Methods to improve gene signal: Application to cDNA microarrays

Rashi Gupta

Department of Mathematics and Statistics
University of Helsinki
Finland
&
DNA sequencing and genomics laboratory
Institute of Biotechnology
University of Helsinki
Finland

Academic dissertation

To be presented, with the permission of the Faculty of Science,
University of Helsinki, for public criticism in Auditorium B123, Exactum,
(Gustaf Hällströmin katu 2b)
on 17 April, 2009, at 12 o'clock noon .

HELSINKI 2009

Supervisors:
Professor Elja Arjas
Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Docent Petri Auvinen
DNA sequencing and genomics laboratory
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Reviewers:
Professor Jukka Corander
Department of Mathematics
Åbo Akademi University
Turku, Finland

Professor Arnoldo Frigessi
Section of Biostatistics
Institute of Basic Medical Research
University of Oslo
Oslo, Norway

Opponent:
Docent Sampsa Hautaniemi
Computational Systems Biology Laboratory
Institute of Biomedicine and Genome-Scale Biology Research Program
Biomedicum Helsinki
University of Helsinki
Helsinki, Finland

Email: rashi1@live.com

ISBN 978-952-10-5437-2 (paperback)
ISBN 978-952-10-5438-9 (PDF, http://ethesis.helsinki.fi)
Helsinki University Printing House
Helsinki 2009

Abstract

Microarrays are high throughput biological assays that allow the screening of thousands of genes for their expression. The main idea behind microarrays is to compute for each gene a unique signal that is directly proportional to the quantity of mRNA that was hybridized on the chip. A large number of steps and errors associated with each step make the generated expression signal noisy. As a result, microarray data need to be carefully pre-processed before their analysis can be assumed to lead to reliable and biologically relevant conclusions.

This thesis focuses on developing methods for improving gene signal and further utilizing this improved signal for higher level analysis. To achieve this, first, approaches for designing microarray experiments using various optimality criteria, considering both biological and technical replicates, are described. A carefully designed experiment leads to signal with low noise, as the effect of unwanted variations is minimized and the precision of the estimates of the parameters of interest are maximized. Second, a system for improving the gene signal by using three scans at varying scanner sensitivities is developed. A novel Bayesian latent intensity model is then applied on these three sets of expression values, corresponding to the three scans, to estimate the suitably calibrated true signal of genes. Third, a novel image segmentation approach that segregates the fluorescent signal from the undesired noise is developed using an additional dye, SYBR green RNA II. This technique helped in identifying signal only with respect to the hybridized DNA, and signal corresponding to dust, scratch, spilling of dye, and other noises, are avoided. Fourth, an integrated statistical model is developed, where signal correction, systematic array effects, dye effects, and differential expression, are modelled jointly as opposed to a sequential application of several methods of analysis.

The methods described in here have been tested only for cDNA microarrays, but can also, with some modifications, be applied to other high-throughput technologies.

Keywords: High-throughput technology, microarray, cDNA, multiple scans, Bayesian hierarchical models, image analysis, experimental design, MCMC, WinBUGS.

Contents

Preface

This work was carried out during the years 2003-2008 at the Department of Mathematics and Statistics, University of Helsinki and at the DNA sequencing and genomics laboratory, Institute of Biotechnology, University of Helsinki, Finland. The work was funded by Graduate school in Computational Biology, Bioinformatics and Biometry (ComBi) for the duration 2003-2006, by SYSBIO Research Programme for the year 2007 and by University of Helsinki for the duration February-April, 2008.

Firstly, I wish to thank Petri for granting permission to work in his lab when I was novice in the field of microarrays. He taught me the basics of microarray and explained me the working of wet lab. It was Petri who pointed me towards the possibilities of using multiple scans for correcting gene signal, which over the time evolved as multiple publications towards this thesis. I would also like to thank him for providing me valuable data for all my publications.

I would sincerely like to thank Elja for his guidance towards regularly improving the models. I am extremely grateful to him for correcting my articles an infinite number of times and making me learn what I know today. I am also thankful to him for being extremely understanding and allowing me the possibility to work from distance.

I would like to express my special thanks to Andrew Thomas for his encouragement and valuable suggestions during model implementations. I would like to thank Pasi and Tarja for making my stay in Helsinki comfortable by taking care of all administrative issues. Finally I would like to thank my colleagues and co-authors both at the Department of Mathematics and Statistics and at DNA sequencing and genomics laboratory for providing a healthy work environment. I would like to thank all my friends at CIMO for giving me company during five long years and a big thanks to Ahlam without whom I would have not completed this work.

I would like to thanks my reviewers Jukka Corander and Arnoldo Frigessi for their valuable suggestions and extremely good comments.

I am extremely grateful to GOD who gave me the strength to keep going in spite of all odds. Warm thanks to my family-in-law for their support. I am grateful to my mum and dad for always being there for me and for their love, care and most honest prayers. Thanks Rajat for always being concerned and a big hug to Kittu who is my most valuable possession and my stress-buster.

Finally, I would like to thank Manish without whose support I would not have completed this work. I also want to thank him for long overseas calls and for designing a very colourful cover page for this book. This book is as much yours as it is mine.

Rashi
Delhi, March 2009

List of original publications

The thesis consists of an introduction and the following five publications.

I R. Gupta, S. Ruosaari, S. Kulathinal, J. Hollmén and P. Auvinen. Microarray image segmentation using additional dye—An experimental study. Molecular and Cellular Probes, 21: 321–328, 2007.

II R. Gupta, P. Somervuo, S. Kulathinal, P. Auvinen. Optimal designs for microarray experiments with biological and technical replicates. Recent Advances In Linear Models and Related Areas, Springer, 2008.

