Microarray Data Analysis Using Probabilistic Methods

Liu, Xuejun

2006

MIMS EPrint: **2006.403**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

MICROARRAY DATA ANALYSIS USING PROBABILISTIC METHODS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2006

By
Xuejun Liu
School of Computer Science

Contents

Word count 41363

List of Tables

List of Figures

Abstract

Affymetrix microarrays are currently the most widely used microarray technology. Due to the complexity of microarray experiments, the experimental data is very noisy. Many summarization methods have been developed to provide gene expression levels from Affymetrix probe-level data. Most of the currently popular methods do not provide a measure of uncertainty for the estimated expression level of each gene. The use of probabilistic models can overcome this limitation. This thesis extends a previously developed probabilistic model, mgMOS, to obtain an improved model, multi-mgMOS. This new model provides improved accuracy and is more computationally efficient than other alternatives. It also provides a level of uncertainty associated with the measured gene expression level. This probe-level measurement error provides useful information to help in the downstream analysis of gene expression data.

In order to show the advantage of the probe-level probabilistic model, the obtained uncertainty is propagated in two downstream analyses of gene expression data. One is detecting differential gene expression, another is clustering. A Bayesian hierarchical model is proposed to include probe-level measurement error into the detection of differential gene expression from replicated experiments and a standard model-based clustering method is augmented to incorporate probe-level measurement error. Due to the inclusion of the probe-level measurement error, the downstream probabilistic models become more complicated or intractable. In order to perform inference with these augmented models efficiently, various inference approximation approaches are compared in this thesis, including Maximum a Posteriori, Laplace approximation, a variational method and Markov chain Monte Carlo. Results from both benchmark data sets and a real-world data set demonstrate that the incorporation of the probe-level measurement error improves the performance of the downstream probabilistic analysis.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

First of all I would like to thank my supervisor Magnus Rattray for his help and guidance, and leading me into such an interesting area.

I would also like to thank my advisor Andy Brass for helpful discussions on the biological application of my work.

The work in this thesis received continuous help from Neil Lawrence and Marta Milo. I am very grateful to their help and encouragement.

I also wish to thank Guido Sanguinetti for useful discussions and Leo Zeef for suggestions on the implemented software.

I used a mouse time-course data set throughout this thesis. I would like to thank Bogi Andersen and Kevin Lin for providing me with such a useful data set and the related biological background knowledge, and helping with interpreting my analysis results.

I would also like to thank every member in the AI group. They gave me useful suggestions on many occasions. Particularly, this thesis received a thorough reading from Richard Pearson for which I'm very grateful.

I gratefully acknowledge the Overseas Scholarship Scheme from the university for paying my tuition fee and a studentship from the school for providing me with the living maintenance.

Finally, I am very grateful to my husband Hongqiang Lu, my parents and my sisters for their love, support and encouragement throughout my life.

Chapter 1

Introduction

1.1 Genomic Scale Biology

1.1.1 The Human Genome Project

The Human Genome Project was launched formally in 1990 to create a reference of the entire human DNA sequence and identify all genes in the human genome. The approach of the human genome project is sequencing, which is examining the order of the basic building blocks of DNA (A, C, G and T in abbreviation) along the human genome. The information provided by the human genome project is expected to contribute to systems biology which aims to understand all processes in cells, and in their development from genes up to phenotypes. Genomes are of fundamental importance in the sense that they will revolutionise all biology and biomedicine.

The project was completed in 2003 and discovered 20,000-25,000 estimated human genes (International Human Genome Sequencing Consortium, 2004). Along with the process of human genome sequencing, the genomes of many other species are now sequenced to provide helpful information used in different area, such as some viruses and bacteria, Baker's yeast, fruit fly, mouse, and so on. As the whole genome of human and other organisms is completely sequenced, the next task is to reveal how genes make life, what their functions are, what the relationship between them is and how the changes in genes affect their functions. To fulfil this task, it is necessary to understand the biological system of tens of thousands of genes.

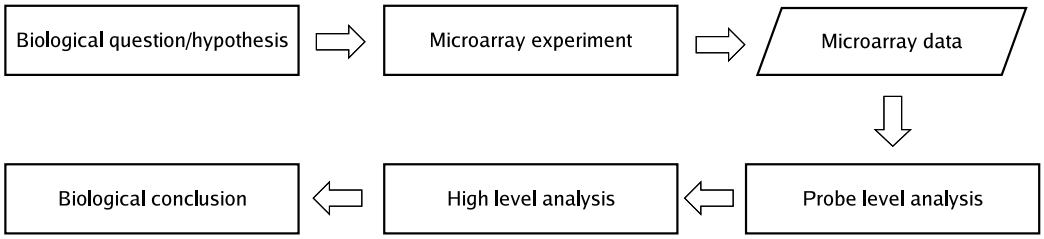

Figure 1.1: The process of biological discovery involving microarray technology.

1.1.2 Microarray Technology

Microarray technology offers the ability to take global views of biological processes by providing a systematic way to study DNA and RNA variation (Schena et al., 1995; Lockhart et al., 1996). The central dogma of the molecular biology states that genetic information flows from DNA to messenger RNA (mRNA) and from RNA to proteins which perform gene functions (Crick, 1970), and this process is called gene expression. The amount of RNA in this process indicates the level of gene expression. Microarrays measure the gene expression level on a genomic scale by examining the amount of mRNA in cell cultures or tissues and they provide insight into gene function by quantitatively studying gene expression. Over the last decade microarrays have become an increasingly important tool in modern biomedical and life sciences research.

Affymetrix is one of the leading manufacturers of microarrays and its product GeneChip® is hugely popular (Lockhart et al., 1996). Using synthetic oligonucleotide technology, Affymetrix microarrays are high-density DNA probe arrays which contain millions of probe sequences. The expression level of each gene is measured by multiple probes on the microarray (see Chapter 2). This thesis focuses on the analysis of data produced by Affymetrix microarrays.

1.1.3 Process of Biological Study Using Microarrays

A typical microarray experiment is usually motivated by a biological question, like "which genes show changes in expression between a healthy tissue and a diseased one?", or a more specific biological hypothesis, like "a certain group of genes are responsible for the development of a particular disease". The correct answer to the question or the verification of the hypothesis will help with the diagnosis and treatment of the particular disease. In order to answer the biological question, a microarray experiment can be conducted following the experiment protocol (see

Section 2.1.2). The general procedure of discovering biological knowledge using microarrays is shown in Figure 1.1.

Microarrays quantitatively measure gene expression on a large scale. Enormous amounts of gene expression data are generated from microarray experiments. In order to obtain meaningful information for the organism being studied, multiple levels of analyses are performed on the primary data. The first stage of the analysis is probe-level analysis which summarises the raw data to obtain a single expression value for each gene from the experimental data. The probe-level analysis should provide reliable measurements of gene expression levels which can be used in the high level analysis. The next stage of analysis (high level analysis) performs various tasks on the measured gene expression resulting from the probe-level analysis (Quackenbush, 2001; Slonim, 2002).

The high level analysis depends on the biological questions which motivate the microarray experiment. It could simply be detecting differential gene expression, which is the most basic aim of a microarray experiment involving two or more microarrays. Alternatively, people may be interested in revealing the patterns of gene expression and unknown gene functions across multiple conditions or time points in a developmental process. Gene expression patterns can be discovered by "bottom up" approaches and gene functions can be predicted by "top down" approaches (Bassett et al., 1999). The "bottom up" approaches include principal component analysis and clustering. These methods are applied on data solely and do not involve any previous knowledge. Genes which have similar functions are usually organised into the same cluster or stay close together in the visual representation resulting from principal component analysis. With known biological information, unknown gene functions can be revealed by examining the known classes they fall in by "top down" analysis approaches (Pan, 2006). Inferring gene regulatory networks is another important goal of high level analysis of gene expression data (Segal et al., 2003; Friedman, 2004). Genes are interacting with each other during the cell's life. Inferring the interaction network of genes is important for the understanding of the genetic mechanism of living organisms.

At the final stage of the discovery process, biological conclusions are drawn based on the results obtained from the high level analysis. As a result, either new biological knowledge is discovered or the original hypothesis is falsified. Every step in the discovery process is vital to subsequent steps. Within a carefully designed and performed experiment, analysis of the experimental data plays an

important role in making sound biological conclusions.

1.2 Probabilistic Microarray Data Analysis

The microarray experiment is a complicated multi-step procedure and variability can be introduced at every experimental stage. The variability can be intrinsic to the biological system and can also come from various sources during the experimental process. The variability caused by the experiment itself is referred to as technical variability. The goal of a microarray experiment is to measure the biological changes in the cell or tissue under investigation. Technical variability, however, can obscure these biological changes. In order to reduce the technical variability, replicated experiments are usually conducted (Lee et al., 2000). However, due to the high cost of the experiments, a small number of replicates is often used (usually three or four).

Since variability exists in the whole procedure of the microarray experiment, the estimated gene expression level is associated with a level of uncertainty which reflects the significant sources of variability. Probability is a natural representation of uncertainty and is suitable for describing the noisy nature of gene expression data. Among the numerous methods for gene expression data analysis, probabilistic approaches (Baldi and Long, 2001; Segal et al., 2001; Yeung et al., 2001; Ghosh and Chinnaiyan, 2002; Friedman, 2004) have proved to be useful in numerous applications.

The uncertainty of the gene expression measurement can be obtained from the probe-level analysis and propagated into the downstream probabilistic analysis to achieve a more reasonable interpretation of the experimental data (Sanguinetti et al., 2005; Wang et al., 2006). However, for the probe-level analysis most popular methods (Affymetrix, 2002; Li and Wong, 2001a; Irizarry et al., 2003; Zhang et al., 2003; Wu et al., 2004) are not able to provide the credibility interval associated with the gene expression measurement. Existing probabilistic models (Milo et al., 2003; Hein et al., 2005; Milo et al., 2004) are capable of providing credibility intervals but it is argued in this thesis that these are not sufficiently accurate or are too computationally expensive to apply in practice.

For the downstream analysis, current probabilistic models use only a single point estimate to represent a gene's expression level and therefore completely ignore much of the available evidence about the source of experimental variability.

For example, in approaches to detect differential gene expression, like t-tests, people have observed that the limited number of replicates leads to a significant underestimate of signal variance (Baldi and Long, 2001; Delmar et al., 2005). People therefore devised many methods to obtain a more reasonable estimate of the uncertainty in the gene expression level (Baldi and Long, 2001; Delmar et al., 2005; Medvedovic et al., 2004; Lin et al., 2004). However, few of these approaches consider the uncertainty in probe-level measurements.

1.3 Aim of the Thesis

The multiple probes for each gene used in Affymetrix technology provide information redundancy in the microarray experimental data. This rich information can be used to measure experimental variability along with the measurement of gene expression. By using probabilistic methods in probe-level analysis, it is possible to associate gene expression measurements with levels of uncertainty to characterise the noisy nature in the experimental data, especially for low expressed genes whose measured expression levels are usually dominated by noise (Milo et al., 2003; Hein et al., 2005; Milo et al., 2004). It is also possible to propagate the probe-level variance into the high level probabilistic analysis to obtain more reasonable results by considering the variance of observed data. It is argued in this thesis that the probe-level measurement error can be propagated through many of these methods in order to make the most efficient use of available data.

The aims of this thesis are:

1. To develop an improved probe-level model which accounts for Affymetrix probe-level data more reasonably and is computationally efficient enough for practical applications. The model should provide an accurate measure of expression level with an associated level of uncertainty.

2. To augment the probabilistic methods in the high-level analysis to incorporate gene expression measurement uncertainty.

3. To test the hypothesis that by including gene expression measurement uncertainty obtained from probe-level analysis, the high level analysis of microarray data will be improved.

1.4 Thesis Outline

Based on the aims of the thesis described in Section 1.3, the work in this thesis includes the development of an improved probabilistic probe-level analysis model, multi-mgMOS (Liu et al., 2005), and the propagating of the probe-level uncertainty into two higher level analyses, detecting differential gene expression and clustering. The presentation of the work in this thesis is organised in the following five chapters.

Chapter 2 introduces the basic biological and microarray technology background to help with the understanding of this work. The key concepts and methods of probabilistic inference used in this work are also introduced.

Chapter 3 , which focuses on probe-level analysis, reviews the currently popular probe-level analysis methods, particularly existing probabilistic models, and investigates the development of the new model used in multi-mgMOS. An empirical comparison with other methods is also presented in this chapter.

Chapter 4 includes the extension of a Bayesian hierarchical model to detect differential gene expression by incorporating probe-level uncertainty. When considering the measurement error associated with gene expression measurements, the standard Bayesian hierarchical model becomes intractable. Several approximation methods are compared. The results on a benchmark data set and a real data set show the improvement obtained by including the probe-level measurement error.

Chapter 5 is an example of the augmentation of a model-based clustering method to include probe-level measurement error. The comparison on simulated data sets and a real mouse time-course data set shows the improvement of the augmented standard Gaussian mixture model by incorporating probe-level measurement error.

Chapter 6 gives the conclusion of this work and proposes possible directions for future research.

Chapter 2

Background

This chapter introduces the foundations of microarray technology and describes the key concepts and methods in probabilistic inference used in this thesis.

2.1 Biological Background

2.1.1 Protein-coding Genes

Cells are the basic functional elements of life. Chromosomes, large segments of DNA (deoxyribonucleic acid), which carry the instructional information for directing cell functions are contained in the nucleolus of cells. The biochemical building blocks that make up DNA are nucleotides which consists of three different biochemical components: a base, a sugar and a phosphate. Each nucleotide contains one of four bases which are known as adenine (A), guanine (G), cytosine (C) and thymine (T). DNA is in a double helix structure. Nucleotides are joined together to build a linear sequence for each strand of the double helix. The two strands are held together by base pairing between nucleotides within the two different strands. A is always paired with T and C is always paired with G. The chemical process by which the double strands are formed from two complementary strands is called hybridisation, or binding, and this is the essential idea behind microarray technology.

Analogously, it is the amino acids that act as the biochemical building blocks used to make proteins which perform most life functions. There are 20 different amino acids. Genetic information is stored in DNA sequences. The fundamental unit of genetic information is called a codon which contains three successive

nucleotides along a DNA sequence to specify an amino acid. There are 64 possible combinations of the four nucleotides (4^3). Of the 64 possibilities, 61 codons specify the 20 amino acids (in the canonical code) and the remaining 3 combinations code the stop signal indicating protein termination.

RNA (ribonucleic acid) is the intermediary during the conversion of genetic information from DNA to proteins. RNA is also made up of nucleotides and has a similar linear structure to DNA, but differs from DNA in several ways. The sugar contained in RNA nucleotides is different from that in DNA nucleotides and the thymine base is replaced by uracil (U). Also, nearly all RNA molecules are single stranded rather than in the double helix structure of DNA. Messenger RNA (mRNA) is the carrier of genetic information from the nucleus, where DNA is located, to the cytoplasm, where protein is synthesized. Each mRNA corresponds to a specific gene contained in DNA sequences.

The process of genetic information flowing from DNA into RNA and from RNA into protein is the process of gene expression and also is known as the central dogma of molecular biology (Crick, 1970). There are three key steps in the process of gene expression: transcription, splicing and translation. Figure 2.1 shows the steps of transcription and translation.

During the process of transcription, single-stranded mRNA is synthesized by treating one strand of the DNA double helix as a template in the nucleus. The synthesis is made by the complementary base paring mechanism of nucleotides. A is transcribed from T, U is from A, and C is from G, and vice versa. There are short DNA sequences called regulatory elements that control the expression of genes. The two main regulatory elements are the promoter and the enhancer. A promoter indicates the start site of the transcription for a gene. An enhancer regulates the transcriptional efficiency of a specific gene. Regulation occurs through the action of the activator and repressor proteins binding to enhancers. The binding of activators to enhancer sites makes the corresponding genes produce more mRNA molecules and the binding of repressors decreases the amount of produced mRNA molecules (see Brown (2002) for a detailed account of gene regulation). The amount of produced mRNA molecules, therefore, indicates the level of gene expression. Measuring the abundance of mRNA molecules is the main task of microarray technology.

For multicellular organisms, such as humans, primates, insects and so on, genes contain two types of sequence segments, coding segments and non-coding

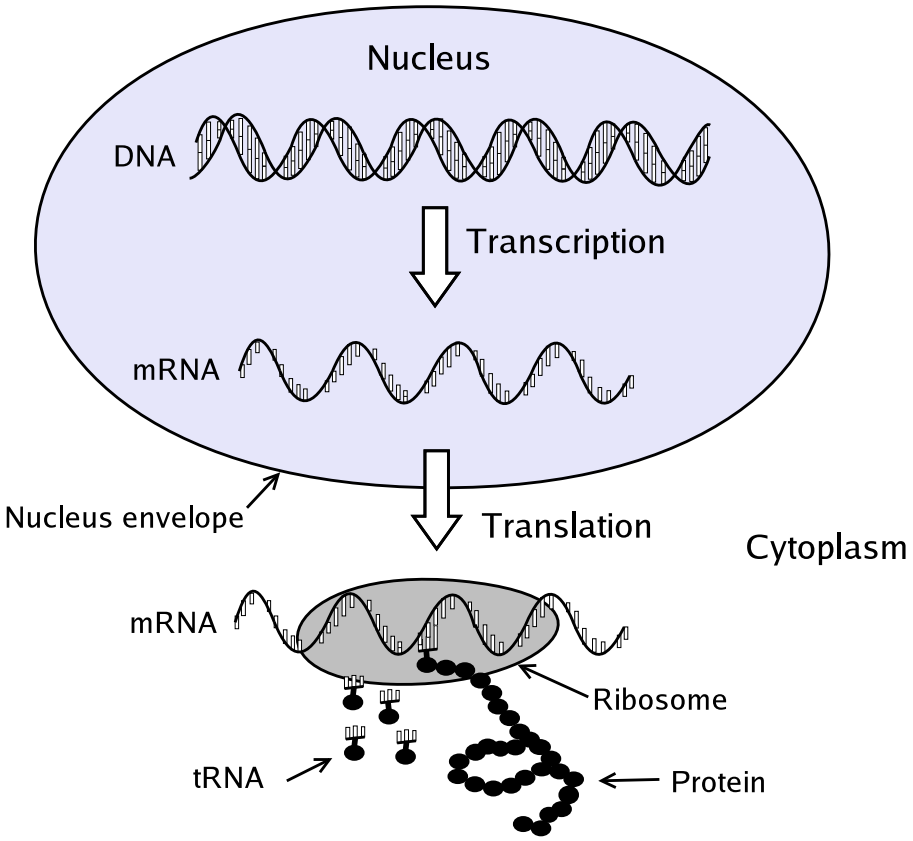

Figure 2.1: The two main steps of gene expression, transcription and translation.

segments. Coding segments are called exons and carry the genetic information to synthesize proteins, while non-coding segments are called introns and do not carry genetic information for protein synthesis. Genes contain interleaving exons and introns. At the splicing step, introns are removed from mRNA leaving a shorter sequence of coding mRNA.

At the translation step, mRNA leaves the nucleus and enters the cytoplasm to start the synthesis of proteins. The mRNA is read by the ribosomes which are large cytoplasmic structures. The ribosome recognises the start codon (AUG) and begins the translation by adding the first amino acid in the protein chain. Translation proceeds codon by codon. At each codon, an amino acid carried by tRNA is added to the growing protein chain by codon recognition. The translation terminates when it encounters one of the three stop codons (UAA, UGA and

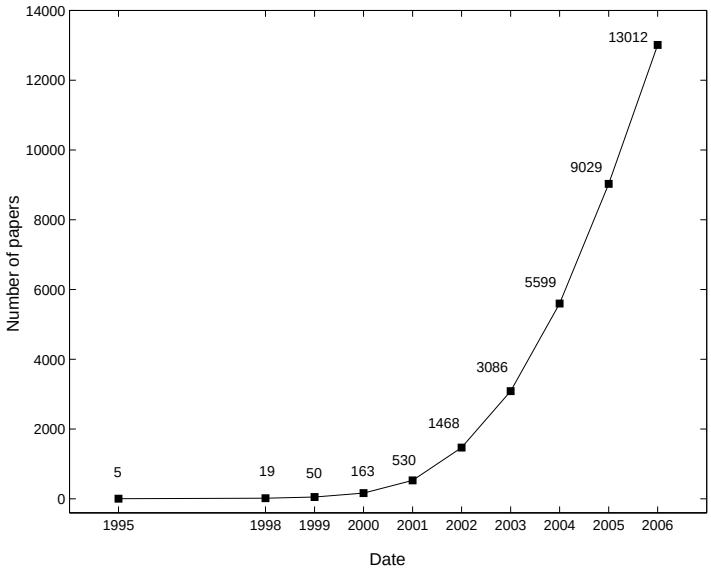

Figure 2.2: The number of research papers related to microarray published in recent years. Data are obtained from PubMed (http://www.pubmed.gov) which is a service of the U.S. National Library of Medicine that includes over 16 million citations for biomedical articles back to the 1950s.

UAG) in the mRNA sequence. Newly synthesized proteins fold into the correct three-dimensional structure and perform biological functions of various types, e.g. as enzymes, antibodies, transcription factors and so on.

The genome is the entire DNA content of a cell, including the nucleotides, genes and chromosomes. The human genome is estimated to contain around 20,000-25,000 protein-coding gene loci (International Human Genome Sequencing Consortium, 2004). Different organisms contain different genomes which configure diverse life forms. Each organismal cell contains the same genome structure, while different tissues of the same organism possess various functions and appearance. This is due to the different gene expression resulting in different cellular processes. The modification of gene sequences or gene expression levels can also lead to uncontrolled cell growth and disease. People are therefore interested in globally exploring the hidden genomic world. The microarray is a tool which enables the rapid and quantitative analysis of gene expression on a genomic scale.

2.1.2 Microarray Technology

Microarrays measure gene expression levels on a genomic scale simultaneously by monitoring the abundance of the intermediary mRNA (Schena et al., 1995;

Lockhart et al., 1996). The large number of published papers in recent years demonstrates that microarray technology is an extremely valuable tool in biomedical and life sciences research. Figure 2.2 shows the quickly increasing number of research articles with "microarray" in their abstracts published since 1995.

A microarray is "an ordered array of microscopic elements on a planar substrate that allows the specific binding of genes or gene products" (Schena, 2003). The key idea of microarray technology is binding or hybridisation, which is the chemical process where the two complementary strands of DNA or RNA combine to form a double strand under certain conditions. The "microscopic elements" are single stranded nucleotide sequences, which are called probes, fixed to the surface of microarrays. "Genes or gene products" are mRNA or total RNA molecules isolated from the biological specimens. They are called targets in microarray terminology. The conception of probes and targets here is consistent with nomenclature in Phimister (1999). Targets are fluorescently labeled and mixed in solution. The mixture of targets, known as sample, is then hybridised to the microarray. The RNA sequences of the targets will bind to their complementary probes. After a certain time allowed for hybridisation, the arrays are washed to get rid of the extra sample and the arrays are scanned to obtain a two dimensional image. The intensity at each probe position indicates the amount of RNA molecules bound to the specific probe and provides a quantitative measurement of the expression level of the related gene.

A typical microarray experiment involves the following steps:

1. Isolate RNA from the tissue of interest and prepare fluorescently labeled targets.

2. Hybridise the labeled targets to the microarray.

3. Wash, process and scan the microarray.

4. Process the resulting image to obtain a quantitative measurement of the intensity for each probe.

Among currently available microarray technologies, there are two widely used classes, cDNA microarrays first developed at Stanford (Schena et al., 1995) and synthetic oligonucleotide microarrays mainly produced by Affymetrix (Lockhart et al., 1996). In this thesis, we focus on Affymetrix microarrays which are often referred to as chips or arrays in abbreviation.

2.1.3 Synthetic Oligonucleotide Microarrays

In Affymetrix microarray technology, probes are oligonucleotides which are short single-stranded nucleotide sequences (DNA or RNA) and usually include 25 nucleotides. According to the available sequence information, probes are chemically synthesised from DNA and RNA building blocks, nucleotides, at a specific location on the surface of arrays (Lockhart et al., 1996). The precise location where each probe is synthesised is called a feature. One single high-density Affymetrix array with typical size $1.28cm \times 1.28cm$ contains millions of features. At each feature position, the probe is present in millions of copies in order to capture the unknown amount of target molecules with the complementary sequence in the sample.

Apart from the synthetic oligonucleotides, another speciality of Affymetrix microarray technology is the redundancy in the probe design (Figure 2.3). The concept of redundancy is embodied in two aspects. One is that each gene corresponds to multiple probes on the array, another is that arrays contain pairs of probes for each of the RNA sequences being monitored. For each gene, the reference sequence comes from its related spliced mRNA which contains only exons and flanking RNA. A subset of exon-specific probes is specifically chosen in order to detect the spliced mRNA in samples. The set of probes related to a particular gene is called a probe-set.

There are two types of hybridisation occurring on the array during the binding of targets to probes, specific hybridisation and non-specific hybridisation (Lockhart et al., 1996; Southern et al., 1999), also known as specific binding and non-specific binding respectively. Specific hybridisation means that the double-stranded molecule is formed from two perfectly complementary strands, one from probe sequences and another from target sequences. Non-specific hybridisation, sometimes also called cross-hybridisation in the literature, refers to the hybridisation happening between two strands which are not perfectly complementary. Each probe on the array that is perfectly paired with its target sequence is called a perfect match (PM) probe. In order to identify the non-specific hybridisation, for each PM probe on the array there is a mismatch (MM) probe which has the identical nucleotide sequence as the PM probe except that the middle nucleotide is changed to the complementary one. For example, A is changed to T and C is changed to G, and vice versa. There are 11-20 PM/MM probe-pairs contained in each probe-set. By design the MM probe detects the non-specific hybridisation

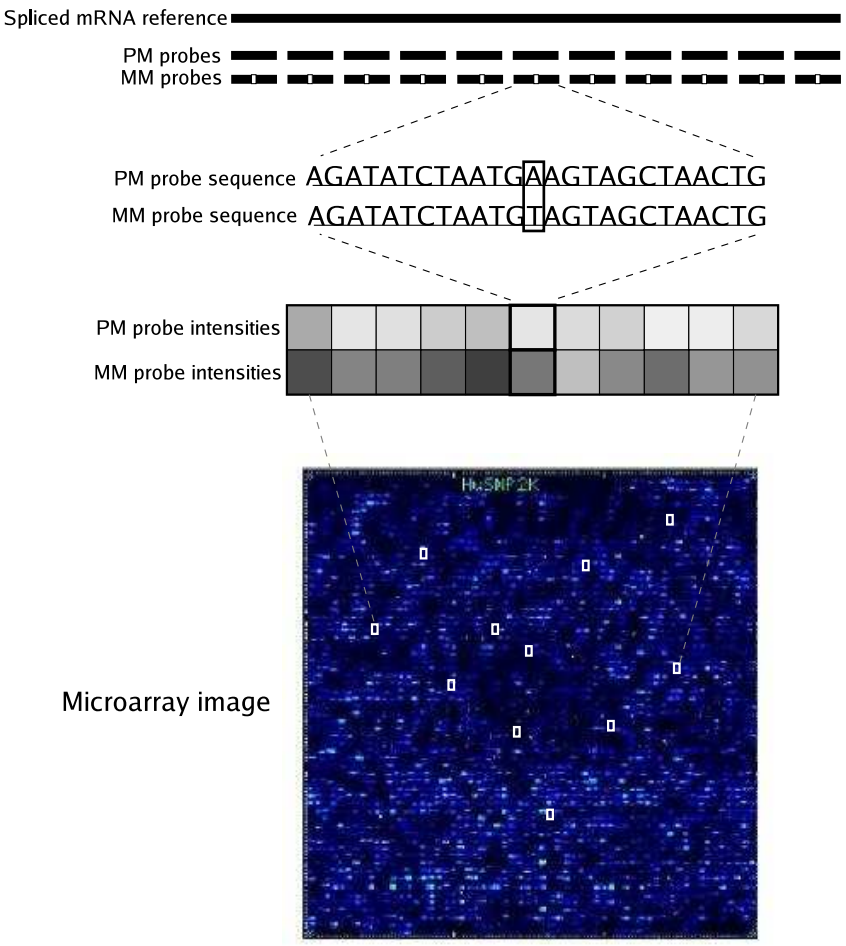

Figure 2.3: The probe design and output of Affymetrix microarrays.

on its complementary PM partner, since with only one base replaced the MM probe has a similar efficiency of binding to the non-specific target sequences as the PM counterpart. The MM probes therefore serve as internal controls for hybridisation specificity. In order to avoid minor defects in the hybridization image, probes are scattered throughout the surface of the arrays.

Figure 2.4 shows the procedure of an Affymetrix microarray experiment. During the experiment, fluorescent labeled RNA molecules are fragmented and hybridised to the array. The degree of hybridisation is assessed by monitoring fluorescent emission using a laser scanner. For each chip, a two dimensional image is created with each probe being identified by its coordinates on the array and measured for its fluorescent intensity. The measured intensity values represent the expression level of the related gene and coordinates on the array are stored in a cell intensity file (*.CEL) as the final results of the experiment. Each chip

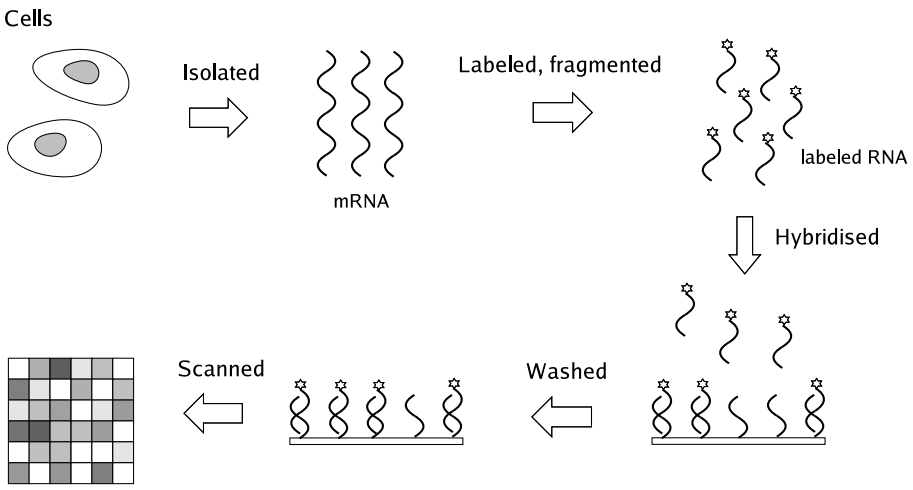

Figure 2.4: The procedure of Affymetrix microarray experiment.

corresponds to a CEL file. For each type of chip related to each particular organism, Affymetrix provides an array layout description file (*.CDF). The CDF file describes the design of a chip defining which probes belong to which probe-sets. By looking up the CDF file the intensity values for each probe-set can be extracted. The relationships between probe-sets and genes are also provided by the manufacturer in the documentations. The probe-level analysis methods which are described in Chapter 3 begin with the intensity measurements in CEL files.

2.2 Probabilistic Inference

Microarray experiments are associated with high variability and microarray data are inherently noisy. Probabilistic models provide powerful tools to deal with the uncertainty associated with this noisy data (Baldi and Brunak, 2001; Durbin et al., 1998). The Bayesian framework is widely used to handle complicated problems due to its flexibility and generality in probabilistic inference (Gelman et al., 2004). In this thesis, we use Bayesian models to quantify the uncertainty in microarray data by fitting data with parameters and hierarchical structures.

2.2.1 Data Likelihood

The aim of probabilistic inference is to reveal the implications behind the observed data by constructing a model that describes the origin of the observed data. Probabilistic theory is very useful in constructing models for data.

The experimental data, $D = \{x_i\}$ where $i = 1, 2, \ldots, n$, are the n observed values for the random variable X. A probabilistic model with unknown parameters θ can be constructed to model the experimental data by assuming the probability density function of the random variable is $P(X|\theta)$ given parameters θ. If the observed data is assumed to be independently and identically distributed (i.i.d.), the likelihood function is

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) \ . \tag{2.1}$$

In some problems, apart from the observed variables X there maybe unobserved variables, $\mathcal{H} = \{h_i\}$, called latent or hidden variables. To obtain the likelihood of a model containing latent variables, these variables should be marginalised first,

$$P(D|\theta) = \prod_{i=1}^{n} \sum_{\mathcal{H}} P(x_i, \mathcal{H}|\theta) \ . \tag{2.2}$$

By maximising the likelihood function in (2.1) or (2.2) with respect to the parameters θ, the maximum likelihood (ML) estimate of θ, $\hat{\theta}$, can be obtained. For convenience, we usually maximise the log-likelihood $\mathcal{L}(\theta) = \log P(D|\theta)$. The resulting model with estimated parameters $\hat{\theta}$ then describes the observed data with a high likelihood.

2.2.2 Bayesian Inference

In the maximum likelihood approach, the parameters are determined by finding the maxima of the likelihood function. In the Bayesian approach, the unknown parameters θ themselves are treated as random variables coming from the prior distribution $P(\theta)$. The prior distribution expresses our degree of belief about the value of θ before viewing the observed data.

After specifying the prior distribution $P(\theta)$, the data D is observed and is used to calculate the posterior distribution of θ, $P(\theta|D)$. The posterior distribution of θ is made up of both the prior information and the observed data. It is then used to construct the estimator of the unknown parameters θ. The relationship between the likelihood, the prior and the posterior distribution of θ is represented by Bayes' theorem,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \ . \tag{2.3}$$

The denominator, called the marginal likelihood or evidence, does not depend on θ and can be considered as a constant with fixed data D. This constant ensures the posterior distribution integrates to one over the allowed parameter space. If we omit the constant in the denominator, we can obtain the unnormalised posterior density,

$$P(\theta|D) \propto P(D|\theta)P(\theta) \ . \tag{2.4}$$

To provide a more flexible model or to model hierarchical data which includes multiple groups, hierarchical models can be defined by specifying the prior density with further unknown parameters, $P(\theta|\phi)$. The parameters ϕ are called hyper-parameters in a hierarchical model and have their own prior distribution, $P(\phi)$. The joint prior distribution is

$$P(\theta, \phi) = P(\theta|\phi)P(\phi) \ , \tag{2.5}$$

and the joint posterior distribution, ignoring the normalisation constant, is

$$\begin{aligned} P(\theta, \phi|D) &\propto P(D|\theta, \phi)P(\theta, \phi) \\ &= P(D|\theta)P(\theta, \phi) \ . \end{aligned} \tag{2.6}$$

The simplification of $P(D|\theta)$ in (2.6) comes from the fact that the distribution of the data depends only on θ.

In the computation of Bayesian models, conjugacy, which means the posterior distribution follows the same parametric form as the prior distribution, is very important for computational convenience. For example, the beta prior distribution is a conjugate family for the binomial data likelihood. The advantages of a conjugate prior are twofold: computational convenience and interpreting prior information as additional data. When there is no prior information about the values of parameters, a uniform or "flat" prior distribution can be used to make inference based on available data. However, a uniform distribution for ϕ implies a non-uniform distribution for any non-linear monotone transformation of ϕ. So whether the prior is uniform will depend on the choice of parameterisation of the likelihood function and therefore it is not always clear whether a flat prior is non-informative or natural (Bernardo and Smith, 1994).

2.2.3 Approximations to the Posterior

Maximum a Posteriori

A crude point estimator of parameters can be obtained by maximum a posteriori (MAP), i.e. finding the posterior mode in (2.4) or (2.6),

$$\theta_{MAP} = \arg\max_{\theta} h(\theta) \qquad \text{where} \quad h(\theta) = \log\left(P(D|\theta)P(\theta)\right) \quad , \tag{2.7}$$

is usually maximised. If a flat prior is used, the MAP estimate is equal to the ML estimate maximising (2.1). This point estimate is reasonable since it gives the single most likely choice for the parameters, and the computation of it is convenient with numerous optimisation algorithms available. However, it may not be representative especially when the posterior distribution of the parameters has low probability density or the estimate is at the boundary of the parameter space. The MAP estimator can be used as a reasonable starting point for more accurate methods or as a comparable reference to check the validity of results from other approaches.

