test for their effects at the probe level in a systematic manner. Finally, this approach combines both the normalization and statistical testing steps.

Before data analysis, a log base 2 ($\log_2$) transformation for individual PM and MM is needed to improve the normality assumption and increase the fit of an additive model. Recently, several studies (Irizary et al., 2003) comparing different ways of using PM and MM intensities have concluded that it is better to exclude mismatch (MM) information from the analysis because of its exaggerated variance, and the reduced efficiency to translate the mathematical subtraction to a biological subtraction. To adjust for gross array-level effects, the global normalization centering the logged values so that they have zero mean is required.

In the mixed model setting, an important issue is to decide whether effects are ‘fixed’ or ‘random’. Fixed effects are those effects with a well-defined, finite number of levels and only those finite levels are of interest in the experiment. Random effects are considered to be drawn from an infinite population having some probability distribution, usually normal. For random effects, the linear mixed model estimates the parameters of this probability distribution (mean and variance components in the normal case). For Microarray data, typically cell line, treatment and probe effects are considered to be fixed, but effects impacting arrays may be considered random, reasoning that they are the accumulation of small experimental sources of noise. Putting all these together, the following linear mixed model serves as an initial template for data from a single gene:

$$y_{ijkl} = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + e_{ijkl}$$

where y_{ijkl} is the transformed and centered expression measurement of the i^{th} cell line applying the j^{th}treatment at k^{th}probe in the l^{th}replicate. The symbols L, T, LT, P, LP, TP and A in the formula represent cell line, treatment, cell-treatment interaction, probe, cell line-probe interaction, treatment-probe interaction, and array effects respectively. The $A_{l(ij)}$ s are assumed to be independent and identically distributed normal random variables with mean 0 and variance σ_a^2 and e_{ijkl}’s are assumed to be independent identically distributed normal random variables with mean 0 and variance σ^2, and independent of $A_{l(ij)}$’s. Both variance components are unknown and often maximum likelihood based methods were used for their estimation. PROC MIXED of SAS offers several options for estimating these parameters.

Further, the Bayesian approach can be used to estimate the unknown parameters of the mixed linear model. In fact, there is a huge literature on the Bayesian implementation of the mixed linear model (Lindley and Smith, 1972; Box and Tiao, 1972; Dempfle, 1977; Gianola and Fernado, 1992). Basically, the Bayesian formulation is based on two sources of information, one is provided by the collected data set and the other is the prior information or degree of belief that the researcher had about the parameters of the model before the data was collected. These two sources of information are combined to generate the joint distribution necessary for all Bayesian statistical inferences. If $\mathbf{y}$ is a sample of observed data and $\boldsymbol{\theta}$ is a vector of unknown parameters in the model, the joint density of $\boldsymbol{\theta}$ and $\mathbf{y}$ is given by:

$$f(\boldsymbol{\theta},\mathbf{y}) = f(\boldsymbol{\theta} \mid \mathbf{y}) f(\mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

where $f(\boldsymbol{\theta})$ and $f(\mathbf{y})$ are the marginal densities of $\boldsymbol{\theta}$ and $\mathbf{y}$, respectively and $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the conditional density of the data given the parameters of the assumed model, known as the likelihood function.

From the previous formula, it is easy to write that:

$$f(\boldsymbol{\theta} \mid \mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) / f(\mathbf{y})$$

Given that the marginal density $f(\mathbf{y})$ does not depend on the vector $\boldsymbol{\theta}$,

$$f(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

The latest formula is a representation of Bayes theorem and shows clearly that the posterior density of the parameters vector combines the data information, $f(\mathbf{y} \mid \boldsymbol{\theta})$ or likelihood, and the accumulated information about the parameters before the data was collected, known in the Bayesian formulation as the prior, $f(\boldsymbol{\theta})$.

In general, the Bayesian formulation does not require more principles than those previously exposed. The difficulty of the inference process depends on the complexity of the obtained joint distribution. Although theoretically simple it is based only on a series of integrations of the joint posterior distribution, the Bayesian inference is very complex to carry out analytically, except in a very special cases (few parameters, normality, …etc.). As a result, several approximations have been proposed over time, such as Gauss-Hermite quadratic rules, Laplacian approximation (Shun, 1995) and more recently the Markov Chain Monte Carlo techniques.

In the gene expression literature, the Bayesian approach was successfully used as an alternative to the fold change method (Baldi and Long, 2001; Newton et al., 2001). However, very few, if any, comparisons between the frequentist and Bayesian implementations of mixed linear model for analysis of microarray data have been conducted. Furthermore, it is widely recognized that the Bayesian and frequentist approaches yield similar results (at least point estimates) when the data is highly informative. However, such a condition is not satisfied in the majority of gene expression experiments, where less than a dozen arrays are involved.

REFERENCES

Affymetrix, Inc. Human Genom U133 Set product specification. http://www.affymetrix.com/products/arrays/specific/hgu133.affx.

Baldi, P., and A. D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics, 17: 509-519

Bioconductor: http://www.bioconductor.org

Bolstad, B.M., R.A. Irizarry, M. Astrand, and T.P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19:185-193.

Box, G. and C. Tiao.1973. Bayesian inference in statistical analysis. John Wiley and Sons. New York.

Brazma, A., M. Vingron. 2001 Minimum inoformation about a microarray experiment (MIAME) –toward standards for microarray data. Nature genetics, 29:365-371

Carr, K. M., J. Bittner, M.Trent. 2003. Gene-expression profiling in human cutaneous melanoma. Oncogene, 22:3076-3080.

Cheng, V. G., 1999. Making and reading microarrays. Nature Genetics Supplement. 21: 15-19

Cui, X., and G. A.Churchill. 2003 Statistical tests for differential expression in cDNA microarray Genome Biol., 4:21- 227

Dempfle L., 1977. Relatin entre BLUP (Best Linear Unbiased Prediction et estimateurs Bayesiens. Ann. Geneti. Sel. Anim., 1:7-32.

Dieckgraefe, B. K., W. F. Stenson, J. R. Korzenik, P. E. Swanson, and C. A. Harrington. 2000. Analysis of Mucosal Gene Expression in Inflammatory Bowel Disease by Parallel Oligonucleotide Arrays. Physiol. Genomics, 4:1-11.