III R. Gupta, E. Arjas, S. Kulathinal, A. Thomas and P. Auvinen. Bayesian hierarchical model for estimating gene expression intensity using multiple scanned microarrays. EURASIP Journal on Bioinformatics and Systems Biology, Article ID 231950, 2008.

IV R. Gupta, P. Auvinen, A. Thomas and E. Arjas. Bayesian hierarchical model for correcting signal saturation in microarrays using pixel intensities. Statistical Applications in Genetics and Molecular Biology, 5, Article 20, 2006.

V R. Gupta, D. Greco, P. Auvinen, E. Arjas. Bayesian integrated modeling of expression data: A case study on RhoG. (Submitted).

Author’s contributions

Publication I: Rashi Gupta had the main responsibility in formulating the method, paper writing and comparison with other existing methods. Salla Ruosaari was responsible for method implementation and paper writing. Sangita Kulathinal provided insight into the statistical analysis. Jaakko Hollmén did the proof reading, and Petri Auvinen provided the initial idea and conducted the experiment.

Publication II: Rashi Gupta had the main responsibility of searching the literature, formulating the problem and paper writing. Panu Somervuo was responsible for implementation and paper writing. Sangita Kulathinal helped in formulating the problem and paper writing. Petri Auvinen conducted the experiment.

Publication III: Rashi Gupta and Elja Arjas were jointly responsible for model construction and paper writing. Rashi Gupta was also responsible for model implementation and carrying out the analysis. Sangita Kulathinal helped in model implementation and analysis. Andrew Thomas helped in implementation. Petri Auvinen gave the initial idea and conducted the experiment.

Publication IV: Rashi Gupta and Elja Arjas were jointly responsible for drafting the problem, model construction and paper writing. Rashi Gupta was also responsible for implementation, and carrying out the analysis. Andrew Thomas gave valuable suggestion for improving the computational speed of the model. Petri Auvinen conducted the experiment and provided the data required for the analysis.

Publication V: Rashi Gupta was responsible for model construction, implementation, functional analysis and paper writing. Dario Greco helped in the functional analysis. Petri Auvinen provided the data and validated the results. Elja Arjas provided valuable insights in the model construction and helped in paper writing.

List of abbreviations and symbols

Cy3	Cyanine3
Cy5	Cyanine5
TIFF	Tagged image file format
LOWESS	Local weighted scatter plot smoother
PCR	Polymerase chain reaction
RT-PCR	Reverse transcription polymerase chain reaction
PMT	Photomultiplier tube
RNA	Ribonucleic acid
TIFF	Tagged image file format
mRNA	Messenger RNA
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
SAGE	Serial analysis of gene expression
A/D	Analog to digital

1. Introduction

Biomedical and biological research is in the middle of a significant transition and is basically driven by two important factors: the massive increase in the amount of DNA sequence information and the development of the technologies that enable researchers to exploit this information. Therefore a very large number of discoveries, analyses, and new experiments are being made. In the last few years, almost 700 bacterial species and more than 20 eukaryote organisms, of which about half are fungi, have had their genome completely sequenced, and work on many more is in progress [1]. Unfortunately, the huge amount of DNA sequence information do not provide direct answers to questions such as what genes do, how a cell works, or how are diseases caused? This is where functional genomics has its important role to play. The goal of functional genomics is to make use of the vast amounts of data produced by genomic projects (such as genome sequencing projects) in order to describe gene and protein functions, and their interactions. It focuses on the dynamic aspects such as gene transcription, translation, and protein-protein interactions. Functional genomics uses new technologies to take full advantage of this large and rapidly increasing sequence information. Among these tools, the most versatile and powerful are high-density arrays of oligonucleotides or complementary DNAs.

These arrays have been in use for biological experiments for many years [2, 3, 4, 5, 6, 7]. Traditionally, they consisted of fragments of DNA, often with an unknown sequence, spotted on porous membrane. The arrayed DNA fragments often come from cDNA, genomic DNA or plasmid libraries. Recently, the use of glass as a substrate and fluorescence for detection, together with the development of new technologies for synthesizing or depositing DNA on glass slides at very high densities, have allowed the miniaturization of DNA arrays and this has resulted in a significant increase in the experimental efficiency and information content [8, 9, 10, 11, 12, 13].

One of the most important applications for DNA arrays so far has been the monitoring of gene expression, in other words, monitoring the abundance of mRNA. The transcription of genomic DNA to produce mRNA is the first step in the process of protein synthesis and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. It could be argued that if mRNA is only an intermediate molecule in the production of protein products, why measure mRNA when in fact proteins are responsible for most biological activities? One reason is that protein-based approaches are generally more difficult, less sensitive, and have a lower throughput than RNA-based ones. But more importantly, mRNA levels are immensely informative about cell state and the activity of genes, and for most genes, changes in mRNA abundance are related to changes in protein abundance.

Gene expressions are basically used in attempts to understand the function of genes, and to know when, where and to what extent a gene is expressed. This information further leads to an understanding of the activity and of the biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms, broader cellular functions, and biochemical pathways. In the context of human health, the knowledge gained from expression measurements can help determine the genes causing a particular disease and consequences of the disease, how drugs and drug candidates work in cells and organisms, and which gene products may be appropriate targets for therapeutic intervention.

Apart from DNA arrays, there are other ways to measure mRNA abundance, gene expression and changes in gene expression. Some of them are: northern blots, polymerase chain reaction after reverse transcription of RNA (RT-PCR), nuclease protection, cDNA sequencing, clone hybridization, differential display [14], subtractive hybridization, cDNA fragment fingerprinting, and serial analysis of gene expression (SAGE) [15]. It is important to emphasize that these new, parallel approaches do not replace the conventional methods. Standard methods such as northern blots, RT-PCR are basically used in a more targeted fashion to follow-up on the genes, pathways and mechanisms implicated by the array results.