Markov Chain Monte Carlo

The most frequently used approach to obtain accurate Bayesian inference for complicated (e.g. non-conjugate) models is Markov chain Monte Carlo (MCMC). Random draws from the posterior distribution of model parameters can be used to summarise Bayesian inference. The key idea behind MCMC is drawing sequential samples from the approximated posterior distribution of parameters and adjusting the distribution for each draw. At each time t, the new sample θ^t depends only on the draw at time $t-1$. For example, with the initial parameter values, θ^0, θ^1 is drawn from $P(\theta|\theta^0, D)$ and θ^2 is drawn from $P(\theta|\theta^1, D)$. The sequence of random variables, $\theta^1, \theta^2, \cdots$, is called a Markov chain in the sense that the distribution of θ^t given all previous θ's depends only on the previous one, θ^{t-1}. As the sampling goes on, the approximated posterior distribution, $P(\theta|\theta^t, D)$, will converge to the target true posterior, $P(\theta|D)$.

There are two widely used MCMC algorithms, the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) and the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). The Gibbs sampler is based on

iterative sampling from the conditional distribution of parameters. For a parameter vector $\theta = (\theta_1, \cdots, \theta_d)$, where θ_i is a subvector, at iteration t there are d steps of sampling. At each step, only one subset of the parameter vector is drawn from the conditional distribution $P(\theta_i|\theta_1^t, \cdots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \cdots, \theta_d^{t-1}, D)$ given all other subvectors at their latest values. If the conditional distributions have standard form, it is convenient to sample directly from them.

For those conditional distributions which are not in standard form, the Metropolis-Hastings algorithm can be used instead. This algorithm uses an acceptance/rejection rule to make a random walk which eventually converges to the target distribution. With a starting draw θ^0, the acceptance/rejection rule at iteration t can be depicted as follows:

1. Draw a proposal θ^* from a proposal distribution, $J(\theta^*|\theta^{t-1})$.

2. Calculate the ratio,

$$r = \frac{P(\theta^*|D)/J(\theta^*|\theta^{t-1})}{P(\theta^{t-1}|D)/J(\theta^{t-1}|\theta^*)} \ . \tag{2.8}$$

3. Set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise.} \end{cases} \tag{2.9}$$

The key fact is that

$$\frac{P(\theta^*|D)}{P(\theta^{t-1}|D)} = \frac{P(D|\theta^*)P(\theta^*)}{P(D|\theta^{t-1})P(\theta^{t-1})} \ , \tag{2.10}$$

so the denominator in (2.3), which is very difficult to compute in many cases, cancels.

In practice, the Gibbs sampler and Metropolis algorithm can be combined to approximate the posterior distribution in a complicated Bayesian model. For the parameter subvectors which have standard distribution form, the Gibbs sampler can be used, and for those which do not, the Metropolis-Hastings algorithm can be used in the iterative updates.

The difficulty of MCMC implementation is the assessment of convergence of the iterative simulation. The recommended approach in Gelman et al. (2004) is to assess convergence based on multiple sequences with overdispersed starting points. For each parameter of interest, θ, suppose there are m paralled simulated

sequences, each of length n after discarding simulations at a burn-in stage. Let θ_{ij} denote the ith draw in the jth sequence. The between- and within-sequence variances, B and W, can be calculated respectively by

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\theta})^2, \quad W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \ , \tag{2.11}$$

where

$$\bar{\theta}_j = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j, \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2 \ . \tag{2.12}$$

The convergence of the simulation is monitored by estimating the root ratio of between- and within-sequence variances,

$$\hat{R} = \sqrt{\frac{B}{W}}, \tag{2.13}$$

which is called the potential scale reduction by Gelman et al. (2004). The simulation is considered to converge when $\hat{R}$ of every parameter is close to 1.

Laplace Approximation

When MCMC is too computationally expensive for some problems, the Laplace method can be used to approximate the posterior distribution. If the posterior in (2.4) is highly peaked at the maximum, a useful approximation of the distribution is a Gaussian centered at the MAP estimate, θ_{MAP}. Expanding (2.7) at θ_{MAP}, we have

$$\begin{aligned} P(\theta|D) &\simeq \exp\left(h(\theta_{MAP}) + \frac{1}{2}(\theta - \theta_{MAP})^T \mathrm{V}(\theta - \theta_{MAP})\right) \\ &\propto \exp\left(-\frac{1}{2}(\theta - \theta_{MAP})^T(-\mathrm{V})(\theta - \theta_{MAP})\right) \\ &\propto \mathcal{N}(\theta; \theta_{MAP}, (-\mathrm{V})^{-1}) \ , \end{aligned} \tag{2.14}$$

where $\mathcal{N}(\mu, \Sigma)$ represents a Gaussian distribution with mean μ and covariance matrix Σ and V is the Hessian matrix with elements defined as

$$\mathrm{V}_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} h(\theta)\bigg|_{\theta=\theta_{MAP}} \ . \tag{2.15}$$

Variational methods

The Laplace approximation is only accurate if the posterior is well approximated by a Gaussian centered at the MAP solution. Other approximate Bayesian methods have been developed. A popular choice which can be applied to hierarchical or latent variable models is the variational approximation. The posterior can be obtained by marginalising over the distribution of hyperparameters,

$$P(\theta|D) = \int P(\theta, \phi|D) d\phi \ . \tag{2.16}$$

However, for some problems the integral is not tractable. In this case, the ML or MAP estimate of ϕ, ϕ_{ML} or ϕ_{MAP}, is useful to obtain $P(\theta|D, \phi_{ML})$ or $P(\theta|D, \phi_{MAP})$. The ML or MAP estimate can be calculated by maximising the marginal likelihood,

$$\mathcal{L}(\phi) = \log P(D|\phi) = \log \int d\theta \, P(D|\theta, \phi) P(\theta|\phi) \ , \tag{2.17}$$

or the marginal posterior, when there is prior information about ϕ,

$$h(\phi) = \log(P(D|\phi)P(\phi)) = \log \int d\theta \, P(D|\theta, \phi) P(\theta|\phi) + \log P(\phi) \ . \tag{2.18}$$

If the integral in (2.17) or (2.18) is tractable, the maximisation can be carried out by standard numerical methods. If it is intractable, the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) combined with a variational method (Ghahramani and Beal, 2001; Jordan et al., 1999) can be used to optimise a lower bound on $P(D|\phi)$ and work out the posterior distribution $P(\theta|D, \phi)$ (Beal, 2003). Taking the ML estimate as an example, the distribution $Q(\theta)$ over θ and Jensen's inequality are used to get a lower bound on $\mathcal{L}(\phi)$,

$$\begin{aligned} \mathcal{L}(\phi) &= \log \int d\theta \, Q(\theta) \frac{P(D|\theta, \phi) P(\theta|\phi)}{Q(\theta)} \\ &\geq \int d\theta \, Q(\theta) \log \frac{P(D|\theta, \phi) P(\theta|\phi)}{Q(\theta)} \\ &= \int d\theta \, Q(\theta) \log P(D|\theta, \phi) P(\theta|\phi) - \int d\theta \, Q(\theta) \log Q(\theta) \ . \end{aligned} \tag{2.19}$$

The lower bound is known as the Kullback-Leibler (KL) divergence (except for an additive constant) which measures the discrepancy between $Q(\theta)$ and $P(\theta|\phi, D)$. If there are no constraints on the form of $Q(\theta)$, optimising the KL-divergence

results in

$$\begin{aligned} Q(\theta) &= P(\theta|\phi, D) \\ &\propto P(D|\theta, \phi)P(\theta|\phi) \ . \end{aligned} \tag{2.20}$$

One can use an EM algorithm to optimise (2.19) with respect to $Q(\theta)$ and ϕ iteratively. Starting from some initial hyper-parameters ϕ^0:

$$\textbf{E-step:} \quad Q(\theta)^{t+1} = P(\theta|\phi^t, D) \tag{2.21}$$

$$\textbf{M-step:} \quad \phi^{t+1} = \arg\max_{\phi} \int d\theta \, Q(\theta)^{t+1} \log P(D|\theta, \phi)P(\theta|\phi) \ . \tag{2.22}$$

In the case of MAP learning, the M-step is

$$\textbf{M-step:} \ \phi^{t+1} = \arg\max_{\phi} \Big(\int d\theta \, Q(\theta)^{t+1} \log P(D|\theta, \phi)P(\theta|\phi) + \log P(\phi) \Big) \ . \tag{2.23}$$

When the EM algorithm converges, $Q(\theta)$ provides an estimate of the posterior distribution of parameter θ given ϕ_{ML} or ϕ_{MAP}.

However, in practice it is difficult to determine the form of $Q(\theta)$. In order to make progress, $Q(\theta)$ is approximated by a function which factorises across disjoint subsets of θ. Take two disjoint subsets, θ_1 and θ_2, of θ for example,

$$Q(\theta) = Q(\theta_1)Q(\theta_2) \ . \tag{2.24}$$

Substituting (2.24) for $Q(\theta)$ in (2.19) and optimising with respect to $Q(\theta_1)$ and $Q(\theta_2)$ results in

$$Q(\theta_1) \propto \exp\Big[\int d\theta_2 Q(\theta_2) \log P(D|\theta_1, \theta_2, \phi^t)P(\theta_1, \theta_2|\phi^t) \Big] \tag{2.25}$$

$$Q(\theta_2) \propto \exp\Big[\int d\theta_1 Q(\theta_1) \log P(D|\theta_1, \theta_2, \phi^t)P(\theta_1, \theta_2|\phi^t) \Big] \ . \tag{2.26}$$

The E-step in (2.21) is replaced by a sub-loop in which $Q(\theta_1)^{t+1}$ and $Q(\theta_2)^{t+1}$ are optimised iteratively by (2.25) and (2.26). The M-step in (2.22) is adjusted accordingly as

$$\phi^{t+1} = \arg\max_{\phi} \int d\theta_1 d\theta_2 \, Q(\theta_1)^{t+1} Q(\theta_2)^{t+1} \log P(D|\theta_1, \theta_2, \phi)P(\theta_1, \theta_2|\phi) \ . \tag{2.27}$$

It can be seen from (2.25) that $Q(\theta_1)$ depends on the expectations under

$Q(\theta_2)$. When these expectations are intractable, an importance sampler can be used to handle this intractability (Lawrence et al., 2004; Vermaak et al., 2003). Supposing that $Q(\theta_i)$ can be factorised as the following form

$$Q(\theta_i) = p(\theta_i) f(\theta_i) \ , \tag{2.28}$$

where $p(\theta_i)$ is the density function of θ_i which has a standard form and $f(\theta_i)$ is a function of θ_i, and drawing n samples of θ_i from $p(\theta_i)$, the estimated expectation of a function $g(\theta_i)$ under $Q(\theta_i)$ is

$$\int d\theta_i Q(\theta_i) g(\theta_i) \approx \sum_{k=1}^{n} \omega_k g(\theta_i^k), \text{ where } \omega_k = \frac{f(\theta_i^k)}{\sum_k f(\theta_i^k)} \ . \tag{2.29}$$

2.2.4 Model Selection

Suppose there are a set of candidate models $\mathcal{M}_m$, $m = 1, \cdots, M$, and corresponding model parameters θ_m. If the models are fitted based on maximisation of a log likelihood $\mathcal{L}(\theta_m)$, we can choose the best model from them according to the Akaike information criterion (AIC, Akaike (1973)) or the Bayesian information criterion (BIC, Schwartz (1978)). The formula of AIC for model $\mathcal{M}_m$ is

$$\text{AIC}_m = -2\mathcal{L}(\hat{\theta}_m) + 2d_m \ , \tag{2.30}$$

where d_m is the number of free parameters to be estimated in model $\mathcal{M}_m$ and $\hat{\theta}_m$ is the ML or MAP estimate for parameters θ_m. The calculation of BIC is given by

$$\text{BIC}_m = -2\mathcal{L}(\hat{\theta}_m) + d_m \log n \ , \tag{2.31}$$

where n is the number of data points. The best model is selected by minimising AIC_m or BIC_m.

Despite its similarity with AIC, BIC is slightly different and is motivated by the Bayesian approach to model selection. Assuming the prior distribution for the parameters of each model $\mathcal{M}_m$ is $P(\theta_m|\mathcal{M}_m)$, the posterior distribution of a given model is

$$\begin{aligned} P(\mathcal{M}_m|D) &\propto P(\mathcal{M}_m) P(D|\mathcal{M}_m) \\ &\propto P(\mathcal{M}_m) \int d\theta_m P(D|\theta_m, \mathcal{M}_m) P(\theta_m|\mathcal{M}_m) \ , \end{aligned} \tag{2.32}$$

where $P(\mathcal{M}_m)$ is the prior probability of choosing $\mathcal{M}_m$. The probability ratio between $\mathcal{M}_m$ and $\mathcal{M}_l$ is

$$\frac{P(\mathcal{M}_m|D)}{P(\mathcal{M}_l|D)} = \frac{P(\mathcal{M}_m)}{P(\mathcal{M}_l)} \frac{P(D|\mathcal{M}_m)}{P(D|\mathcal{M}_l)} \ . \tag{2.33}$$

The prior over models is usually assumed uniform, so the second ratio of evidence for different models, which is called Bayes' factor, shows how well the observed data were predicted by $\mathcal{M}_m$ compared to $\mathcal{M}_l$. The difficulty is to evaluate the evidence $P(D|\mathcal{M}_m)$. Using the Laplace method in (2.14), the evidence of $\mathcal{M}_m$ can be approximated as the following

$$\begin{aligned} P(D|\mathcal{M}_m) &\propto \int d\theta_m P(D|\theta_m, \mathcal{M}_m) P(\theta_m|\mathcal{M}_m) \\ &\simeq \int d\theta_m \exp\left(h(\hat{\theta}_m) + \frac{1}{2}(\theta_m - \hat{\theta}_m)^T V (\theta_m - \hat{\theta}_m)\right) \\ &= \exp\left(h(\hat{\theta}_m)\right) (2\pi)^{d_m/2} |-V^{-1}|^{1/2} \ , \end{aligned} \tag{2.34}$$

where $h(\cdot)$ and V are defined in (2.7) and (2.15) respectively. Thus

$$\log(P(D|\mathcal{M}_m)) \simeq h(\hat{\theta}_m) + \frac{d_m}{2}\log(2\pi) + \frac{1}{2}\log|-V^{-1}| \ . \tag{2.35}$$

Followed by some other simplifications (Ripley, 1996), the evidence in (2.35) can be simplifed as

$$\log(P(D|\mathcal{M}_m)) \simeq \mathcal{L}(\hat{\theta}_m) - \frac{d_m}{2}\log n \ , \tag{2.36}$$

which is proportional to the BIC score in (2.31). It is also feasible to use (2.35) directly for model selection.

Chapter 3

Probabilistic Probe-level Analysis

This chapter describes the development of an improved probabilistic probe-level model, multi-mgMOS. After introducing the background information and related work, multi-mgMOS is proposed and a comparison with other existing methods is given, followed by the conclusion.

3.1 Affymetrix Probe Characteristics

The aim of probe-level analysis is to summarise the gene expression level from a set of PM and MM probe intensity values as shown in Figure 3.1. As the first stage of microarray data analysis, the probe-level analysis should provide reliable gene expression values for the high level analysis. However, this work is challenging due to the noisy nature of the probe-level data. The PM/MM probes are designed so that PM probes measure the amount of specific hybridisation and MM probes are a control for background and non-specific hybridisation. From previous work on the characteristics of Affymetrix probe-level data, the following key observations are obtained:

1. In spike-in studies it is observed that PM and MM intensities both increase with concentration, although the increase in the MM intensities is less than that of the PM intensities (Chudin et al., 2002; Irizarry et al., 2003). This means both PM and MM probes measure the true specific signal to some extent (see Figure 3.2).

2. The response of spike-in genes to increasing transcript concentration is not always linear, especially for the low and high intensity probes (Chudin et al.,

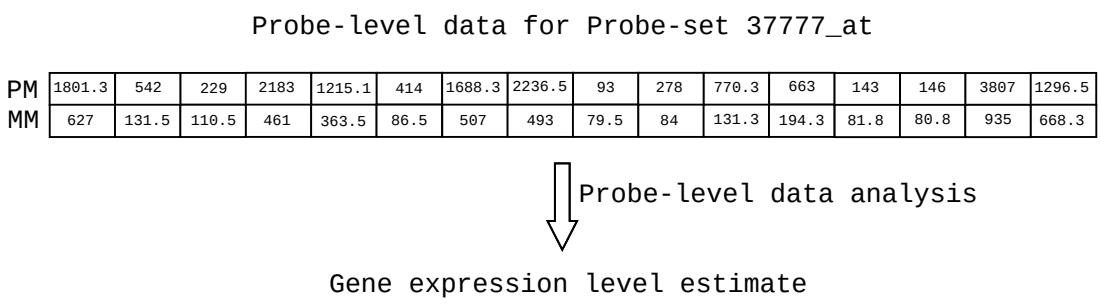

Figure 3.1: The process of probe-level data analysis. The probe data is from probe-set 37777_at of Affymetrix HG-U95a spike-in data set (Appendix A.1) at concentration 32pM. This probe-set contains 16 probe-pairs. In a spike-in study mRNA extracts are spiked in at known concentrations.

2002; Antonellis et al., 2001) as shown in Figure 3.2. At the lower end of transcript concentrations, the intensity of the PM probe is close to its corresponding MM probe intensity and the nonlinearity of the response is largely due to background effects (Chudin et al., 2002). The scanner effects are likely to be more pronounced at the higher target concentrations, as the response of the scanner photomultiplier is no longer in the linear range (Antonellis et al., 2001). From Figure 3.3, which shows the density of PM and MM intensities for 25 chips in a real mouse data set, most probe intensities lie between 6 and 10 on a log scale. In other words, most probes are associated with genes that are relatively low expressed. The correction of the non-linear response of probes is therefore more important for the lower concentration targets than the higher concentration targets.

3. About 10-30% of probe pairs (called negative probe pairs) have higher MM probe intensity than their corresponding PM probe, especially in the low intensity value probe-sets (Naef et al., 2002). An example can be seen clearly in the upper-left plot in Figure 3.2. The curve of MM probe is higher than the PM curve in the lower end. When the amount of mRNA is large, PM intensities are typically significantly larger than the corresponding MM intensities. However, it can be seen from Figure 3.3 that most probes are relatively low expressed and when the mRNA concentration is low the PM and MM intensities of the same probe-pair are very close to each other. This is reasonable but makes the summarisation of expression values for low expressed genes more difficult since stochastic error due to non-specific binding and experimental variability dominates the signal in this regime.

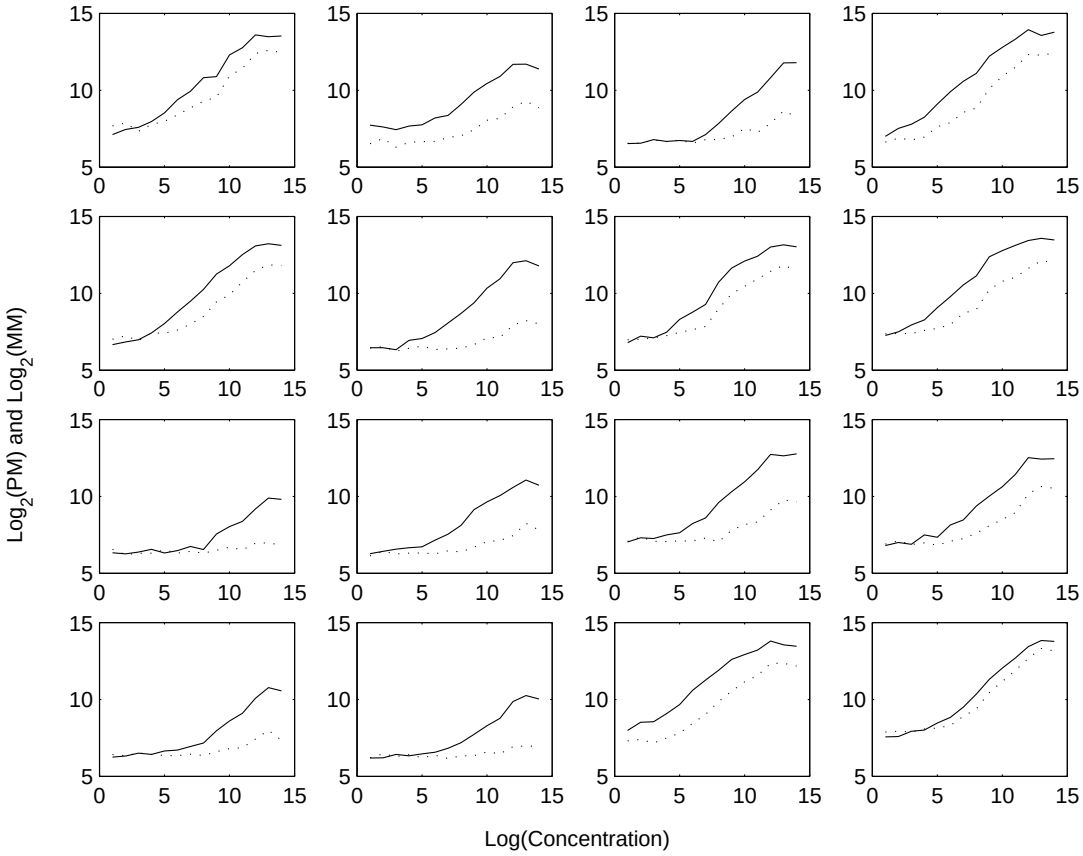

Figure 3.2: The logarithm of the intensity of the 16 probe pairs of the spike-in gene 37777_at in Affymetrix HG-U95a spike-in data set (Appendix A.1) versus the logarithm of transcript concentrations. The solid line represents PM intensities and the dotted line represents MM intensities. A linear relationship between concentration and intensities is observed in the median range of the concentrations.

4. PM and MM probe intensities, and also differences between them, vary in probe-specific ways (Li and Wong, 2001a; Irizarry et al., 2003; Wu et al., 2004; Hein et al., 2005) as shown in Figure 3.4. The similarity of nucleotide content between the PM probe and MM probe is thought to make the two probes have correlated sensitivity to signal. The sensitivity of probes varies within the probe-set and the pattern of the probe effect stays invariant across different conditions. The variation due to probe effects is larger than the variation across arrays (Li and Wong, 2001a). It can also be seen in Figure 3.1 that there is high variability for the probe intensities within the same probe-set.

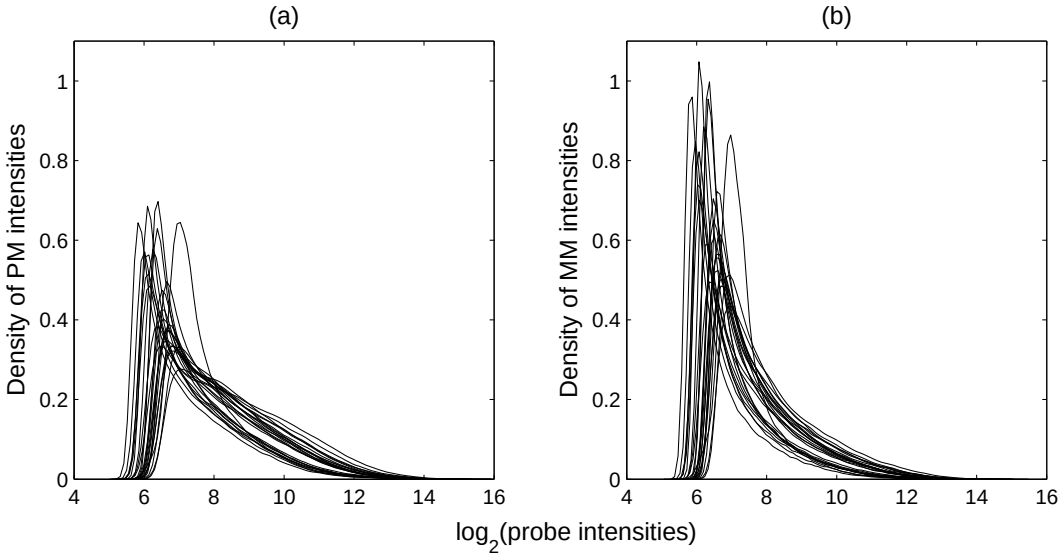

Figure 3.3: Density of probe intensities for the 25 chips in the mouse time-course data set (Appendix A.3). (a) is for PM intensities, (b) is for MM intensities.

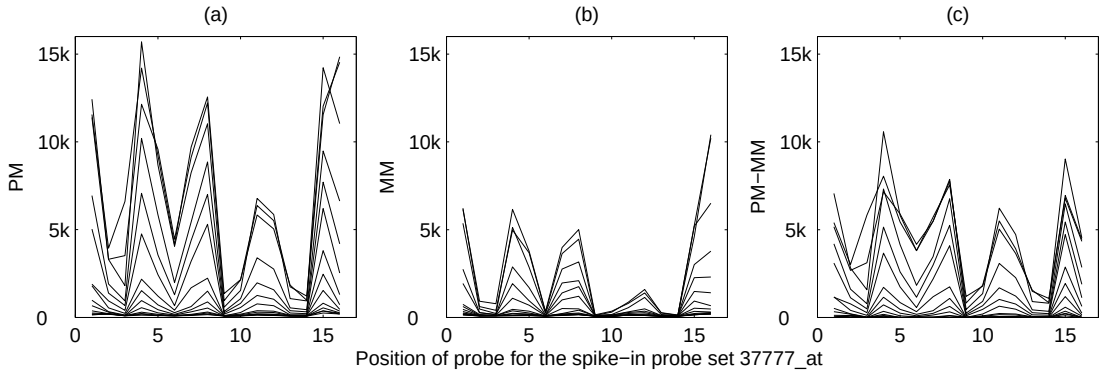

Figure 3.4: Probe intensity patterns for spike-in probe-set 37777_at at 14 different concentrations in Affymetrix Latin Square spike-in data set (Appendix A.1). (a) is for PM intensity, (b) is for MM intensity and (c) is for the difference between the intensities of PM and MM probes. PM, MM and PM-MM have high probe-specific effects.

3.2 Related Work

The noisy nature of microarray experiment data has motivated the development of numerous algorithms for estimating gene expression values. These algorithms perform various transformations on the probe intensities, trying to enable researchers to compare microarray outcomes of different genes quantitatively across separate arrays. A variety of computational approaches are used in these algorithms.

3.2.1 Popular Statistical Methods

Statistical methods have been developed to summarise gene expression values. It is difficult to account for all characteristics of probe data described in Section 3.1 in a single model. Therefore, most models handle the experimental data in several steps such as background correction, normalization and summarization.

Affymetrix Microarray Suite Software version 5.0, MAS 5.0 (Affymetrix, 2002), overcomes the large proportion of negative probe-pairs by modifying the intensities of MM probes to make the difference between PM and MM probe intensities always positive. A robust average of the probe-pair intensity differences for each probe-set is then calculated as the expression level of the probe-set. Finally, a global scaling normalisation is performed on the estimated expression values so that the signal for each chip has the same average value.

The model-based expression index, MBEI (Li and Wong, 2001a), normalises probe data based on an invariant set of probes that belong to non-differentially expressed genes. In order to find the invariant set, it is assumed that the intensity ranks of a probe from an invariant gene in two arrays are similar. A piecewise linear running median line is then calculated and used as the normalisation curve. MBEI models probe effects as a probe-specific multiplicative component in the computation of the expression level on the raw intensity scale. The existence of negative probe pairs is prone to make MBEI produce negative expression values. In order to reduce the negative expression values MBEI excludes all negative probe pairs or uses only PM probe intensities as an option for the users (Li and Wong, 2001b). The software MBEI is available from http://www.biostat.harvard.edu/complab/dchip.

The robust multi-array average, RMA (Irizarry et al., 2003), uses only PM values to fit a linear model on the log intensity scale. Before fitting the model, the true signal and the background is modeled to follow an exponential and

Gaussian distribution respectively. The true signal is then normalised using the quantile normalisation algorithm (Bolstad et al., 2003). In quantile normalisation, the probes for each array are sorted according to their intensities. The average intensity of probes over different arrays which have the same rank is assigned to these probes, so that the normalised probe intensities for each array have exactly the same density. The probe affinity effect is modeled as an additive component in the linear model at the summarisation step. The modified version of RMA, GCRMA (Wu et al., 2004), assumes the non-specific hybridisation tends to be directly related to it GC-content and removes the non-specific background signal based on a model using GC-content. GCRMA is implemented as an R package *gcrma* and is available from http://www.bio-conductor.org.

Most of these statistical methods only measure the gene expression level and do not provide a level of uncertainty of this measurement. To account for the uncertainty, MAS 5.0 calculates p-values to determine whether the transcript of a particular gene is present in single array analysis or to tell whether the expression level of a gene on one chip changes with respect to the other chip. However, there is no approach to account for this uncertainty in further more complicated downstream analysis, such as detecting differential gene expression with replicates available or clustering of gene expression. MBEI is able to provide a standard error (SE) of the estimated gene expression level (Li and Wong, 2001a), but it is calculated by fixing the probe effects as constants and this thus ignores the variability in the estimation of probe effects. As for the application of the obtained SE, a statistical test is constructed for finding differential gene expression between two single chips and bootstrap resampling is used in hierarchical clustering. No approach is provided to propagate this estimate of uncertainty into more general analyses.

3.2.2 Probabilistic Models

BGX (Bayesian gene expression index)

BGX (Hein et al., 2005) is a probabilistic model derived from a fully Bayesian hierarchical model. It allows for the binding of a fraction of the specific signal to the MM probes. For probe-set g on replicate r at condition c, it assumes the observed intensity of PM probe j, y_{gjcr}, is a sum of specific signal, s_{gjcr}, and background and non-specific signal, h_{gjcr}, and the corresponding observed MM

intensity, m_{gjcr}, is modelled as a fraction, $\phi \in (0,1)$, of the specific hybridization and non-specific hybridization. The first level of BGX is summarized below:

$$\begin{aligned} y_{gjcr} &\sim \mathcal{N}(h_{gjcr} + s_{gjcr}, \tau^2_{cr}) \\ m_{gjcr} &\sim \mathcal{N}(h_{gjcr} + \phi s_{gjcr}, \tau^2_{cr}) \ , \end{aligned} \tag{3.1}$$

where τ^2_{cr} is the chip-specific variance. The fraction of specific hybridization binding to the MM probes is shared for all probes across all chips. The second level of BGX is

$$\begin{aligned} \log(s_{gjcr}+1) &\sim TN(\mu_{gc}, \sigma^2_{gc}) \\ \log(h_{gjcr}+1) &\sim TN(\lambda_{cr}, \eta^2_{cr}) \ , \end{aligned} \tag{3.2}$$

where TN is a truncated Gaussian distribution to account for the fact that $\log(s_{gjcr}+1)$ is always positive. The distribution of the specific signal is probe-set specific. The distribution of the background and non-specific signal is chip-specific. The model is worked out by a computationally intensive Markov chain Monte Carlo (MCMC) method (see Section 2.2.3).

The advantages of BGX are that it uses both PM and MM probes for extracting specific hybridization intensities, while at the same time allowing for gene and probe specific background correction in the non-specific hybridization term. The problems with BGX are mainly in two aspects. The model does not account for the different binding energies of nucleotide pairs (Zhang et al., 2003; Naef and Magnasco, 2002) and the slow MCMC implementation of the model, especially for low expressed genes which need more iterations for the burn-in step in order to obtain accurate estimates. It is not clear which genes should be considered as low expressed, so this makes it difficult to predict the correct burn-in. The software BGX is available from http://www.bgx.org.uk/software.html.

gMOS and mgMOS

The gamma model for oligonucleotide signal, gMOS (Milo et al., 2003), is a probabilistic model assuming an underlying gamma distribution for the PM and MM probe intensities. For a probe-set g in one chip, the model assumes,

$$y_{gj} = s_{gj} + h^1_{gj} \tag{3.3}$$

$$m_{gj} = h_{gj}^2 , \tag{3.4}$$

for the PM and MM intensity of probe j in probe-set g. The specific binding signal for the jth probe pair s_{gj} is assumed to follow a gamma distribution, $s_{gj} \sim \mathrm{Ga}(\alpha_g, b_g)$, where $\mathrm{Ga}(a, b)$ represents the gamma distribution with parameters a and b,

$$p(s_{gj}) = \frac{b_g^{\alpha_g}}{\Gamma(\alpha_g)} s_{gj}^{\alpha_g - 1} \exp(-b_g s_{gj}) , \tag{3.5}$$

where $\Gamma(\cdot)$ is the gamma function. The background and non-specific signal h_{gj}^i for both the PM and MM intensity are different random variables but assumed to be drawn from the same gamma distribution, $h_{gj}^i \sim \mathrm{Ga}(a_g, b_g)$, with the same inverse scale parameter b_g as s_{gj}. If gamma distributed random variables $X_1, X_2, \cdots, X_n$ have parameters $(\alpha_1, \theta), (\alpha_2, \theta), \cdots, (\alpha_n, \theta)$, then $\sum_i X_i$ is distributed as gamma with parameters $(\sum_i \alpha_i, \theta)$. Since the intensity of the jth PM probe is the sum of s_{gj} and h_{gj}^1, it also follows a gamma distribution. Therefore, gMOS can be described as

$$\begin{aligned} y_{gj} &\sim \mathrm{Ga}(a_g + \alpha_g, b_g) \\ m_{gj} &\sim \mathrm{Ga}(a_g, b_g) . \end{aligned} \tag{3.6}$$

The parameter a_g accounts for the background and non-specific signal and α_g accounts for the specific hybridisation.

In gMOS, the PM and MM probe intensities are independently sampled from two gamma distributions. Milo et al. (2004) improve the original gMOS to the modified gMOS (mgMOS) by modelling the correlation between PM and MM intensities within a probe-set. mgMOS assumes PM and MM intensities are drawn from a joint probability density

$$P(y_{gj}, m_{gj}) = \int \mathrm{d}b_{gj}\, p(b_{gj}) P(y_{gj}, m_{gj} | a_g, \alpha_g, b_{gj}) , \tag{3.7}$$

where $b_{gj} \sim \mathrm{Ga}(c_g, d_g)$. The b_{gj} are latent variables reflecting the different binding affinity of probes within the probe-set. This modified distribution accurately captures the correlated changes in the binding affinity of probe-pairs within the probe-set due to the similar content of the PM and MM probe sequences within the probe-pair.

mgMOS has been shown to be an efficient and accurate model (Milo et al.,

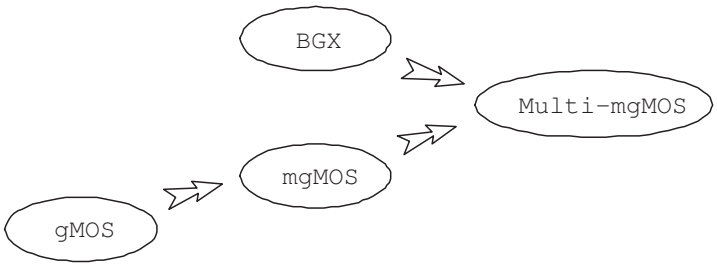

Figure 3.5: Relationships between probe-level probabilistic models.

2004). However, two significant problems remain with the model:

1. The existing model is a single chip model and does not account for the fact that the latent variables b_{gj} are modelling the same information for every chip in the data set. The estimated b_{gj} for the same probe-pair on the same type of chip may differ and cannot reflect an intrinsic characteristic of probe sequences.

2. The intensities of MM probes are taken as background and non-specific hybridisation, so they cannot account for presence of true signal in the MM probes. This causes improper estimates of the non-specific binding from MM intensities.

3.3 Multi-mgMOS

3.3.1 Model

To overcome the limitations of BGX and mgMOS a multiple chip model, multi-mgMOS, is proposed. Relationships between multi-mgMOS and its previous models are shown in Figure 3.5. multi-mgMOS assumes

$$\begin{aligned} y_{gjc} &= s^1_{gjc} + h^1_{gjc} \\ m_{gjc} &= \phi s^2_{gjc} + h^2_{gjc} \ , \end{aligned} \tag{3.8}$$

where y_{gjc} and m_{gjc} represent respectively the PM and MM intensities of the jth probe-pair in the gth probe-set on the cth chip. s_{gjc} and h_{gjc} are the corresponding specific binding signal and non-specific background signal. Similar to the approach in BGX, a fraction of the true signal, ϕ, is allowed to bind to MM probes.