Douglas, B., D. Kell, M. Robert, and J. Draper. 2001. Genomic Computing. Explanatory Analysis of Plant Expression Profiling Data Using Machine Learning. Plant Physiology, 126: 943-951.

Dror, R. O., J. G. Murnick, N. J. Rinaldi, V. D. Marinescu, and R. A. Young. 2003. Bayesian estimate of transcript levels using a general model of array measurement noise J. comp. Biol., 10: 433-452.

Dudoit, S., R. Gentleman, R. Irizarry and Y. H. Yang. 2002. Bioconductor short course Presentation

Duggan, D. J., 1999. Expression profiling using cDNA Microarray. Nature Genetics Supplement, 21:10-14

Getz, G., E. Levine, and E. Domany. 2000. Coupled Two-way Clustering Analysis of Gene Microarray Data. Proc. Natl. Acad. Sci., 97:12079-12084.

Gianola, D., J. L. Foulley, R. L. Fernardo, C. R. Henderson, and K. A. Weigel. 1992. Estimation of heterogeneous variances using empirical Bayes methods: Theoretical considerations. J. of Dairy Sci., 75:2805-2823.

Hubbell, E., 2001 Estimating signal with next generation Affymetrix software, In: Gene logic Workshop on low level analysis of Affymetrix GeneChip data. http:// www.stat.berkeley. edu/users/terry/zarray/Affy/GL_workshop/genelogic2001.html.

Irizary, R.A., C. B. Hobbs, Y. B. Barclay, K. Antonellis, U. Scherf, and T. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4:249-264.

Irizary, R A., B. M. Bolstad, F. Collin, L. M. Cope, C. B. Hobbs, and T. P. Speed. 2003. Summaries of affymetrix genechip probe level data. Nucleic Acids Research, 31: 15.

Lindley, D.V., and A. F. Smith. 1972. Bayes estimate for the linear model (with discussion). J. R. Statist. Soc., 34:1-44.

Lipshutz, R. J. et al., 1999. High density synthetic oligonucleotide arrays. Nature Genetics Supplement, 21:20-24

Newton, M. A., C. M. Kenziorski, C. S. Richmond, F. R. Blattner, and K.W. Tsui. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. J. Comp. Biol., 8:37-52.

Rocke, D. M., B. Durbin. 2001. A model for measurement error for gene expression arrays. J. Comp. Biol., 8:557-69.

Schena, M., 1999. DNA Microarrays: A Practical Approach, Oxford University Press

Shun, Z., and P. McCullaugh.1995. Laplace approximation of high-dimensional integrals. J. Roy. Stat. Soc., 57:749-760.

Southern, E., K. Mir, and S. Shchepinov. 1999. Molecular interactions on microarrays. Nature Genetics Supplement, 21:5-9.

Waston, A., A. Mazumder, M. Stewart, and S. Balausubramanian.1998. Technology for microarray analysis of gene expression. Current Opinion in Biotechnology, 9: 609-614.

Wolfinger, R.D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J Comp. Biol., 8:625-37.

Wu, Z., and R. A. Irizarry. 2004. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. Proceedings of RECOMB

Yang, Y. H., S. Dudoit, P. Lu, and T. P. Speed. 2002. Normalization for cDNA Microarry Data: a robust composite method addressing single multiple slide systematic variation. Nucleic Acids Research, 30: e15.

Zhu, G., P. Spellman, T. Volpe, P. Brown, D. Botstein, T. Davis, B. Futcher. 2000. Forkhead Genes Regulate the Cell Cycle and Pseudohyphal Growth. Nature, 406:90-94.

CHAPTER 3

COMPARISON BETWEEN FREQUENTIST AND BAYESIAN IMPLEMENTATION OF MIXED LINEAR MODEL FOR ANALYSIS OF MICROARRAY DATA [1]

[1] Qin, X., R. Rekaya and G. Datta. 2004. To be submitted to Genetic Selection Evolution

ABSTRACT

Microarray data obtained from Affymetrix chips are proving to be extremely useful for gene expression analysis. The Mixed linear model is becoming a widely accepted tool for analysis of microarray data. In this study, a simulation was carried out to compare the performances of a frequentist and Bayesian implementation of mixed linear model for analysis of expression data. Eight arrays, with 10,000 genes each, equally divided between two treatments levels were simulated following a pre-existing real data structure. Four simulation scenarios with varying variances ratio (ratio between array variance and residual variance) of 0.25, 0.50, 0.75 and 1.0 were implemented. The mixed linear model used in the simulation and analysis included treatment and probes as fixed effects and array and error term as random effects. In order to minimize the inherent Monte Carlo error, 5 replicates were carried out for each simulation scenario. The results indicate that both methods performed exceptionally well in identifying highly differentially expressed genes with a success rate of 0.96 and 0.98 for the frequentist and Bayesian approach, respectively. However, the Bayesian approach was far more superior in clustering the most important genes into their correct significance classes. In fact, 139 out 145 most important genes (98%) were correctly classified by the Bayesian approach versus 74 (51%) genes using the frequentist approach. With respect to the false positive and negative cases, both procedures performed similarity with a slight superiority for the Bayesian approach.

Keywords: Microarray, Mixed linear model, Bayesian analysis

INTRODUCTION

Research groups from diverse fields have become actively involved in designing and analyzing gene expression data from microarray experiments. Specifically, oligonucleotide technology as provided by the Affymetrix system is proving to be an extremely valuable tool for studying gene expression patterns. The methods developed for utilizing the GeneChips provide the potential for obtaining enormous amounts of data in a relatively short period of time. Most of the research done with gene expression data so far has focused on the development of visualization tools, using statistical techniques such cluster analysis. These have proven somewhat useful for identifying gene clusters and for the prediction of biochemical pathways involved. However, the technology requires the development of appropriate and meaningful statistical methods for analyzing and interpreting the large amount of data obtained.