This thesis is structured in two parts. The purpose of the first part is to present both the biological and methodological background for understanding the basics behind a typical microarray experiment. The second part consists of five publications, which present novel methods for improving gene expression signal and eventually use this improved signal for finding differentially expressed genes.

The structure for the first part of the thesis is as follows: Chapter 2 presents some concepts of molecular biology required to understand the nature of data resulting from microarray experiments. Chapter 3 presents a brief introduction on how arrays are produced, spotted, hybridized and eventually scanned. It also presents a detailed discussion on scanning and image processing. Chapter 4 deals with the importance of experimental design and data pre-processing. Chapter 5 highlights the methods for selecting differentially expressed genes and lists software/methods that can be used for annotating the interesting genes. Chapter 6 gives a brief introduction about Bayesian hierarchical modelling and the software “WinBUGS” used for analyzing the data in Publications III, IV and V.

2. Biological Background

Here we review the basic concepts of modern molecular biology, which are needed before understanding the purpose and the nature of data resulting from microarrays. We confine our discussion to those topics that are most essential. A detailed discussion about cell biology can be found, *e.g.*, in [16].

2.1 Organisms and cells

A cell, the structural and functional unit of all living organisms, contains the hereditary information necessary for regulating cell functions and for transmitting information to the next generation of cells. The cell theory, first developed in 1839 by Matthias Jakob Schleiden and Theodor Schwann, states that all organisms are composed of one or more cells. Some organisms, such as bacteria, are unicellular (consisting of a single cell) and other organisms, such as humans, are multicellular.

Cells can be classified in three domains: Eukaryota, Eubacteria, and Archaea where Eubacteria and Archaea are the split of the prokaryotes based on genetic differences. The major difference between prokaryotes and eukaryotes is that eukaryotic cells contain membrane-bound compartments in which specific metabolic activities take place. The most important difference is the presence of nucleus in the eukaryotic cells, which is absent in prokaryotic cells. All cells, whether prokaryotic or eukaryotic, possess DNA, the hereditary material of genes, and RNA, containing the information necessary to build various proteins such as enzymes, the cell's primary machinery.

2.2 Gene

A gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions. The physical development and phenotype of organisms can be thought of as a product of genes interacting with each other and with the environment.

In cells, genes consist of a long strand of DNA that contains a promoter, which controls the activity of a gene, and coding and non-coding sequences. A coding sequence determines what the gene produces, while non-coding sequences can regulate the conditions of gene expression. The regions encoding gene products are called exons in eukaryotic cells. When a gene is active, the coding and non-coding sequences are copied in a process called transcription, producing an RNA copy of the gene's information. This RNA can then direct the synthesis of proteins via the genetic code.

Genes of eukaryotic organisms can contain regions called introns that are removed from the messenger RNA in a process called splicing. In eukaryotes, a single gene can encode multiple proteins, which are produced through the creation of different arrangements of exons through alternative splicing. In prokaryotes, introns are less common and genes often contain a single uninterrupted stretch of DNA that codes for a product. Prokaryotic genes are often arranged in groups called operons with promoter and operator sequences that regulate transcription of a single long RNA. This RNA can contain multiple coding sequences.

2.3 Physical definition of gene

The vast majority of living organisms encode their genes in long strands of DNA. DNA is most commonly recognized as two paired chains of chemical bases, spiralled into a double helix structure. The double helix structure of DNA is presented in Figure 1. There are four different kinds of bases in DNA: adenine (A), cytosine (C), guanine (G) and thymine (T). The order in which the bases occur determines the information stored in the region of DNA being looked at. The bases are divided into two classes: purines (A and G) and pyrimidines (C and T). When a base is attached to a sugar it is referred to as nucleoside. If a phosphate group is attached to this nucleoside then it is referred to as nucleotide. The nucleotide is the basic repeat unit of a DNA strand.

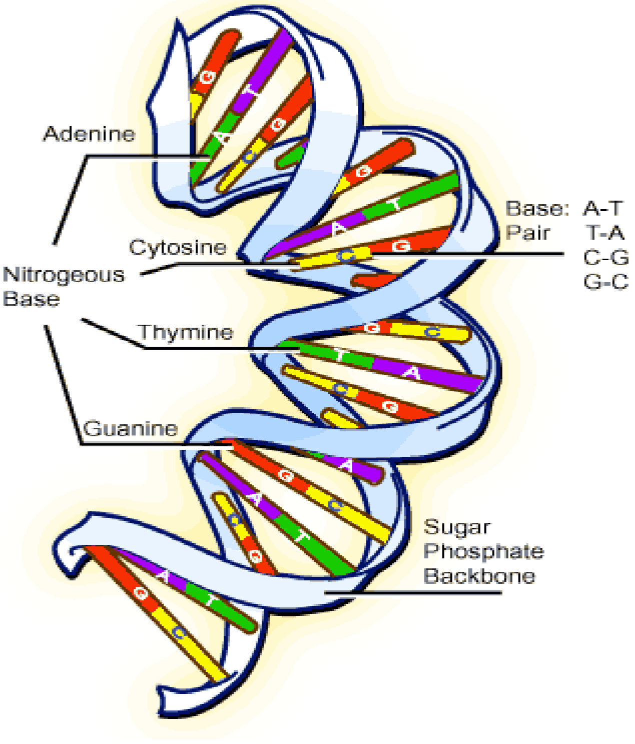

Figure 1: *The DNA double helix structure (Source: http://www.scq.ubc.ca/a-monks-flourishing-garden-the-basics-of-molecular-biology-explained/)*

The double helix structure of DNA is due to the hydrogen bonding that occurs between laterally opposed bases. The chemical structure of the bases is such that adenine (A) specifically binds to thymine (T) and cytosine (C) specifically binds to guanine (G). Since no other interactions are possible between any other combinations of base pairs, it is said that A is complementary to T and C is complementary to G. Two strands are called complementary if for any base on one strand, the other strand contains this base complement. Two complementary single-stranded DNA chains that come into close proximity react to form a stable double helix in a process known as hybridization or annealing.