The true signal s_{gjc} and the non-specific signal h_{gjc} are assumed to follow gamma distributions

$$s^i_{gjc} \sim \text{Ga}(\alpha_{gc}, b_{gj}) \tag{3.9}$$
$$h^i_{gjc} \sim \text{Ga}(a_{gc}, b_{gj}) \ . \tag{3.10}$$

Both s_{gjc} and h_{gjc} have the same inverse scale parameter b_{gj} which is a latent variable. The parameter b_{gj} is also assumed to follow a gamma distribution, $b_{gj} \sim \text{Ga}(c_g, d_g)$, with parameters c_g and d_g which are both probe-set specific. This scale parameter is shared across chips for each probe-pair and therefore captures the sequence-dependent nature of the binding affinity. Parameters α_{gc} and a_{gc} are related to the amount of specific and non-specific signal binding to the probe-set g of chip c. It is possible to share the parameter α_{gc} across replicates as in BGX (Hein et al., 2005). However, in most cases here a different α parameter for each chip is used since this is more robust to outliers and between-chip experimental variation.

For $\phi = 0$, y_{gjc} and m_{gjc} are simple combinations of gamma distributed variables representing signal and noise. For $\phi > 0$, the distribution of ϕs_{gjc} is

$$\phi s_{gjc} \sim \text{Ga}(\alpha_{gc}, b_{gj}/\phi) \ . \tag{3.11}$$

If the distribution of ϕs_{gjc} is the gamma in (3.11), there is no standard distribution for m_{gjc} in (3.8). An approximated distribution of ϕs_{gjc} is then adopted,

$$\phi s_{gjc} \sim \text{Ga}(\phi\alpha_{gc}, b_{gj}) \ , \tag{3.12}$$

which ensures the same mean effect of true signal binding to each MM probe. The approximation in (3.12) is reasonable and is useful to keep the model tractable.

Using the additive property of gamma distributed random variables, multi-mgMOS can be described as

$$\begin{aligned} y_{gjc} &\sim \text{Ga}(a_{gc} + \alpha_{gc}, b_{gj}) \\ m_{gjc} &\sim \text{Ga}(a_{gc} + \phi\alpha_{gc}, b_{gj}) \ . \end{aligned} \tag{3.13}$$

The shape parameters of the two gamma distributions are different and are comprised of two parts: the background term a_{gc} and the true specific hybridisation

signal term α_{gc} which are probe-set and chip specific. The parameter ϕ is then shared by all probes. In practice this is an approximation as it can be observed that the parameter ϕ varies between probe-pairs. In Section 3.4 this assumption is relaxed and ϕ is allowed to be probe-specific.

3.3.2 Parameter Estimation

The log-likelihood of the observed PM and MM intensities for each probe-set g is

$$\begin{aligned} & \mathcal{L}_g(\boldsymbol{a}_g, \phi, \boldsymbol{\alpha}_g, c_g, d_g) = \log P(\boldsymbol{Y}_g, \boldsymbol{M}_g) \\ = & \sum_j \log \int \mathrm{d}b_{gj} P(b_{gj}|c_g, d_g) \prod_c P(y_{gjc}, m_{gjc}|a_{gc}, \phi, \alpha_{gc}, b_{gj}) \\ = & \sum_j \log \left[\frac{d_g^{c_g}\Gamma(q_g)}{\Gamma(c_g) w_{gj}^{q_g}} \prod_c \frac{y_{gjc}^{a_{gc}+\alpha_{gc}-1} m_{gjc}^{a_{gc}+\phi\alpha_{gc}-1}}{\Gamma(a_{gc}+\alpha_{gc})\Gamma(a_{gc}+\phi\alpha_{gc})} \right] , \end{aligned} \tag{3.14}$$

where $\boldsymbol{\alpha}_g = [\alpha_{gc}]$, $\boldsymbol{a}_g = [a_{gc}]$, $\boldsymbol{Y}_g = [y_{gjc}]$, $\boldsymbol{M}_g = [m_{gjc}]$, $q_g = \sum_c (2a_{gc} + (1+\phi)\alpha_{gc}) + c_g$ and $w_{gj} = \sum_c (y_{gjc} + m_{gjc}) + d_g$. The parameters $a_{gc}, \phi, \alpha_{gc}, c_g$ and d_g can be estimated iteratively using maximum likelihood. Firstly, with fixed ϕ, a_{gc}, α_{gc}, c_g and d_g are fitted for each probe-set, then using the fitted a_{gc}, α_{gc}, c_g and d_g, ϕ is estimated. This process is iterated until all parameters reach stable values.

It is found that the model has a flat likelihood for a range of parameters. In order to make the model parameters uniquely identifiable, the empirical knowledge of ϕ, which can be estimated from spike-in data, is adopted. It is assumed that for highly expressed spike-in genes the background and non-specific hybridisation can be ignored. Therefore, in (3.13) a_{gc} is set to be zero and one finds $\phi \simeq \langle m_{gjc} \rangle / \langle y_{gjc} \rangle$, which means the difference between $\log(MM)$ and $\log(PM)$ for each probe-pair is approximately equal to the constant $\log(\phi)$ shown in Figure 3.6 (a). Using the experimental data from all known spike-in genes whose spiked concentrations are above 50 pM the fitted log-normal distribution for ϕ is obtained and shown in Figure 3.6 (b). A log-normal prior for ϕ is introduced to obtain the maximum a posteriori (MAP) estimate of ϕ. The posterior distribution of ϕ is

$$P(\phi|\{y_{gjc}, m_{gjc}\}) \propto P(\{y_{gjc}, m_{gjc}\}|\phi) P(\phi) \ . \tag{3.15}$$

The logarithm of the posterior probability of ϕ is then (ignoring an irrelevant

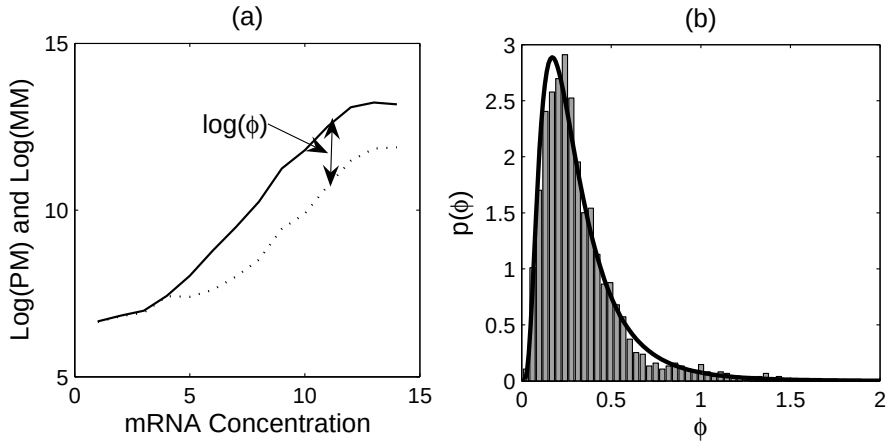

Figure 3.6: (a) The logarithm of the intensity of one probe pair of the spike-in gene 37777_at in Affymetrix Latin Square spike-in data set (described in Section A.1) versus the logarithm of transcription concentrations. The solid and dotted lines represent PM and MM intensities respectively. (b) The histogram and fitted log-normal distribution of ϕ, which measures the fractional amount of specific binding to the MM probe, estimated from the highly expressed spike-in genes.

constant term)

$$h(\phi) = \sum_g \mathcal{L}_g(\phi) + \log(P(\phi)) \ . \tag{3.16}$$

This is the quantity that is maximised to estimate ϕ.

3.3.3 Distribution of Gene Expression Level

Once the parameters have been estimated the distribution of signal s_{gjc} for probe-pair j in probe-set g of chip c is

$$\begin{aligned} P(s_{gjc}|\hat{\alpha}_{gc}, \hat{c}_g, \hat{d}_g) &= \int \mathrm{d}b_{gj} P(s_{gjc}|\hat{\alpha}_{gc}, b_{gj}) P(b_{gj}|\hat{c}_g, \hat{d}_g) \\ &= \frac{\Gamma(\hat{c}_g + \hat{\alpha}_{gc}) \hat{d}_g^{\hat{c}_g} s_{gjc}^{\hat{\alpha}_{gc}-1}}{\Gamma(\hat{\alpha}_{gc})\Gamma(\hat{c}_g)(\hat{d}_g + s_{gjc})^{\hat{c}_g + \hat{\alpha}_{gc}}} \ , \end{aligned} \tag{3.17}$$

where $\hat{\alpha}_{gc}, \hat{c}_g$ and $\hat{d}_g$ are the ML estimates of α_{gc}, c_g and d_g respectively. The expected log true probe signal and the variance of log signal for the gth probe-set are respectively given by,

$$\begin{aligned} \langle \log(s_{gjc}) \rangle &= \log(\hat{d}_g) + \Psi(\hat{\alpha}_{gc}) - \Psi(\hat{c}_g) \ , \\ \mathrm{Var}[\log(s_{gjc})] &= \Psi'(\hat{\alpha}_{gc}) + \Psi'(\hat{c}_g) \ , \end{aligned} \tag{3.18}$$

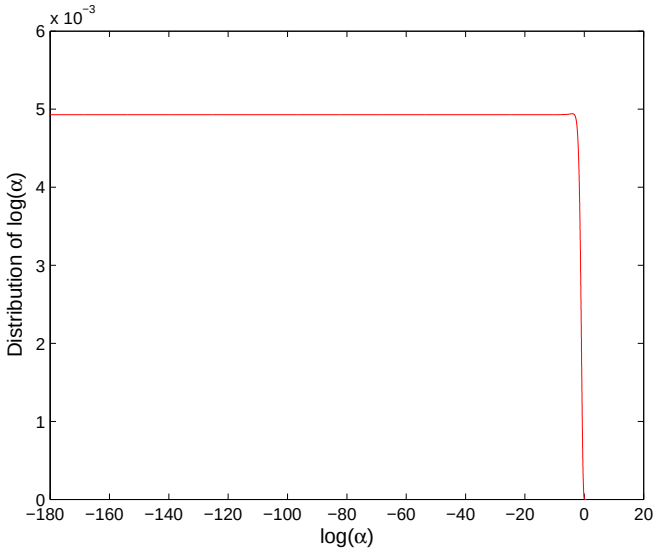

Figure 3.7: Posterior probability density of estimated $\log(\alpha)$ for spike-in probe-set 37777_at in Affymetrix Latin Square spike-in data set (described in Section A.1) at zero concentration. The density is flat and the plot is cut at -180 to aid the clarity.

where Ψ is the digamma function which is the derivative of the logarithm of the gamma function and Ψ' is the first derivative of the digamma function.

In (3.18) c_g and d_g are characteristic of the probe-set for a specific type of chip, while $\langle \log(s_{gjc}) \rangle$ varies with α_{gc} over different chips in the data set. So the posterior distribution of $\log(s_{gjc})$ is of interest given α_{gc} and fixing c_g and d_g at the ML estimates, $P(\log(s_{gjc})|\alpha_{gc}, \hat{c}_g, \hat{d}_g)$.

3.3.4 Approximation to the Posterior Distribution of α_g

The posterior distribution of α_{gc} is found to be unimodal and the value of α_{gc} is always positive, so it was initially thought that it would be reasonable to obtain the Laplace approximation of the posterior distribution of $\log(\alpha_{gc})$. However, for low expressed genes the density of $\log(\alpha_{gc})$ has very low probability mass as shown in Figure 3.7 and the Gaussian approximation is not suitable for this case. A truncated Gaussian is thus used to directly approximate the posterior distribution of α_{gc} and it will be shown that this provides a reasonable approximation.

MAP Approximation

Parameters $\boldsymbol{\alpha}_g$ are assumed to be independent of each other so that the posterior distribution of $\boldsymbol{\alpha}_g$ factorizes as a product of independent distributions for each α_{gc}. At the ML solution of (3.14), $\hat{\alpha}_{gc}$, which is equal to the MAP estimate under a uniform prior on α_{gc}, assume

$$\begin{aligned} \mathcal{L}'_g(\hat{\alpha}_{gc}) &= \left.\frac{d\mathcal{L}_g(\alpha_{gc})}{d\alpha_{gc}}\right|_{\hat{\alpha}_{gc}} \\ \mathcal{L}''_g(\hat{\alpha}_{gc}) &= \left.\frac{d^2\mathcal{L}_g(\alpha_{gc})}{d\alpha_{gc}^2}\right|_{\hat{\alpha}_{gc}} , \end{aligned} \tag{3.19}$$

where $\mathcal{L}_g(\alpha_{gc})$ is the log likelihood function of α_{gc} given other parameters fixed at their modal values. Expanding $P(\alpha_{gc}|D)$ at the ML estimates, the following can then be obtained for $\alpha_{gc} > 0$,

$$\begin{aligned} P(\alpha_{gc}) &\propto \exp\left(\mathcal{L}'_g(\hat{\alpha}_{gc})(\alpha_{gc} - \hat{\alpha}_{gc}) + \frac{1}{2}\mathcal{L}''_g(\hat{\alpha}_{gc})(\alpha_{gc} - \hat{\alpha}_{gc})^2)\right) \\ &= TN(\alpha_{gc}; \mu_{gc}, \sigma^2_{gc}) \ . \end{aligned} \tag{3.20}$$

The gradient term is only non-zero when $\hat{\alpha}_{gc}$ is zero and then the mean μ_{gc} is below zero. From (3.20), the mean and variance of the approximating truncated Gaussian are

$$\mu_{gc} = \left(\mathcal{L}'_g(\hat{\alpha}_{gc})\right)\sigma^2_{gc} + \hat{\alpha}_{gc} \tag{3.21}$$

$$\sigma^2_{gc} = \left(-\mathcal{L}''_g(\hat{\alpha}_{gc})\right)^{-1} . \tag{3.22}$$

Since α_{gc} is positive, the Gaussian in (3.20) left truncated at zero is used to approximate the posterior distribution of α_{gc}. The normalisation constant C for the truncated Guassian is

$$C = \frac{2}{1 - \text{erf}\left(-\frac{\mu_{gc}}{\sigma_{gc}\sqrt{2}}\right)} , \tag{3.23}$$

where $\text{erf}(\cdot)$ is the error function which is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \text{d}t \ \exp(-t^2) \ . \tag{3.24}$$

Laplace Approximation

If the dependence of the components of $\boldsymbol{\alpha}_g$ is considered, a multivariate Gaussian can be used to approximate the posterior distribution of $\boldsymbol{\alpha}_g$,

$$\begin{aligned} P(\boldsymbol{\alpha}_g) &\propto \exp\left(\mathcal{L}'_g(\hat{\boldsymbol{\alpha}}_g)^T(\boldsymbol{\alpha}_g - \hat{\boldsymbol{\alpha}}_g) + \frac{1}{2}(\boldsymbol{\alpha}_g - \hat{\boldsymbol{\alpha}}_g)^T H^g(\boldsymbol{\alpha}_g - \hat{\boldsymbol{\alpha}}_g))\right) \\ &= \mathcal{N}(\boldsymbol{\alpha}_g; \boldsymbol{\mu}_g, \Sigma_g) \ , \end{aligned} \tag{3.25}$$

where the element of the Hessian matrix H^g is

$$H^g_{ij} = \left.\frac{\partial^2 \mathcal{L}_g(\boldsymbol{\alpha}_g)}{\partial \alpha_{gi} \partial \alpha_{gj}}\right|_{\boldsymbol{\alpha}_g = \hat{\boldsymbol{\alpha}}_g} \ . \tag{3.26}$$

The mean of the approximated multivariate Gaussian is calculated by

$$\boldsymbol{\mu}_g = \Sigma_g \mathcal{L}'_g(\hat{\boldsymbol{\alpha}}_g) + \hat{\boldsymbol{\alpha}}_g \ , \tag{3.27}$$

where the covariance matrix Σ_g is calculated from

$$\hat{\Sigma}_g = (-\hat{H}^g)^{-1} \ . \tag{3.28}$$

The components of $\boldsymbol{\alpha}_g$ are constrained to be positive and therefore a truncated multivariate Gaussian should be used. However, it is not possible to find an analytical expression for the truncated multivariate distribution due to the difficulty of normalisation. Instead, the marginal distribution of α_{gc} is obtained by assuming a standard Gaussian and this marginal distribution is normalised. The correlation between chips is only significant for highly expressed genes (see Figure 3.10), where the truncation is irrelevant, so this approximation works well in practice.

MCMC

In order to verify the goodness of the MAP approximation and Laplace approximation, standard Markov chain Monte Carlo (MCMC) can be applied to approximate the true posterior distribution of $\boldsymbol{\alpha}_g$. Given the assumption of the uniform

prior on α_{gc} and fixing other parameters at the modal values, the posterior distribution of $\boldsymbol{\alpha}_g$ is

$$\begin{aligned} P(\boldsymbol{\alpha}_g|D) &\propto \prod_j \int \mathrm{d}b_{gj} P(b_{gj}|c_g, d_g) \prod_c P(y_{gjc}, m_{gjc}|a_{gc}, \phi, \alpha_{gc}, b_{gj}) \\ &\propto \prod_j \left(\frac{\Gamma(q_g)}{w_{gj}^{q_g}} \prod_c \frac{y_{gjc}^{\alpha_{gc}} m_{gjc}^{\phi\alpha_{gc}}}{\Gamma(a_{gc}+\alpha_{gc})\Gamma(a_{gc}+\phi\alpha_{gc})} \right) . \end{aligned} \tag{3.29}$$

There is no standard form for the distribution of α_{gc}, so the Metropolis algorithm is adopted to update α_{gc} iteratively. The proposal distribution is a truncated Gaussian with the mean being the current value of α_{gc} and the variance being proportional to the variance in (3.22) for MAP approximation,

$$\alpha_{gc}^t \sim TN(\alpha_{gc}^{t-1}, k^2 \hat{\sigma}_{gc}) , \tag{3.30}$$

where the scale k is set to 2.4 empirically (Gelman et al., 2004). The initial value of $\boldsymbol{\alpha}_g$ is the ML estimate, $\hat{\boldsymbol{\alpha}}_g$. At each Metropolis-Hasting random walk, if a non-positive value of α_{gc}^t is drawn from the Gaussian $\mathcal{N}(\alpha_{gc}^{t-1}, k^2\hat{\sigma}_{gc})$, it is rejected. Once a positive value is drawn, the acceptance/rejection rule in (2.8) and (2.9) is used.

The posterior distribution of α_{gc} at different concentrations for spike-in gene 37777_at in Affymetrix Latin Square spike-in data set is shown in the upper panel of Figure 3.8. For the Laplace approximation the marginal distribution of α at each concentration is shown. Presumably Laplace and MCMC methods obtain broader densities than MAP approximation since they include the variability of α_{gc} across all chips. For the lower and median concentration (columns 1-3) both the MAP and Laplace approximations are close to simulations from the MCMC method. For high concentration, which is related to highly expressed genes (column 4), the difference between different methods gets larger due to higher correlation between chips. Note that modes of $P(\langle \log(s_{gjc}) \rangle)$ from the two approximated methods are the same, but they are different from MCMC. This is because the two Gaussian approximations of MAP and Laplace to the posterior of α_{gc} have the same modes and the densities are symmetrical, and this leads to the same mode under the transformation of (3.18). However, $P(\alpha_{gc})$ from MCMC is right skewed, so the mode of $P(\langle \log(s_{gjc}) \rangle)$ changes and is different from the two Gaussian approximated methods.

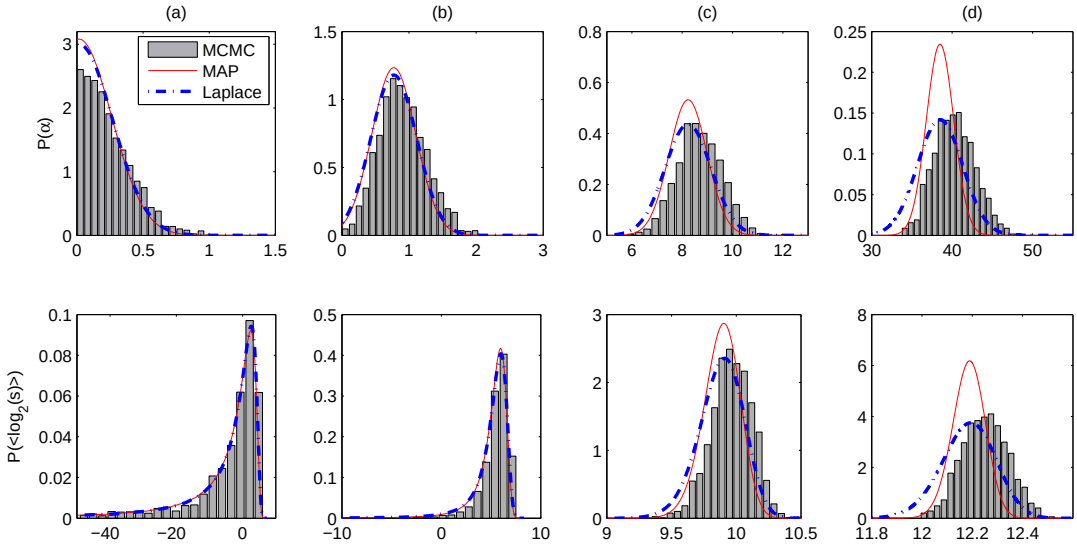

Figure 3.8: Posterior probability density function of estimated α (upper panel) and log expression levels (lower panel) for spike-in gene 37777_at in the Affymetrix Latin Square spike-in data set (described in Section A.1) at concentration (a) 0, (b) 2, (c) 32 and (d) 512 pM. The thin solid lines are MAP approximation in (3.20), the dash-dotted lines are Laplace approximation in (3.25) and the histograms are from the MCMC method.

With the posterior distribution of α_{gc}, it is then straightforward to calculate the percentiles and credibility intervals. The expression in (3.18) and the percentiles of α_{gc} can be used to calculate the percentiles of $\langle \log(s_{gjc}) \rangle$ since these are invariant under the transformation of (3.18). The approximated distribution of measured expression level at various concentrations of the spike-in gene 37777_at in the Affymetrix spike-in data set is shown in the lower panel of Figure 3.8. Similar to the distribution of α, the difference between distributions from the three methods increases with the concentration.

It can be seen from Figure 3.8 that the difference between the two different approximations of the distribution of α_{gc}, MAP and Laplace, is not obvious for the genes at lower and medium concentrations. Figure 3.9 shows the estimated $\langle \log(s) \rangle$ calculated in (3.18) for 25 chips of the mouse data set. It can be seen that the logged expression level for most genes lies below 10, which we refer to as low and medium expressed, so the cases like the fourth column in Figure 3.8 are not common. Moreover, the MAP approximation method assumes the independence of the gene expression level on each chip while the Laplace and MCMC method

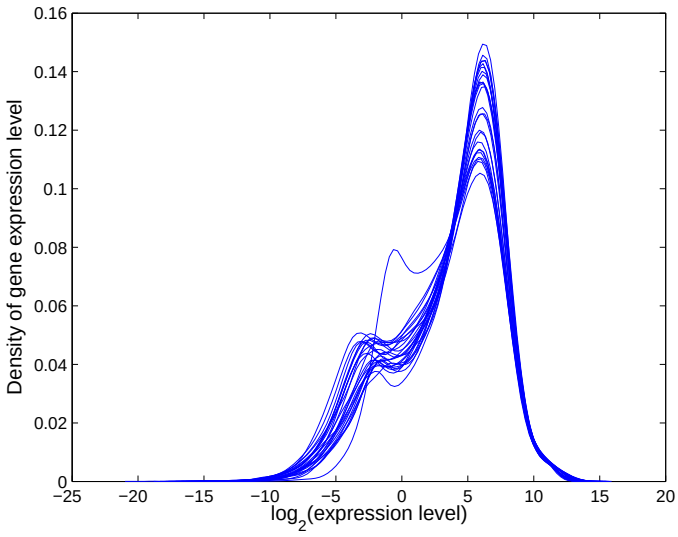

Figure 3.9: Density of estimated expression level from multi-mgMOS for 25 chips in the mouse time-course data set (described in Appendix A.3).

consider the correlation between expression level measured by different chips. In the downstream analysis, if the Laplace method is used to measure the uncertainty of gene expression measurements, the correlation between different chips should be considered in the downstream analysis model and this will introduce extra complexity into the model. In the further downstream analysis, where possible, the correlation between chips is ignored to simplify the model since the difference between the two approximation methods is negligible for most genes.

Under the circumstance where the variability of α_{gc} across chips is ignored, the goodness of the MAP approximation is examined by comparing it with the numerically calculated histogram as shown in Figure 3.10. The dash-dotted lines are calculated from $P(\alpha_{gc}|\{\boldsymbol{Y}_g, \boldsymbol{M}_g\})$ under the uniform prior for α_{gc} when fixing the other parameters at the modal values. The lower panel shows the corresponding density of $\langle \log(s_{gjc}) \rangle$ under the transformation of (3.18). It can be seen that the MAP approximation is very close to the numerically calculated density. Therefore the MAP approximation is used in the rest of the thesis.

3.3.5 Approximation of the Distribution of $\langle \log(s) \rangle$

In the downstream probabilistic analysis of microarray data, such as Bayesian methods, it is useful to provide a Gaussian approximation of the estimated log

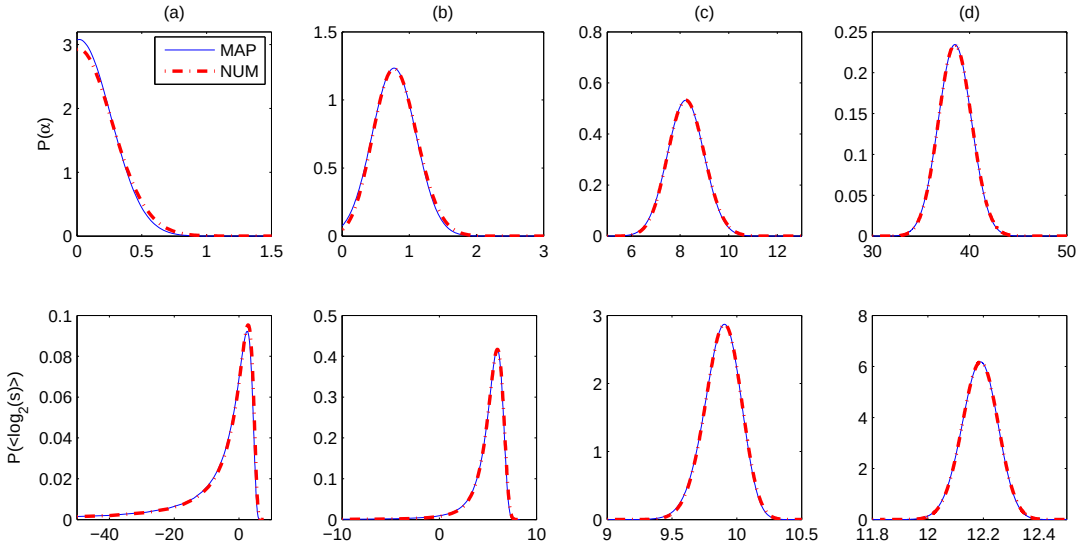

Figure 3.10: Posterior probability density function of estimated α (upper panel) and log expression levels (lower panel) for spike-in gene 37777_at in the Affymetrix Latin Square spike-in data set (described in Section A.1) at concentration (a) 0, (b) 2, (c) 32 and (d) 512 pM. The thin solid lines are MAP approximation in (3.20) and the dash-dotted lines are from numerically calculated histograms.

expression level since it is straightforward to plug a Gaussian noise into other probabilistic models. The indices of genes and conditions are omitted for brevity and the Gaussian approximation is assumed

$$\langle \log(s_j) \rangle \sim \mathcal{N}(\mu_s, \sigma_s^2) \ . \tag{3.31}$$

The delta method (Oehlert, 1992) is used to obtain such a Gaussian distribution. From (3.18) the estimated log expression level, $\langle \log(s_j) \rangle$, is a function of the random variable α,

$$g(\alpha) = \langle \log(s_j) \rangle = \log(d) + \Psi(\alpha) - \Psi(c) \ . \tag{3.32}$$

When ignoring the index of genes and conditions, the distribution of α is the Gaussian with mean μ and variance σ^2 in (3.20) truncated at zero. The mean of the truncated Gaussian, $\hat{\alpha}$, and the variance $\hat{\sigma}^2$ are

$$\hat{\alpha} \ = \ C \left[\frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2} \right) + \frac{\mu}{2} - \frac{\mu}{2} \mathrm{erf}\left(-\frac{\mu}{\sqrt{2}\sigma} \right) \right]$$

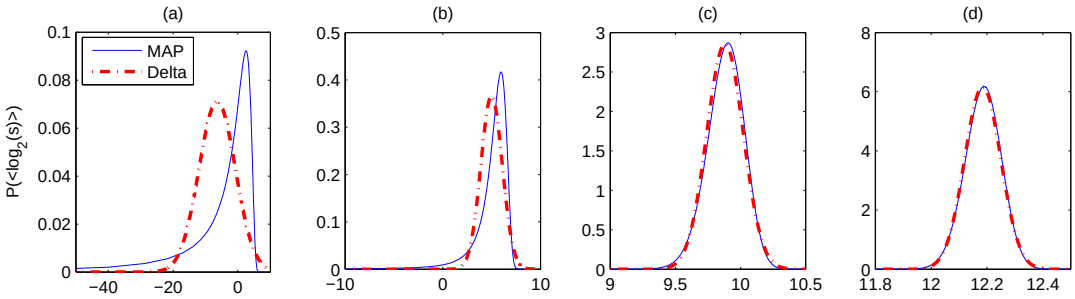

Figure 3.11: Approximated posterior probability density function of estimated log expression levels for spike-in gene 37777_at in Affymetrix Latin Square spike-in data set (described in Section A.1) at concentration (a) 0, (b) 2, (c) 32 and (d) 512 pM. The thin solid lines are from the MAP approximation and the thick dash-dotted lines are from the delta approximation.

$$\hat{\sigma}^2 \;=\; C\left[\frac{1}{2}\left(\sigma^2+(\mu-\hat{\alpha})^2\right)\left(1-\mathrm{erf}\left(-\frac{\mu}{\sqrt{2}\sigma}\right)\right)+\frac{\sigma}{\sqrt{2\pi}}(\mu-2\hat{\alpha})\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right]\;,$$

where C is the normalisation constant defined in (3.23). Using the second order Taylor series expansion of $g(\alpha)$ about $\hat{\alpha}$, the mean of $g(\alpha)$ can be obtained by,

$$\mu_s \approx g(\hat{\alpha}) + \frac{\hat{\sigma}^2}{2}\left.\frac{\partial^2 g(\alpha)}{\partial \alpha^2}\right|_{\alpha=\hat{\alpha}} \;. \tag{3.33}$$

Using the first order Taylor series expansion, the approximated variance of $g(\alpha)$ is

$$\sigma_s^2 \approx \hat{\sigma}^2\left(\left.\frac{\partial g(\alpha)}{\partial \alpha}\right|_{\alpha=\hat{\alpha}}\right)^2 \;. \tag{3.34}$$

Figure 3.11 shows the approximation of gene expression level from the delta method compared to the MAP approximation calculated according to the invariant percentiles of the posterior of α and $\langle \log(s_j)\rangle$ under the transformation of (3.18). For the low expressed genes, the long tail of the distribution to the left is chopped off, and for high expressed genes, the delta method obtains a very good approximation to the MAP result.

3.3.6 Implementation

This model is the original version described in Liu et al. (2005) and implemented in the R package, *mmgmos*, for public use of the model. This package uses the fast C program donlp2 (Spellucci, 1998) for parameter optimisation. Since the

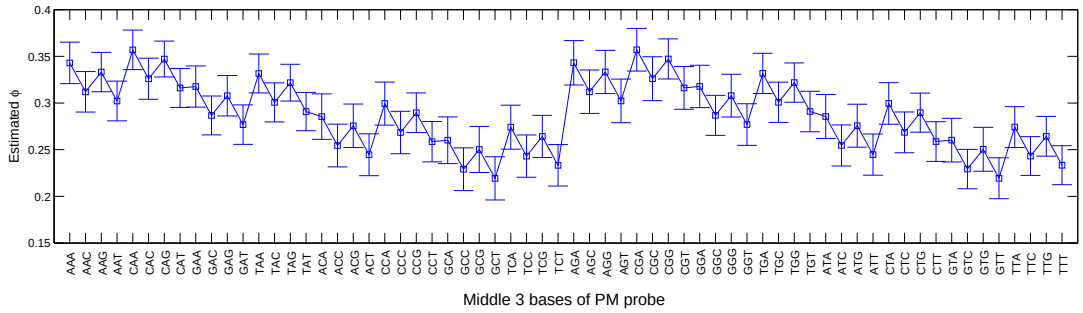

Figure 3.12: Estimated ϕ for each of 64 possible middle triple bases of PM probes from all known spike-in genes which are spiked in at high concentration (above 50 pM) in the data sets in Appendix A.1 and A.2.

publication of Liu et al. (2005), the software has been updated several times. The main updates are setting ϕ zero when there is no empirical distribution available and adding a global scaling normalisation option after the gene expression values have been calculated.

3.4 Possible Improvements of ϕ Estimate

In Section 3.3, the single value of ϕ is shared for all probe-pairs. As shown in Figure 3.6 (b) ϕ varies over probe-pairs. The single value may not be representative of the sensitivity to true signal for different MM probes. In this section, two possible ways are discussed to include more information in order to make ϕ more representative.

3.4.1 Estimating ϕ for Different Probe Content

It has been shown in previous studies that the GC-content of a probe affects its binding efficiency and therefore its associated intensity (Naef and Magnasco, 2002; Zhang et al., 2003). Since the PM and MM probes have a mismatch central base, the intensity difference for a PM probe and its corresponding MM probe is largely due to the central triple bases of a probe. It is thus assumed that ϕ is mostly determined by the middle triple bases. Using probe intensities at high concentration (above 50 pM) of all known spike-in genes in data sets described in Appendices A.1 and A.2, and assuming that the background is negligible at high concentrations, the value of ϕ ($\sim MM/PM$) for probes with the same

middle triple bases is modelled using a generalised linear model (Dobson, 1990). Figure 3.12 shows the calculated ϕ for each of 64 possible middle triple bases of PM probes and associated 95% confidence levels. Probe-pairs with different central bases have different ϕ. For the probe-pairs with central bases A and G, the estimated ϕ is larger than the probe-pairs with middle bases C and T. This shows that there is more true signal binding to MM probes if the corresponding PM probes have A or G as the middle base. The obtained values of ϕ in Figure 3.12 can be directly plugged into the model of multi-mgMOS in (3.13) according to the different central triple bases of each PM probe.

Zhang et al. (2003) find similar results by considering the average of log ratio of PM and MM intensities and explain this phenomenon by the different stacking energy related to the central three bases in a probe-pair. The result here is consistent with their results. The results in Naef and Magnasco (2002) also show that the MM intensities of probe-pairs with A or G as PM middle base are in general higher than the probe-pairs with C or T as middle base of the PM probe. The explanation in Naef and Magnasco (2002) is that in the experiment base C and T carry the fluorescent labels and these labels interfere with binding of the target to the probes causing the various brightness difference between PM and MM probes. The findings in Naef and Magnasco (2002) are also consistent with the values of ϕ shown in Figure 3.12.