The Affymetrix arrays utilize probe pair sets for each transcript of interest, comprised of perfect match (PM) and mismatch (MM) probes. Initial studies detailing the production of the Affymetrix arrays have outlined a clear rationale behind using these probe pair sets (Lockhart et al., 1996). Recent studies (Chu et al., 2002; Irizarry et al 2003) have individually identified two primary facts of Gene Chip Analysis. The first factor to be taken into consideration involves the relevant data that must be acquired from the GeneChips for utilization in the analysis, while the second factor involves the appropriate statistical modeling approaches to be used in analyzing the collected data. These studies outlined alternatives to the data analysis approaches that Affymetrix had recommended with its Microarray Suite (MAS 5.0) software (Irizary et al., 2003). Basically, two approaches were proposed as alternatives. The first one recommends the use of summarized probe level data for eventual comparison between arrays (Li and Wong, 2001; Irizarry et al., 2003; Bolstad et al., 2003). The second approach relies on the use of the probe level data on a gene-by-gene basis (Chu et al., 2002, Wolfinger et al. 2001) and the mixed linear model was proposed for the analysis of such data. Although both modeling approaches can weed out sources of variability between arrays, the mixed linear approach is more flexible as it can handle different experimental designs and sources of variation in a clear and straightforward manner.

There is an extensive literature on the Bayesian and frequentist implementation of mixed linear models in virtually every area of scientific research. However, there have been very few, if any, comparisons between the performances of both implementations for analysis of microarray data. Furthermore, the limited number of arrays (less than a dozen) in the majority of microarray experiments makes such comparisons more relevant. Using the fold change approach, several authors (Baldi and Long, 2001) showed that the Bayesian approach compared favorably to a simple fold change or a straight t-test and helped in a statistically consistent way, partially overcomes deficiencies related to low replication. In this study, a comparison between a frequentist and a Bayesian implementation of mixed linear models using probe level data is conducted. The correspondence between the two lists of differentially expressed genes obtained offers a simple, yet very objective way of accessing the adequacy as well as the efficiency of the analysis.

MATERIAL AND METHODS

Simulation

A simulation study was conducted to investigate the adequacy of a Bayesian approach via Markov Chain Monte Carlo (MCMC) methods to detect genes that are differently expressed. The simulation was conducted following a simple mixed linear model with the array being the random effect. Mathematically, the model can be expressed as:

$$\log_2(PM_{ijkl}) = t_i + P_j + a_k + e_{ijkl}$$

where $\log_2(PM_{ijkl})$: normalized $\log_2$ perfect match intensities

t$_i$: is the treatment effect i $(i = 1,2)$

P_j: is the probe effect j ($j = 1,2,3,...,20$)

a_k : is the random effect of the array k (k=1,2,…,8)

e_{ijkl}: is the residual term

Further, the following assumptions were made about the distribution of both random effects in the model:

$$\mathbf{a} \sim N(\mathbf{0}, \mathbf{I}\sigma_a^2)\,;\; \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

where $\mathbf{I}$ is the identity matrix with the appropriate dimension and σ_e^2 and σ_a^2 are the within and between array variation, respectively.

Data sets were simulated using the following combinations of the model parameters:

1. The magnitude of the array variation as a proportion of the residual variance (0.25, 0.50, 0.75, 1.0)
2. The percentage of differentially expressed genes (10%)

In total, four simulation scenarios were implemented. For each scenario, five replicates were simulated. The simulation was conducted following an existing structure of a real gene expression experiment generated at the Department of Animal and Dairy Science (Rao et al., 2004). This consisted of 8 arrays with 10,000 genes each. For each gene and depending of the model parameter combination, values were assigned to the within and between array variation. In all case, both variances were assumed to be non zero. The array effects were generated from a normal distribution with mean zero and variance equal to the already specified value for σ_a^2. Depending on the percentage of genes differentially expressed, the status of every gene being on or off was assigned randomly. If the status of a gene was off then both treatment levels was set equal to zero $(t_1 = t_2 = 0)$ otherwise, t_1 was set to zero and t_2 were sample from a uniform distributions U[0.3, 3] or U[-3, -0.3] with equal probabilities. In other words, we assumed that differentially expressed genes are equally likely to be up or down regulated. The boundaries of the uniform distributions were chosen in a way such that the observed fold change for gene differentially expressed range between 1 and 8 folds. The probe effects were generated from $N(0.7,0.09)$. The residual terms were sampled from a normal distribution with zero mean and variance equal to the already specified value for σ_e^2. Finally, the $\log_2$ intensity for every observation in the data set was calculated as the sum of the assigned values for all effects in the model.

Analysis of the simulated data

Each data set was analyzed using proc mixed of SAS and a full Bayesian approach via the Gibbs Sampling. In both cases, the simulation model was fitted. For the SAS analysis, the Restricted Maximum Likelihood (REML) method was used to estimate the variance components. The estimate statement was used to estimate the contrast between the two treatment levels as well as the associated standard deviation. The critical t value and the associate p-value were calculated using the appropriate degrees of freedom. For the Bayesian implementation, all needed conditional distributions were in closed form, being normal for the position parameters and scaled inverted chi square distributions for the dispersion parameters (σ_u^2 and σ_e^2) and the marginal posterior distributions of all parameters were easily obtained by successive sampling for their respective conditional distributions. Furthermore, quantities of interest such as the probability of treatment effect being significant, greater than a specific fold change, or the probability of being between two specific fold changes were computed and used to assess if the gene was differentially expressed. Furthermore, the genes detected by the analysis as being differentially expressed were contrasted against the true differentially expressed genes (determined during the simulation process).

Bayesian implementation

Based on the assumptions made during the simulation, the conditional distribution of the data given the model parameters was assumed to be normal:

$$\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{a}, \sigma_e^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2) \qquad [1]$$

where $\mathbf{y}$ is nx1 vector of $\log_2$ intensities, $\boldsymbol{\beta}$ is a px1 vector of systematic effects that includes the treatment and probe effects and $\mathbf{a}$ is the vector of the array effect of order qx1. Further, $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices with the appropriate dimensions, n is the total number of observations in data set, p is the sum of the treatment and probe levels and q=8 (number of arrays).

To complete the Bayesian formulation, prior information has to be specified to all unknown parameters in the model. It was assumed a prior that:

$$\boldsymbol{\beta} \sim N(\mathbf{0}, 10^4)$$

$$\mathbf{a} \mid \sigma_a^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_a^2)$$

$$\sigma_e^2 \sim U[0,1]$$

$$p(\sigma_a^2) \propto \frac{1}{\sigma_a^2}$$

where N(.,.) is a normal distribution with the specified mean and variance and U(.) is the uniform distribution.