2.4 Expression of genetic information

The flow of genetic information is from DNA to mRNA to proteins. This one-way process is described as the central dogma of molecular biology, see Figure 2. To make products from gene, the information in the DNA is first copied, base to base, into a similar kind of information carrier, called a transcript, or RNA. In eukaryotic cells, the RNA copy of the gene sequence acts as a messenger, taking information from the nucleus and transporting it into the cytoplasm of the cell. Once in the cytoplasm, the messenger RNA is translated into the product of the gene, a protein. The sequence of protein is defined by the original sequence of the DNA bases found in the gene.

In most cases, RNA is an intermediate product in the process of manufacturing proteins from genes. However, for some gene sequences, the RNA molecules are the actual functional products. The DNA sequences from which such RNAs are transcribed are known as non-coding DNA or RNA genes. Some viruses store their entire genomes in the form of RNA, and contain no DNA at all. Because they use RNA to store genes, their cellular hosts may synthesize their proteins as soon as they are infected and without the delay in waiting for transcription. On the other hand, RNA retroviruses, such as HIV, require the reverse transcription of their genome from RNA into DNA before their proteins can be synthesized. The process of producing a functional molecule of either RNA or protein is called gene expression, and the resulting molecule is called a gene product.

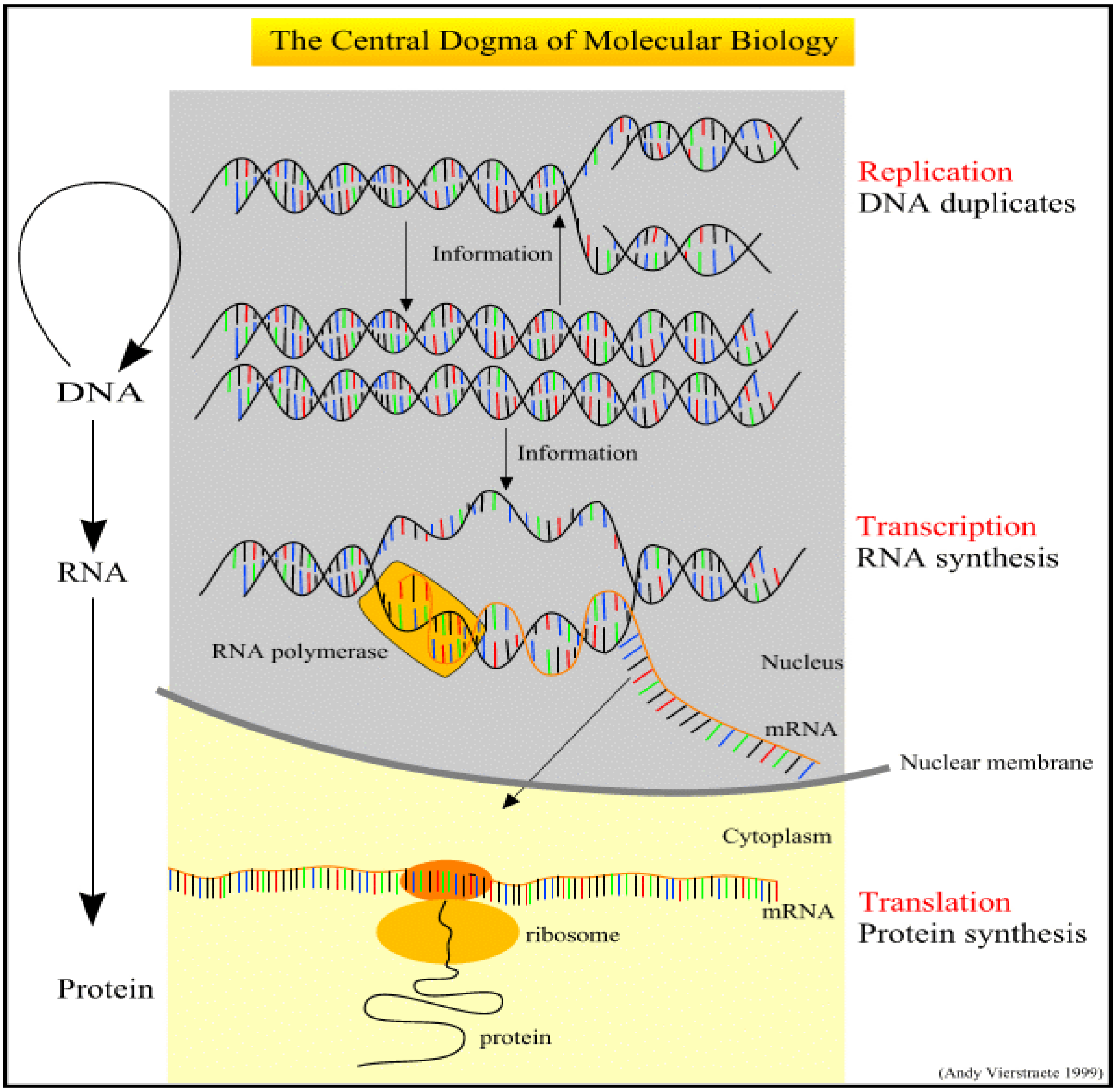

Figure 2: *Central dogma of molecular biology (Source:http://users.ugent.be/~avierstr/principles/centraldogma.html)*

3. cDNA microarray

3.1 Experimental aspects

cDNA microarray is based upon the mutual and specific affinity of the complementary strands of DNA. The technique is applicable under laboratory settings because it miniaturizes the quantity of information contained within a genome. The number of genes on the array can range from 500 to as many as 40,000 genes. Once the desired genes are chosen, individual clones of each must be obtained. Universal primers are used for the polymerase chain reaction (PCR) amplification of each gene either from a plasmid preparation or the bacterial vector itself.