3.4.2 Integrating the Histogram of ϕ into multi-mgMOS

Suppose the fraction of true signal binding to MM probe of probe-pair j is ϕ_j and the true value of ϕ_j should vary across different probe-pairs. Instead of treating ϕ_j as a single value, the empirical distribution of ϕ_j, $P(\phi_j)$ in Figure 3.6 (b) can be introduced into the model as the prior and approximated by a histogram. Assuming ϕ_{jk} is the central point of the kth bin of the histogram of $P(\phi_j)$ and $n(\phi_{jk})$ is the area of the kth bin, the joint distribution of probe-pair j is

$$\begin{aligned}
& P(y_{gjc}, m_{gjc}|a_{gc}, \alpha_{gc}, c_g, d_g) \\
= & \int \mathrm{d}\phi P(\phi_j) \int \mathrm{d}b_{gj} P(b_{gj}|c_g, d_g) \prod_c P(y_{gjc}, m_{gjc}|a_{gc}, \phi_j, \alpha_{gc}, b_{gj}) \\
\approx & \sum_{k=1} n(\phi_{jk}) \int \mathrm{d}b_{gj} P(b_{gj}|c_g, d_g) \prod_c P(y_{gjc}, m_{gjc}|a_{gc}, \phi_{jk}, \alpha_{gc}, b_{gj}) \\
= & \sum_k \frac{n(\phi_{jk}) d_g^{c_g} \Gamma(q_{gk})}{\Gamma(c_g) w_{gj}^{q_{gk}}} \prod_c \frac{y_{gjc}^{a_{gc}+\alpha_{gc}-1} m_{gjc}^{a_{gc}+\phi_{jk}\alpha_{gc}-1}}{\Gamma(a_{gc}+\alpha_{gc})\Gamma(a_{gc}+\phi_{jk}\alpha_{gc})} ,
\end{aligned} \tag{3.35}$$

where $q_{gk} = \sum_c (2a_{gc} + (1 + \phi_{jk})\alpha_{gc}) + c_g$ and $w_{gj} = \sum_c (y_{gjc} + m_{gjc}) + d_g$. This approach also considers the variability of ϕ_j, but the sum in the joint distribution in (3.35) introduces additional computational complexity into the model.

3.5 Results and Discussion

In this thesis, multi-mgMOS I is used to denote the model of multi-mgMOS in (3.13) which optimises ϕ from the data, multi-mgMOS II to represent the model using pre-estimated ϕ according to the middle triple bases in PM probes and multi-mgMOS III the variant integrating the histogram of ϕ in (3.35). During the developmental stage, these models are implemented in Matlab using the optimisation toolbox SNOPT (Gill et al., 2002).

3.5.1 Performance on Spike-in Data Sets

Data sets

The simplified GeneLogic spike-in data set (Appendix A.2), denoted as Data set A, is used to compare the three variants of multi-mgMOS with other alternative probabilistic models BGX (Hein et al., 2005) and mgMOS (Milo et al., 2004). The performance of other popular statistical methods on data set A is shown in Hein et al. (2005). Data set A is the same one used by Hein et al. (2005) to evaluate BGX. It is a subset of the GeneLogic spike-in data set and includes only 1011 probe-sets for each chip. Data set A includes six chips for conditions a and k and each condition has three replicates. According to differences between spiked in concentrations in the two conditions, ranks of the difference between expression levels for the 11 spike-in genes are shown in Table 3.1.

In order to further demonstrate the accuracy of the multi-mgMOS variants, they are also compared with the most popular statistical models, MAS 5.0 (Affymetrix, 2002), MBEI (Li and Wong, 2001a) and GCRMA (Wu et al., 2004) using a larger data set B. Data set B is a subset of the Affymetrix Latin Square spike-in data set (Appendix A.1). It includes 14 chips for the 14 conditions (conditions a–m and q) out of replicate group 1521. This data set includes all of the 12,626 probe-sets on each chip.

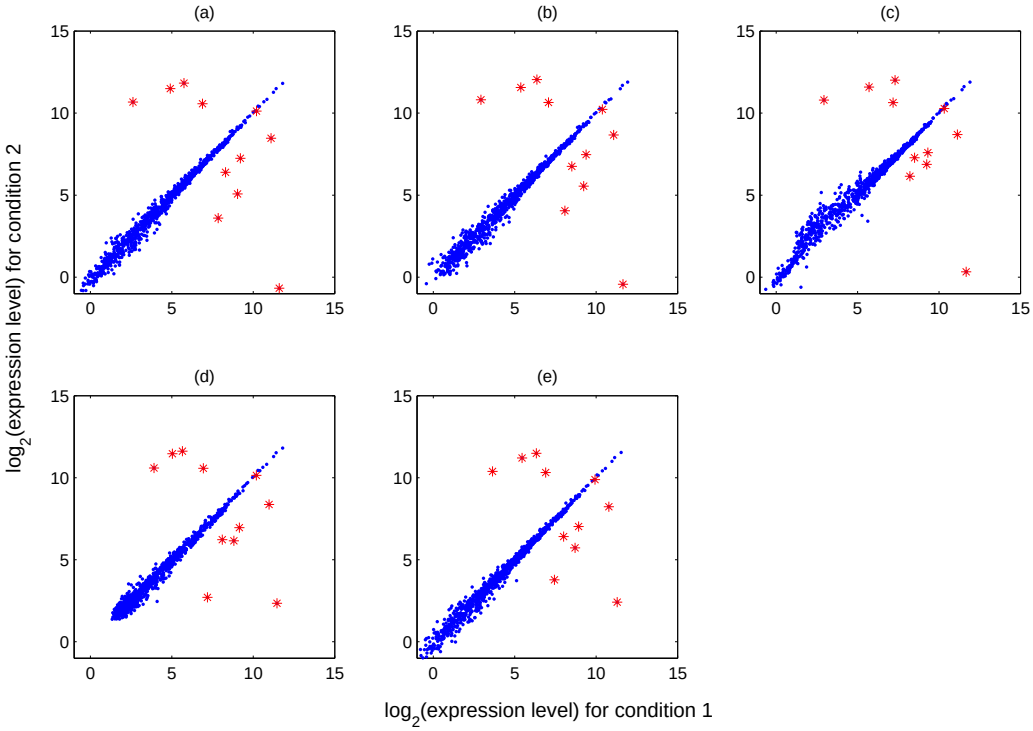

Figure 3.13: Scatter plots of gene expression measures of the two conditions in data set A from (a) multi-mgMOS I, (b) multi-mgMOS II, (c)multi-mgMOS III, (d) BGX and (e) mgMOS. For mgMOS and multi-mgMOS, the mean estimated gene expression levels for each condition over the three replicates are used.

Spike-in gene	2	1	3	5	6	9	7	8	4	10	11
True rank	1	2	3/4	3/4	5	6	7	8	9	10	11
BGX	1	2	3	6	4	8	9	7	862	10	5
mgMOS	1	2	3	6	4	8	9	7	756	10	5
multi-mgMOS I	1	2	3	7	4	8	9	6	737	10	5
multi-mgMOS II	1	2	3	7	4	8	9	6	479	10	5
multi-mgMOS III	2	1	3	5	4	6	13	8	870	43	11

Table 3.1: The ranks of the 11 spike-in genes in data set A, with respect to the degree of differences between expression levels under two conditions, obtained with the different probabilistic methods. All models rank 10 of 11 spike-in genes in the top 10 except multi-mgMOS III.

Comparison of BGX with mgMOS and multi-mgMOS

Figure 3.13 shows scatter plots of gene expression values from BGX, mgMOS and multi-mgMOS I–III for data set A. The dotted points represent the non-spike-in genes and the star points represent the spike-in genes. Since the expression levels of non-spike-in genes in the two samples is identical, the dotted points should follow the diagonal. From Figure 3.13 multi-mgMOS III is obviously worse than the other four models. For non-spike-in genes, the correlation coefficient between gene expression levels for the two conditions estimated with BGX, mgMOS and multi-mgMOS I–III are 0.9919, 0.9926, 0.9934, 0.9916 and 0.9855 respectively, demonstrating that multi-mgMOS I is most consistent for these genes. The spike-in genes are spiked in at different concentrations under the two conditions, so the star points should be away from the diagonal. Except for one spike-in gene which is spiked-in inappropriately in the experiment, all the other 10 spike-in genes lie away from the diagonal. Table 3.1 shows results for the estimated ranks of 11 spike-in genes from the three models. All the models rank 10 of 11 spike-in genes in the top 10 and show similar performance except multi-mgMOS III which failed for spike-in genes 7 and 10, although multi-mgMOS I and II seem to show slightly worse performance in identifying the rank of spike-in genes 5 and 8. The results on this small data set are rather inconclusive and it is difficult to distinguish between the three methods. The difference in performance of these models is investigated on the larger data set B below.

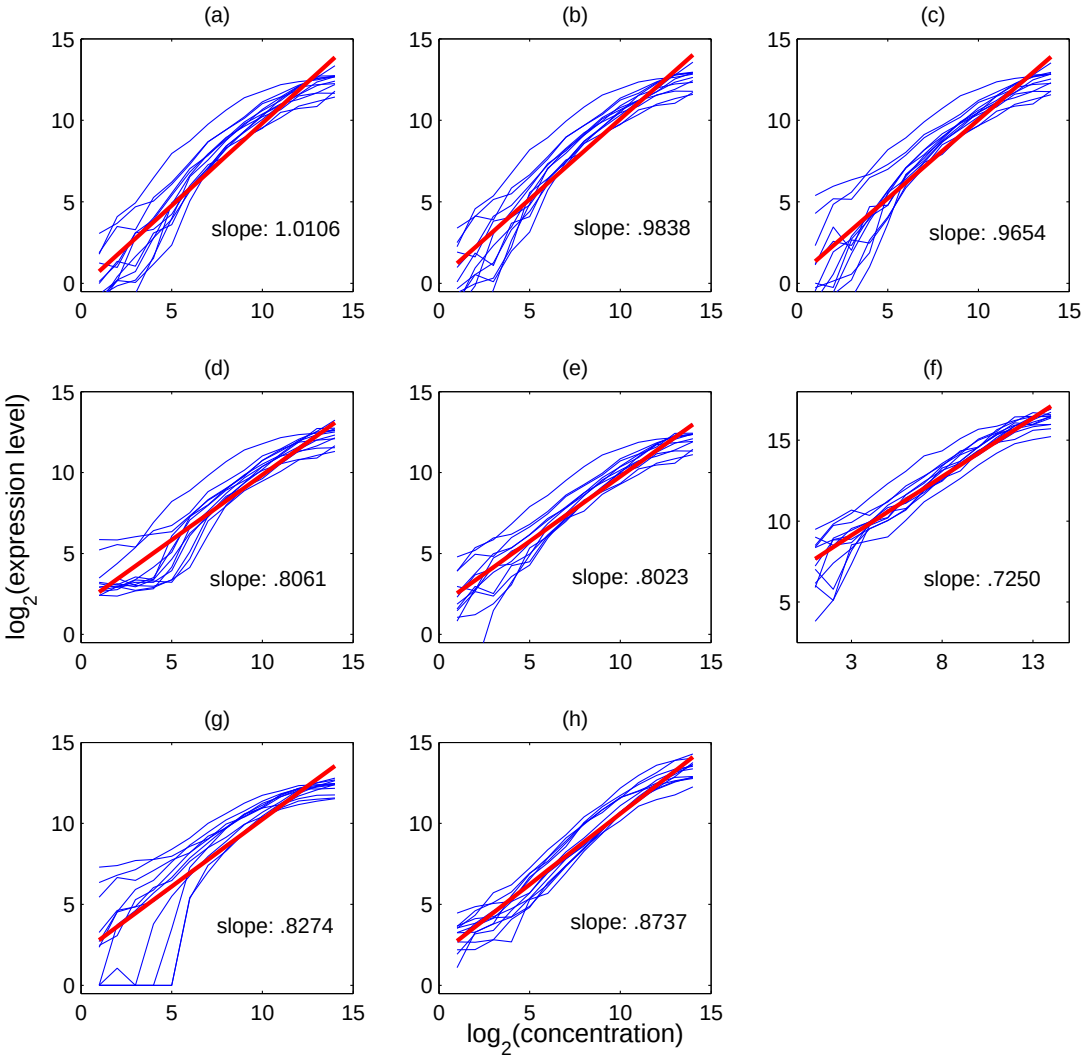

Figure 3.14: Curves of logarithm of gene expression values obtained from different methods for the 12 spike-in genes in the data set B against the log transformation of transcription concentrations: (a) multi-mgMOS I, (b) multi-mgMOS II, (c) multi-mgMOS III, (d) BGX, (e) mgMOS, (f) MAS 5.0, (g) MBEI and (h) GCRMA. The slope of the average fitted straight line is added at each plot. The ideal slope of curves is one.

Sensitivity to Variation in Concentration

Figure 3.14 shows curves of log expression values for 12 spike-in genes in data set B from eight methods against log transformed concentrations which are scaled to (1,14). Following instructions of Affymetrix, two spike-in genes, 407_at and 36889_at, in data set B are excluded due to the poor performance of certain probe-pairs. For variants of multi-mgMOS and mgMOS the negative log expression levels are truncated at -0.5 and negative values obtained from other methods are truncated at zero. Ideally curves for spike-in genes should have a slope of one since the difference in the concentration should result in an identical difference in measured expression level. The slope of the average fitted line for each method is shown in each plot in Figure 3.14. For highly expressed spike-in genes, all methods obtain similar results, but for low expressed spike-in genes, the slopes of curves of multi-mgMOS I and II, mgMOS and GCRMA are closest to one, showing high sensitivity to the variation in concentrations. For two genes multi-mgMOS III obtains relatively high expression values at the lower expressed end. BGX estimates the non-specific hybridisation signal equally over the whole chip without taking probe effects into consideration. This seems unreasonable in practice since PM and MM share very similar oligo sequences. Consequently, BGX has poor performance in the low expressed area where non-specific binding has the strongest effect. The inability of mgMOS to share probe-specific effects across chips results in reduced accuracy for some genes. This can be observed from some spike-in genes at low concentrations where the expression measurements seem higher than true concentrations. GCRMA uses the GC content of probes in order to obtain improvement for the lower end, but the slope is still slightly less than one. The same problem exists for other popular statistical methods which get relatively large expression measures for those weakly expressed genes.

From the comparison above, multi-mgMOS III does not perform as well as multi-mgMOS I and II and it especially obtains spurious results on data set A. It is therefore not considered in the remainder of the thesis.

3.5.2 Performance on a Real Data set

For more practical assessment of the new probe-level analysis method, the real mouse time-course data set which is PCR validated (see Appendix A.3) is used to show the performance of the proposed models, multi-mgMOS I,II. For the mouse

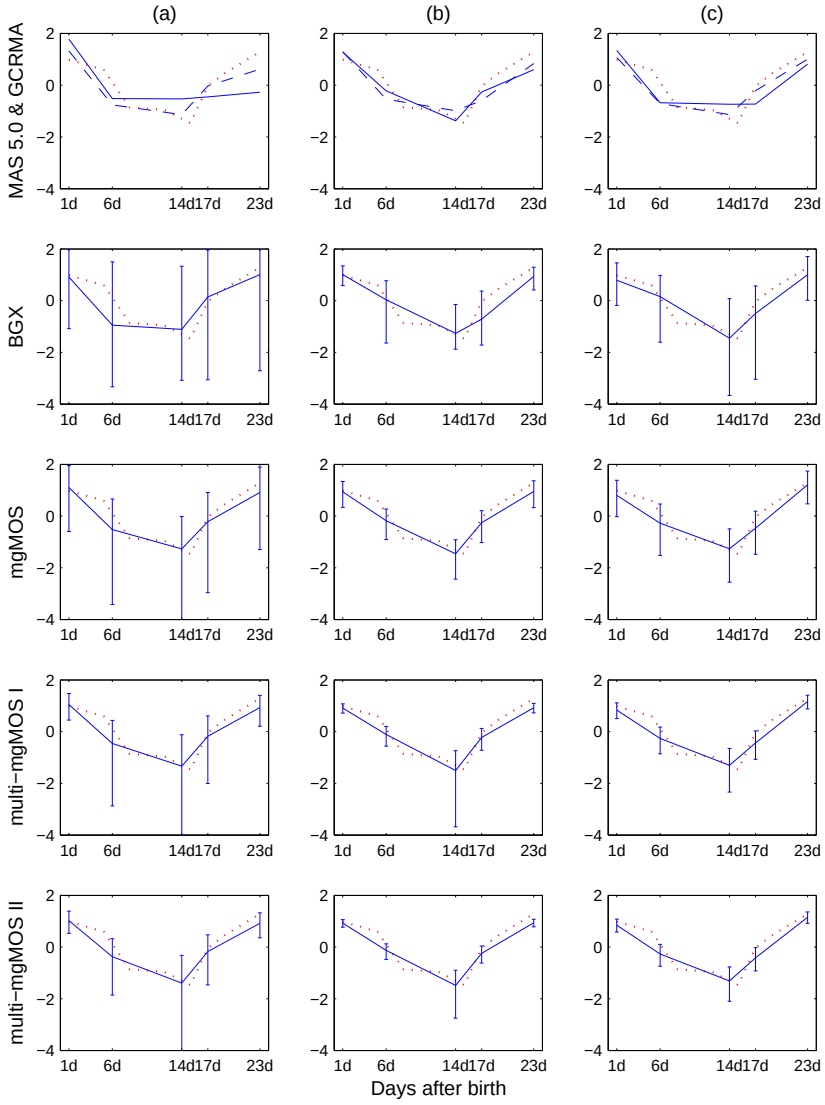

Figure 3.15: Temporal profile of gene Dab2 which contains three probe-sets in column (a), (b) and (c) from mouse data set using five models: MAS 5.0, GCRMA, BGX, gMOS and multi-mgMOS I,II. The qr-PCR profile is the dotted line in each plot. The first row shows the results from MAS 5.0 (dashed lines) and GCRMA (solid lines). The second, third and fourth rows show the profile from BGX, mgMOS and multi-mgMOS respectively. The 5-95% credibility intervals are also plotted for each time point for the probabilistic models BGX, mgMOS and multi-mgMOS. The credibility intervals are truncated at 2.0 and -4.0 in order to make plots clear.

data set, the profiles of the eight PCR confirmed genes are obtained from MAS 5.0, GCRMA, BGX, mgMOS and multi-mgMOS I,II. As an example, Figure 3.15 shows results from the six models for three probe-sets measuring the expression of gene Dab2 in the first hair-growth synchronous cycle. The data from one randomly selected replicate is shown. In order to show the profile pattern for each probe-set, the measured expression level is normalised to have zero mean and unit standard deviation across the five time points. For the first probe-set, GCRMA does not correctly identify the anti-hair-growth pattern as shown in the leftmost plot of the first row. MAS 5.0, BGX, mgMOS and multi-mgMOS I,II obtain more reasonable cycle patterns for all three probe-sets. However, BGX is less confident in capturing the cycle pattern for the first probe-set due to the large credibility intervals for expression levels at day 17 and 23.

Table 3.2 shows the Root Mean Square Error (RMSE) of estimated profiles for eight hair-cycle associated genes. The RMSE to qr-PCR data is shown in the top row of Table 3.2 and measures the difference between the estimated profiles and the corresponding qr-PCR results for three common time points (day 1, 17 and 23) in the first hair-growth cycle for all three mice. This is computed using all probe-sets for the eight genes. It is found that mgMOS and multi-mgMOS I obtain the best values of RMSE and the reason for this becomes apparent when looking in detail at the profiles from each probe-set. Hair-growth patterns obtained from mgMOS and multi-mgMOS are consistent with the qr-PCR results for all eight genes. Profiles from BGX are significantly different from the corresponding qr-PCR data for two probe-sets associated with genes Elf5 and Wnt11. Profiles from MAS 5.0 are inconsistent with qr-PCR profiles for two probe-sets related to gene Fbln1. Profiles from GCRMA are inconsistent for one and two probe-sets respectively from genes Dab2 and Fbln1. multi-mgMOS II obtains very similar profiles to multi-mgMOS I.

There are three PCR confirmed genes (Junb, Dab2 and Fbln1) that have multiple probe-sets, and they have 2, 3 and 4 probe-sets respectively. Profiles of these nine probe-sets for three mice are used to calculate the RMSE to the same gene probe-set (bottom row of Table 3.2) and this shows that multi-mgMOS I identifies the most consistent quantities for probe-sets associated with the same gene.

It was found that the performance of the new method was most impressive on this real data set and it is believed that it is important to validate new methods

RMSE	I	II	mgMOS	BGX	MAS 5.0	GCRMA
To qr-PCR	0.601	0.605	0.601	0.721	0.656	0.694
To same gene	0.233	0.237	0.245	0.274	0.360	0.370

Table 3.2: The Root Mean Square Error (RMSE) of profiles from multi-mgMOS I,II, mgMOS, MAS 5.0 and GCRMA for hair-growth associated genes in the mouse data set. The first row is the RMSE to profiles from qr-PCR data and the second row is the RMSE between probe-sets measuring the same gene. I, II are the abbreviation of multi-mgMOS I and II.

for normalisation and probe-level analysis on real experimental data as well as on spike-in data. Most spike-in data have the unrealistic property that almost all genes have identical expression levels in different experiments. This property may be better suited to some methods than others. For example, the quantile normalisation (Bolstad et al., 2003) used in RMA and GCRMA works under the assumption that gene expression levels have the same distribution in different experiments. This assumption is especially well-suited to the analysis of artificial spike-in data sets in which the distribution of expression levels between experiments is almost identical. It is unclear how well this assumption holds in general.

3.5.3 Model Selection

mgMOS, multi-mgMOS I and multi-mgMOS II have quite similar performance in terms of accuracy, however there are often quite substantial differences in terms of the inferred posterior signal distribution and corresponding error bars. Therefore one would like to determine which model has the most statistical support by using standard model selection methods (see Section 2.2.4). Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be computed to select the most appropriate model. However, sharing b_{gj} across chips makes the dimension of data points large and leads to the BIC score providing an inaccurate approximation to the model evidence. The inverse of the Hessian matrix in the approximated evidence in (2.35) is also difficult to compute due to its huge dimension in multi-mgMOS I. The formula of AIC in (2.30) is therefore used to select between the models of mgMOS and multi-mgMOS I. Because multi-mgMOS II uses the extra spike-in data to pre-estimate ϕ, it is not comparable with mgMOS and multi-mgMOS I. The values of AIC are divided by the number

Data set	mgMOS	multi-mgMOS I
A	2184	2092
B	5130	4488
Mouse	9790	8865

Table 3.3: Results of AIC model selection criteria per gene for mgMOS and multi-mgMOS I on the three data sets.

of genes in each data set in order to avoid producing very large numbers as shown in Table 3.3. For the two models, mgMOS and multi-mgMOS I, the number of free parameters, d_m, is

$$\begin{aligned} d_{\text{mgMOS}} &= 4 \times \text{no of genes} \times \text{no of chips} \\ d_{\text{multi\_mgMOS I}} &= \text{no of genes} \times (2 \times \text{no of chips} + 2) + 1 \ . \end{aligned} \quad (3.36)$$

According to the results of AIC, multi-mgMOS I provides a better explanation of these three data sets. This may explain why, in Figure 3.15, multi-mgMOS I typically obtains more confident results when compared to mgMOS. The selection between multi-mgMOS I and II is discussed in Section 3.6.

3.5.4 Computational Efficiency

A major advantage of multi-mgMOS over BGX is that the likelihood can be written in closed form and with the use of an efficient optimisation package SNOPT (Gill et al., 2002) the parameters can be obtained much faster than BGX. The different computation times for BGX, mgMOS and multi-mgMOS I and II for the three data sets used in this study are shown in Table 3.4. As a multiple chip model, multi-mgMOS is expected to perform better on relatively large data sets and its computational efficiency makes it applicable in practice. multi-mgMOS II is faster than the original version multi-mgMOS I because it does not require the ϕ parameter to be estimated.

3.5.5 Credibility of Expression Measures

In order to demonstrate the advantages of the probabilistic approach, some probabilistic quantities of interest in the results, such as the credibility in both the expression measures and the signal log-ratios, are shown in this section.

Model	Prog. language	Data set A	Data set B	Mouse Data set
BGX	C++	70 mins	32.5 hours	70.5 hours
mgMOS	Matlab & Fortran	12 mins	7 hours	12.5 hours
I	Matlab & Fortran	4 mins	80 mins	5 hours
II	Matlab & Fortran	2 mins	40 mins	50 mins

Table 3.4: The computation time of BGX, mgMOS and multi-mgMOS I and II on different data sets. Computation time is obtained on a 1.8GHz AMD Opteron machine. BGX used 32,768 sweeps after a burn-in of 8,192 sweeps as suggested in Hein et al. (2005).

One spike-in gene 37777_at from data set B is randomly selected and the probability density function of estimated expression levels at concentrations 0, 8 and 512 pM respectively is shown in the lower panel of Figure 3.8. The upper panel is the related posterior distribution of α. As the concentrations increase, the most likely expression level increases and the variance of the expression measurement decreases to obtain more and more confidence in the estimated expression levels. Figure 3.16 shows 2.5-97.5% credibility intervals of expression levels for 12 spike-in genes in data set B from multi-mgMOS I. As concentration increases, more confidence with the expression estimates is obtained.

A key use of microarrays is identifying differentially expressed genes from different experimental samples. With the proposed model it is easy to carry out this task. For data set A with technical replicates whose sample comes from the same fragmented complex cRNA, it is assumed that the variation between chips is low enough to share α across replicates (methods where α is not shared will be discussed in Chapter 4). Using the delta method (see Section 3.3.5) to approximate the posterior distribution of $\langle \log(s_{gjc}) \rangle$, the distribution of the difference between the expression levels for each gene in data set A is obtained. Figure 3.17 shows the median and 5-95% credibility of the differences between the estimated expression levels for all genes (a) and 52 genes (b) under two conditions from data set A, using multi-mgMOS I. For the spike-in genes, except gene 1004 which is not spiked in properly, the credibility intervals do not include zero which means the spike-in genes are differentially expressed with high probabilities. The credibility intervals of all but five non-spike-in genes include zero. With increased credibility intervals, the error bars of the false positives under 5-95% credibility intervals embrace zero gradually while the true positives remain significant. Notice that in some cases the credibility intervals are very large. This is because the log-ratio

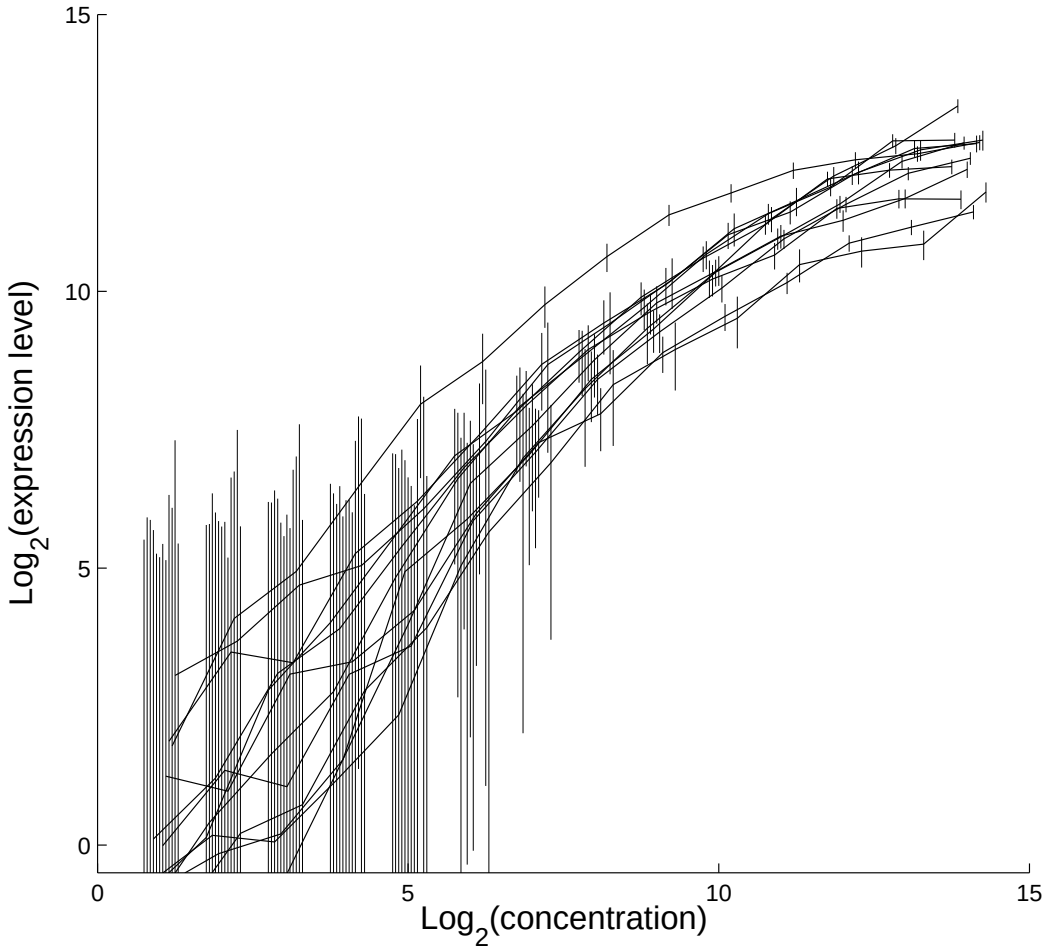

Figure 3.16: 2.5-97.5% credibility intervals of expression levels for 12 spike-in genes in data set B from multi-mgMOS I. The expression levels and the credibility intervals are truncated at -0.5 to aid clarity. As an aid to the eye, the thin grey bars used to illustrate the credibility intervals have been slightly displaced along the horizontal axis.

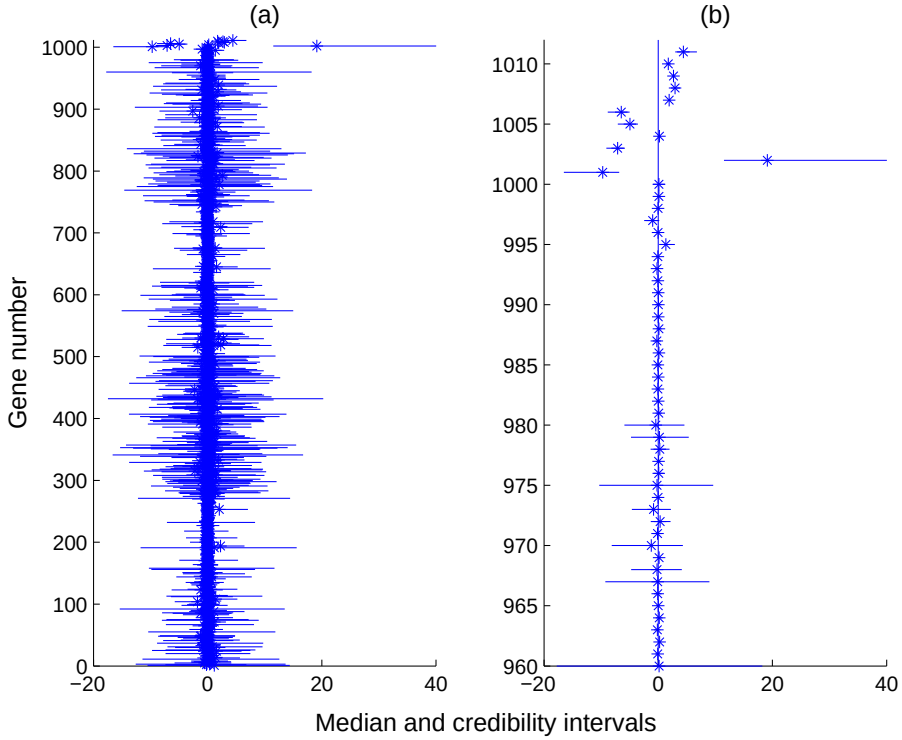

Figure 3.17: Median and 5-95% credibility intervals of the log-ratio between expression levels of data set A under two conditions for (a) all genes and (b) genes from 960 to 1011 from multi-mgMOS I. Median values are represented by star points and credibility intervals are represented by horizontal solid lines.

of signal is essentially unidentifiable for very low expressed genes. However, this does not present a problem for the method since a significant log-ratio is inferred when the credibility intervals do not include zero. Applications to identifying differential expression will be discussed in greater detail in Chapter 4.

3.5.6 Results on Affycomp

The multi-mgMOS I method was submitted to the benchmark, Affycomp, described by Cope et al. (2004) on 15 June, 2005. This is a popular benchmark for evaluating probe-level analysis methods. At the time of submission the method ranked top of all methods in 4 out of 15 criteria in the original assessment (Table B.1 and B.2) and top in 3 out of 14 criteria in the two newer assessments (Table B.3–B.6).

The model showed excellent sensitivity to variation in concentration which is consistent with the results in Section 3.5.1. The evaluation methods in this benchmark use a simple fold-change rule to identify differentially expressed genes and do not consider approaches that make use of credibility intervals, such as those proposed in Section 3.5.5. This leads to an underestimate of the area under ROC curve for probabilistic models. A median posterior estimate of $\langle \log(s) \rangle$ is used as the point estimate of log concentration and this estimator has a low bias. However, for weak signals this estimator is naturally associated with a large variance and this leads to some large point estimates of fold-change for weakly expressed genes. These are considered "false positives" by Cope et al. (2004) but the method as described in Section 3.5.5 would reject most of these large fold-changes as insignificant by taking their associated credibility intervals into account. The point estimate of signal must be combined with a credibility interval in order to get sensible results for weakly expressed genes. One could reduce the typical size of these large fold-changes by, for example, using $\langle \log(s + c) \rangle$ for some positive constant c as a signal estimate. This would reduce variance but at the cost of increased bias and reduced accuracy in measuring concentration. Therefore, criteria that are solely based on point estimates of fold-change are inappropriate for the evaluation of probabilistic methods.

In Chapter 4 the probability of positive log-ratio is used to rank genes and this is a more sensible ranking for probabilistic methods. However, the current version of Affycomp (Irizarry et al., 2006) only allows submitters to provide point estimates of expression level.

3.6 Conclusion

A new probabilistic model was presented for probe-level analysis of Affymetrix microarray data. The model showed competitive performance compared with other models on commonly used benchmarks and shows very impressive results on a real time-course data set. The likelihood function can be written in closed form and the computation is therefore very fast which allows the potential application to large data sets. Probe effects are shared across all chips of the same type and this improves the accuracy as well as the model's support. Moreover, as a probabilistic model multi-mgMOS provides a measure of confidence for the inferred true signal and this will be very useful in downstream analyses, especially those adopting Bayesian methods.

According to the comparison on the two spike-in data sets and the real mouse time-course data set, multi-mgMOS I is better than other methods in terms of accuracy and sensitivity to concentration changes. The difference between the three variants of multi-mgMOS is the different way of modelling the parameter ϕ, which is the fraction of specific hybridisation detected by MM probes. The model of multi-mgMOS II obtains very similar results to multi-mgMOS I. However, the pre-estimated ϕ related to the middle triple base of probes is obtained from a limited number of spike-in genes and under a limited number of conditions, and it may not be representative of the true sequence specific ϕ. The other variant, multi-mgMOS III, plugs in the histogram of empirical ϕ which is obtained by the limited number of spike-in genes and therefore is probably not reliable enough. It seems to be more appropriate to use it as a prior on ϕ as in multi-mgMOS I.

From the comparison of computational efficiency in Table 3.4, the computation of multi-mgMOS II is very fast compared with other probabilistic models. However, the pre-estimated ϕ comes from the spike-in data of HG-U95a chips. It is not clear whether this result is suitable for other types of chips. Therefore, multi-mgMOS I is more reasonable since it uses the empirical information of ϕ obtained from data of the HG-U95a chip as a prior and optimises it using the new observed experimental data. The parameter b_{gj} in (3.13) is probe-specific which already accounts for the specificity of the GC-content of a probe. So multi-mgMOS I is more robust in practice especially for non-human chips although it needs more time to optimise ϕ than multi-mgMOS II. multi-mgMOS I has been implemented in an R package, *mmgmos*, to provide the public use of the model

for the community. In the remainder of the thesis, multi-mgMOS refers to multi-mgMOS I.

Chapter 4

Detecting Differential Gene Expression

In this chapter, the probe-level measurement error of gene expression level is propagated into a hierarchical Bayesian model in order to detect differential gene expression. After the background information and related work on detecting differential gene expression are introduced, the augmented hierarchical model is proposed. The performance of this model is tested on a spike-in data set and a real mouse time-course data set.