The joint posterior distribution of all parameters is easily obtained as the product of the density in equation [1] and the densities of all priors.

$$p(\boldsymbol{\beta},\mathbf{a},\sigma_a^2,\sigma_e^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta},\mathbf{a},\sigma_e^2)p(\boldsymbol{\beta})p(\mathbf{a} \mid \sigma_a^2)p(\sigma_a^2)p(\sigma_e^2) \qquad [2]$$

Finally, the fully conditional distributions of all parameters required for the implementation of the Gibbs sampler were derived from equation [2] by taking those terms that are function of the parameter of interest and treating all the rest as nuisance parameters. Let $\mathbf{W} = [\mathbf{X}\ \ \mathbf{Z}]$and $\boldsymbol{\theta} = (\boldsymbol{\beta}',\mathbf{a}')'$. The conditional distribution of the position parameters, assuming that the inverse of the coefficients matrix exist, is given by:

$$\boldsymbol{\theta} \mid \sigma_a^2,\sigma_e^2,\mathbf{y} \sim N(\hat{\boldsymbol{\theta}},\mathbf{C}^{-1}\sigma_e^2)$$

where $\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}\sigma_{\mathrm{a}}^{-2} \end{bmatrix}$, $\mathbf{R} = \mathbf{I}\sigma_e^2$ and $\hat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{y}$

For the dispersion parameters, $\sigma_a^2$and σ_e^2, their respective conditional distributions were inverted scaled chi-square distributions with the following parameters

$$\sigma_a^2 \mid \boldsymbol{\theta},\sigma_e^2,\mathbf{y} \sim (\mathbf{a}'\mathbf{a})\chi_q^{-2}$$

$$\sigma_e^2 \mid \boldsymbol{\theta},\sigma_a^2,\mathbf{y} \sim (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})\chi_{n-2}^{-2}$$

For all analysis, convergence was assessed using methodology presented by Raftery and Lewis (1992). The required length of the burn-in period was always less then 2,000 iterations for all parameters. Thus, 25,000 iterations of the sampler were run with a conservative 5,000 iterations discarded as burn-in; all remaining 20,000 iterations were retained without thinning for post Gibbs analysis.

One of the big advantages of the Bayesian approach implementation via Markov Chain Monte Carlo (MCMC) such as the Gibbs sampler is the flexibility for computing statistics of interests, harder to compute with other methods, in a very simple manner. For example, the probability of the difference between two treatment levels being greater than a specified value or being within a given interval can be computed as a by-product of the sampling process. As an illustration, the probability that absolute difference between two treatment levels is greater than a specific value can be computed as follow:

$$p(|t_1 - t_2| > c) = \frac{\#\text{ of samples with } |t_1 - t_2| > c}{\text{total number of samples}}$$

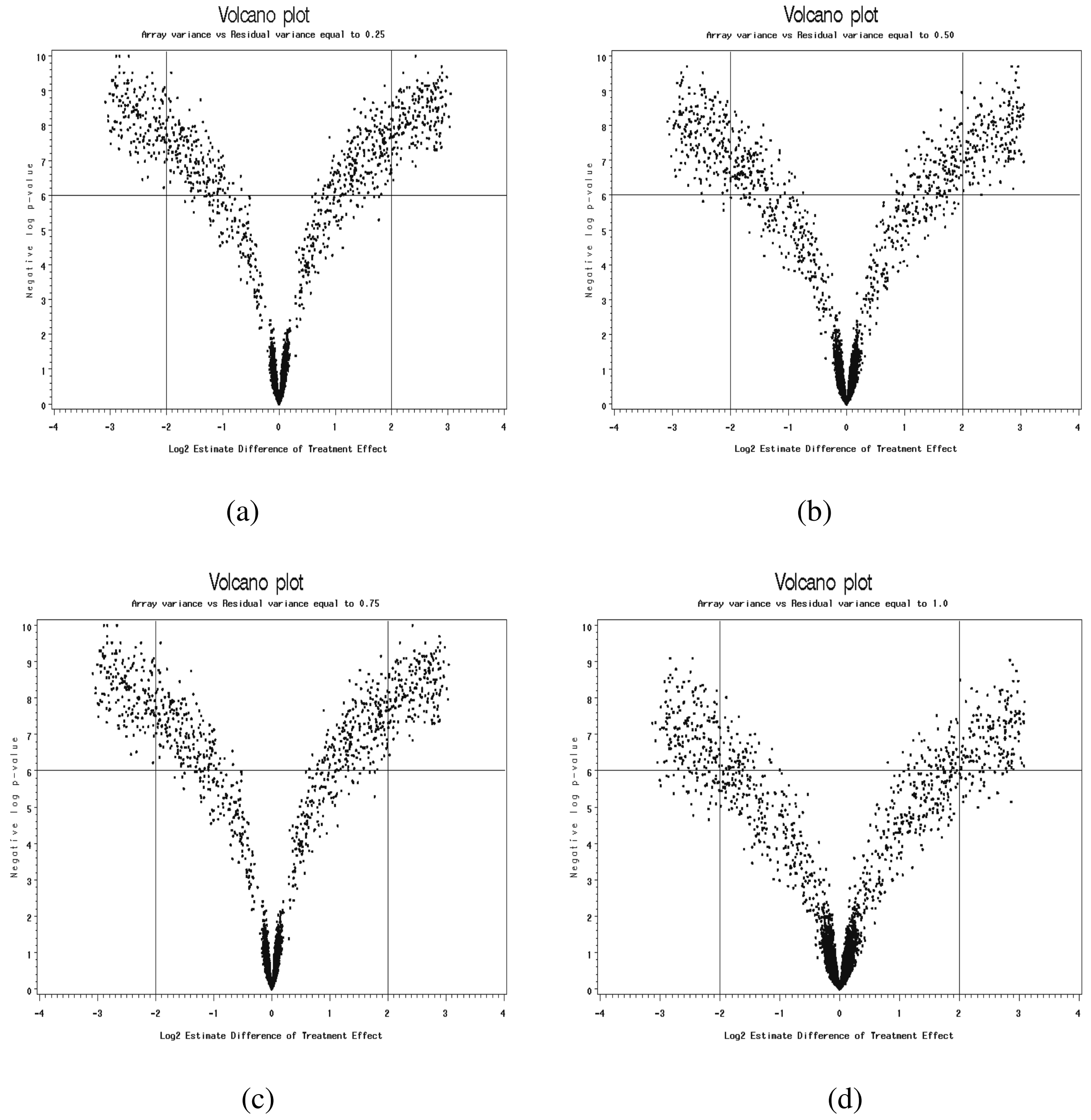
where t_1 and t_2 are the two treatment levels and c is the specified value.