Purified PCR products are then individually spotted, usually in duplicate, onto a glass slide. Printing is usually done in one of three ways: photolithography, mechanical micro spotting, or ink jetting [17, 18]. Photolithography uses light to covalently synthesize the DNA strands to the slide, mechanical spotting uses spotting pins and capillaries to transport DNA to the surface of the glass slide, and ink jetting uses electric current to dispense the appropriate amount of purified DNA on the glass slide.

Printing is a delicate procedure and should be performed in a controlled environment. Surface chemistry, temperature, and humidity play an important role in the spot formulation. Variation in such conditions may lead to non-uniformities among the spots. Other problems could be fluctuations in the amount of target spotted by the same pin, or variation in the geometry of different pins. Sometimes while printing, a pin touches the surface strongly thereby damaging the spots and preventing a good hybridization in the centre. Superficial tension of the liquid can cause spots with tails. Dust is yet another potential problem for printing and therefore most arrayers are enclosed in a glass container to prevent dust deposition during printing.

To prepare a sample for hybridization, the first step is to purify high-quality mRNA or total RNA from the cellular contents. The experimenter is then faced with several challenges: (i) mRNA accounts for only a small fraction (less than 3%) in a cell, (ii) the more heterogeneous the cell (*e.g.*, cells of solid tumours), the more difficult it is to isolate mRNA specific to the study, and (iii) captured mRNA degrades very quickly. Amplification methods can be utilized for small RNA amounts [19, 20]. To stop mRNA degradation, it is immediately reverse-transcribed into more stable cDNA (for cDNA microarrays) or cRNA (for oligonucleotide arrays, cRNA is synthetic RNA produced by transcription from a single-stranded DNA template).

In order to detect which cDNAs are bound to the microarray, each sample is labelled with a reporter molecule that flags its presence. The reporter molecules currently used in microarray experiments are fluorescent dyes that fluoresce when exposed to a specific wavelength of light. Two methods are currently used for labelling: direct and indirect labelling [21, 22]. In the direct labelling method, extracted RNA is reverse-transcribed into cDNA and labelled with fluorochromes such as Cy3 (cyanine 3) and Cy5 (cyanine 5). Alternatively, in the indirect labelling method, amino-allyl conjugated nucleotides are incorporated into the first-strand cDNA, followed by chemical coupling of the fluorochromes. Both types of labelling will generally introduce a dye bias effect into the expression signal, with the bias caused by indirect labelling being smaller than that resulting from direct labelling.

The labelled targets are poured onto the microarray and allowed to diffuse uniformly. The array is sealed in a hybridization chamber and incubated at a specific temperature for enough time to allow the hybridization reactions to complete. The experimental conditions should ensure that all areas of the array are exposed to the same amount of labelled sample. Two single-stranded DNA molecules will bind with high affinity if they have precisely matching (complementary) sequences, and with significantly lower affinity if they have an imperfect match. The microarray is removed from the hybridization chamber and thoroughly washed to eliminate any excess-labelled sample. Finally the microarray is dried using a centrifuge or by blowing with clean compressed air.

The microarray is scanned to determine the amount of labelled sample bound to each spot. The emitted light is captured by a scanner that records its intensity. Although the scanner is only supposed to pick up light emitted by the target cDNAs bound to their complementary spots, the scanner will inevitably also pick up light from various other sources, including the labelled sample hybridizing non-specifically to the glass slide, unwashed labelled sample adhering to the slide, various chemicals used up in processing the slide, and even the slide itself. This extra light from the slide is called background. The end product of scanning is a gray scale image usually stored in the 16-bit tagged image format (TIFF), then resulting in intensity measurements which range from 0 to 2^{16}-1.

3.2 Creation of scanned microarray images

A microarray scanner performs an area scan of the slide and converts each hybridized array into a digital image. The scanned region is divided into equally sized pixels and the laser generates excitation light, which is focused on a small portion of the hybridized array. Fluorescent molecules in this area absorb the excitation photons generated by the laser and emit fluorescent photons. These emitted photons are gathered by the detector. The detector in a scanner converts the emission photons into electric current. A common type of detector is a photomultiplier tube (PMT). A PMT converts each photon into several electrons. The amount of amplification can be adjusted by varying the PMT's voltage input. Finally, an analog to digital (A/D) converter is used to convert the electrons into a sequence of digital signals. The digitalizing process produces for each pixel a signal that represents the total fluorescence in the region corresponding to that pixel [23, 24, 25].

For a typical microarray experiment, the scanner produces two TIFF images, one for each fluorescent dye. To measure the abundance of the two fluorescent dyes for each spot, the scanners are designed to generate excitation light at different wavelengths and detect different emission wavelengths. The commonly used dyes are Cy3 and Cy5, with corresponding ranges of emission being 510-550 nm and 630-660 nm, respectively. A sequential scanner will first scan the glass slide with one wavelength and then scan it at the other wavelength. Alternatively, a dual scanner has two lasers and two detectors, and it scans the slide at two wavelengths simultaneously.

Various types of noise can affect the final signal produced by the scanner, *e.g.*, photon noise, dust on the slide, treatments of the glass slide, noise while amplification, and digitalizing. A perfect scanned image should only reflect the measures of the fluorescent intensities for the dye of interest. However, in practice, we have an imperfect system and the scanned image is a combination of the desired fluorescent signal and of the undesired noise.