4.1 Background

Finding differentially expressed genes is a fundamental objective of a microarray experiment. Due to the high variability in microarray data, replicate arrays are often used to obtain improved accuracy and reproducibility. Because of the high cost of the experiment, the number of chips for each condition is usually small, typically 2-4 replicates. There are two main reasons that make the detection of differential gene expression difficult in this context:

1. Microarray experiments are associated with low precision probe-level measurements, especially for weakly expressed genes (probe-level measurement error).

2. The small number of replicates makes it difficult to obtain an accurate variance estimate for each gene across replicates (between-replicate variance).

Each gene on the Affymetrix array has 11-20 probe-pairs and these provide useful information about probe-level measurement error which can be estimated as described in Chapter 3. In Section 3.5.5 differentially expressed genes are found by including credibility intervals. Technical replicates were used there in which the replicate sample comes from the same fragmented complex cRNA. It was therefore reasonable to assume that between-replicate variance is negligible and all variability is technically caused by the experiment itself. However, in practice biological replicates are usually used. In this case, the replicate sample comes from a different specimen and biological variance exists among replicates in addition to technical variability. Thus, the between-replicate variance which is related to biological variability cannot be neglected.

4.2 Related Work

The simplest method to detect differential gene expression is by ranking based on the fold change (FC) or ratio in gene expression means between two conditions, 1 and 2 for example. For each gene, let $\bar{m}_1$ and $\bar{m}_2$ denote the mean logged gene expression over replicates under the two conditions. The difference between the two means $\bar{m}_1 - \bar{m}_2$ is called the log-ratio. Ranking genes according the log-ratio implicitly assumes equal expression variance for every gene. A widely used alternative method is a t-test. The t-statistic is defined as

$$t = \frac{\bar{m}_1 - \bar{m}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} , \tag{4.1}$$

where σ_i^2 and n_i, $i = 1, 2$, is the variance of the expression and the number of replicates respectively under the two conditions.

The primary drawback of the t-test is that the variance of each condition may be inaccurate due to the small number of replicates and an underestimate will result in over-confidence. Many different approaches have been devised to address this problem, such as the widely used methods Cyber-T (Baldi and Long, 2001) and SAM (Tusher et al., 2001), and some newly devised methods, K5M (Najarian et al., 2004), varMixt (Delmar et al., 2005) and DEDS (Yang et al., 2005). These methods are based on single point estimates of gene expression values that can be obtained from popular probe-level analysis methods for Affymetrix arrays. Cyber-T implements a regularized t-test and SAM adds a regularising constant

to the gene-specific standard deviation. K5M combines K-means clustering and an EM algorithm to optimise a mixture model of t-test statistic scores. varMixt uses groups of genes to estimate the expression variance of individual genes. DEDS integrates different statistics to achieve robust statistical properties.

Cyber-T is a popular probabilistic model and uses a similar hierarchical model to the approach proposed in this chapter. For each gene g, a Gaussian distribution of logged expression level across replicates is commonly assumed (Baldi and Long, 2001; Delmar et al., 2005; Najarian et al., 2004; Krohn et al., 2005),

$$x_{gij} \sim \mathcal{N}(\mu_{gj}, \lambda_{gj}^{-1}) \ , \tag{4.2}$$

where i is the index of replicate and j is the index of condition. The parameter μ_{gj} is the mean logged expression level under condition j and λ_{gj} is the inverse of the between-replicate variance (this form is used because it is more natural to place a prior on λ_{gj}). Each gene is treated independently on the array and each individual gene is focused on, so the index of the gene, g, is omitted in the following equations. Since a dependence between μ_j and λ_j^{-1} is observed in microarray data, Cyber-T uses the following conjugate prior on μ_j and λ_j,

$$\begin{aligned} \mu_j &\sim \mathcal{N}(\mu_0, \frac{1}{\lambda_j \lambda_0}) \\ \lambda_j &\sim \mathrm{Ga}(a_0, b_0) \ . \end{aligned} \tag{4.3}$$

The advantage of this model, with parameters $\theta = \{\mu_j, \lambda_j\}$ and hyperparameters $\phi = \{\mu_0, \lambda_0, a_0, b_0\}$, is that it can be solved in closed form since the conjugate prior is used. The point estimator of the posterior distribution of θ is combined with a t-test to provide an inference approach which performs better than simple t-test or fold change methods and partly compensates for the lack of replication.

The shortcoming of these methods is that they do not consider the probe-level measurement error, and it is shown here that further improvement can be obtained by including this information.

4.3 Methods

The model in (4.2) can be augmented to take the probe-level measurement error ϵ_{ij} into consideration. Suppose x_{ij} is the true expression level for replicate i

under condition j. The observed expression level $\hat{x}_{ij}$ can be expressed as $\hat{x}_{ij} = x_{ij} + \epsilon_{ij}$. A zero-mean Gaussian measurement noise is assumed, $\epsilon_{ij} \sim \mathcal{N}(0, \nu_{ij}^{-1})$, with variance ν_{ij}^{-1} obtained from the probe-level analysis described in Chapter 3. After including the probe-level measurement variance in the model (4.2) the distribution of the observed gene expression level is

$$\hat{x}_{ij} \sim \mathcal{N}(\mu_j, \lambda_j^{-1} + \nu_{ij}^{-1}) \ . \tag{4.4}$$

The probabilistic probe-level analysis obtains a logged gene expression level $\hat{x}_{ij}$ for each gene on each single chip, together with a measurement variance ν_{ij}^{-1} which is treated as a known parameter. The parameters $\theta_j = \{\mu_{gj}, \lambda_{gj}\}$ are to be estimated from this data.

4.3.1 Likelihood and Prior

A full Bayesian approach is considered for the combination of replicate signals. With the absence of probe-level measurement variances it has been widely accepted that μ_j and λ_j^{-1} should be dependent variables. However, it is unclear whether this remains true when probe-level measurement error is accounted for. A *prior* assumption is therefore made that μ_j and λ_j^{-1} are independent and a Gaussian prior over μ_j is selected,

$$\mu_j \sim \mathcal{N}(\mu_0, \eta_0^{-1}) \ , \tag{4.5}$$

where μ_0 and η_0 are hyperparameters. In practice, there are tens of thousands of genes to be processed in each experiment. In order to make the model simple for efficient computation, noninformative hyperpriors on μ_0 and η_0 are used.

The parameter λ_j^{-1} is assumed to be shared across different conditions and measures the gene specific variability. The common variance is denoted as λ^{-1}. In practice many experiments consider only a small number of conditions with a small number of replicates. To reduce the effect of the small size of data set, a conjugate gamma prior over λ is selected,

$$\lambda \sim \mathrm{Ga}(\alpha, \beta) \ , \tag{4.6}$$

where $\mathrm{Ga}(\cdot)$ denotes the Gamma distribution with hyperparameters α and β.

The different contributions to the variance play similar roles to the contributions in Rocke and Durbin's two-component model (Rocke and Durbin, 2001). In their model the additive measurement noise dominates for weakly expressed genes, similar to the probe-level measurement variance ν_{ij}^{-1}, while the multiplicative component in their model corresponds to the replicate variance parameter λ^{-1} which dominates for highly expressed genes. The difference here is that the internal probe replication on Affymetrix arrays is used to estimate ν_{ij}^{-1}.

A hierarchical model is now obtained with parameters $\theta = \{\mu, \lambda\}$ and hyperparameters $\phi = \{\mu_0, \eta_0, \alpha, \beta\}$, where μ represents the set of all μ_j. The posterior distribution $P(\mu_j|D, \phi)$, which summarises the information about the mean expression value of replicates under each condition, is of interest. With the assumption of the independence of observations, the likelihood of the observed data D is

$$\begin{aligned} P(D|\theta) &= \prod_{j=1}^{c} \prod_{i=1}^{r_j} P\left(\hat{x}_{ij}|\mu_j, \lambda\right) \\ &\propto \prod_{j=1}^{c} \prod_{i=1}^{r_j} p_{ij}^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\hat{x}_{ij} - \mu_j\right)^2 p_{ij}^{-1}\right), \end{aligned} \tag{4.7}$$

where $p_{ij} = (\lambda^{-1} + \nu_{ij}^{-1})^{-1}$, c is the number of conditions and r_j is the number of replicates under condition j. The prior on parameters is given by

$$\begin{aligned} P(\theta|\phi) &= P(\lambda|\alpha, \beta) \prod_{j=1}^{c} P(\mu_j|\mu_0, \eta_0^{-1}) \\ &\propto \lambda^{\alpha-1} \exp\left(-\frac{\eta_0}{2} \sum_j \left(\mu_j - \mu_0\right)^2 - \beta\lambda\right) \quad . \end{aligned} \tag{4.8}$$

4.3.2 Parameter Estimation

With the introduction of measurement error in (4.4), Bayesian inference becomes intractable. Various parameter estimation methods are used here and compared in terms of efficiency and accuracy. MAP is used for crude estimation, a variational method for approximate Bayesian inference and MCMC for more accurate Bayesian inference.

MAP Approximation

With the selected priors on θ in (4.8) there is no closed form solution for the marginal likelihood $P(D|\phi)$. By maximising the joint posterior distribution

$$P(\theta, \phi|D) \propto P(D|\theta)P(\theta|\phi), \tag{4.9}$$

the posterior mode θ^*, ϕ^* can be found. Based on MAP solutions a rough estimate of the posterior $P(\theta, \phi|D)$ can be obtained (Gelman et al., 2004). For simple analyses, the conditional posterior of μ_j, $P(\mu_j|\lambda^*, \phi^*, D)$, given other parameters fixed at their modal values, is

$$\begin{aligned} P(\mu|\lambda^*, \phi^*, D) &\propto P(D|\mu, \lambda^*)P(\mu|\lambda^*, \phi^*) \\ &= \prod_j \mathcal{N}\left(\mu_j; \frac{\sum_i p^*_{ij}\hat{x}_{ij} + \eta^*_0\mu^*_0}{\sum_i p^*_{ij} + \eta^*_0}, \left(\sum_i^{r_j} p^*_{ij} + \eta^*_0\right)^{-1}\right) \quad . \end{aligned} \tag{4.10}$$

where $p^*_{ij} = \left((\lambda^*)^{-1} + \nu_{ij}^{-1}\right)^{-1}$. The crude estimates are calculated simply and efficiently (see Gelman et al., 2004, p. 276), but discard the variability in the parameters.

Variational Inference

In order to account for variability in parameters, the EM algorithm combined with a variational method is used to optimise a lower bound on $P(D|\phi)$ and work out an approximation to the posterior distribution $P(\theta|D, \phi)$.

The log marginal likelihood, $\mathcal{L}(\phi)$, of data D is a function of ϕ,

$$\mathcal{L}(\phi) = \log P(D|\phi) = \log \int d\theta \; P(D|\theta, \phi)P(\theta|\phi) \quad . \tag{4.11}$$

The integral is intractable, so the EM algorithm in (2.21) and (2.22) is used to get a lower bound on $\mathcal{L}(\phi)$.

At the E-step the posterior distribution $P(\theta|D, \phi^t)$ is not tractable, however by introducing constraints on the form of $Q(\theta)$ the distribution can be approximated. In variational inference a factorisation constraint on distribution Q is often used. Instead of optimising (2.19) over the whole Q, one can optimise with respect to the distribution of disjoint subsets of θ. For this model, it has been assumed that λ is independent of μ_j a priori, so a reasonable approximation should be to

assume $Q(\theta)$ factorises as

$$Q(\theta) = Q(\lambda)Q(\mu) \ . \tag{4.12}$$

The factorised Q distribution is substituted into (2.19) and optimised with respect to $Q(\mu)$ and $Q(\lambda)$ respectively. The optimal expressions for $Q(\mu)$ and $Q(\lambda)$ are

$$\begin{aligned} Q(\mu) &\propto \exp \int d\lambda \, Q(\lambda) \log P(D|\theta, \phi^t) P(\theta|\phi^t) \\ &= \prod_j \mathcal{N}\left(\mu_j; \frac{\sum_i \langle p_{ij} \rangle \hat{x}_{ij} + \eta_0^t \mu_0^t}{\sum_i \langle p_{ij} \rangle + \eta_0^t}, \left(\sum_{i=1}^{r_j} \langle p_{ij} \rangle + \eta_0^t\right)^{-1}\right) \ , \end{aligned} \tag{4.13}$$

$$\tag{4.14}$$

$$\begin{aligned} Q(\lambda) &\propto \exp \int \left(\prod_{j=1}^{c} d\mu_j \, Q(\mu_j)\right) \log P\left(D|\theta, \phi^t\right) P\left(\theta|\phi^t\right) \\ &= \mathrm{Ga}\left(\lambda; \alpha^t, \beta^t\right) \prod_{ij} \mathrm{Ga}\left(p_{ij}; \frac{3}{2}, \frac{1}{2}\langle (\hat{x}_{ij} - \mu_j)^2 \rangle\right) \\ &= \mathrm{Ga}\left(\lambda; \alpha^t, \beta^t\right) f(\lambda) \ , \end{aligned} \tag{4.15}$$

where $\langle \cdot \rangle$ denotes the expectation of a function with respect to $Q(\lambda)$ or $Q(\mu)$ and $f(\lambda)$ denotes $\prod_{ij} \mathrm{Ga}\left(p_{ij}; \frac{3}{2}, \frac{1}{2}\langle (\hat{x}_{ij} - \mu_j)^2 \rangle\right)$. In the variational approach, when there is no standard form for the Q distribution, importance sampling can be used to obtain the required expectations (see Section 2.2.3).

At each E-step the distributions $Q(\mu)$ and $Q(\lambda)$ are calculated iteratively until the parameters have converged when parameter changes are small enough per iteration. When the whole EM algorithm in (2.22) converges, $Q(\mu)$ in (4.13) is the approximated posterior distribution $P(\mu|D, \phi)$ of mean gene expression level. From the density function of $Q(\mu)$ in (4.13) it can be seen that the μ_j's are independent of each other. For condition j, the mean and variance of μ_j is

$$\langle \mu_j \rangle = \frac{\sum_i \langle p_{ij} \rangle \hat{x}_{ij} + \eta_0^t \mu_0^t}{\sum_i \langle p_{ij} \rangle + \eta_0^t} \tag{4.16}$$

$$\mathrm{Var}[\mu_j] = \left(\sum_{i=1}^{r_j} \langle p_{ij} \rangle + \eta_0^t\right)^{-1} \ . \tag{4.17}$$

From (4.17) it can be seen that as the number of replicates increases, more confidence in the mean expression level is obtained since the inverse variance grows

with the number of replicates.

MCMC

Intractable hierarchical Bayesian models are usually solved by random sampling from the posterior distribution of model parameters. Standard MCMC is applied to summarise the Bayesian inference of the model by assuming a flat uniform prior on μ_0 and flat Gamma priors on λ and η_0 with shape and inverse scale both 0.001. The joint posterior distribution of all parameters is

$$P(\mu, \lambda, \mu_0, \eta_0|D) \propto P(\lambda)P(\mu_0)P(\eta_0)\prod_j P(\mu_j|\mu_0, \eta_0)\prod_i\prod_j P(\hat{x}_{ij}|\mu_j, \lambda) \ . \quad (4.18)$$

Gibbs Sampler The conditional posterior distribution of each μ_j, μ_0 and η_0 is in standard form. Gibbs sampling can therefore be used to update these parameters. The conditional posterior distributions of these parameters are,

$$\begin{aligned} P(\mu_j|\lambda, \mu_0, \eta_0, D) &\propto P(\mu_j|\mu_0, \eta_0)\prod_i P(\hat{x}_{ij}|\mu_j, \lambda) \\ &= \mathcal{N}\left(\frac{\sum_i p_{ij}\hat{x}_{ij} + \eta_0\mu_0}{\sum_i p_{ij} + \eta_0}, (\sum_i p_{ij} + \eta_0)^{-1}\right) \ , \quad (4.19) \\ P(\mu_0|\mu, \lambda, \eta_0, D) &\propto P(\mu_0)\prod_j P(\mu_j|\mu_0, \eta_0) \\ &= \mathcal{N}\left(\frac{1}{c}\sum_j \mu_j, \frac{1}{c\eta_0}\right) \ , \quad (4.20) \end{aligned}$$

$$\begin{aligned} P(\eta_0|\mu, \lambda, \mu 0, D) &\propto P(\eta_0)\prod_j P(\mu_j|\mu_0, \eta_0) \\ &= \mathrm{Ga}\left(\frac{c}{2} + 0.001, \frac{1}{2}\sum_j (\mu_j - \mu_0)^2 + 0.001\right) \ , \quad (4.21) \end{aligned}$$

where c is the number of conditions.

The starting point of μ_j is

$$\mu_j^0 = \frac{1}{r_j}\sum_i \hat{x}_{ij} \ , \quad (4.22)$$

and the starting point of μ_0 is

$$\mu_0^0 = \frac{1}{c}\sum_j \mu_j^0 \ . \tag{4.23}$$

Given the starting points of μ_j and μ_0, the starting point η_0 can be drawn from the density in (4.21).

Metropolis Algorithm The conditional distribution of λ is

$$\begin{aligned} P(\lambda|\mu,\mu_0,\eta_0,D) &\propto P(\lambda)\prod_{ij} P(\hat{x}_{ij}|\mu_j,\lambda) \\ &\propto \lambda^{0.001-1}\prod_{ij} p_{ij}^{\frac{1}{2}} \exp\left(-0.001\lambda - \frac{1}{2}\sum_{ij} p_{ij}\left(\hat{x}_{ij}-\mu_j\right)^2\right) \ , \end{aligned} \tag{4.24}$$

which has no standard form, so the Metropolis algorithm is adopted to update $\log(\lambda)$ using a Gaussian proposal distribution, $\mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$, with μ_λ being the current value of $\log(\lambda)$ and the initial variance $\sigma_{\lambda^0}^2$ coming from the second derivative of the logged conditional posterior distribution in (4.24) with respect to $\log(\lambda)$,

$$\sigma_{\lambda^0}^2 = k^2\left(-\frac{d^2}{d\left(\log\left(\lambda\right)\right)^2}\log\left(P\left(\lambda|\mu,\mu_0,\eta_0,D\right)\right)\Big|_{\lambda^0}\right)^{-1} \ , \tag{4.25}$$

where λ^0 is the starting value for λ. The optimal scale k is set to 2.4 (Gelman et al., 2004). During the first half of the simulation, $\sigma_{\lambda^0}^2$ can be increased or decreased to tune the simulation efficiency. In the second half, the shape of the Gaussian should be fixed to avoid converging to the wrong distribution.

Given the starting point of μ, μ_0 and η_0, the starting point of λ is obtained at the mode of the density in (4.24). At time t, μ^t, μ_0^t and η_0^t are drawn from the distributions in (4.19), (4.20) and (4.21) respectively, a proposal $\tilde{\lambda}$ from the Gaussian proposal distribution is sampled, and the ratio of the densities can be calculated,

$$r = \frac{P(\tilde{\lambda}|\mu^t,\mu_0^t,\eta_0^t,D)}{P(\lambda^{t-1}|\mu^t,\mu_0^t,\eta_0^t,D)} \ . \tag{4.26}$$

Then one can set

$$\lambda^t = \begin{cases} \tilde{\lambda} & \text{with probability } \min(r,1) \\ \lambda^{t-1} & \text{otherwise.} \end{cases} \tag{4.27}$$

4.3.3 Significance of Differential Expression

Once the posterior distribution $P(\mu|D,\phi)$ is obtained, it is possible to compute the significance of differential expression between any two conditions. Taking a treatment and a control (indicated by 1 and 2 respectively), for instance, the probability of positive log-ratio (PPLR), $P(\mu_1 > \mu_2|D,\phi)$, can be calculated by

$$P(\mu_1 > \mu_2|D,\phi) = \int_0^{+\infty} d(\mu_1 - \mu_2)\, P(\mu_1 - \mu_2|D,\phi) \ . \tag{4.28}$$

Equation (4.28) gives the posterior probability of increased expression in a treatment compared with a control. One can find the up-regulated genes by setting a level of confidence, like an α-level in a conventional statistical test. Down-regulated genes can also be found using a similar equation to (4.28) by calculating the integral of $P(\mu_1 - \mu_2|D,\phi)$ over $(-\infty, 0)$.

4.3.4 Implementation and Computation Time

During the developmental stage of the model, all three parameter estimation approaches were implemented in Matlab. For importance sampling in the variational method, 1000 samples were drawn once at each EM iteration. At each E-step, $Q(\mu)$ and $Q(\lambda)$ were considered to have converged when parameters change by less than 10^{-6} per iteration. For the implementation of MCMC 10 parallel sequences are simulated, each of length 200. The convergence of the iterative simulation is monitored by estimating the potential scale reduction $\hat{R}$ (Gelman et al., 2004) for all parameters. If $\hat{R} > 1.1$, another 200 samples for each sequence are drawn. After discarding the first half of the simulations and mixing the remaining simulated sequences, the length of the effective simulations are between 1000 and 3000 for most genes.

To process the golden spike-in data set (described in Appendix A.4), the MAP approximation, variational inference and MCMC take around 5 minutes, 4 hours and 50 hours respectively using a 1.8 GHz AMD Opteron machine with 512M RAM. It is concluded that MCMC is too computationally expensive to use in practice although it is useful to compare with other methods as a gold standard. Variational inference is a good compromise between computational efficiency and accuracy. MAP estimation is very fast, but less accurate (see Section 4.4.2), so it is recommended for crude inference. For more accurate and applicable inference, one can choose to use the variational method. The MAP estimation

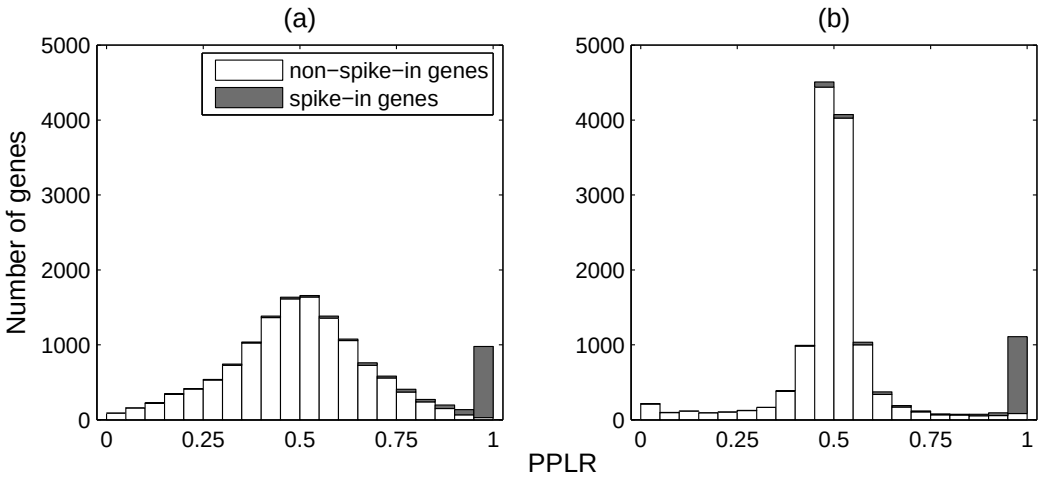

Figure 4.1: Histogram shows probability of positive log-ratio (PPLR) between (a) C1 and S1, and (b) replicated condition C and S in the golden data set. The histogram is a stack of non-spike-in genes and spike-in genes. The white is for non-spike-in genes and the shade is for the spike-in genes.

and variational inference have been implemented in an R package, *pplr*, for public use of the model.

4.4 Results and Discussion

The new method is compared with one of the most popular approaches, Cyber-T (Baldi and Long, 2001), to show the improvement obtained by including probe-level measurement variance. Two data sets are used to perform this comparison. One is a wholly defined spike-in data set called the golden spike-in data set (see Appendix A.4). The other is the real-world mouse time-course data set (Appendix A.3) which was used to identify hair-cycle associated genes by Lin et al. (2004).

4.4.1 Making Use of Measurement Error

Before the Bayesian hierarchical model is applied on replicated data, the usefulness of the measurement error from multi-mgMOS in a single chip experiment is shown here. Chip C1 and S1 are selected from the golden data set as the control and treatment chips. Since all spiked-in genes in S1 are up-regulated, the equation in (4.28) is used to calculate PPLR in S1 compared with C1 using the measured gene expression values and variances from multi-mgMOS. The histogram of PPLR is shown in Figure 4.1 (a). The PPLRs of most spike-in genes

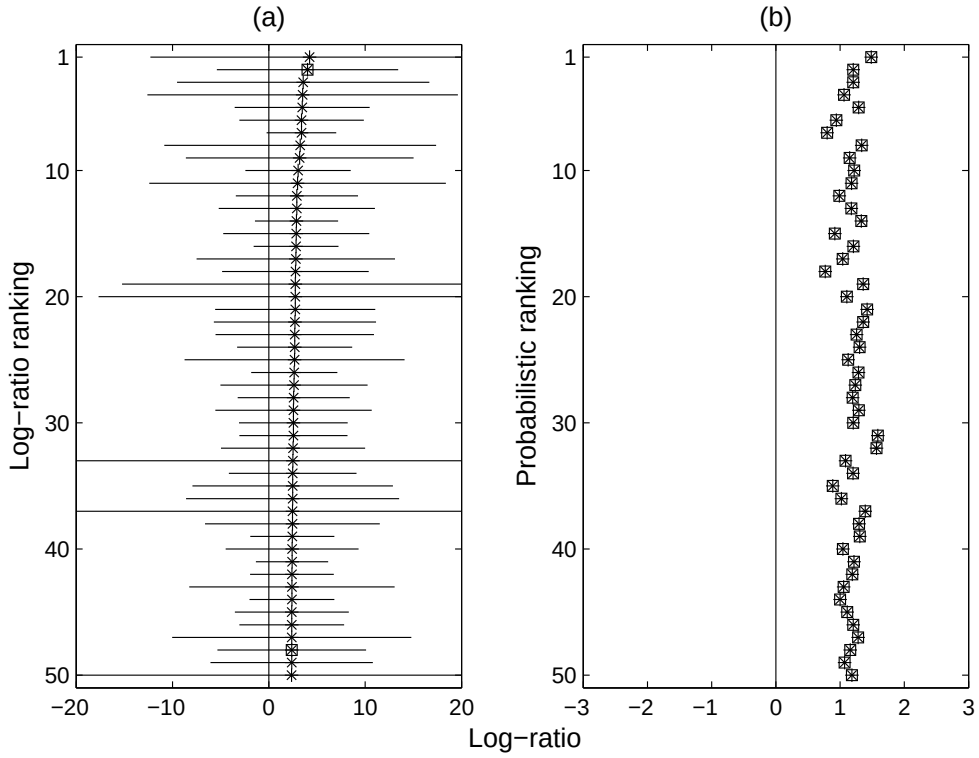

Figure 4.2: 5-95% credibility intervals of positive log-ratio between S1 and C1 in the golden data set. The left figure shows the top 50 most significantly differentially expressed genes ranking by log-ratio and the right is the ranking by the probability of positive log-ratio (PPLR) between S1 and C1. Stars represent the mean of log-ratio. Spike-in genes are indicated by a square box. Credibility intervals are small in (b) and cannot be seen. Without considering the measurement error, log-ratio ranking (on the left) obtains a much larger false positive rate than PPLR ranking (on the right).

are close to 1 which shows the high confidence of the increasing gene expression in S1. This is consistent with the golden data where all spike-in genes are up-regulated. Most invariant genes have PPLR close to 0.5 and there is no obvious evidence for these genes to be differentially expressed. Figure 4.2 shows the top 50 differential expression ranking between chip C1 and S1 by (a) simple fold-change and (b) PPLR. Without considering the measurement variance, multi-mgMOS obtains a large number of false positives by fold-change ranking. After taking the credibility intervals into account, there are now no false positives in the top 50 positions.

For the golden spike-in data set, the true differentially expressed genes are known, so for any given cut-off value of the test statistic (such as fold change,

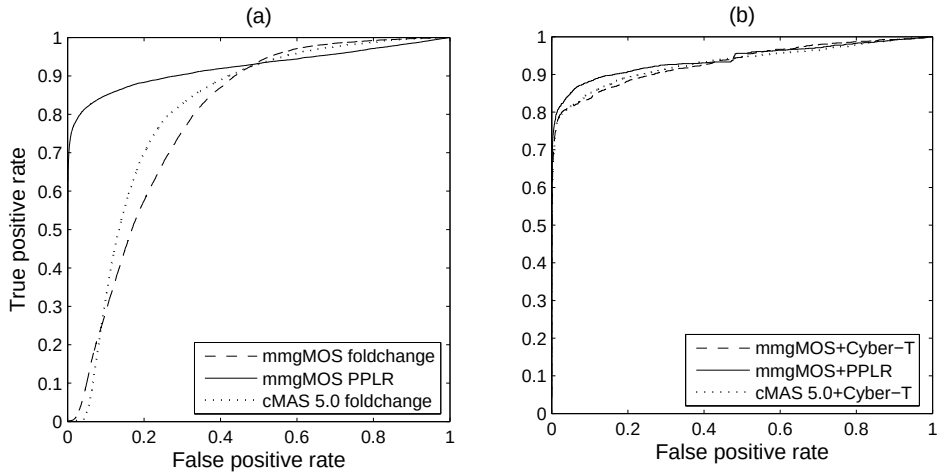

Figure 4.3: ROC curves for (a) all nine possible single chip-pairs, and (b) replicated conditions in golden data set. Three methods are used in each case, (a) simple fold-change of multi-mgMOS and cMAS 5.0 gene expression, and probability of positive log-ratio (PPLR) of multi-mgMOS gene expression measurements together with measurement error, (b) multi-mgMOS combined with Bayesian hierarchical model and Cyber-T respectively, and cMAS 5.0 with Cyber-T. Curves from Cyber-T are obtained using a Bayesian t-test and the curve from the Bayesian hierarchical model is plotted by calculating PPLR between sample S and sample C.

p-value and so on), the true positive rate and the false positive rate can be calculated. By moving this cut-off value along the whole range of the test statistic, one can obtain a set of true and false positive rate pairs. Based on these true and false positive rate pairs, a Receiver Operator Characteristic (ROC) curve can be drawn. If the true differentially expressed genes are ideally identified by the ranking based on the test statistic, the area under the ROC curve (AUC) is one. Otherwise, AUC is less than one. The ROC curve is a useful tool to judge rankings based on different test statistics and it can be used to compare methods of detecting differentially expressed genes. Methods with higher AUC are often considered better although the shape of the curve is also informative.

The golden data set has three replicates for each of the C and S condition. The average ROC curves for all nine possible single chip-pairs between condition C and S are plotted in Figure 4.3 (a). As well as the two methods using multi-mgMOS, as shown in Figure 4.2, a combined method suggested in Choe et al. (2005) is also included. This method is denoted as cMAS 5.0 here because the major procedures in this method (background correction and PM adjustment) come from MAS

5.0 (Affymetrix, 2002). The results in Choe et al. (2005) show that cMAS 5.0 performs best among current statistical methods, including RMA(Irizarry et al., 2003) and GCRMA(Wu et al., 2004), on the golden spike-in data set. The same form of loess normalisation for all methods in Choe et al. (2005) is used here based on the 2,532 invariantly expressed probe-sets, in order to make results comparable. The area under ROC curves are 0.9226, 0.8062 and 0.7869 for PPLR of multi-mgMOS, fold-change of cMAS 5.0 and fold-change of multi-mgMOS respectively. Notice that at the upper-right part of ROC curves PPLR obtains a slightly lower true positive rate than the other two methods. In practice people are often more concerned with obtaining a low false positive rate. For a reasonable number of false positives, on the left of the ROC curves, PPLR is much better than the other two alternatives. These results show that the uncertainty of the estimated expression level helps in detecting differential gene expression where there is only a single chip for each condition.

4.4.2 Combining Replicates

In practice, people usually use replicates to estimate a level of uncertainty with the estimated gene expression level. The proposed Bayesian hierarchical method includes the measurement error of replicates to improve the estimation of the uncertainty of gene expression measurements.

Comparison of Estimation Accuracy

MCMC simulation obtains reliable results when it converges. The results from MCMC are used as a gold standard to evaluate the accuracy of MAP estimation and variational inference. Figure 4.4 shows the distribution of signal log ratio between days 14 and 1 for two probe-sets in the mouse time-course data set. The distribution shown in (a) is for one probe-set of gene Dab2 which is a known hair-growth associated gene. The estimated expression levels of this probe-set are variable over different time points. MAP approximation and variational inference obtain similar accuracy for this probe-set. The probe-set shown in (b) is randomly selected and is not obviously hair-growth related, so the signal log ratio is not as high as for the probe-set in (a). For this probe-set, the variational method is closer than the MAP approximation to MCMC. In general, the posterior estimates of the variational approach are found to be closer than those provided by the MAP

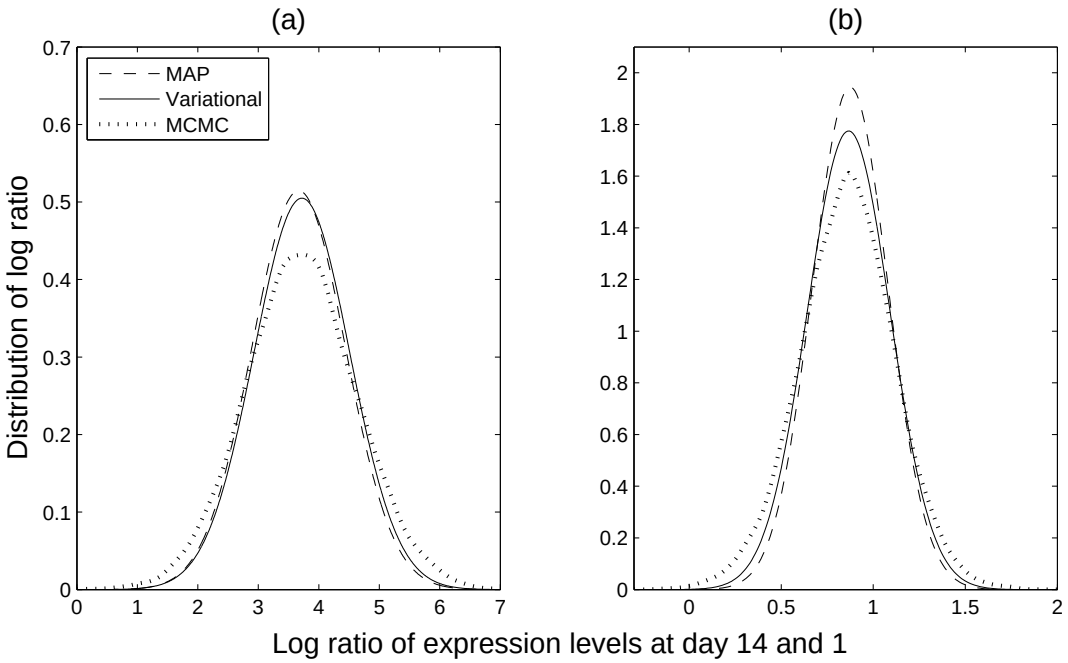

Figure 4.4: Distribution of log ratio of expression level between days 14 and 1 for (a) one probe-set of gene Dab2 and (b) a randomly selected probe-set in the mouse time-course data set. Three different parameter estimation methods are used: MAP estimation, variational approximation and MCMC.

approximation to MCMC results. The variational method is therefore used in the following examples.