RESULTS AND DISCUSSION

The simulated data was first analyzed using proc mixed of SAS. Data of each gene (96 to 128 observations) were analyzed separately. The estimate statement was used for estimating the difference at the log base two of the perfect match (PM) intensity for the two treatment levels. The p-value of the corresponding t-test of the estimate statement was used as a criterion for selecting the most differently expressed genes. Given that five replicates were performed to reduce the inherent Monte Carlo error, the significant level was computed based on the five simulated data sets. To do so, the average of the estimated treatment differences and associated standard deviations were used to compute the critical t value and subsequently the p-value for each gene. The resulting t-test p-values were used to identify the most differentially expressed genes.

The Volcano plot is the easiest and most effective way of presenting the results graphically. It combines both the differences between the two treatment levels and the associated p-values. It is a scatter plot of the negative $\log_{10}$ of the p-values versus the $\log_2$ of the estimated treatment difference ($t_1 - t_2$). For the four simulation scenarios, the correspondent volcano plots are presented in Figure 3.1. As expected, plots have the well recognized ‘V’ shape indicating that genes with large fold change due to the treatment effect tend to have a lower p-value. However, such relationship is not a one to one mapping. As a results, it is not infrequent that genes with large fold changes have low significance levels and vise versa. Given the large number of statistical tests performed, it is necessary to account for the multiple testing to reduce the percentage of false positive cases. Although several methods exist to deal with such problem, we decided to use restrictive criteria at least for two reasons: a) to reduce to the maximum the number of false positive cases and b) to reduce the number of highly differentially expressed genes. The latest is important, at least for practical reasons, because very few laboratories can study and understand the function of large number of genes mostly for economical reasons. Hence, it seems reasonable to focus in the most important genes (top genes). Therefore, the Bonferroni correction at 1% level was used in this study. Given that 10,000 genes and two treatment levels were used in the simulation, the p-value cut off point was set to $10e^{-6}$ (horizontal line in figure 3.1 indicate the negative $\log_{10}$ of the cutoff value). The vertical lines in the same figure indicate the 4 fold changes in the treatment effect estimates. The genes of interest or "top" genes are those residing in the top left and right of the plot. Those in the left represent the under regulated genes or genes with reduced expression after being subjected to the treatment. In the opposite side, are the upper regulated genes those expressions have increased as a result of the treatment.

To evaluate the adequacy of the statistical analysis and its ability of detect the truly differentially expressed genes, the list of genes in the top left and right boxes of the volcano plots have to be contrasted against the list of true differentially expressed genes determined during the simulation process. Table 3.1 presents the highly differentially expressed genes identified by the statistical analysis and their distribution in each class of true fold change for the four simulation scenarios. In all cases, the highly differentially expressed genes resulting from the statistical analysis have a true fold change greater than 2 and the majority of them (over 95%) have a fold change greater than 4. These results indicate that no non-differentially expressed gene (true fold change less than 2) was misidentified as differentially expressed. However, this result has to be interpreted with caution as it indicates just a part of the whole picture as it will be described in the next paragraph.



(a)

(b)

(c)

(d)

Figure 3.1 Volcano plots for the four simulation scenarios based on the ratio between the array and residual variances: a) 0.25, b) 0.50, c) 0.75 and d) 1.0

It is not sufficient merely that the highly differentially expressed genes determined by the statistical analysis correspond to true genes of interest, but it is also required that all genes of interest must be identified. Table 3.2 presents the number of genes with true fold change greater than 4 which were not identified by the statistical analysis as genes highly differentially expressed (false negative cases). The number of those genes varies depending on the ratio between the array variance and the residual variance. It was higher (53) when the ratio was large (1.0). Further, the number of false negative cases decreases with the decrease of the variances ratio.

Table 3.1. Most differentially expressed genes based on the volcano plot and their distribution into the different true fold change classes.

	Ratio[1]			
True Difference True fold change Number of genes	0.25	0.5	0.75	1
	N=354[2]	N=360[2]	N=348[2]	N=318[2]

Abs (T1-T2) = 0
Fold = 1
Ng = 9,000
	0	0	0	0

0 < Abs (T1-T2) ≤ 1 1< Fold ≤ 2 Ng =236	0	0	0	0
1 < Abs (T1-T2) ≤ 2 2< Fold ≤ 4 Ng = 404	6	10	10	11
2<Abs (T1-T2) ≤ 2.58 4< Fold ≤ 6 Ng = 215	204	205	193	162
Abs (T1-T2)≥ 2.58 Fold ≥ 6 N=145	144	145	145	145

[1]: Ratio between array variance and residual variance

[2]: The total number of highly differentially expressed genes

For all scenarios, the number of false negative cases was not negligible indicating that some genes of interest to biological process will not be identified, especially if the variances ratio is large. Furthermore, the number of important genes being non-identified by the statistical analysis increases when the fold change cutoff point is reduced. In fact, for a true fold change of 2 or greater the number of genes of interest not identified by the analysis ranges from 205 to 258.