Yet another crucial problem that arises while scanning is signal saturation. Since the sensitivity level of the microarray scanner is adjustable, it plays a crucial role in getting reliable measurements from both the weakly and highly expressed spots/genes present on the hybridized array. The scanner's sensitivity is raised to a certain level to ensure that the intensity levels of weakly expressed genes exceed the intrinsic noise level of the scanner and that they are measurable. This, however, can lead to problems caused by signal censoring for the highly expressed genes.

Publications III and IV aim at improving the quality of intensity measurements by first producing three images with different scanner sensitivities, and then obtaining three different data sets of expression values. A novel Bayesian latent intensity model is applied to estimate the suitably calibrated true expression of genes, by using the three different sets of expression measurements.

3.3 Processing the scanned images

The scanned images of the hybridized array are black and white and are usually stored as high-resolution TIFF files. For visualization, most of the available softwares create a composite image by overlapping the two images corresponding to the individual channels. To allow a visual assessment of the relationship between the quantities of mRNA corresponding to a given gene in the two channels, the software normally uses different artificial colours for each of the two channels. Typically, red and green colours are used and the composite image produced by overlapping the red and green images will have spots with colours from green through yellow to red. Green/red spot implies that the spot/gene is over-expressed/under-expressed in the sample labelled with green/red dye. A yellow spot implies that the spot is expressed equally in both samples and a black spot implies that the spot is not expressed in either of the samples. Figure 3 displays one such composite image.

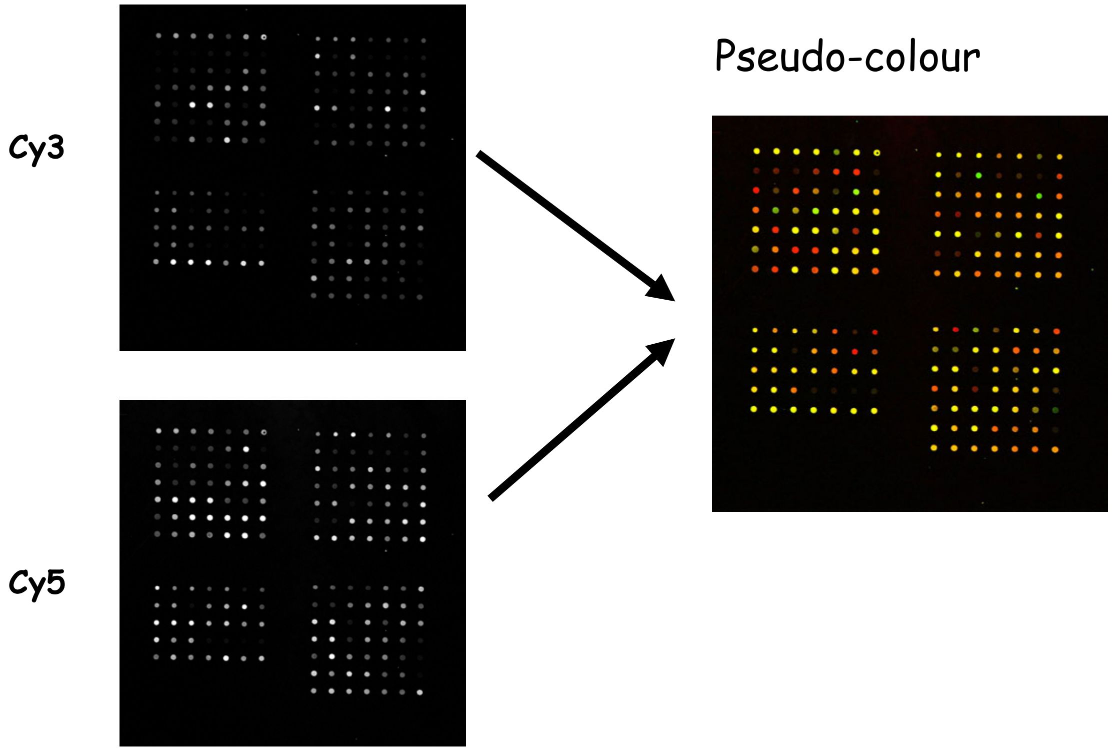

Figure 3: *A pseudo image is obtained by overlapping the two channels*

Spots within a microarray slide are divided into sub-grids. These sub-grids are arranged on the slide into rows and columns, and are usually separated from their neighbouring grids by small spaces. Each sub-grid is created by one pin of the printing head. A spot can be localized on the array by specifying its location in terms of the sub-row, sub-column, row, and column. Figure 4 displays a microarray slide with 32 sub-grids.

The processing of scanned microarray images can be divided into:-

1) Spot finding or gridding
2) Segmenting
3) Quantification or intensity extraction
4) Spot quality assessment