Performance on an Artificial Data set

Figure 4.1 (b) shows the histogram of PPLR between replicated condition C and S in the golden spike-in data set. In addition to more spike-in genes moving close to one, non-spike-in genes get tighter around 0.5 compared with the histogram of PPLR between C1 and S1 in Figure 4.1 (a). More confidence is therefore obtained in the up-regulated genes and the invariant genes. ROC curves (Figure 4.3 (b)) are plotted on the golden spike-in data set to show the ability of the proposed combination method compared with the widely used approach Cyber-T which does not consider the probe-level measurement variance. From Figure 4.3 (b) it can be seen that the inclusion of measurement error does improve the ability of multi-mgMOS to detect differentially expressed genes. The area under ROC curves for the Bayesian hierarchical method and Cyber-T on results from multi-mgMOS are 0.9431 and 0.9310 respectively, and 0.9306 for cMAS 5.0 combined

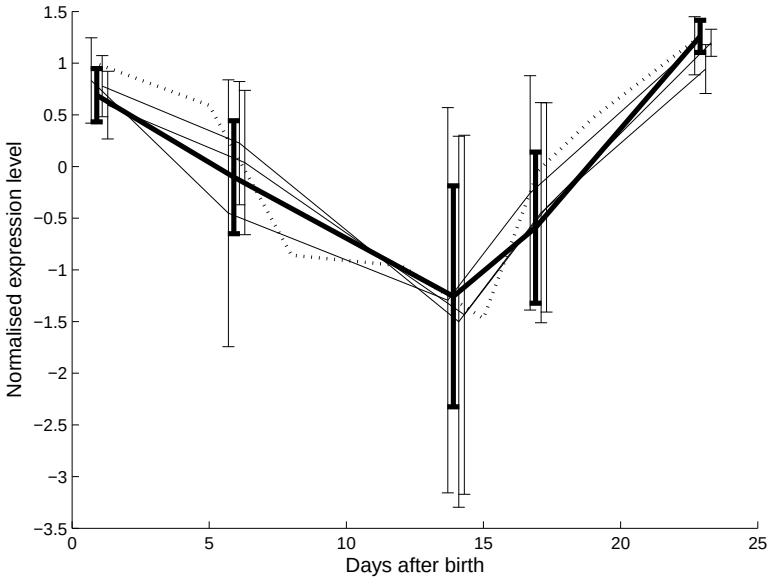

Figure 4.5: Temporal profile of one probe-set of gene Dab2 in the mouse time-course data set. Thin solid lines are the estimated results from multi-mgMOS for three replicate chips and the thick solid line is the combined result from the hierarchical Bayesian model. At each time point 2.5-97.5% credibility intervals are shown. The dotted line is the quantitative real-time PCR profile.

with Cyber-T. PPLR obtains the best result. More results from popular statistical methods combined with Cyber-T on the golden spike-in data set are presented in Choe et al. (2005).

Performance on a Real Data set

In order to examine the method in an application on a real-world experiment, the method is applied to the mouse time-course data set. Figure 4.5 is the temporal profile of one probe-set related to gene Dab2 in this data set. The first hair-growth cycle is shown, which is covered by five time points (day 1, 6, 14, 17 and 23). It can be seen that the combined signal obtains more confidence for each condition compared with the signal from each individual replicate and remains consistent with the qr-PCR profile.

The qr-PCR data in the mouse time-course data set has three measurements at each time point. These replicate values can be processed using Cyber-T at those time points that were also measured by microarray data (day 1, 17 and 23). Using the obtained statistical values of qr-PCR data, the PPLR calculated from the proposed method is validated and compared with the results from Cyber-T

using expression measurements obtained from cMAS 5.0 and the popular probe-level processing method GCRMA (Wu et al., 2004). The results generated from Cyber-T are associated with a p-value which does not have the same meaning as for PPLR and cannot be compared with PPLR directly. In order to make them comparable, different credibility levels are selected so that methods obtain a similar number of significant genes. For PPLR the significance level is set at 0.06, while the p-values for the other methods are set at 0.01. The number of significant genes at the different credibility levels for different methods are shown in the lower part of Table 4.1. A global scaling normalisation is used for results from multi-mgMOS and cMAS 5.0 at the probe-set level, and GCRMA uses quantile normalisation at the probe level, in order to obtain the best result for each method. The eight PCR validated genes have 14 associated probe-sets shown in Table 4.1. The differential gene expression between two pairs of time points, (day 1, 17) and (day 17, 23), is found. Day 1 of gene Crisp1 is excluded due to the absence of replicate measurements of qr-PCR data at this time point. There are 27 tests altogether in this data. From these 27 tests PPLR obtains two inconsistent results with qr-PCR, while cMAS 5.0+Cyber-T and GCRMA+Cyber-T obtain three. Results from the new method therefore show more consistency with results from qr-PCR data, although this is a relatively small set of comparisons.

4.5 Conclusion

This chapter has presented a Bayesian hierarchical model using probe-level measurement error in order to improve the detection of differential gene expression. The introduction of an additional variance term makes Bayesian hyper-parameter estimation intractable. Three different computation methods, MAP approximation, a variational method and MCMC, are compared to solve the intractability in the model. The MAP approximation is very efficient, but cannot provide accurate inference. MCMC is accurate, but is computationally expensive and not applicable for large data sets. The variational inference is relatively efficient and provides a good approximation to the MCMC results. This model is described in Liu et al. (2006) and for public use of the method, the MAP approximation and the variational inference are implemented in an R package, *pplr*.

This approach makes full use of information in microarray experimental data

and performs relatively efficient and accurate inference based on this rich information. Results on a spike-in data set and a real time-course data set show that the inclusion of probe-level measurement error improves the accuracy of finding differentially expressed genes, especially when there are few replicate chips for conditions.

Probe-set ID	Time points	qr-PCR p-value	0.01	multi-mgMOS PPLR	0.06	cMAS 5.0 p-value	0.01	GCRMA p-value	0.01
160647_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
93122_at	17,23	0.0	S	0.0	S	0.0	S	0.0	S
103283_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
102362_i_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
102363_r_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
103048_at	1,17	0.0	S	0.001	S	0.0	S	0.0	S
	17,23	0.409	N	0.079	N	0.002	S	0.278	N
103490_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
98044_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.080	N	0.045	N	0.154	N
98045_s_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
104633_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
94307_at	1,17	0.0	S	0.002	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.0	S
94308_at	1,17	0.0	S	0.002	S	0.0	S	0.002	S
	17,23	0.0	S	0.174	N	0.686	N	0.769	N
94309_g_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.029	S	0.008	S	0.0	S
161628_r_at	1,17	0.0	S	0.0	S	0.0	S	0.0	S
	17,23	0.0	S	0.0	S	0.0	S	0.273	N
Significant genes	1,17	-		3660		3566		3599	
	17,23	-		4288		4034		4450	

Table 4.1: Finding differential gene expression among eight qr-PCR validated genes in a mouse time-course data set. qr-PCR data, cMAS 5.0 and GCRMA expression measurements are processed by Cyber-T to obtain p-values and multi-mgMOS estimates are processed by the Bayesian hierarchical model to calculate the probability of positive log-ratio (PPLR). For Cyber-T results a credibility level at $\alpha = 0.01$ is set. 'S' stands for significant differential expression, and 'N' not significant. For comparison with p-value, $min(PPLR, 1-PPLR)$ is shown in the table. A comparable credibility for PPLR, $\alpha = 0.06$, is used which means genes which have at least 94% probability of signal change are considered as significantly differentially expressed. The numbers of significant genes for different methods at different credibility levels are shown in the lower part of the table.

Chapter 5

Propagating Uncertainty in Model Based Clustering

In this chapter a method is developed for propagating the probe-level measurement error into gene expression clustering using standard model-based methods. After background information and related work on model-based clustering are introduced, the augmented standard Gaussian mixture model is proposed. The performance of this model is tested on two simulated data sets and a real mouse time-course data set.

5.1 Introduction

In addition to the detection of differential gene expression, clustering is another important method in the analysis of gene expression data. By clustering, the large number of genes are divided into a smaller number of categories according to their expression patterns that may reflect their similar function or common regulation. By exploring and studying the obtained gene clusters, the function of unknown genes can be inferred from other known genes in the same cluster. Unsupervised clustering is currently the most frequently used approach for exploring gene function. The properties of the data is directly inferred without the help of the correct response since the truth of the data is unknown.

There are many unsupervised algorithms which have been applied to cluster gene expression data, including the most popular hierarchical clustering (Eisen et al., 1998) and k-means (Tavazoie et al., 1999) based on similarity measures, and self-organising maps (Tamayo et al., 1999). Most of these algorithms are

largely heuristically motivated and rely on the similarity measures and the specific data they work on. It is hard to say which one is generally better than others (D'haeseleer, 2005) and these methods lack the capability to deal with the variability in the gene expression data in a principled way. Furthermore, there is no formal way to determine the number of clusters for these algorithms. Probabilistic models provide a principled alternative to these heuristic-based methods. In particular, model-based approaches have been proposed to cluster gene expression data in a probabilistic way (Fraley and Raftery, 2002b; Yeung et al., 2001; Siegmund et al., 2004; Lin et al., 2004). Probabilistic models also adopt model selection methods to determine the number of clusters (see Section 2.2.4). The advantage of model-based approaches over heuristic methods has been demonstrated by Yeung et al. (2001).

In spite of the clear advantages of model-based methods, existing methods do not consider the probe-level measurement error associated with gene expression levels and discard this rich information about variability. This may lead to biologically irrelevant clusters, especially due to the noisy nature of the data. The model developed in Chapter 3, multi-mgMOS, provides accurate gene expression measurements along with the associated uncertainty in this measurement. It has been shown in Chapter 4 and by Sanguinetti et al. (2005) that the probe-level measurement error can be propagated through the downstream probabilistic analysis, thereby improving the performance of the analysis. This chapter describes an approach to propagating probe-level measurement error into model-based clustering to improve performance over current standard clustering methods.

5.2 Methods

5.2.1 Mixture Model

The mixture model is a useful tool for revealing the inherent structure of data. In a mixture model with K components, the data is generated by

$$p(x_i) = \sum_{k=1}^{K} P(k)p(x_i|k;\theta_k) \ , \tag{5.1}$$

where $P(k)$ denotes the probability of selecting the kth component with parameters θ_k and $\theta = \{\theta_1, \theta_2, \cdots, \theta_K, P(k)\}$ is the complete parameter set of the mixture

model. The parameters k are latent variables determining which cluster the data belongs to.

Mixture models are usually solved by maximum likelihood using an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). With the initialised parameters at $t = 0$, the values of parameters can be determined iteratively through an E-step and M-step:

- E-step: Compute
$$P^t(k|x_i) = P(k|x_i; \theta^t) \tag{5.2}$$
for each data point x_i and each component k.

- M-step:
$$\theta^{t+1} = \arg\max_{\theta} \sum_i \sum_k P^t(k|x_i) \log(p(x_i|k; \theta_k)P(k)) \tag{5.3}$$
with constraint $\sum_k P(k) = 1$.

5.2.2 Standard Gaussian Mixture Model

For mixture component distributions from the exponential family, like the Gaussian, both steps are exactly tractable. In a Gaussian mixture model, each component k is modeled by a Gaussian distribution with mean μ_k and covariance matrix Σ_k,

$$\begin{aligned} p(x_i|k; \theta_k) &= \mathcal{N}(s_i|\mu_k, \Sigma_k) \\ &= \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right) \quad , \end{aligned} \tag{5.4}$$

where $|\cdot|$ denotes determinant and p is the dimension of the data. As well as changing the number of components in the mixture, the covariance matrix Σ_k can be constrained to determine the flexibility of the model. The most constrained model is parameterised by $\Sigma_k = \sigma^2 I$ with only one free parameter in the covariance matrix for all components. The unconstrained model has full rank Σ_k with $p(p+1)/2$ free parameters in the covariance matrix for each component. All representations of the covariance matrix are explored by Banfield and Raftery (1993). Allowing the number of free parameters in the covariance matrix to vary leads to various models accommodating varying characteristics of data. All of these models are implemented in MCLUST (Fraley and Raftery, 2002a).

5.2.3 Propagating Measurement Uncertainty into a Gaussian Mixture Model

From a probabilistic probe-level model, such as multi-mgMOS, for each data point x_i one can obtain the measurement error, β_i, which is also a vector with each element as the variance of the measured expression level on each chip if one assumes the independence of the gene expression measurement on each chip. When the mixture model accounts for the measurement error of each data point, β_i, the Gaussian component can be augmented as

$$p(x_i) = \sum_{k=1}^{K} P(k)p(x_i|k; \mu_k, \Sigma_k + \text{diag}(\beta_i)) \ , \tag{5.5}$$

where $\text{diag}(\beta_i)$ represents the diagonal matrix whose diagonal entries starting in the upper left corner are the elements of β_i. Ideally, the covariance matrix should be of full rank to obtain the largest flexibility of the model. However, this will increase the complexity of the model. Since in (5.5) the additive measurement error $\text{diag}(\beta_i)$ accounts for inherent variability in the data, especially for extremely noisy gene expression data, the unequal volume spherical model (VI) described in Yeung et al. (2001) with the covariance $\Sigma_k = \sigma_k^2 I$ is adopted. This model allows the spherical components to have different variances which accounts for the variability within different gene function groups. Therefore, in this model the gene-specific variance β_i is known and obtained from a probabilistic probe-level analysis model, and the function-specific variance σ_k^2 is to be estimated from the mixture model via the EM algorithm. The parameters are denoted $\theta_k = \{\mu_k, \sigma_k^2\}$ for Gaussian component k and $\theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$ for all components, where K is the number of components. The augmented clustering model in (5.5) is denoted PUMA–CLUST (Propagating Uncertainty in Microarray Analysis – CLUSTering).

Using the simple K-means algorithm, one can obtain the initial parameters θ^0 for all components. Equal probability of the component prior is also assumed for the initial value of $P(k)$, $P^0(k)$. At the E-step, for each data point x_i the posterior of the mixture is calculated by

$$\begin{aligned} P^t(k|x_i) &= P^t(k|x_i; \theta^{t-1}) \\ &= \frac{P(x_i|\theta_k^{t-1})P^{t-1}(k)}{\sum_k P(x_i|\theta_k^{t-1})P^{t-1}(k)} \ . \end{aligned} \tag{5.6}$$

At the M-step, the component prior and the parameters of components are optimised,

$$P^t(k) = \frac{1}{N}\sum_{n=1}^{N} P^t(k|x_i) \tag{5.7}$$

$$\theta^t = \arg\max_{\theta} \sum_i \sum_k P^t(k|x_i)\log(p(x_i|\theta_k)P^t(k)) \ . \tag{5.8}$$

The parameter θ cannot be solved analytically in (5.8) due to the incorporation of β_i in the variance terms. However, with fast optimisation methods available such as SNOPT (Gill et al., 2002) and donlp2 (Spellucci, 1998), it is easy to calculate the optimal parameters numerically at the M-step.

5.2.4 Model Selection

In Section 5.2.3 the covariance matrix of the Gaussian mixture model is specified and the parameters are worked out via an EM algorithm for a given K. In practice the most appropriate number of clusters should also be determined. In mixture models, the Bayesian Information Criterion (BIC, described in Section 2.2.4) is usually used to decide the appropriate number of clusters. For model m with the number of clusters K, the calculation of BIC is

$$\mathrm{BIC}_m = -2\log(p(D|\hat{\theta}_m) + d_m \log(n) \ , \tag{5.9}$$

where d_m is the number of free parameters to be estimated in model m, n is the number of genes and $\hat{\theta}_m$ is the estimated parameters θ_m obtained by the EM algorithm. For the unequal volume spherical model (VI), the number of free parameters is $d_m = K(p+2) - 1$. The model with optimum K has the minimum BIC value.

5.3 Results and Discussion

The performance of the extended Gaussian mixture model on two simulated data sets and a real-world mouse time-course data set (see Appendix A.3) are examined. The simulated data sets are generated to reflect the noise commonly seen in real microarray experiments. The extended mixture model is compared with the standard Gaussian mixture model implemented in MCLUST (Fraley and Raftery,

2002a), which includes all variants of standard Gaussian mixture models in terms of the representation of the covariance matrix. However, these models do not take the probe-level measurement error into consideration.

The performance of different clustering methods on data sets with known structures can be evaluated by using the adjusted Rand index (Milligan and Cooper, 1986). The adjusted Rand index (Hubert and Arabie, 1985) is a technique for measuring the similarity of two clusterings on a data set and it is widely used by the clustering research community (Yeung et al., 2001, 2003; Bolshakova and Azuaje, 2003; Medvedovic et al., 2004). The adjusted Rand index lies between 0 and 1, and is calculated based on whether pairs are placed in the same or different clusters in two partitionings. A higher adjusted Rand index means better agreement between two clusterings.

For the simulated data sets, since the true structure of the data is known, one can use the adjusted Rand index to evaluate the different partitioning ability of the extended mixture model which incorporates the probe-level measurement error and the standard mixture model. For the real mouse time-course data set, biological interpretation is used to examine the different clusterings from the two clustering methods.

5.3.1 Clustering on Simulated Data Sets

Simulated Periodic Data

Periodic patterns are often observed in real-world time-course microarray data (Lin et al., 2004; Tu et al., 2005). However, the true structure of the real data sets is unavailable. Simulated periodic data with the noise coming from the real data can be generated instead. Similar to the methods used by Yeung et al. (2003) and Medvedovic et al. (2004), the simulated data is generated by the following four steps.

At the first step, the logged gene expression within each known group is generated. There are six groups and 600 genes in the data set. Each group has 100 genes. The first four groups have a periodic sine pattern. The expression of gene i in group q, $q = 1, 2, 3, 4$, is generated by

$$x_{qij} = A_i \sin(2\pi j/10 - \pi q/2) + S \ , \tag{5.10}$$

where $j = 1, 2, \cdots, J$ and J is the number of conditions or time points. A_i is

a random scaling factor which is sampled from U$(0, 7)$, where U represents the uniform distribution. S is a shifting factor which is set as 7. The assignment of A_i and S is to make the gene expression level lie between 0 and 14 which is the normal range of the logged gene expression level from real data sets. The gene expression levels of group 5 and group 6 are generated by linear functions

$$x_{qij} = jA_{qi}/J \text{ and } x_{qij} = -jA_{qi}/J + S \ , \tag{5.11}$$

respectively, where A_{qi} is sampled from $U(0, 14)$ and $S = 14$ when $q = 6$ so as to ensure that the simulated expression level lies within the accepted logged expression range.

The simulated data from the first step follows perfectly the same sine wave within the same group except for a different magnitude. However, in practice there is biological and technical noise in the experiment distorting the true sine wave (see Section 1.2). At the second step, the real mouse data set (see Appendix A.3) is used to obtain the combined noise of biological and technical sources which is related to the variance of observed gene expression level from replicated experiments. The mouse data set has three or four replicates for each condition. Using the gene expression summaries from MAS 5.0, the combined noise can be obtained from Cyber-T (see Section 4.2). Since the gene expression level is correlated with its variance, the combined noise, σ^2_{qij}, is sampled from a subset of variances calculated from Cyber-T whose corresponding expression levels are close to x_{qij}. Thus, the final simulated expression level, $\hat{x}_{qij}$, is

$$\hat{x}_{qij} = x_{qij} + \epsilon_{qij} \ , \tag{5.12}$$

where ϵ_{qij} is drawn from $\mathcal{N}(0, \sigma^2_{qij})$. When $J = 10$, the simulated expression level for group three is shown in Fig 5.1 (a). It can be seen that there is more noise for the lower expressed genes than the highly expressed ones which is commonly observed in real data sets.

At the third step, in order to show the clustering improvement by including probe-level measurement error the corresponding probe-level variance of the simulated expression level is sampled from the real mouse data set processed by multi-mgMOS. Similar to the second step, since the gene expression level is highly correlated with its measurement error, the standard deviation for each simulated expression value, $\hat{\sigma}_{qij}$ is sampled from a subset of standard deviation calculated

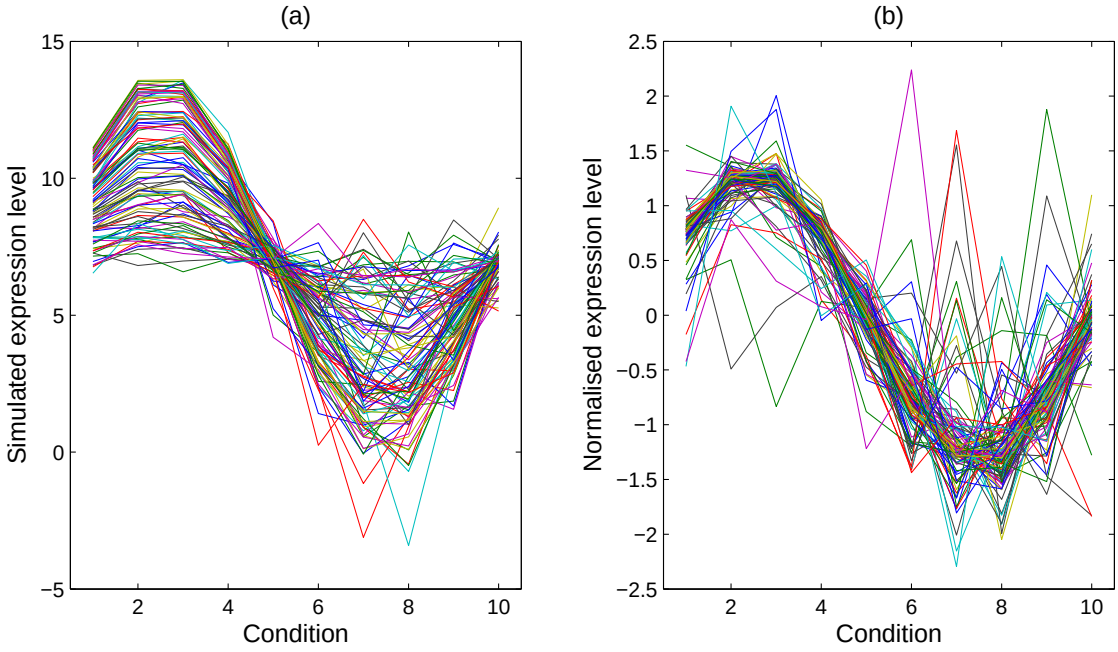

Figure 5.1: Simulated expression profiles for one group under 10 conditions. (a) is on log scale and (b) is the normalised profiles with zero mean and standard deviation one.

from multi-mgMOS whose corresponding expression levels are close to $\hat{x}_{qij}$. Fig 5.2 (a) shows the scatter plot of the sampled standard deviation against the simulated expression level for one randomly selected condition. It can be seen that the variance of the gene expression for the low expressed genes is generally larger than that for the highly expressed genes which is commonly observed in real data sets.

At the final step, the simulated expression level for each gene over all conditions is normalised by subtracting the mean expression level and dividing by the standard deviation such that the profile of each gene has zero mean and standard deviation one. The simulated standard deviation is also divided by the standard deviation of the expression level to show the corresponding measurement error of the normalised expression level. The normalised profile is shown in Fig 5.1 (b) when $J = 10$.

Since the true partition of the simulated data set is known, the agreement of the clustering results from different methods with the true partition can be assessed by the adjusted Rand index. The true number of groups, six, is selected arbitrarily for both MCLUST and PUMA-CLUST. Three sets of data sets are

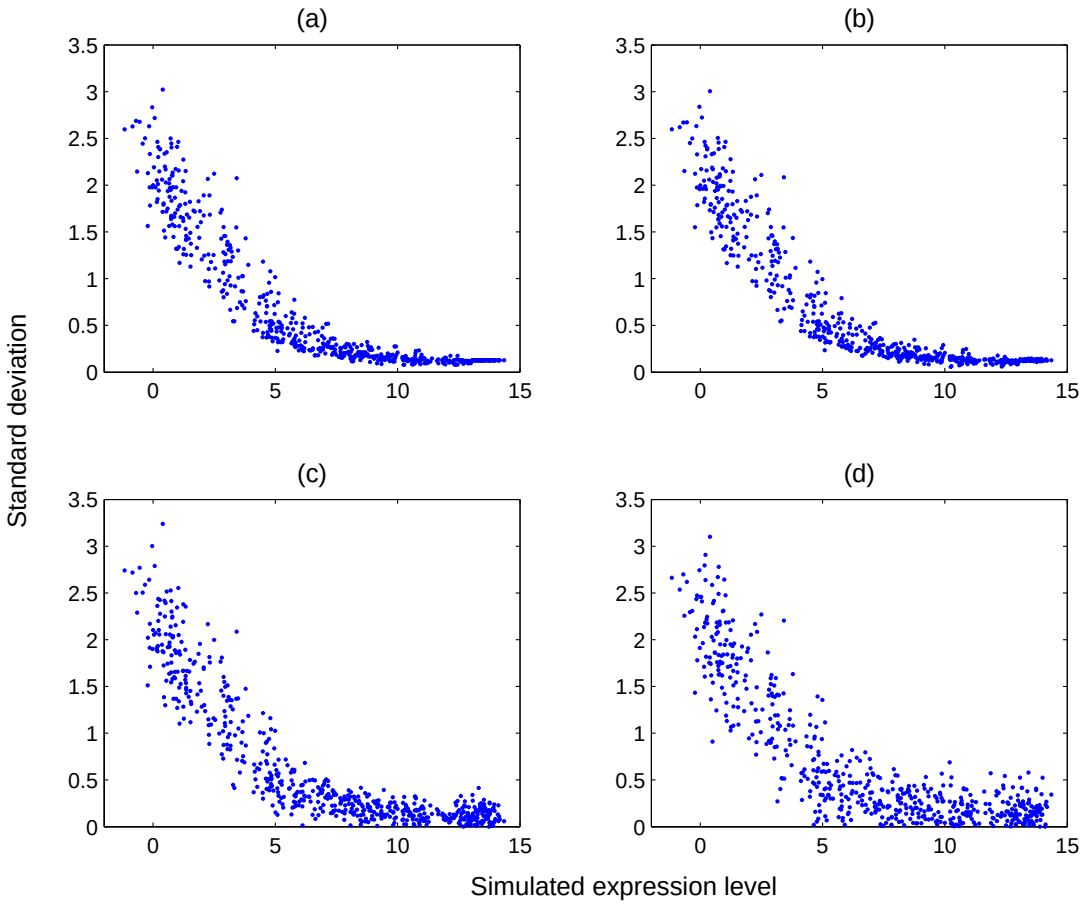

Figure 5.2: Scatter plots of standard deviation against the simulated gene expression level. The standard deviation in (a) is the sampled from multi-mgMOS results of the mouse data set, the standard deviation is randomly changed by adding a noise drawn from (b) $\mathcal{N}(0, 0.01)$, (c) $\mathcal{N}(0, 0.1)$ and (d) $\mathcal{N}(0, 0.2)$.

generated to evaluate the different performance of PUMA-CLUST with MCLUST with number of conditions 10, 20 and 30. For each set, 10 random simulated data sets are randomly generated. The average adjusted Rand index from PUMA-CLUST and MCLUST are shown in the first column of Fig 5.3. For the three sets of simulated data sets, PUMA-CLUST obtains obviously better performance compared with MCLUST. This shows the improvement of model-based clustering by including the variance of the each data point.

Including a Noise Group

In a real-world microarray data set, there are usually a certain fraction of genes whose expression levels are mainly random noise. These genes do not belong to any pattern group in the data set (Medvedovic et al., 2004). To assess the performance of PUMA-CLUST on this kind of data set, a group of random noise is added into the previously simulated data sets. The first generating step of the gene expression level for group seven is

$$x_{qij} = A_{qi} \ , \tag{5.13}$$

where A_{qi} is sampled from $U(0, 14)$. The following steps of the simulation are the same as those for the former six groups. Three sets of simulated data sets with 10 randomly generated data sets for each set are also sampled and the average adjusted Rand index for three cases with condition 10, 20, and 30 are shown in the second column of Fig 5.3. The number of groups for both MCLUST and PUMA-CLUST is arbitrarily assigned to seven. From the three plots it can be seen that the performance of the clustering from both PUMA-CLUST and MCLUST decreases with the inclusion of the group of noise, but PUMA-CLUST still outperforms MCLUST over all three noise levels with the three different data dimensions.

Testing the Robustness to Misspecified Technical Variance

The uncertainty of the gene expression in the simulated data sets generated above is sampled from multi-mgMOS results from the real mouse data set. It was assumed that the level of uncertainty is known but in practice it would be estimated using multi-mgMOS. An amount of noise was therefore added into the sampled standard deviation, $\hat{\sigma}_{qij}$ to test robustness to errors in estimating the measurement error variance. For the cases of six-group data sets and seven-group data sets, three kinds of random noise are added by sampling from $\mathcal{N}(0, 0.01)$, $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(0, 0.2)$. If the error-added variance is negative, its absolute value is used. The scatter plots of the error-added standard deviation against the simulated gene expression are shown in Fig 5.2 (b) – (d). Fig 5.3 gives the average adjusted Rand index of the clustering results from PUMA-CLUST on the error-added standard deviation for various cases. In the case of PC.01, the added noise is quite small so that the clustering results of PC.01 are very close

to the clustering results on the original simulated data. As the added noise gets higher, the performance of PUMA-CLUST decreases. This shows that clustering is most accurate when the measurement uncertainty is known, but is quite robust to errors in the estimate.

5.3.2 Clustering on a Real Mouse Time-course Data set

The improved performance of the new model, PUMA-CLUST, over the standard Gaussion mixture model on simulated data sets was shown in the previous section. Here, the performance of PUMA-CLUST is evaluated on the real periodic mouse data set (see Appendix A.3) compared to the standard mixture model, MCLUST.

Both PUMA-CLUST and MCLUST are applied on the first five time points which belong to the synchronised cycle and include 15 chips. For MCLUST the raw mouse data set is processed using the popular and accurate probe-level method GCRMA. For PUMA-CLUST the raw data is processed by multi-mgMOS as descried in Chapter 3. The clustering is performed on the 2,461 potential hair cycle-associated genes which were selected by Lin et al. (2004). The obtained expression level for each probe-set from both probe-level methods are normalised to have zero mean and standard deviation one. The calculation of BIC in Section 5.2.4 is used to determine the number of clusters for both methods. The calculated BICs at various number of clusters for the two methods are shown in Figure 5.4. It can be seen that the optimal BIC for PUMA-CLUST is obtained at K=22 and the optimal BIC for MCLUST is obtained at K=30. In order to make different clustering methods comparable, the number of clusters for each method should be the same. Therefore, the 22-cluster and the 30-cluster cases are compared separately. The 22 clusters obtained from PUMA-CLUST and MCLUST are shown in Figure 5.5 and Figure 5.6, and the 30 clusters obtained are shown in Figure 5.7 and Figure 5.8, respectively. For visualisation, the average expression level at each time point over replicates is shown for both the gene profile and the cluster center.

To assess if biologically-relevant clusters are created using the two methods, gene-annotation (GO) enrichment analysis is systematically performed for the individual clusters using DAVID 2006 (The Database for Annotation, Visualization and Integrated Discovery, Dennis et al. (2003)). The GO enrichment analysis allows the direct assessment of the biological significance for gene clusters found based on the enrichment of genes belonging to a specific GO functional category.

A meaningful GO enrichment analysis is to examine enriched categories of GO Biological Process at term level 5 and to select an enrichment cutoff at a p-value of 0.05.

From the GO enrichment analysis on the 22-cluster results for the two methods, PUMA-CLUST produced more clusters (21 of 22) with at least one enriched GO category in comparison to MCLUST (17 of 22), as shown in Figure 5.9 (a). A visual inspection of these MCLUST clusters without an enriched GO category indicates that four out of five of these clusters (Cluster #1,6,8,15) contain heterogeneous temporal expression profiles (i.e. not tightly clustered). Since the number of enriched GO categories found varies greatly among clusters (shown in Figure 5.10 (a)), the average number (13.1) of categories among the 22 PUMA-CLUST clusters is only slightly greater than the average among the MCLUST clusters (11.5). Hence, a more meaningful indicator of the distribution differences is the median number of categories: PUMA-CLUST clusters (14) and MCLUST clusters (7). The same enrichment analysis method was repeated using the 30 clusters from both methods, and the results still clearly indicate that the PUMA-CLUST method results in more biologically-meaningful clusters than the MCLUST method. Using 30 clusters, all clusters generated by PUMA-CLUST have at least one enriched GO categories, in comparison to only 21 out of 30 clusters created by MCLUST as shown in Figure 5.9 (b). The median number of enriched categories for PUMA-CLUST and MCLUST are 8 and 1, respectively, as shown in Figure 5.10 (b). Based on this GO enrichment analysis, it is evident that the PUMA-CLUST method generated more biologically-relevant clusters than the MCLUST method.

The MCLUST results on MAS5.0 and multi-mgMOS gene expression measurements were also tried and the performance was similar to results presented here with GCRMA, indicating that the probe-level summary method did not make a big difference. The improved performance of PUMA-CLUST is therefore due to the inclusion of probe-level measurement error.

5.4 Conclusion

This chapter demonstrates the usefulness of the measurement error in model-based clustering of gene expression data. A Gaussian mixture model with an

unequal volume spherical covariance matrix is augmented to incorporate probe-level measurement error. Results from simulated data sets and a real mouse time-course data set show that the inclusion of probe-level measurement error results in biologically meaningful clustering of gene expression data. The augmented clustering model has been implemented in an R package, *pumaclust*, for public use of the method.

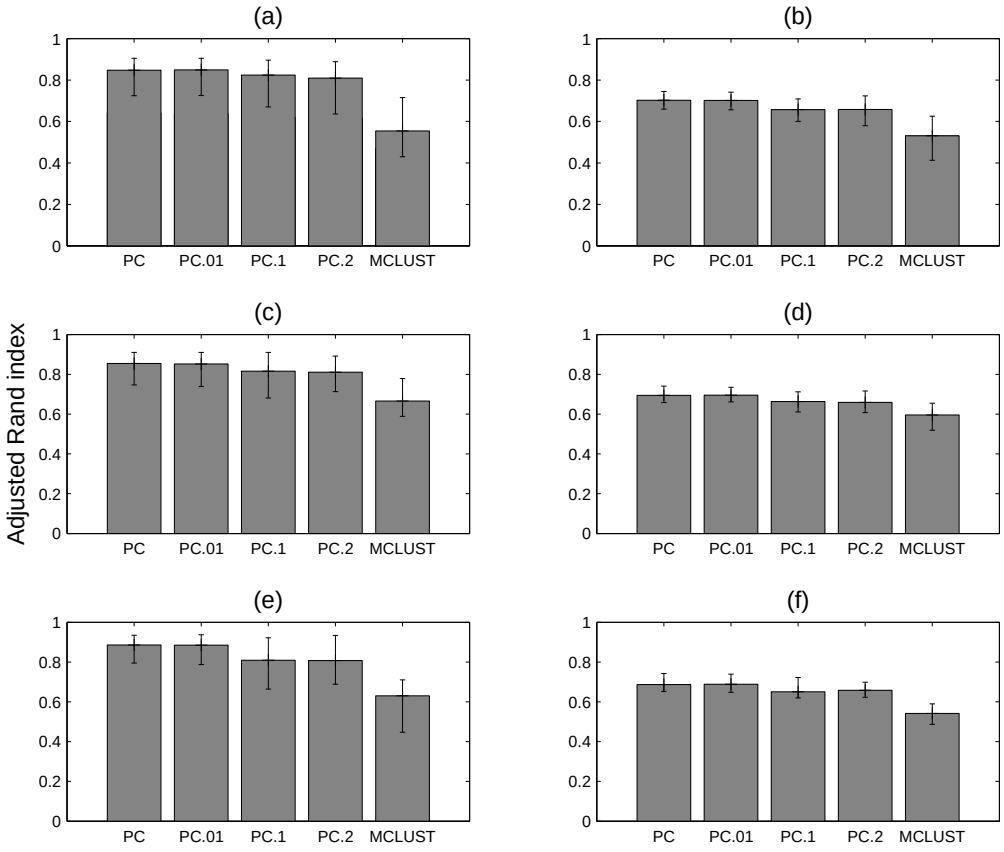

Figure 5.3: The average adjusted Rand index of the clustering results from PUMA-CLUST and MCLUST on the simulated data. The first column is for the six-group data set and the second column is for the seven-group data set with one noise group added. The upper panel shows results on data sets with 10 conditions, the middle panel is for 20 conditions and the lower panel is for 30 conditions. PC represents PUMA-CLUST results on the original simulated data. PC.01, PC.1 and PC.2 represent the PUMA-CLUST results on the data sets with added noise drawn from $\mathcal{N}(0, 0.01)$, $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(0, 0.2)$ respectively. The average adjusted Rand index is calculated over 10 simulated data sets for each plot and the range of the adjusted Rand index of each case is shown by error bars.