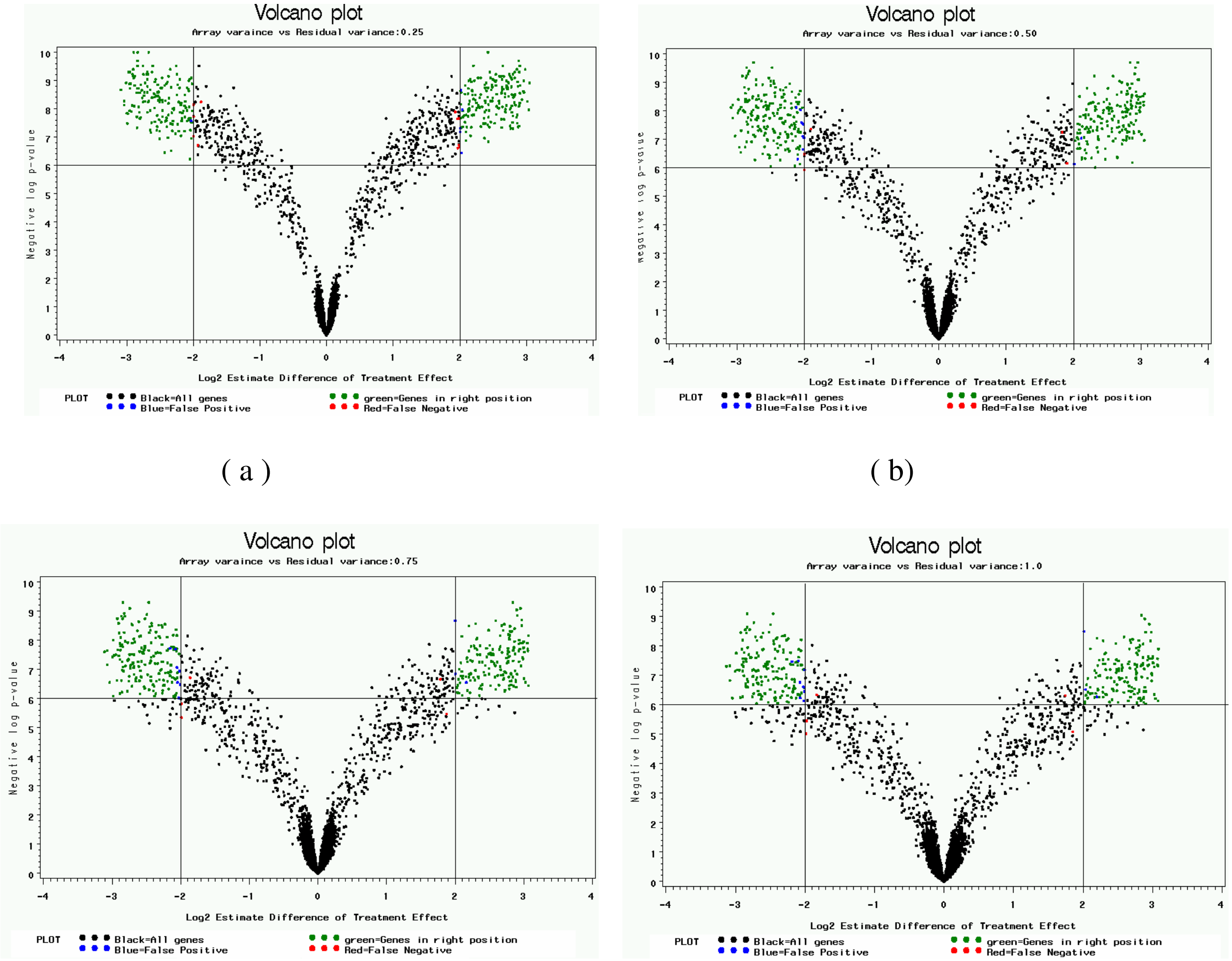
Table 3.2 Number of false negative cases for genes with true fold change greater than 2 and 4.

	Ratio[1]			
True Fold change	0.25	0.5	0.75	1
ABS (t_1-t_2) > 2, 4 fold	12	10	22	53
ABS (t_1-t_2) > 1, 2 fold	215	205	205	258

[1]: Ratio between array variance and residual variance

A more detailed examination of true genes of interest non being identified as well as less important genes being identified as highly differentially expressed (false positive cases) is graphically presented in Figure 3.2. It clearly noticeable that both the genes truly non-differentially expressed that are detected as differentially expressed (blue dots) and truly differentially expressed genes that were not identified as such (red dots) are very close to the edges of upper left and right boxes indicating only a small bias.

Although the small misclassification of some genes based on the mixed linear model analysis, the results could be very useful and can be used for sample classification or as a diagnostic tool with minor consequences. However, if the objective of the experiment is to determine the list of the most important genes for the study of biological functions or pathways for drug discovery for example, only a small number of genes can be looked at with great details. As a consequence, it is crucially important that the ranking of highly differentially expressed genes has to be maintained so that the biologist or geneticist will have access to the correct short list of most influential genes for further experimentation. For example, if only 145 genes were to be picked for a more detailed study out of the 354 highly differentially expressed genes determined during the analysis when the variances ratio was 0.25 (results for the other three scenarios are presented in Appendix A), such list will include 74, 61 and 10 genes from the true highest fold change, the 4 to 6 true fold change and 2 to 4 true fold change classes, respectively as indicated in Table 3.3. This result indicates that almost half of the most important genes will



(a) (b)

(c) (d)

Figure 3.2: Distribution of false positive and false negative cases for the four simulation scenarios based on variances ratio a) 0.25, b) 0.50, c) 0.75 and d) 1.0

not be selected and will be replaced by less interesting genes. Furthermore, this result indicates that selection of differentially expressed genes based only on the p-values can lead to misleading conclusions. In fact, this pronounced re-ranking of most influential genes could explain, in part, the large discrepancy and low reproducibility of several microarray experiment results. A detailed examination of those genes that were not identified correctly when the decision was based on the p-values indicates that 18 of them have their associate array variance wrongly estimated (set to zero).

Table 3.3: Distribution of top identified genes into the different true fold change classes (Variances ratio = 0.25).

True Difference			Rank[1]		
True fold change					
Number of genes	1 ~ 145	146 ~ 360	361 ~ 764	765 ~ 1000	> 1000
Abs (T1-T2) = 0 Fold = 1 Ng = 9,000	0	0	0	4	8996
0 < Abs (T1-T2) ≤ 1 1< Fold ≤ 2 Ng =236	0	0	39	193	4
1 < Abs (T1-T2) ≤ 2 2< Fold ≤ 4 Ng = 404	9	69	287	39	0
2<Abs (T1-T2) ≤ 2.58 4< Fold ≤ 6 Ng = 215	61	90	64	0	0
Abs (T1-T2) ≥ 2.58 Fold ≥ 6 N=145	75	56	14	0	0

[1] Rank based on p-values

This result is disturbing given that no gene was simulated with array variance equal to zero. Although those zero estimates of the arrays variance are in part the result of the small number of arrays (8) involved in the simulation, such number is not infrequent in microarray experiments. In fact, the vast majority of expression experiments involve less than a dozen arrays.