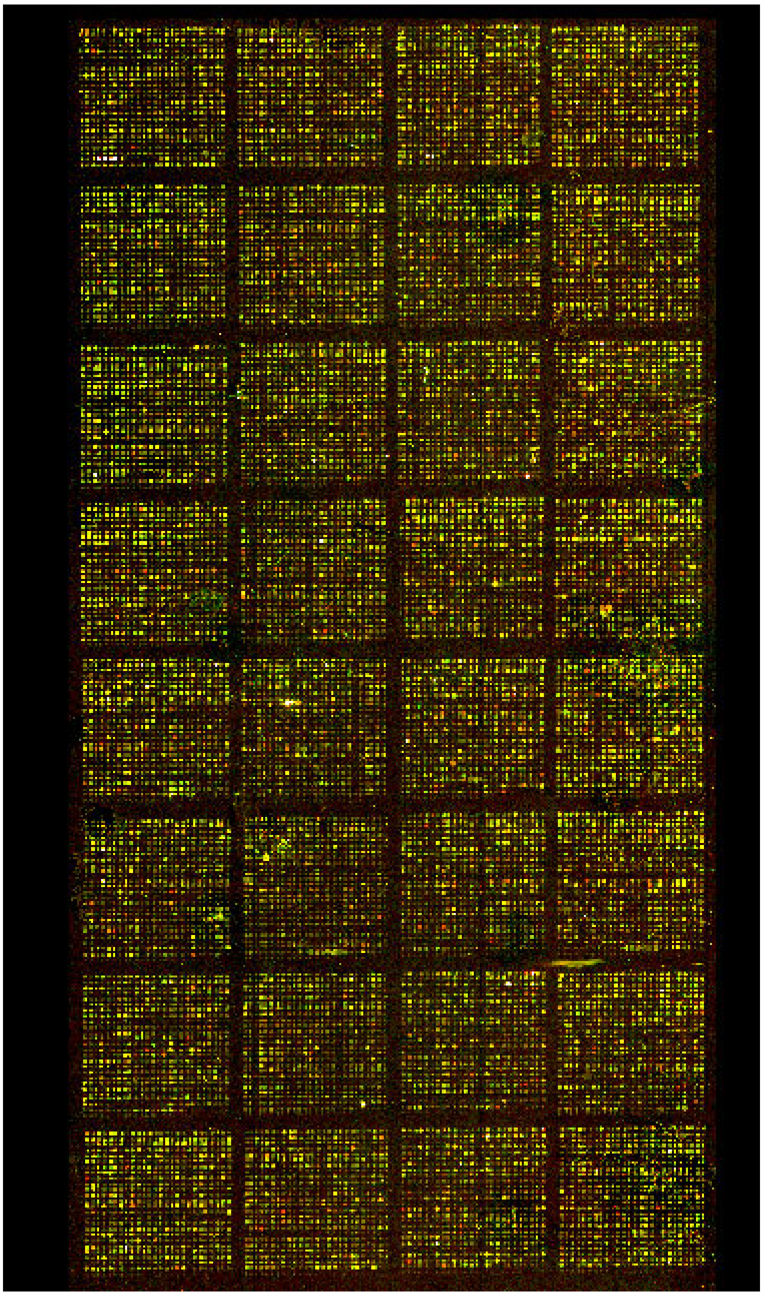

Figure 4: *A microarray with 32 sub-grids*

The number of spots on the array, the pattern according to which they are printed, and their sizes are all known in advance. Therefore image processing should be fairly simple, but unfortunately this is not the case. In reality, the exact location of each grid may vary from slide to slide. Furthermore, individual spots within a sub-grid may be severely misaligned. Reasons contributing to imperfect gridding are mainly caused by problems while spotting, such as hybridization inconsistencies, and by the necessity to print dense arrays. As a result, the first step in image processing is finding the exact positions of the spots, which can sometimes be rather far from their expected location.

The spot finding operation aims to locate the spots in images and estimate the size of each spot. There are various levels of sophistication in the algorithm for finding spots, depending upon the degree of human intervention in performing the operation.

Image segmentation is the next step that aims at deciding which pixels forming the spot should be considered for calculating the signal and which pixels should be considered for calculating the background signal. Figure 5 presents a segmented microarray image, where the foreground and background pixels have been separated. Segmentation also aims at identifying the pixels that are just noise artefacts. Several algorithms have been proposed for segmentation *e.g.*, pure spatial-based segmentation, intensity based segmentation, and Mann-Whitney segmentation [23].

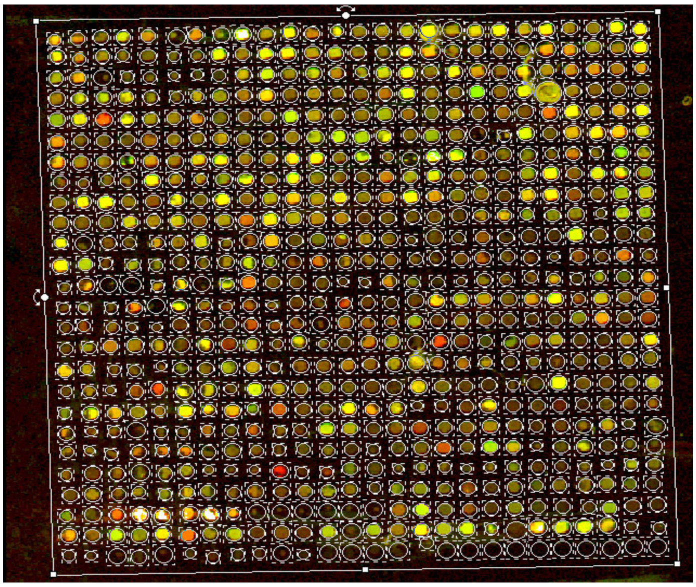

Figure 5*: An image of an array after segmentation algorithm has been applied. Foregrounds have been identified by circles, and corresponding backgrounds by the areas outside the circles, but still within the squares.*

The final goal of image processing is to compute a value that hopefully is directly proportional to the quantity of mRNA that was hybridized on the chip. Such a value is computed using the spot quantification procedure. The purpose of spot quantification is to combine pixel intensity values into a quantitative measure that can be used to represent the expression level of a gene deposited on a given spot. Typically, spot quantification is done by taking the mean, median, or mode of intensities of pixels of the spot.

Spot quality assessment is an important feature that must be considered while image processing. Uniformity, signal to noise ratio, shape, and diameter of the spot are features that determine the quality of a spot. Information regarding low quality spots should be collected and data corresponding to these spots should not be used for inference. In Publication I, we use an additional dye known as SYBR green RNA II for segmentation and for quality assessment.