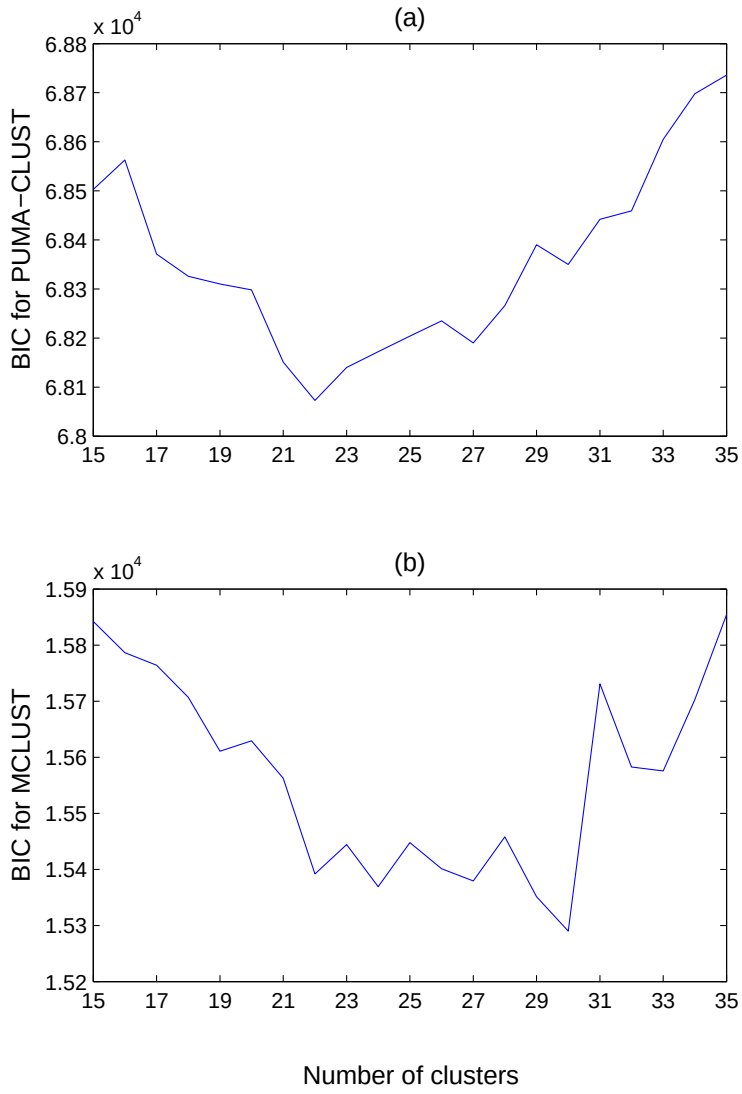

Figure 5.4: BIC for (a) PUMA-CLUST and (b) MCLUST at various number of clusters on the 2,461 potential hair growth-associated genes from the mouse time-course data set. PUMA-CLUST obtains the minimum BIC at K=22 and MCLUST obtains at K=30.

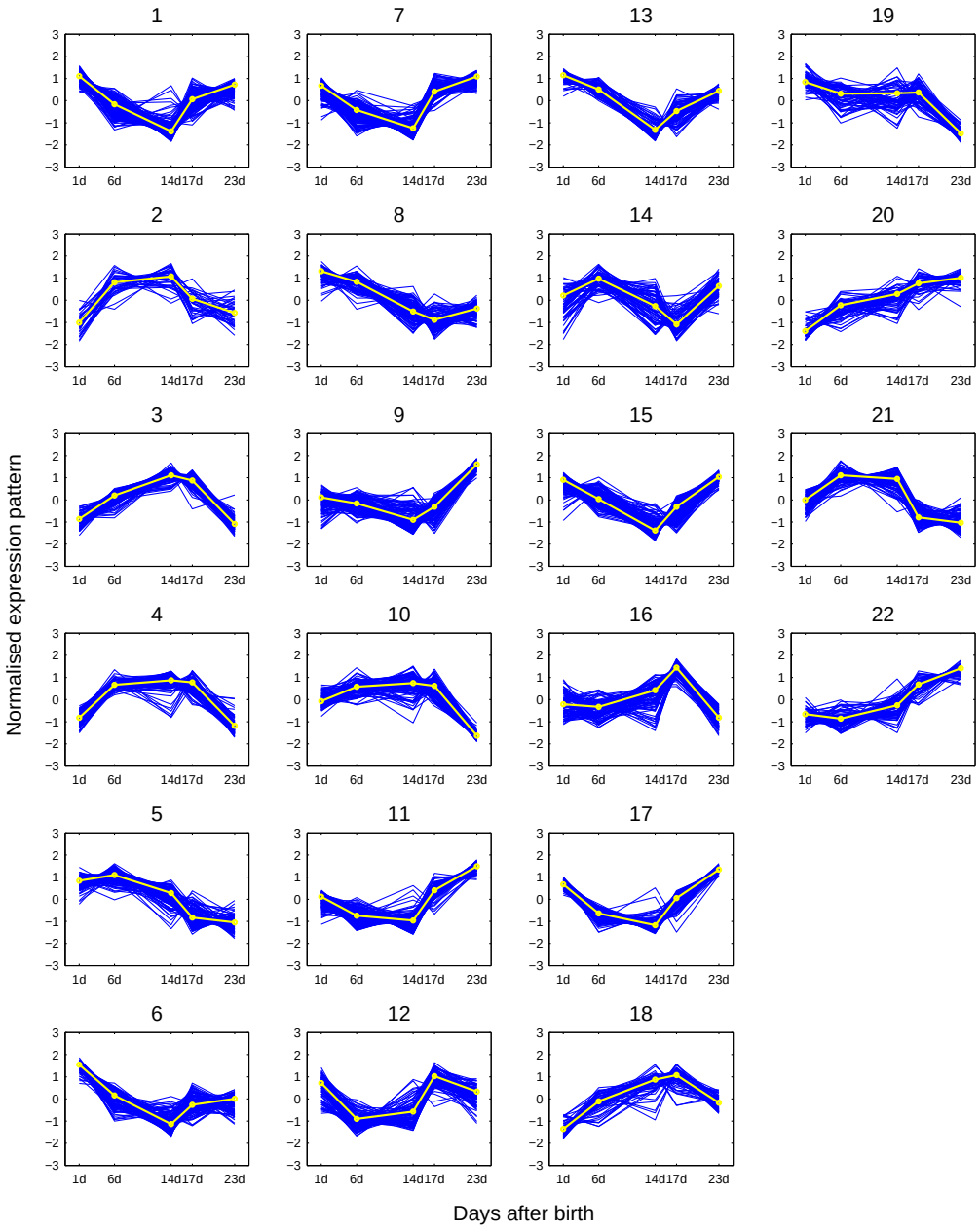

Figure 5.5: Expression pattern clusters of the expression patterns from PUMA-CLUST on the 2,461 potential hair-growth-associated genes of the mouse time-course data set when K=22. The expression pattern for each probe-set is shown in dark line for five time points. The light line on each plot is the clustering center for each group. At each time point, the expression value is the average of the three replicated measurements.

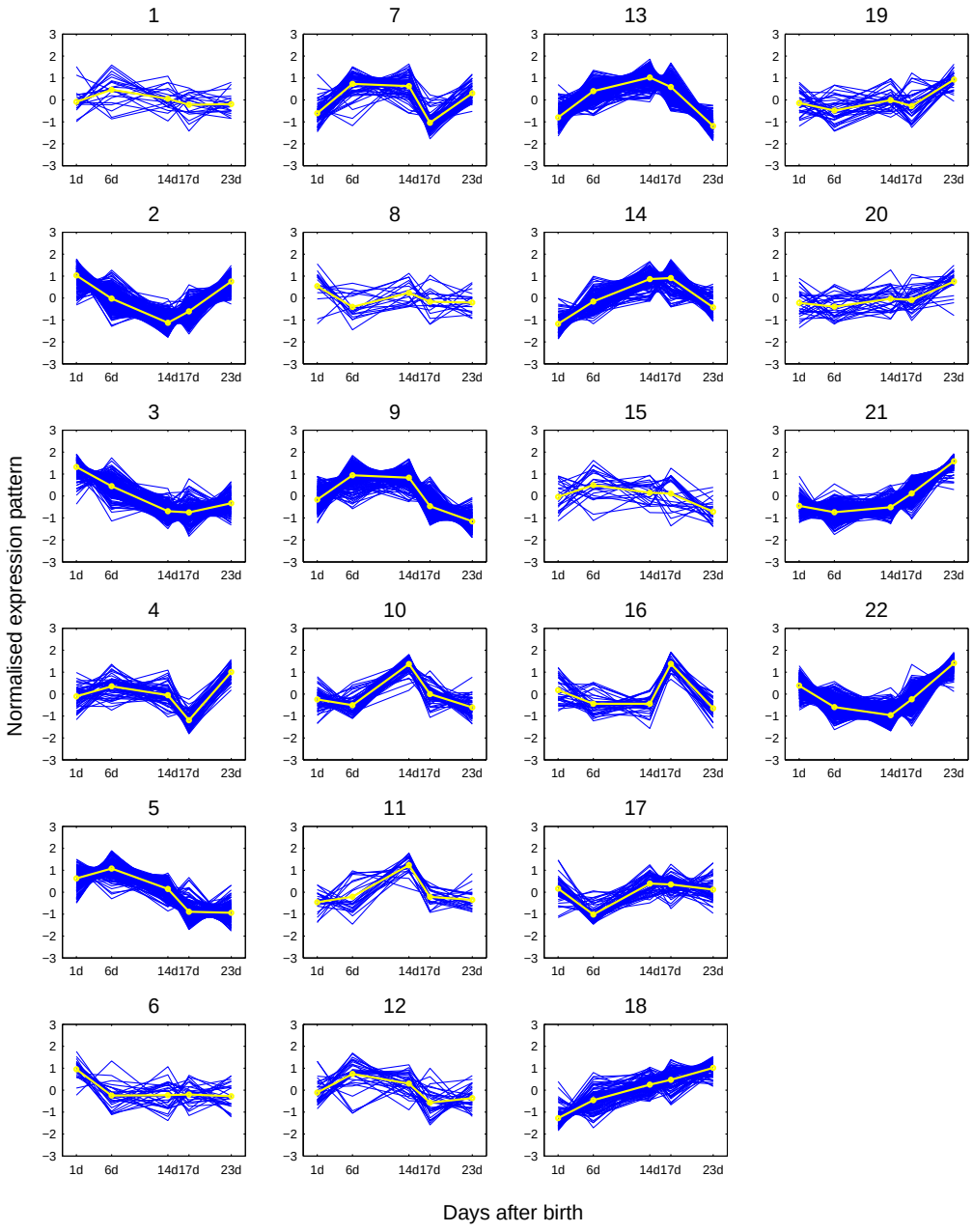

Figure 5.6: Expression pattern clusters of the expression patterns from MCLUST on the 2,461 potential hair-growth-associated genes of the mouse time-course data set when K=22. The expression pattern for each probe-set is shown in dark line for five time points. The light line on each plot is the clustering center for each group. At each time point, the expression value is the average of the three replicated measurements.

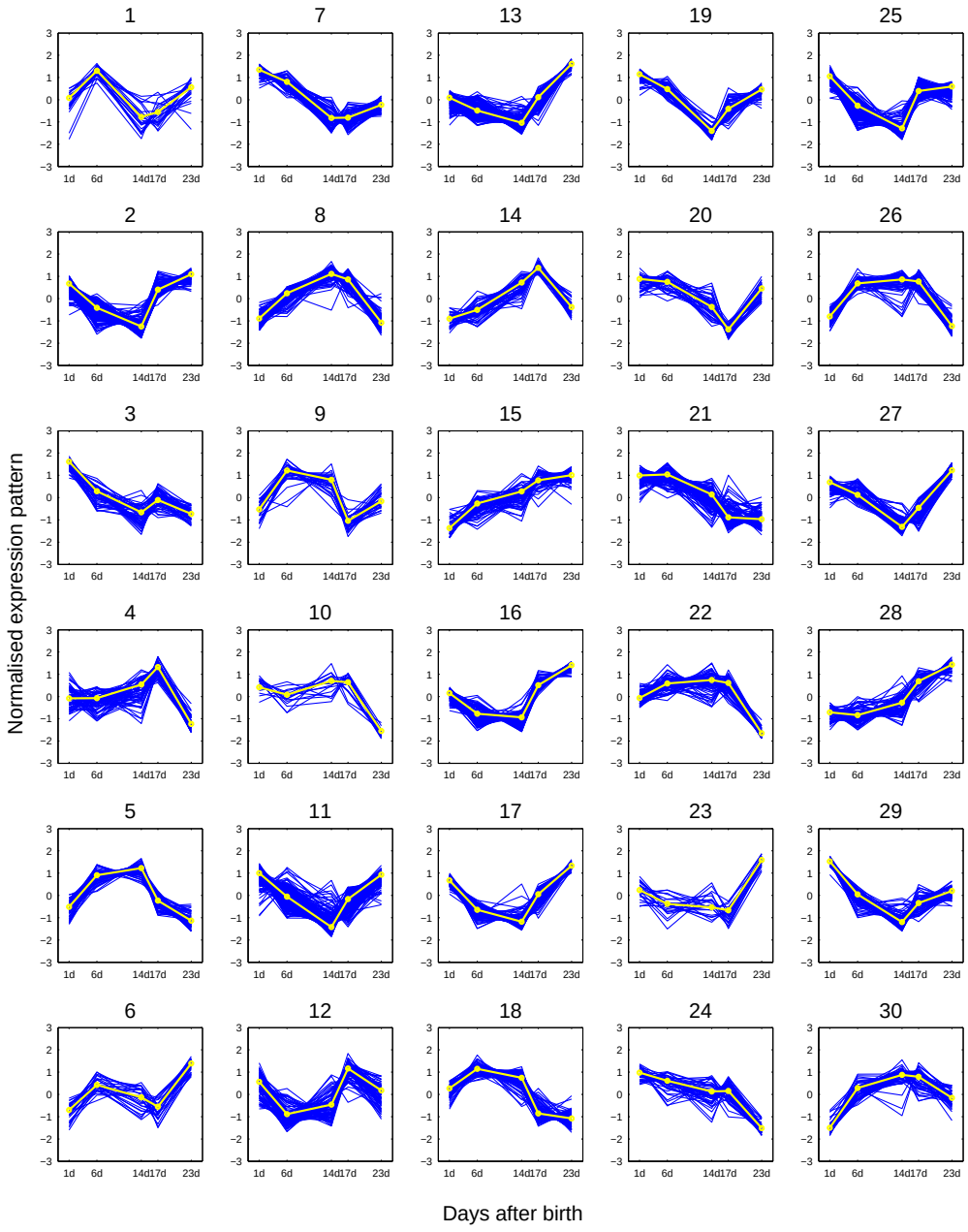

Figure 5.7: Expression pattern clusters of the expression patterns from PUMA-CLUST on the 2,461 potential hair-growth-associated genes of the mouse time-course data set when K=30. The expression pattern for each probe-set is shown in dark line for five time points. The light line on each plot is the clustering center for each group. At each time point, the expression value is the average of the three replicated measurements.

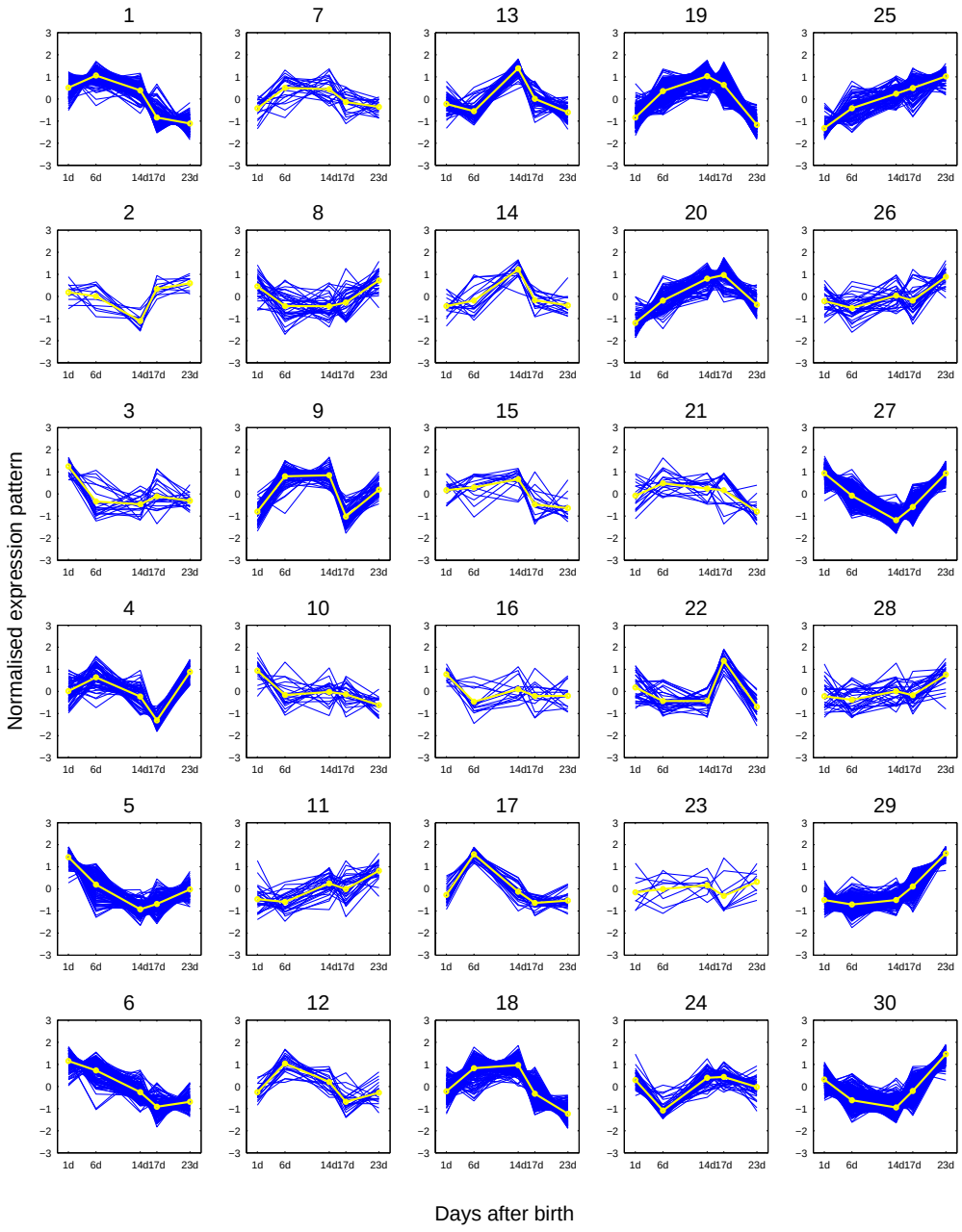

Figure 5.8: Expression pattern clusters of the expression patterns from MCLUST on the 2,461 potential hair-growth-associated genes of the mouse time-course data set when K=30. The expression pattern for each probe-set is shown in dark line for five time points. The light line on each plot is the clustering center for each group. At each time point, the expression value is the average of the three replicated measurements.

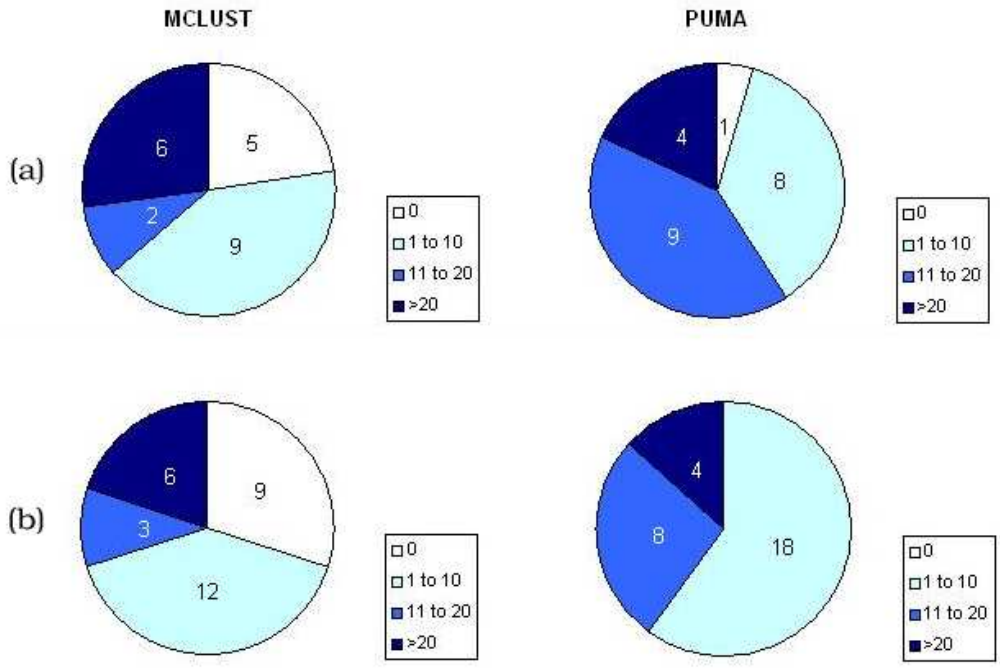

Figure 5.9: Comparison of the number of clusters found with the indicated ranges of enriched categories for MCLUST and PUMA-CLUST clusters using (a) 22 clusters and (b) 30 clusters. For both comparisons, the enriched categories were found using GO Biological Process term level 5, enrichment cutoff at p-value of 0.05, and mouse (*Mus Musculus*) as the population background.

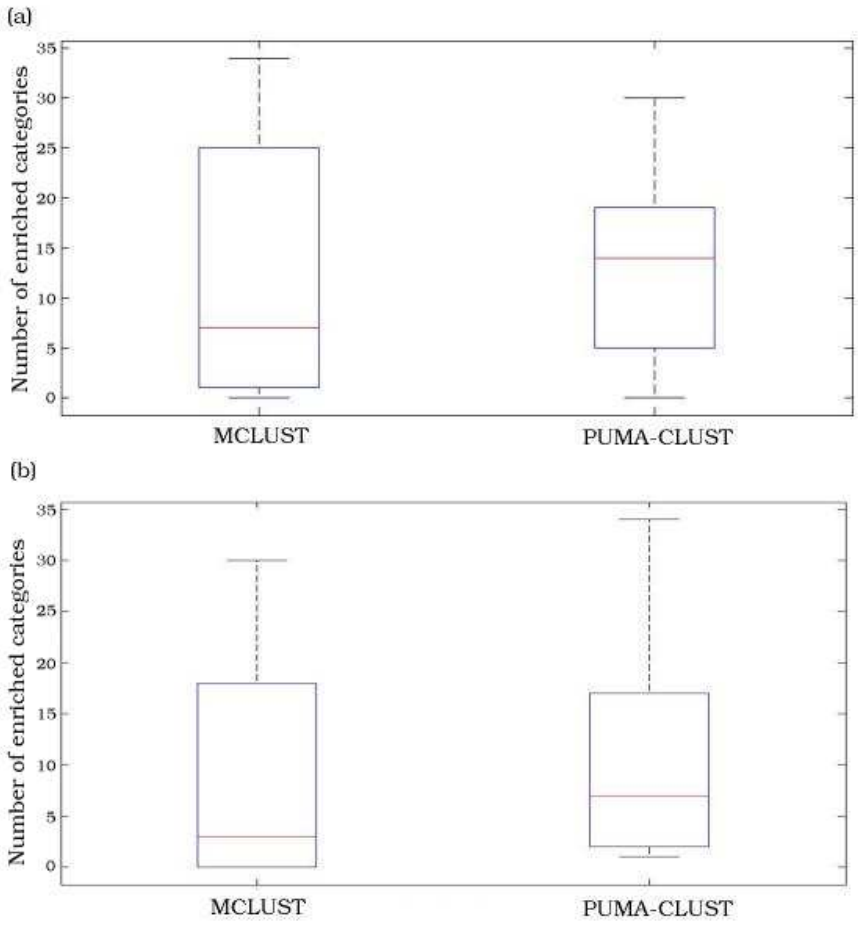

Figure 5.10: Boxplot of the number of enriched categories for MCLUST and PUMA-CLUST clusters using (a) 22 clusters and (b) 30 clusters. The boxes show the lower quartile, median, and upper quartile values. The dotted lines show the extent of the rest of the data. The number of enriched categories for MCLUST has larger variance than that for PUMA-CLUST.

Chapter 6

Conclusion

6.1 Thesis Summary

Microarray data analysis is challenging due to the huge amount of variability existing in the process of the experiment. This hinders the discovery of biological knowledge from microarray experiments. Many methods have been proposed to handle the noise in microarray data at various levels of the analysis. For the probe-level analysis of microarray data, most methods provide only a single point estimate of gene expression level discarding the uncertainty of this measurement. Consequently, existing methods for the downstream analysis do not consider the measurement error of expression level and this may result in inaccurate biological conclusions. Because of their natural representation of uncertainty, probabilistic models were developed in this thesis to handle the noise associated with microarray data.

In Chapter 3, a probabilistic probe-level analysis model, multi-mgMOS, was proposed based on the previously developed models, gMOS and mgMOS. Compared with the existing probabilistic probe-level model, BGX, the likelihood function of multi-mgMOS can be written in a closed form and thus the ML estimate is calculated more efficiently than the MCMC implementation of BGX. Probe-specific effects are an important characteristic of probe-level microarray data. As a multi-chip model, multi-mgMOS shares the probe effect across all chips of the same type and obtains more accurate results than its previous single chip versions, gMOS and mgMOS. The uncertainty associated with the estimated gene expression level is also shown to be useful in the downstream analysis.

In Chapter 4, a Bayesian hierarchical model was presented to detect differential gene expression. The uncertainty obtained from multi-mgMOS was included in the model. Due to the variance added to each data point the model is intractable. Three methods were used to the handle the intractability of the model: MAP approximation, a variational method and MCMC. The MAP approximation cannot obtain accurate results and MCMC is too computationally demanding. The variational method is therefore adopted to calculate the model since it obtains relatively fast computation and more accurate results than MAP estimation. The new model is compared with the popular method Cyber-T on a spike-in data set and a real time-course data set. Results showed that the incorporation of the probe-level measurement error improves the accuracy of detecting differential gene expression.

In Chapter 5, it was demonstrated that the measurement error improved the performance of the model-based clustering of gene expression data. A Gaussian mixture model with unequal volume spherical covariance matrix is augmented to incorporate probe-level measurement error. Unlike the intractability of the Bayesian model in Chapter 4, the computation of the augmented Gaussian mixture model is straightforward by using the standard EM algorithm for mixture models. Results from simulated data sets and a real mouse time-course data set show that the inclusion of probe-level measurement error leads to biologically meaningful clustering of the gene expression data.

In summary, the redundancy in the probe-level microarray data makes it possible to obtain a level of uncertainty for the gene expression measurement using an improved probabilistic model, multi-mgMOS. The inclusion of uncertainty for each data point in the probabilistic models (a Bayesian hierarchical model and a mixture model) for data analysis obtains better results when the data is noisy. However, a disadvantage is that the models often become more complicated or intractable since there is an additional term to handle the uncertainty. In these cases, advanced machine learning techniques can be used to solve this problem. With the various inference approximation approaches and advanced optimisation methods available, the augmented models in this thesis are computed and improve upon the performance of the original models which do not include the probe-level measurement error. The software implemented in this thesis is available from http://umber.sbs.man.ac.uk/resources/puma/.

6.2 Future Work

6.2.1 Improving the Computation Time of multi-mgMOS

Although the computational efficiency of multi-mgMOS over the other previously developed alternatives is shown to be favorable in Chapter 3, for very large data sets, especially for those involving non-human chips, the method still takes time due to the iteration for the optimisation of ϕ. The reason for this is that the prior distribution of ϕ is obtained from the spike-in data set of the human chip and it reflects the characteristics of this type of chip. For other types of chip, there are no spike-in data sets available to estimate a proper prior for ϕ and ϕ therefore needs more iterations to be estimated. In order to accelerate the computation speed of the software, the current version of the R package *mmgmos* sets ϕ zero when non-human chips are used to avoid the large number of iterations for the estimation of ϕ.

It is obvious that setting ϕ to an arbitrary value affects the accuracy of results. In the long term, ϕ should be pre-calculated using well designed spike-in data sets or using the data set which is under study. For example, using the filtered high expressed probe-pairs the prior distribution of ϕ can be obtained. This prior should be more appropriate for the data set of interest and accelerate the iteration for estimating ϕ. Since the performance of multi-mgMOS II is very close to multi-mgMOS I in many cases, but more computationally efficient, one can also pre-calculate the value of ϕ using the data under study according to the middle three bases in each probe-pair. A similar strategy has been used in GCRMA where sequence-specific probe effects are computed and PDNN (Zhang et al., 2003) where the sequence-specific hybridisation energy is pre-calculated. This approach can also be adopted into multi-mgMOS to accelerate the computation.

6.2.2 Improving Background Correction of multi-mgMOS

In the assumption of multi-mgMOS in (3.8), h_{gjc} is the non-specific binding and background of the jth probe-pair. It is assumed in (3.10) that h_{gjc} follows a gamma distribution with probe-specific inverse scale parameter b_{gj}. Therefore, h_{gjc} is assumed to be probe-specific. It is reasonable to assume that the non-specific hybridisation is probe-specific. However, the overall background is chip-specific and should be independent of the probe sequence. It can be seen from

Figure 3.2 that for high expressed genes the specific and non-specific signal is much higher than the background signal, so the effects of the background is negligible. For the weakly expressed genes, the effects of the background is more important. Treating background and the non-specific hybridisation in the same way might lead to inaccurate expression measurements, especially for those lower expressed genes.

From the "No RNA" experiment in Wu et al. (2004) where the sample had no target RNA and all observed intensities represent the background optical noise, the overall logged background fits well a Gaussian distribution with mean 5 and variance 0.1. Due to the small variance of the fitted Gaussian, Wu et al. (2004) treats the background as constant and estimates it with the minimum intensity on each chip. The latest version of the R package, *mmgmos*, also subtracts such a constant from the intensity of each probe to eliminate the effect of the background. In the future, a more principled approach can be devised to address this problem.

6.2.3 Improving Data Fit of multi-mgMOS

The improved performance of the new probabilistic model multi-mgMOS has been shown in Chapter 3 in terms of the accuracy of the measured gene expression level. The usefulness of the measurement error associated with the expression level has also been demonstrated in Chapter 4 and 5. However, when the consistency of the model is examined, some deficiencies can be seen, as shown in Figure 6.1. The solid loess smooth line shows the average relationship between probe-level measurement (i.e. technical) error and the expression level. The dashed loess smooth line gives an estimate of the average tendency of total error against the expression level. The total variance includes technical and biological components. Theoretically, the dashed line should be above the solid line. However, this fact holds only for high expressed genes. At the lower end, the average total variance is less than the related technical variance. It is possible that the technical variance for low expressed genes is over-estimated by multi-mgMOS and this leads to the low estimates of the total variance. Alternatively, the mean estimate may be biased and this could result in reduced variance over replicates.

The inconsistency of the model may be explained by Figure 6.2 which shows the contours of the joint distribution of PM and MM intensities for one probe-set after the parameters of multi-mgMOS are estimated. For the low expressed case shown in (a), there are almost no data points within the first contour level which

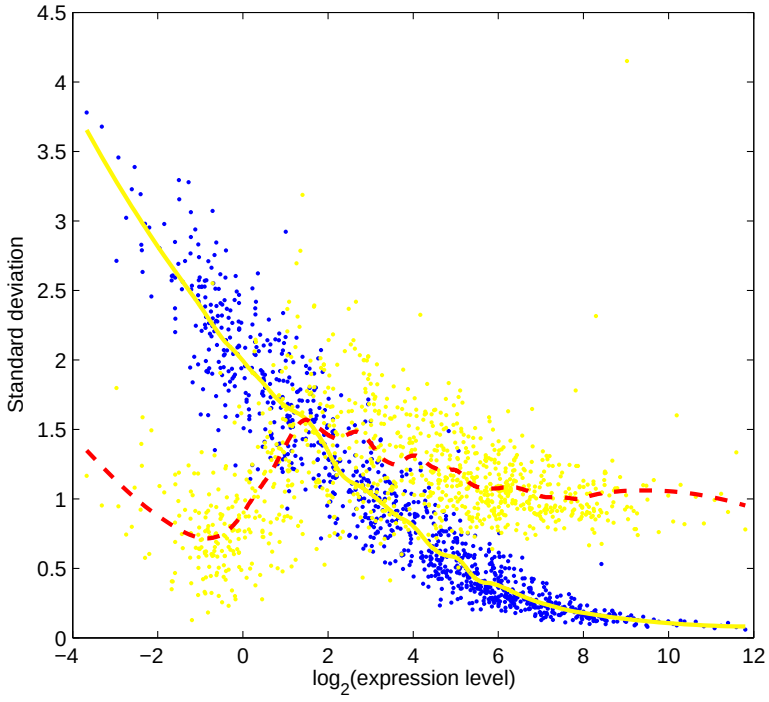

Figure 6.1: The loess fits to the technical error as a function of expression level (solid line) and to the total error of the expression level (dashed line) for data set A in Chapter 3. The blue dots are the average technical error against the average expression level over three replicates for condition 1. The yellow dots are the total error of the expression level against the same expression level as the blue dots.

shows high density of the joint distribution. Also, the fitted joint distribution spreads more than the observed data. It can be seen that the model does not fit the data well in this case. For the high expressed case of the same gene in (b), the fit of the model to the observed data is better than for the low expressed genes. The misfit of the model may be caused by the assumption of the model that PM and MM intensities follow gamma distributions. It is possible that the gamma distributions cannot account well for the observed data on the raw scale. The Gaussian distribution on a log scale may fit the data better but it is then harder to model the joint distribution of PM and MM intensities, since a sum of two log-normal distribution does not lead to a simple distribution.

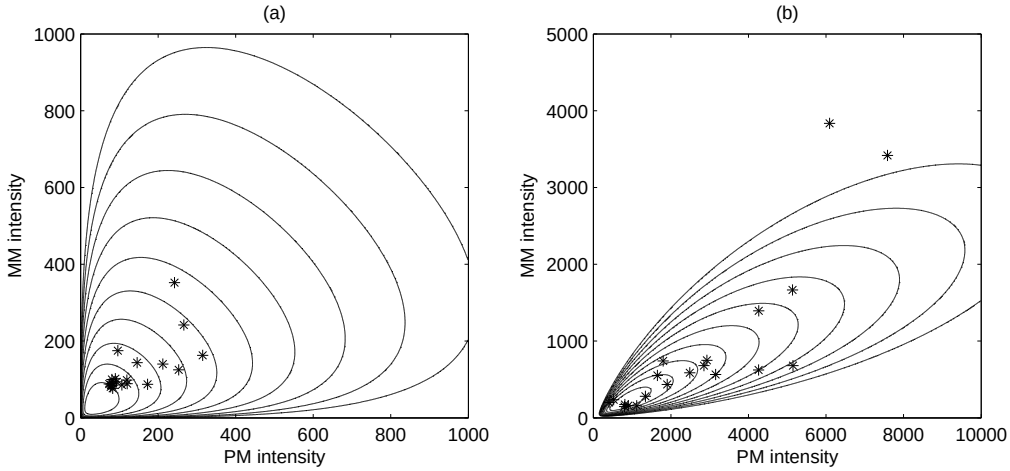

Figure 6.2: Contours of the joint distribution of PM and MM intensities of spike-in gene 1 in data set A of Chapter 3 at concentration (a) 0.5 and (b) 100 pM after the parameters of multi-mgMOS are estimated. The stars show the observed PM and MM intensities.