Bayesian implementation

Contrarily to the frequentist implementation where all information about unknown parameters is in the data, the Bayesian approach combines the data information with an external source of information through the prior. Although a non-informative prior (reference prior) was used for the array variance, all estimates of the latter were different from zero, as expected. In situations where the information content of the data is limited, such as in microarray experiments, the prior information plays a crucial role for having meaningful estimates for the parameters of interest. However, the prior information must be carefully chosen to avoid unrealistic estimates.

A full Bayesian implementation via Markov Chain Monte Carlo (MCMC) methods offers the possibility of calculating quantities of interest in a straightforward manner. In fact, the whole posterior distributions of unknown parameters are easily obtained and can be used to compute several quantities of interest such as point estimates, standard deviations, high density intervals and order statistics. In this study, the probabilities that the differences between the treatment levels being greater (smaller) than a specific value or being within a giving interval were computed as indicated in the material and methods part. The results presented are based on the average of 5 replicates. For each gene, four probabilities were computed: the probability that the difference between the two treatment levels is less than 2 fold change (p_1); the probability that the difference between the two treatment levels is between 2 and 4 fold change (p_2); the probability that the difference between the two treatment levels is between 4 and 6 fold change (p_3); and the probability that the difference between the two treatment levels is greater than 6 fold change (p_4). Genes were assigned to fold change classes based on these four probabilities.

Table 3.4 presents the distribution of genes and their average probability in each of the four fold change classes for variance ratio of .25 (results for the other three scenarios are presented in Appendix B). For the 145 most highly differentially expressed genes, 139 genes or 96% were correctly classified with an average p_4 probability greater than 0.96. Only 6 genes out of 145 were classified in the 4 to 6 fold change when their true fold change was greater than 6. This number is smaller than the one (11) obtained using a frequentist approach when both fold change and p-values were considered. The Bayesian results are much more superior when only the p-values are used for genes selection in the frequentist setup. In fact, only 74 out of the 145 most important genes were correctly classified (Table 3.3) from Frequentist. For genes with true fold change between 4 and 6 and between 1 and 2, the correct classification rate was 91.6% and 98.3%, respectively. For truly non-differentially expressed genes (9000 genes), they were correctly detected as such with probability of one.

Comparison between linear mixed model and Bayesian implementation

The purpose of this study was to evaluate the performances of a frequentist and Bayesian implementation of linear mixed model in the analysis of simulated microarray. The comparison between both approaches was based on four criteria: a) the correct identification of differentially expressed genes, b) the correct classification of differentially expressed genes into their true fold change classes, c) the minimization of false positives cases or non-differentially expressed genes being identified as differentially expressed and d) computation cost. For the first criteria, both methods have performed exceptionally well. In fact, out of the 360 true highly differently expressed genes (true fold change greater than 4), 348 (96.7%) and 350 (97.2%) genes (See Tables 3.3 and 3.4) were correctly identified by the frequentist and Bayesian analyses, respectively. Additionally, even for less highly differentially expressed genes (true fold change between 2 and 4), both methods gave good and similar results. When both methods were compared based on their capacity of clustering differentially expressed genes into their true fold change classes, the Bayesian approach was far superior. Out of the 145 true most highly differentially expressed genes, only 74 or 51% were within the top 145 genes with the smallest p-values (Table 3.3) using the frequentist approach with true variance ratio of 0.25. However, for the same comparison, 139 (96%) genes were correctly classified using the Bayesian approach (Table 3.4). The same trend was observed for less differentially expressed gene (true fold change between 2 and 4) and the other three simulation scenarios. This superiority of the Bayesian approach in classifying

Table 3.4. Distribution of genes and their average probability in each of the five true fold change classes (variances ratio = 0.25)

True difference True fold change Number of genes	P_1	P_2	P_3	P_4

t_1-t_2 = 0
Fold =1
Ng = 9000
	0.9999
	9000
		0
		0
			0
			0
				0
				0

0 <t_1-t_2 <1
Fold = 1~2
Ng = 236
	0.9775
	232
		0.6765
		4
			0
			0
				0
				0

1 ≤ t_1-t_2 ≤ 2
Fold = 2~4
Ng = 404
	0.6496
	11
		0.9475
		388
			0.680
			5
				0
				0

2 < t_1-t_2 < 2.58
Fold = 4~6
Ng = 215
	0
	0
		0.6902
		10
			0.932
			197
				0.692
				8

T_1-t_2 >2.58 Fold > 6 Ng = 145	0 0	0 0	0.704 6	0.96 139

differentially expressed genes is of crucial interest to biologists and geneticist as it helps them focus on real important genes rather than wasting time and money looking at less interesting ones.

Finally, a point worth mentioning is the computational cost of both implementations. For the 10,000 genes in the data set, it took 3 minutes for the frequentist analysis using proc mixed of SAS in a Dell 2650 machine with four processors. However, it took almost 5 hours in the same machine to conduct the Bayesian implementation.

CONCLUSIONS

Mixed linear model offers a general and flexible framework for analysis of microarray data. It replaces the ad-hoc normalization step by a systematic and theoretically sound procedure to account for all sources of variation. In situations where the number of arrays in the experiment is limited, the Bayesian implementation has proven to be superior to the frequentist counterpart, especially in the clustering or ranking of most important genes. Such superiority is of crucial practical interest as it gives biologists and geneticists better opportunities to focus on truly important genes for the biological process under investigation.

REFERENCES

Baldi, P., and A. D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics, 17: 509-519

Bolstad, B. M., R.A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19: 185-93.

Chu, T. M., B. Weir, and R. Wolfinger. 2002. A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci. 176: 35-51.

Irizarry, R.A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T.P. Speed. 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res., 31: 15.

Li, C., and W. H. Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci., 98: 31-37.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M.V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol., 14: 1675-1680.

Raftery, A. E., and S. M. Lewis.1992. One long run with diagnostics: Implementation strategies for Markow chain Monte. Statistical Sci., 7: 493-497.

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol., 8: 625-37.

CHAPTER 4

CONCLUSIONS

The advances at the genome level have produced a massive amount of information and have created the necessity to develop the quantitative methods capable of identifying the underlying biological processes that take place. Microarray technology has raised much interest both in the academic and commercial sector, especially in the field of human medicine. Mixed linear model is becoming a standard tool for the analysis of microarray data. It offers a general and flexible framework for analysis of microarray data. It replaces the ad-hoc normalization step by a systematic and theoretically sound procedure to account for all sources of variation. The frequentist and Bayesian implementations of the mixed linear model gave similar results in identifying highly differentially expressed genes. However, the Bayesian approach was far more superior in clustering the most important genes into their correct significance classes. In fact, 139 of 145 most important genes were correctly classified by the Bayesian approach verses 74 genes using the frequentist approach. Such superiority is of crucial practical interest as it gives biologists and geneticists better opportunities to focus on truly important genes for the biological process under investigation. These results suggest that in situations where the number of arrays in the experiment is limited, the Bayesian implementation seems to have better performances compared to the frequentist counterpart.

APPENDIX A

FREQUENTIST RESULTS FOR VARIANCES RATIOS OF 0.5, 0.75 AND 1.0

Table 3.5: Distribution of top identified genes into the different true fold change classes (Variances ratio = 0.50).

True Difference True fold change Number of genes	Rank[1]				
	1 ~ 145	146 ~ 360	361 ~ 764	765 ~ 1000	> 1000
Abs (T1-T2) = 0 Fold = 1 Ng = 9,000	0	0	0	13	8987
0 < Abs (T1-T2) ≤ 1 1< Fold ≤ 2 Ng =236	2	1	40	180	9
1 < Abs (T1-T2) ≤ 2 2< Fold ≤ 4 Ng = 404	53	80	228	43	0
2<Abs (T1-T2) ≤ 2.58 4< Fold ≤ 6 Ng = 215	47	78	90	0	0
Abs (T1-T2) ≥ 2.58 Fold ≥ 6 N=145	43	56	46	0	0

[1] Rank based on p-values

Table 3.6: Distribution of top identified genes into the different true fold change classes (Variances ratio = 0.75).

True Difference True fold change Number of genes	Rank[1] 1 ~ 145	 146 ~ 360	 361 ~ 764	 765 ~ 1000	 > 1000
Abs (T1-T2) = 0 Fold = 1 Ng = 9,000	0	0	0	27	8973
0 < Abs (T1-T2) ≤ 1 1< Fold ≤ 2 Ng =236	0	2	41	166	27
1 < Abs (T1-T2) ≤ 2 2< Fold ≤ 4 Ng = 404	17	73	271	43	0
2<Abs (T1-T2) ≤ 2.58 4< Fold ≤ 6 Ng = 215	56	89	70	0	0
Abs (T1-T2) ≥ 2.58 Fold ≥ 6 N=145	72	51	22	0	0

[1] Rank based on p-values

Table 3.7: Distribution of top identified genes into the different true fold change classes (Variances ratio = 1.0).

True Difference True fold change Number of genes	Rank[1]				
	1 ~ 145	146 ~ 360	361 ~ 764	765 ~ 1000	> 1000
Abs (T1-T2) = 0 Fold = 1 Ng = 9,000	0	0	0	37	8963
0 < Abs (T1-T2) ≤ 1 1< Fold ≤ 2 Ng =236	0	3	41	155	37
1 < Abs (T1-T2) ≤ 2 2< Fold ≤ 4 Ng = 404	17	74	269	44	0
2<Abs (T1-T2) ≤ 2.58 4< Fold ≤ 6 Ng = 215	55	89	71	0	0
Abs (T1-T2) ≥ 2.58 Fold ≥ 6 N=145	73	49	23	0	0

[1] Rank based on p-values

APPENDIX B

BAYESIAN RESULTS FOR VARIANCES RATIOS OF 0.5, 0.75 AND 1.0

Table 3.8. Distribution of genes and their average probability in each of the five true fold change classes (Variances ratio = 0.50)

True difference				
True fold change	P_1	P_2	P_3	P_4
Number of genes				
t_1-t_2 = 0	0.9999	0	0	0
Fold =1	9000	0	0	0
Ng = 9000				
0 <t_1-t_2 <1	0.974	0.6427	0	0
Fold = 1~2	226	10	0	0
Ng = 236				
1 ≤ t_1-t_2 ≤ 2	0.6936	0.9474	0.674	0
Fold = 2~4	12	384	8	0
Ng = 404				
2 < t_1-t_2 < 2.58	0	0.6983	0.8991	0.7028
Fold = 4~6	0	12	192	11
Ng = 215				
T_1-t_2 >2.58	0	0	0.734	0.9478
Fold > 6	0	0	7	138
Ng = 145				

Table 3.9. Distribution of genes and their average probability in each of the five true fold change classes (Variances ratio = 0.75)

True difference True fold change Number of genes	P_1	P_2	P_3	P_4
t_1-t_2 = 0	0.9999	0	0	0
Fold =1	9000	0	0	0
Ng = 9000				
0 <t_1-t_2 <1	0.9633	0.7065	0	0
Fold = 1~2	226	10	0	0
Ng = 236				
1 ≤ t_1-t_2 ≤ 2	0.6623	0.9374	0.6962	0
Fold = 2~4	18	378	8	0
Ng = 404				
2 < t_1-t_2 < 2.58	0	0.669	0.8858	0.6873
Fold = 4~6	0	18	182	15
Ng = 215				
T_1-t_2 >2.58	0	0	0.7362	0.9302
Fold > 6	0	0	8	137
Ng = 145				

Table 3.10. Distribution of genes and their average probability in each of the five true fold change classes (Variances ratio = 1.0)

True difference True fold change Number of genes	P_1	P_2	P_3	P_4
t_1-t_2 = 0	0.9999	0	0	0
Fold =1	9000	0	0	0
Ng = 9000				
0 <t_1-t_2 <1	0.9614	0.6870	0	0
Fold = 1~2	224	12	0	0
Ng = 236				
1 ≤ t_1-t_2 ≤ 2	0.6623	0.9311	0.683	0
Fold = 2~4	18	376	9	0
Ng = 404				
2 < t_1-t_2 < 2.58	0	0.6747	0.8728	0.693
Fold = 4~6	0	19	180	16
Ng = 215				
T_1-t_2 >2.58	0	0	0.749	0.9191
Fold > 6	0	0	8	137
Ng = 145				