6.2.4 Integrating Nomalisation in Downstream Analysis

In the microarray experiment involving multiple chips, different amounts of RNA sample may be used for different chips due to the variability existing in the sample preparation. This will lead to a different overall brightness for different chips. For this reason, normalisation is usually conducted in order to perform the analysis across different chips. In the description of multi-mgMOS in (3.13), the normalisation can be partly accounted for by the shared parameters b_{gj} and ϕ across chips. However, the data may not be well normalised in some cases. Currently, an additional global scaling normalisation is therefore adopted before the further downstream analysis is performed. The expression values are centered at the same mean or median for each chip and the measurement error of the expression value is adjusted accordingly.

More principally, the normalisation can be integrated in the downstream analysis models. In Chapter 4 and 5, the observed logged expression level $\hat{x}_{gc}$ for gene g under condition c can be expressed as $\hat{x}_{gc} = x_{gc} + \epsilon_{gc}$, where x_{gc} is the true expression level and ϵ_{gc} is the zero-mean Gaussian measurement error. When the normalisation is considered in the downstream analysis, the observed expression level can be expressed as

$$\hat{x}_{gc} = x_{gc} + \delta_c + \epsilon_{gc} \ , \tag{6.1}$$

where δ_c is the chip-specific normalisation factor. Equation (6.1) can be plugged in downstream probabilistic analysis models. By treating δ_c as a random variable, the normalisation can be performed among the further analysis.

6.2.5 Possible Improvement of PUMA-CLUST

In order to obtain confidence in the gene expression measurements, replicates are usually used in experiments. If each chip is treated as an individual condition in the data, the clustering will become increasingly difficult with the increase in the number of chips used in the experiment due to the noisy nature of the data. The approach in Chapter 5 augments the standard model which treats each individual chip as a single independent data point. This standard model does not consider the replicate information where repeated measurements are available for time points. Clustering on repeated measurements has been considered by Yeung et al. (2003), Lin et al. (2004) and Medvedovic et al. (2004), but all of these approaches do not include the probe-level measurement error. So including both probe-level noise and replicate information in the clustering is the focus of future work.

One solution may be to use the Bayesian hierarchical model (described in Chapter 4), which combines replicated measurements for each condition by considering probe-level measurement error, to obtain a single value of gene expression measurement for each condition, $\tilde{x}_{ij}$, along with the measurement error, $\tilde{\beta}_{ij}$, where j indexes the different conditions or time points. As a result of this model, the probe-level measurements, x_{ijr}, are generated from the Gaussian distribution $\mathcal{N}(\tilde{x}_{ij}, \tilde{\beta}_{ij})$, and r is the index of replicates for condition j. We can then cluster on $\tilde{x}_{ij}$ and $\tilde{\beta}_{ij}$ using the augmented mixture model in (5.5).

The multiple step process of the clustering on replicated gene expression data may be questioned in terms of introducing new variability to the original data. Another possible approach is to augment the mixture model in Lin et al. (2004), which is able to deal with replicated measurements within the mixture model,

$$p(x_i|k,\theta_k) = \int_e \mathrm{d}e \prod_{jr} p(x_{ijr}|e_j, s_{kj}^2) p(e_j|\mu_{kj}, \sigma_{kj}^2) \ , \tag{6.2}$$

where e_j are the true expression levels, which are unobserved variables, and s_{kj}^2 is the gene-independent between-replicate variance for condition j. Each Gaussian component in this model has different variance for each condition which reflects

the variability of the data. The complete set of parameters is $\theta = \{\mu, \sigma^2, s^2, P(k)\}$. The model in (6.2) can be augmented to incorporate the probe-level measurement error,

$$p(x_i|k, \theta_k) = \int_e \mathrm{d}e \prod_{jr} p(x_{ijr}|e_j, s_i^2 + \beta_{ijr}) p(e_j|\mu_{kj}, \sigma_k^2) \ . \tag{6.3}$$

Similar to the Bayesian hierarchical model in Chapter 4, the between-replicate variance s_i^2 for each gene is shared over conditions. Since the between-replicate variance is gene-specific, the number of free parameters is much larger than that of the original model in (6.2). In order to work out this model, a modified M-step in the standard EM algorithm in Chapter 5 may be adopted. This strategy is still under research.

6.2.6 Modelling the Measurement Error As an Unknown Variable

The approaches in this thesis are to produce the measured expression and the associated measurement variance in the probe-level analysis, then to include the estimated measurement variance in the high level analysis. Since the measurement error is not known exactly and it is estimated in the low level analysis, it is itself associated with error (as discussed in Section 5.3.1). Therefore, instead of using the estimated measurement uncertainty in the high level analysis, the uncertainty can also be modelled by a prior distribution.

For example, in the hierarchical Bayesian model in (4.4) ν_{ij} is the estimated measurement error associated with the measured expression level x_{ij} for gene i. Considering the estimated error of ν_{ij} , it can be modelled with a prior,

$$\nu_{ij} \sim \mathrm{Ga}(a_i, b_i/x_{ij}) \ , \tag{6.4}$$

where a_i and b_i are the unknown hyper-parameters, and the scaling factor x_{ij} is introduced to model the dependence between ν_{ij} and x_{ij}. Therefore, the prior is a conditional gamma distribution. Obviously, this approach introduces more parameters and model complexity. However, the advantage of the Bayesian framework is that it can deal with such problems. With the various advanced machine learning techniques available, it may then be possible to model the data more appropriately and possibly improve the performance of the current models further.

Bibliography

Affymetrix (2002). *Statistical algorithms reference guide*. Affymetrix Inc, Santa Clara CA.

Akaike, H. (1973). Information theory and extension of the maximum likelihood priniple. In Petrov, B. and Csaki, S., editors, *2nd International Symposium in Information Theory*, pages 267–281, Budapest. Akademia Kiado.

Antonellis, K. J., Beazer-Barclay, Y. D., Elashoff, M., Jelinsky, S. A., Whitley, M. Z., Brown, E. L., and Scherf, U. (2001). *Optimization of an external standard for the normalization of Affymetrix GeneChip arrays*. Gene Logic Inc.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 2nd edition.

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

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.

Bassett, D. E. J., Eisen, M. B., and Boguski, M. S. (1999). Gene expression informatics–it's all in your mine. *Nature Genetics*, 21(Suppl 1):51–55.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley & Sons, Chichester, UK.

Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Process.*, 83:825–833.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Brown, T. (2002). *Genomes*. BIOS Scientific Publishers, Oxford, UK, 2nd edition.

Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6:R16.

Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K., and Kreder, D. E. (2002). Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biology*, 3(1):RESEARCH0005.

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20:323–331.

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature*, 227:561–563.

Delmar, P., Robin, S., and Daudin, J. J. (2005). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, 21:502–508.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

Dennis, G., Jr, Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23:1499–1501.

Dobson, A. (1990). *An Introduction to Generalised Linear Models*. Chapman & Hall, London, UK.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868.

Fraley, C. and Raftery, A. (2002a). Mclust: software for model-based cluster analysis. *J. Classification*, 16:297–306.

Fraley, C. and Raftery, A. (2002b). Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.*, 97:911–931.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Jounal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Ghahramani, Z. and Beal, M. (2001). Graphical models and variational methods. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods — Theory and Practice*, pages 161–177. MIT Press.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286.

Gill, P. E., Murray, W., and Saunders, M. A. (2002). SNOPT: an SQP algorithm for large-scale constrained optimization. *SIAM Journal on Optimization*, 12:979–1006.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully bayesian integrated approach to the analysis of Affymetrix Genechip data. *Biostatistics*, 4:249–264.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, D. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Irizarry, R. A., Wu, Z., and Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22:789–794.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Krohn, K., Eszlinger, M., Paschke, R., Roeder, I., and Schuster, E. (2005). Increased power of microarray analysis by use of an algorithm based on a multivariate procedure. *Bioinformatics*, 21:3530–3534.

Lawrence, N., Milo, M., Niranjan, M., Rashbass, P., and Soullier, S. (2004). Reducing the variability in cdna microarray image processing by bayesian inference. *Bioinformatics*, 20:518–526.

Lee, M. L., Kuo, F., Whitmore, G., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Science USA*, 97:9834–9839.

Li, C. and Wong, W. (2001a). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science USA*, 98:31–36.

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2(8):RESEARCH0032.

Lin, K. K., Chudova, D., Hatfield, G. W., Smyth, P., and Andersen, B. (2004). Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *Proceedings of the National Academy of Science USA*, 101:15955–15960.

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2005). A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644.

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*. in press.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.*, 14(13):1675–1680.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20:1222–1232.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Milligan, G. and Cooper, M. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441–458.

Milo, M., Fazeli, A., Niranjan, M., and Lawrence, N. D. (2003). A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Society Transactions*, 31:1510–1512. part 6.

Milo, M., Holley, M. C., Rattray, M., Niranjan, M., and Lawrence, N. D. (2004). Improving temporal gene expression profiles with probabilistic models. technical report.

Naef, F., Lim, D. A., Patil, N., and Magnasco, M. (2002). DNA hybridization to mismatched templates: a chip study. *Physical Review E65*. 040902.

Naef, F. and Magnasco, M. O. (2002). Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68. 011906.

Najarian, K., Zaheri, M., Rad, A. A., Najarian, S., and Dargahi, J. (2004). A novel Mixture Model Method for identification of differentially expressed genes from DNA microarray data. *BMC Bioinformatics*, 5:201.

Oehlert, G. (1992). A note on the delta method. *Amer. Stat.*, 46:27–29.

Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801.

Phimister, B. (1999). Going global. *Nature Genetics Supplement*, 21:1.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427.

Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, UK.

Rocke, D. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8:557–569.

Sanguinetti, G., Milo, M., Rattray, M., and Lawrence, N. D. (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21:3748–3754.

Schena, M. (2003). *Microarray Analysis*. John Wiley & Sons, Hoboken, New Jersey.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6:461–464.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34:166–176.

Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, 17(Suppl 1):S243–S252.

Siegmund, K., Laird, P., and Laird-Offringa, I. (2004). A comparison of cluster analysis methods using dna methylation data. *Bioinformatics*, 20:1896–1904.

Slonim, D. K. (2002). From pattern to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32(Suppl):502–508.

Southern, E., Mir, K., and Shchepinov, M. (1999). Molecular interactions on microarrays. *Nature Genetics*, 21(1 Suppl):5–9.

Spellucci, P. (1998). A SQP method for general nonlinear programs using only equality constrained subproblems. *Mathematical Programming*, 82:413–448.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 22:2907–2912.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–285.

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310:1152–1158.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

Vermaak, J., Lawrence, N. D., and Prez, P. (2003). Variational inference for visual tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, pages 773–780.

Wang, G., Kossenkov, A. V., and Ochs, M. F. (2006). LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 7:175.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917.

Yang, Y., Xiao, Y., and Segal, M. (2005). Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21:1084–1093.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987.

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, 4:R34.

Zhang, L., Miles, M. F., and Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, 21(7):818–821.

Appendix A

Data Sets

A.1 Affymetrix HG-U95a Spike-in Data Set

This data set was provided by Affymetrix for the purposes of developing and comparing expression algorithms. It was used to develop and validate the Affymetrix MAS 5.0 expression algorithm. This data set contains 14 spiked-in transcripts in 14 experimental groups (conditions) with a Latin Square design. The concentration of the 14 experiment groups in the first experiment is 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 picoMolar (pM). Each subsequent experiment rotates the concentrations by one group. Among the 14 spiked-in transcripts probe-set 407_at is an exception which doesn't follow the original Latin square. Table A.1 shows how the spike-in concentrations are arranged for each of the experimental groups. The data set consists of 59 arrays and there are three replicate arrays for each group except group c for which there was only two replicates. Replicates are divided into three groups, 1521, 1532 and 2353. For example, the three replicates of experimental group a is 1521a, 1532a and 2353a. Affymetrix states that two of the probe-sets, 407_at and 36889_at, have poorly behaving probe pairs and should be excluded from the analysis. This data set is available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx. The known concentrations of the 14 transcripts give the true information of the expression level for the related 14 probe-sets and this is the standard to judge the performance of different expression algorithms.

Group	37777_at	684_at	1597_at	38734_at	39058_at	36311_at	36889_at
a	0	0.25	0.5	1	2	4	8
b	0.25	0.5	1	2	4	8	16
c	0.5	1	2	4	8	16	32
d	1	2	4	8	16	32	64
e	2	4	8	16	32	64	128
f	4	8	16	32	64	128	256
g	8	16	32	64	128	256	512
h	16	32	64	128	256	512	1024
i	32	64	128	256	512	1024	0
j	64	128	256	512	1024	0	0.25
k	128	256	512	1024	0	0.25	0.5
l	256	512	1024	0	0.25	0.5	1
m,n,o,p	512	1024	0	0.25	0.5	1	2
q,r,s,t	1024	0	0.25	0.5	1	2	4

Group	1024_at	36202_at	36085_at	40322_at	407_at	1091_at	1708_at
a	16	32	64	128	0	512	1024
b	32	64	128	256	0.25	1024	0
c	64	128	256	512	0.5	0	0.25
e	128	256	512	1024	1	0.25	0.5
e	256	512	1024	0	2	0.5	1
f	512	1024	0	0.25	4	1	2
g	1024	0	0.25	0.5	8	2	4
h	0	0.25	0.5	1	16	4	8
i	0.25	0.5	1	2	32	8	16
j	0.5	1	2	4	64	16	32
k	1	2	4	8	128	32	64
l	2	4	8	16	256	64	128
m,n,o,p	4	8	16	32	512	128	256
q,r,s,t	8	16	32	64	1024	256	512

Table A.1: The Latin square arrangement of Affymetrix HG-U95a spike-in data set. Concentrations are in picoMolar (pM). Experiments n, o and p are replicates of experiment m and experiments r, s and t are replicates of experiment q.

Gene No	1	2	3	4	5	6	7	8	9	10	11
Group	BioB-5_at	BioB-M_at	BioB-3_at	BioC-5_at	BioC-3_at	BioDn-3_at	DapX-5_at	DapX-M_at	DapX-3_at	CreX-5_at	CreX-3_at
a	0.5	37.5	25	75	100	50	1.5	1	3	2	5
b	1	50	37.5	100	3	75	2	1.5	5	25	12.5
c	1.5	75	50	3	5	100	25	2	12.5	37.5	0.5
e	2	100	75	5	12.5	3	37.5	25	0.5	50	1
e	3	1.5	1	25	37.5	2	12.5	5	50	0.5	75
f	5	2	1.5	37.5	50	25	0.5	12.5	75	1	100
g	12.5	25	2	50	75	37.5	1	0.5	100	1.5	3
h	37.5	5	3	0.5	1	12.5	75	50	1.5	100	2
i	50	12.5	5	1	1.5	0.5	100	75	2	3	25
j	75	0.5	12.5	1.5	2	1	3	100	25	5	37.5
k	100	1	0.5	2	25	1.5	5	3	37.5	12.5	50

Table A.2: GeneLogic Latin square design with complex cRNA from AML cell line. There are 11 experimental groups (conditions) in this experiment. Each group had three replicates except group B which had two replicates.

A.2 GeneLogic AML Spike-in Data

In this data set 11 control cRNAs were spiked into a hybridization mix to prepare 11 samples as described in Table A.2. Each hybridization mixture was hybridized to multiple Affymetrix HG-U95a arrays. The concentrations used were 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50, 75 and 100 pM, arranged in a Latin square experiment. This Latin square experiment was carried out in the presence of complex cRNA prepared from an acute myeloid leukemia (AML) tumor cell line. Each concentration appeared once in each row and column. The probe-set BioC-5_at was not spiked-in appropriately and can be considered as an invariant probe-set. This data set is available at http://www.genelogic.com/newsroom/studies/.

A.3 Mouse Time-course Data Set

This time-course data set profiles the gene expression changes during the hair growth cycle, which is synchronised for the first two cycles following birth. After the two cycles the hair growth is unsynchronised. Lin et al. (2004) used Affymetrix microarray MG-U74Av2 to profile mRNA expression in mouse back skin from eight representative time points to discover regulators in hair-follicle morphogenesis and cycling. The microarray data set utilised a total of 25 chips with each time point consisting of three or four replicates. The first five time

Cycle	Time point	Microarray	qr-PCR
1st synchronous	day 1	√	√
	day 5		√
	day 6	√	
	day 8		√
	day 12		√
	day 14	√	
	day 15		√
	day 17	√	√
	day 19		√
	day 23	√	√
2nd synchronous	day 25		√
	day 29		√
	day 31		√
	day 34		√
	day 37		√
	day 41		√
	day 44		√
asynchronous	week 9	√	
	month 5	√	
	year 1	√	

Table A.3: The time points covered by the microarray experiment and the qr-PCR experiment. The common time points are day 1, 17 and 23 in the first synchronous cycle.

points (day 1, 6, 14, 17 and 23) cover the first synchronised cycle and the last three time points (week 9, month 5 and year 1) belong to the asynchronous cycles. Lin et al. (2004) identified 2,461 hair cycle-associated genes using statistical methods. There are eight genes not previously known to be hair-cycle associated that were identified by their temporal and spatial expression patterns during the hair-growth cycle. A quantitative real-time PCR (qr-PCR) experiment is conducted to confirm the discovery of these eight new hair cycle-associated genes. The qr-PCR experiment includes 15 time points and covers the first two synchronous cycles, and there are three replicates at each time point. The time points covered by the microarray experiment and the qr-PCR experiment are shown in Table A.3. For the first asynchronous cycle there are three time points in common (day 1, 17 and 23) for the two types of experiments. The eight genes and the 14 related probe-sets in the microarray data are shown in Table A.4. This data set is available at http://www.ncbi.nlm.nih.gov/projects/geo/.

Gene symbol	Probe-set ID
Car 6	160647_at
Crisp 1	93122_at
Elf5	103283_at
Junb	102362_i_at, 102363_i_at
Nmyc1	103048_at
Wnt11	103490_at
Dab2	98044_at, 98045_s_at, 104633_at
Fbln1	94307_at, 94308_at, 94309_at, 161628_at

Table A.4: The related probe-sets of the eight discovered hair cycle-associated genes in Lin et al. (2004).

A.4 Golden Spike-in Data set

The golden spike-in data set is designed in Choe et al. (2005) for the purpose of evaluating methods for identifying differential gene expression between two sets of replicated experiments. This data set includes two conditions each of which has three replicate chips. This data set contains a large number of differentially expressed genes with known fold-change, 1.2 to 4-fold. Unlike the spike-in data sets described in Appendix A.1 and A.2 which include small numbers of spike-in genes, 14 and 11 respectively, this spike-in data set provides a large number of true positives to obtain adequate statistics. The two conditions, control and spike, are labelled as 'C' and 'S' respectively. The S sample contains the same cRNAs as the C sample except that 1309 individual cRNAs are present at a defined different increasing concentration in the S sample which leads to 1,331 up-regulated probe-sets. The remaining sample contains 2,551 RNA species presenting at identical concentration in both experiments and this leads to 2,535 invariant probe-sets which can be very good normalisation background. Among the remaining probe-sets, 10,131 are empty probe-sets which have no targets in the sample and 13 are mixed probe-sets which match to more than one clone.

Appendix B

Affycomp Results

Affycomp (Cope et al., 2004) is a web-based benchmark for estimating Affymetrix microarray expression measurements. Tables B.1–B.6 are the copy of entries listed in order of submission for assessment from http://affycomp.biostat.jhsph.edu on 15 June, 2005. The bottom line in each table is the results from multi-mgMOS. Tables B.1 and B.2 show the original 15 assessments of dilution and spike-in hgu95 studies. Tables B.3 and B.4 show the new 14 assessments of the spike-in hgu95 study. Tables B.5 and B.6 are copies of the new 14 assessments of the spike-in hgu133 study. The best result for each entry is shown in bold. For the representation of the score component for each entry, please refer to http://affycomp.biostat.jhsph.edu/ for more information.

N	Method	1	2	3	4	5	6	7	8
0	(perfection)	0	1	1	1	1	1	1	0
1	MAS_5.0	0.29	0.89	0.73	0.85	0.71	0.86	0.36	3108.99
2	RMA	0.09	0.99	0.94	0.87	0.63	0.80	0.82	15.84
3	dChip	0.09	0.99	0.91	0.77	0.53	0.85	0.67	36.91
4	ZAM2NBG	0.07	0.99	0.94	0.72	0.57	0.77	0.84	2.44
5	qn.p5	0.11	0.98	0.56	0.06	0.42	0.50	0.62	20.30
6	vsn_scal	0.08	0.99	0.96	**1.00**	0.77	0.81	0.85	6.69
7	vsn	0.06	0.99	0.96	0.76	0.51	0.81	0.85	0.40
8	RMAVSN	0.09	0.99	0.94	0.89	0.61	0.81	0.83	17.87
9	RMA_NBG	0.04	1.00	0.91	0.56	0.48	0.81	0.85	0.13
10	GSVDmin	0.05	0.98	0.97	0.59	0.50	0.83	0.81	4.87
11	PLIER	0.13	0.09	0.01	0.84	0.71	0.91	0.02	596.77
12	GSVDmod	0.05	1.00	0.97	0.55	0.51	0.85	0.84	0.79
13	PLIER+16	0.08	0.99	0.88	0.64	0.65	**0.91**	0.81	8.42
14	GCRMA	0.09	0.99	0.89	0.72	0.97	0.84	0.82	7.62
15	ChipMan	0.31	0.99	0.94	1.26	0.88	0.82	0.67	183.99
16	ProbePro	0.16	0.70	0.58	0.84	1.45	0.47	0.17	2087.07
17	MMEI	**0.02**	**1.00**	0.92	0.52	0.45	0.80	**0.86**	**0.12**
18	PM	0.05	0.99	0.97	0.53	0.46	0.87	0.84	1.39
19	RMA:GNV	0.09	0.99	**0.98**	0.68	0.62	0.80	0.82	15.86
20	GL	0.05	0.99	0.92	0.56	0.48	0.81	0.83	0.15
21	MAS5+32	0.07	0.98	0.93	0.71	0.60	0.88	0.72	20.56
22	gMOS_v.1	0.32	0.97	0.81	0.64	0.95	0.75	0.54	1358.01
23	rsvd	-	-	-	-	0.66	0.90	0.81	2.63
24	ZL	0.34	0.99	0.04	0.23	0.57	0.65	0.79	21.99
32	gltran	0.04	0.99	0.94	0.64	0.51	0.78	0.84	1.46
33	UM-Tr-Mn	-	-	-	-	0.68	0.87	0.51	1399.72
34	GS_RMA	-	-	-	-	0.63	0.80	0.82	15.86
35	GS_GCRMA	-	-	-	-	0.84	0.91	0.84	6.53
36	gcrma113	-	-	-	-	0.87	0.91	0.85	3.89
41	mgMOS_gs	0.21	0.96	0.88	0.80	0.76	0.82	0.57	1061.31
42	mmgMOSgs	0.23	0.96	0.82	0.86	**1.03**	0.80	0.59	1616.01

Table B.1: Copy of entries 1–8 listed in order of submission for original assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES/0.html on 15 June, 2005.

N	Method	9	10	11	12	13	14	15
0	(perfection)	16	0	1	1	1	0	16
1	MAS_5.0	12.82	2.66	0.69	0.65	0.07	3072.18	3.71
2	RMA	11.98	0.31	0.61	0.36	0.54	1.00	1.71
3	dChip	11.43	0.45	0.52	0.32	0.17	28.64	1.25
4	ZAM2NBG	11.70	0.24	0.57	0.32	0.61	0.57	1.14
5	qn.p5	9.58	0.38	0.43	0.14	0.24	15.75	1.39
6	vsn_scal	12.23	0.23	0.75	0.28	0.66	0.43	3.89
7	vsn	10.83	0.15	0.50	0.19	0.66	0.21	1.11
8	RMAVSN	11.79	0.25	0.60	0.32	0.59	0.50	1.61
9	RMA_NBG	10.45	0.12	0.47	0.15	0.68	**0.11**	1.04
10	GSVDmin	11.09	0.21	0.49	0.24	0.56	2.43	1.00
11	PLIER	12.85	4.03	0.72	0.65	0.02	589.96	3.57
12	GSVDmod	11.19	0.19	0.50	0.24	0.60	0.54	1.11
13	PLIER+16	12.34	0.34	0.65	0.46	0.46	5.07	2.04
14	GCRMA	12.97	0.35	0.92	0.66	0.54	7.07	5.29
15	ChipMan	13.03	0.67	0.87	0.44	0.20	159.86	5.11
16	ProbePro	12.53	15.70	1.33	1.93	0.07	2046.46	4.93
17	MMEI	10.41	**0.12**	0.45	0.16	**0.69**	**0.11**	1.00
18	PM	10.67	0.15	0.45	0.18	0.64	0.68	1.00
19	RMA:GNV	11.99	0.31	0.61	0.36	0.54	1.00	1.71
20	GL	10.42	0.14	0.47	0.16	0.66	**0.11**	1.18
21	MAS5+32	11.76	0.51	0.59	0.33	0.18	19.18	1.68
22	gMOS_v.1	12.75	2.15	0.94	**1.04**	0.10	1319.07	5.36
23	rsvd	12.14	0.35	0.66	0.41	0.49	2.04	2.89
24	ZL	11.98	0.24	0.57	0.36	0.42	37.57	2.36
32	gltran	11.09	0.19	0.50	0.22	0.65	0.68	1.04
33	UM-Tr-Mn	12.56	1.76	0.67	0.54	0.07	1385.00	2.64
34	GS_RMA	11.98	0.31	0.61	0.36	0.54	0.93	1.71
35	GS_GCRMA	13.15	0.41	0.82	0.65	0.58	3.00	4.71
36	gcrma113	13.16	0.37	0.86	0.68	0.61	2.57	4.89
41	mgMOS_gs	13.12	1.37	0.75	0.90	0.07	1028.61	4.14
42	mmgMOSgs	**13.65**	1.79	**1.02**	1.40	0.09	1570.11	**6.71**

Table B.2: Copy of entries 9–15 listed in order of submission for original assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES/0.html on 15 June, 2005.

N	Method	1	2	3	4	5	6	7
0	(perfection)	0.00	0.00	0.00	1.00	1.00	1.00	1.00
1	MAS_5.0	0.63	0.85	4.48	0.86	0.71	0.72	0.80
2	RMA	0.11	0.19	0.57	0.80	0.63	0.29	0.73
3	dChip	0.13	0.20	1.44	0.85	0.53	0.25	0.64
4	ZAM2NBG	0.09	0.16	0.50	0.77	0.57	0.25	0.66
5	qn.p5	0.12	0.22	1.09	0.50	0.42	0.11	0.44
6	vsn_scal	0.09	0.15	0.43	0.81	0.77	0.21	0.82
7	vsn	0.06	0.10	0.29	0.81	0.51	0.14	0.55
8	RMAVSN	0.09	0.16	0.48	0.81	0.61	0.25	0.71
9	RMA_NBG	0.04	0.08	0.24	0.81	0.48	0.12	0.50
10	GSVDmin	0.08	0.13	0.60	0.83	0.50	0.17	0.58
11	PLIER	0.19	0.33	123.27	0.91	0.71	0.76	0.84
12	GSVDmod	0.07	0.13	0.44	0.85	0.51	0.18	0.60
13	PLIER+16	0.13	0.21	0.83	**0.91**	0.65	0.45	0.78
14	GCRMA	0.09	0.16	0.77	0.84	0.97	0.73	1.19
15	ChipMan	0.27	0.33	2.26	0.82	0.88	0.36	1.04
16	ProbePro	0.31	0.47	18.75	0.47	1.45	1.26	1.73
17	MMEI	0.04	0.06	**0.23**	0.80	0.45	0.11	0.49
18	PM	0.05	0.09	0.40	0.87	0.46	0.14	0.52
19	RMA:GNV	0.11	0.19	0.58	0.80	0.62	0.29	0.73
20	GL	0.05	0.08	0.25	0.81	0.48	0.12	0.50
21	MAS5+32	0.14	0.23	1.07	0.88	0.60	0.31	0.68
22	gMOS_v.1	0.29	**0.00**	3.35	0.75	0.95	0.82	1.22
23	rsvd	**0.00**	**0.00**	0.58	0.90	0.66	0.31	0.84
24	ZL	0.22	0.12	0.52	0.65	0.57	0.39	0.67
32	gltran	0.07	0.12	0.42	0.78	0.51	0.19	0.57
33	UM-Tr-Mn	0.32	0.51	2.92	0.87	0.68	0.53	0.82
34	GS_RMA	0.11	0.19	0.57	0.80	0.63	0.29	0.73
35	GS_GCRMA	0.10	0.07	0.79	0.91	0.84	0.51	**1.02**
36	gcrma113	0.08	0.04	0.74	0.91	0.87	0.52	1.06
41	mgMOS_gs	0.36	0.55	2.86	0.82	0.76	0.77	0.89
42	mmgMOSgs	0.40	0.58	3.27	0.80	**1.03**	**1.21**	1.26

Table B.3: Copy of hgu95a entries 1–7 listed in order of submission for newer assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES.hgu/0.html on 15 June, 2005.

N	Method	8	9	10	11	12	13	14
0	(perfection)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	MAS_5.0	0.45	0.69	0.65	0.07	0.00	0.00	0.05
2	RMA	0.47	0.61	0.36	0.51	0.91	0.64	0.60
3	dChip	0.39	0.52	0.32	0.21	0.43	0.16	0.26
4	ZAM2NBG	0.47	0.57	0.32	0.57	0.95	0.74	0.66
5	qn.p5	0.52	0.43	0.14	0.09	0.43	0.46	0.17
6	vsn_scal	**0.70**	0.75	0.28	0.53	0.97	0.86	0.64
7	vsn	0.47	0.50	0.19	0.53	0.97	0.86	0.64
8	RMAVSN	0.46	0.60	0.32	0.52	0.94	0.70	0.62
9	RMA_NBG	0.46	0.47	0.15	0.55	0.97	0.92	0.65
10	GSVDmin	0.41	0.49	0.24	0.39	0.87	0.65	0.51
11	PLIER	0.46	0.72	0.65	0.04	0.00	0.00	0.03
12	GSVDmod	0.42	0.50	0.24	0.47	0.94	0.74	0.59
13	PLIER+16	0.46	0.65	0.46	0.61	0.83	0.46	0.66
14	GCRMA	0.55	0.92	0.66	0.62	0.94	0.59	0.69
15	ChipMan	0.68	0.87	0.44	0.21	0.50	0.24	0.28
16	ProbePro	0.39	1.330	1.93	0.05	0.04	0.00	0.05
17	MMEI	0.46	0.45	0.16	0.57	**0.98**	**0.94**	0.67
18	PM	0.43	0.45	0.18	0.47	0.93	0.83	0.59
19	RMA:GNV	0.47	0.61	0.36	0.50	0.91	0.64	0.60
20	GL	0.46	0.47	0.16	0.58	0.96	0.88	0.67
21	MAS5+32	0.44	0.59	0.33	0.04	0.28	0.08	0.10
22	gMOS_v.1	0.42	0.94	**1.04**	0.07	0.04	0.00	0.06
23	rsvd	0.40	0.66	0.41	0.54	0.93	0.54	0.64
24	ZL	0.45	0.57	0.36	0.65	0.87	0.73	0.70
32	gltran	0.45	0.50	0.22	0.55	0.94	0.80	0.65
33	UM-Tr-Mn	0.42	0.67	0.54	0.11	0.00	0.00	0.08
34	GS_RMA	0.47	0.61	0.36	0.51	0.91	0.64	0.60
35	GS_GCRMA	0.55	0.82	0.65	0.64	0.94	0.56	0.72
36	gcrma113	0.56	0.86	0.68	**0.68**	0.97	0.63	**0.75**
41	mgMOS_gs	0.43	0.75	0.90	0.25	0.04	0.00	0.20
42	mmgMOSgs	0.45	**1.02**	1.40	0.36	0.07	0.00	0.29

Table B.4: Copy of hgu95a entries 8–14 listed in order of submission for newer assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES.hgu/0.html on 15 June, 2005.

N	Method	1	2	3	4	5	6	7
0	(perfection)	0.00	0.00	0.00	1.00	1.00	1.00	1.00
1	MAS_5.0	0.29	0.47	4.01	0.91	0.77	0.58	0.73
2	RMA	0.07	0.13	0.40	0.90	0.68	0.20	0.71
8	RMAVSN	0.02	0.04	0.15	0.89	0.12	0.06	0.13
23	rsvd	0.14	0.12	0.73	**0.94**	0.74	0.31	0.78
25	rsvd.pm	0.06	0.11	0.34	0.89	0.53	0.12	0.53
26	rma-tog	0.07	0.13	0.40	0.90	0.68	0.20	0.71
27	rma-sep	0.18	0.28	0.96	0.90	0.71	0.27	0.72
28	LW1	0.08	0.14	1.18	0.91	0.59	0.19	0.62
29	LW2	0.14	0.25	13.88	0.56	1.08	1.50	0.80
30	rsvd.bgc	0.08	0.14	0.52	0.89	0.58	0.16	0.59
31	cor523	0.02	0.03	0.12	0.88	0.12	0.06	0.13
33	UM-Tr-Mn	0.15	0.25	1.86	0.93	0.70	0.36	0.72
34	GS_RMA	0.07	0.13	0.40	0.90	0.68	0.20	0.71
35	GS_GCRMA	0.07	0.09	0.65	0.93	0.93	0.37	0.96
36	gcrma113	0.06	0.04	0.61	0.91	**1.00**	0.25	1.13
37	rsvd2	0.17	0.28	1.74	0.91	0.75	0.46	0.74
38	W237	0.02	0.04	0.17	0.87	0.12	0.05	0.13
39	RMA_NBG	**0.01**	**0.02**	**0.06**	0.90	0.09	0.02	0.09
40	RMAVSN	0.02	0.04	0.15	0.89	0.12	0.06	0.13
41	mgMOS_gs	0.24	0.34	2.02	0.92	0.81	0.58	0.78
42	mmgMOSgs	0.24	0.34	2.71	0.89	1.03	**0.88**	**0.98**

Table B.5: Copy of hgu133 entries 1–7 listed in order of submission for newer assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES.hgu/0.html on 15 June, 2005.

N	Method	8	9	10	11	12	13	14
0	(perfection)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	MAS_5.0	0.77	0.77	0.64	0.09	0.00	0.00	0.06
2	RMA	0.80	0.68	0.31	0.57	0.91	0.96	0.65
8	RMAVSN	0.10	0.12	0.08	0.46	0.59	0.43	0.49
23	rsvd	0.73	0.74	0.43	0.53	0.73	0.71	0.58
25	rsvd.pm	0.77	0.53	0.16	0.42	0.90	**0.96**	0.54
26	rma-tog	0.80	0.68	0.31	0.57	0.91	0.96	0.65
27	rma-sep	0.84	0.71	0.39	0.38	0.53	0.63	0.42
28	LW1	0.74	0.59	0.25	0.23	0.47	0.55	0.29
29	LW2	0.68	1.08	1.45	0.19	0.00	0.00	0.14
30	rsvd.bgc	0.79	0.58	0.22	0.38	0.80	0.90	0.49
31	cor523	0.10	0.12	0.08	0.54	0.77	0.61	0.60
33	UM-Tr-Mn	0.70	0.70	0.44	0.18	0.10	0.10	0.16
34	GS_RMA	0.80	0.68	0.30	0.56	0.91	0.96	0.65
35	GS_GCRMA	0.96	0.93	0.55	**0.59**	0.87	0.90	**0.66**
36	gcrma113	**0.97**	**1.00**	0.48	0.45	**0.91**	0.92	0.57
37	rsvd2	0.81	0.75	0.52	0.29	0.16	0.21	0.26
38	W237	0.10	0.12	0.07	0.35	0.54	0.39	0.39
39	RMA_NBG	0.10	0.09	0.04	0.54	0.90	0.93	0.63
40	RMAVSN	0.10	0.12	0.08	0.46	0.59	0.43	0.49
41	mgMOS_gs	0.77	0.81	0.68	0.33	0.08	0.12	0.27
42	mmgMOSgs	0.79	1.03	**1.05**	0.37	0.02	0.01	0.28

Table B.6: Copy of hgu133 entries 8–14 listed in order of submission for newer assessment from http://affycomp.biostat.jhsph.edu/AFFY2/TABLES.hgu/0.html on 15 June, 2005.