It is said that the total fluorescent intensity from a spot is proportional to the expression strength under idealized conditions. These idealized conditions are:

1) The preparation of the target cDNA solution is done appropriately, so that the cDNA concentration in the solution is proportional to that in the tissue.
2) The hybridization is done appropriately, so that the amount of cDNA binding on spots is proportional to the target cDNA concentration in the solution.
3) The amount of cDNA deposited on each spot during chip fabrication is constant.
4) Spots are not contaminated.
5) The pixels contributing to the signal are correctly identified.

Whether the first two conditions are satisfied or not should be controlled at the time of designing the experiment. But the last three conditions are usually violated and affect the measurements obtained from image analysis. The amount of DNA deposited during the spotting procedure may vary from time to time and from spot to spot as a result of which the spot size cannot be considered as constant. Spot contamination due to dust, artefacts *etc*. leads to incorrect identification of the pixels contributing to the signal. Thus, it is extremely important to consider the above points while designing and conducting the experiment, and while analyzing the data.

4. Experimental design and data pre-processing

4.1 Experimental design

Each microarray experiment consists of large number of steps. As a result, various sources of error and uncontrolled variability emerge while performing the experiment and affect the expression data generated. Some such sources of error are listed in Table 1. In most cases their effect is clearly visible on the scanned microarray image.

Although all potential sources of error in measuring the signals from microarray experiments are not completely understood, the extent to which these complexities are known should be considered carefully when designing experiments. This would help obtain high quality data and eventually more precise results. In addition, practical constraints should also be considered while designing an experiment, such as the limited number of slides which can be hybridized in any given experiment, limited availability of the mRNA probes, or other cost considerations.

Replication of the biological samples, technical replicates (two RNA samples obtained from each experimental unit), and duplication of the spots, are mandatory elements of every design. It is also crucial that all the details about the experiment, the factors that would influence the experiment and their levels of interest, are known in advance. One should also identify the nuisance factors and divide them into controllable and uncontrollable factors. The factors that can be controlled should be blocked and the ones that cannot be blocked should be randomized. Apart from considering the listed points (above), one must keep a check on the data that are being generated. This will enable the researcher to modify the design if the quality of the generated array is poor.

Usually appropriate designs are not investigated while planning an experiment. Designs that appear simple and easy to execute are chosen without inferring whether the design would minimize the effect of unwanted variation, or maximize the precision of the estimates of interest. In addition, improved technology has allowed the usage of three to four dyes on the same array. On the other hand, a larger number of samples on the same array have further increased the complexity of the analysis of the resulting data. The purpose of Publication II was to describe approaches for optimal planning and designing of microarray experiments for any number of dyes, arrays and conditions.

Factors	***Comments***
mRNA preparation	Tissues, kits and procedure vary
Transcription	Inherent variation in the reaction, type of enzyme used
Pin geometry	Variation in the pins
Target volume	Fluctuations in the volume spotted while printing even for the same pin
Hybridization parameters	Temperature, humidity, time and buffering condition affect the hybridization.
Slide in homogeneities	Slide production parameters lead to variation in slides from different batches.
Non-specific hybridization	cDNA hybridizes to background or to sequences that are not their exact complement
Scanner setting	Scanner settings can cause a shift in the distribution of the pixel intensities
Dynamic range limitations	Limitation of the acquisition device does not allow to measure signal beyond 2^{16}-1
Image alignment	Images of the same array obtained at different wavelengths are not aligned. This leads to different pixels to be considered for the same spot
Spot shape	Irregular spots are hard to segment from their background
Segmentation	Bright contaminations may seem like signal and misguide while segmentation
Spot quantification	Pixel mean or median and area of the spot are few parameters used to quantify spots. Choice of parameter can lead to variation from slide to slide.

Table 1: *Sources of variations in a microarray experiment*

4.2 Data pre-processing

Despite a carefully designed microarray experiment, there are always error-sources that cause variations to the data generated. To reduce the effect of these error-sources, microarray data need to be pre-processed.

General pre-processing techniques

4.2.1 Background correction

The fluorescence of a spot is the cumulative effect of the fluorescence of the background and the fluorescence due to the labelled mRNA. To obtain the value to the amount of mRNA, one needs to subtract the value corresponding to the background. This is the idea behind background correction and it can be done in several ways, by locally correcting the background, performing a sub-grid background correction, background correction using blank spots, or background correction using control spots. A brief description of these methods is presented below.

Local background correction

In local background correction method, the intensity of the background is calculated using the local area around the spot. A measure of central-tendency (*e.g,*, mean, median or mode) is used to calculate the local background and this is subtracted from the intensity of the spot. This method is preferred when background intensity varies considerably from spot to spot and is avoided when the local neighbourhoods of the spots do not contain sufficiently many pixels. This method is usually applied on high-density arrays where the spots are separated by only a few pixels.

Sub-grid background correction

Most current robots print a sub-grid using the same pin so that each sub-grid should be homogenous as far as its shape and the size of the spots are concerned. A sub-grid includes sufficiently many pixels to allow a more reliable estimate of a measure of central tendency while it is still smaller than the whole array and may be flexible enough to compensate for local variations in the background intensity. In this method, a measure of central tendency is calculated for all the spots in a sub-grid. This is a useful approach particularly for high density arrays.

Background correction using blank spots

This method can be used when the design of an array includes a few blank spots, *i.e.,* spot locations where no DNA was deposited. A measure of central tendency is calculated on a number of such blank spots.

Background correction using control spots

The spot intensity depends on the properties of interaction between labelled target and DNA deposited in the spot. Some researchers have concluded that the labelled target may be more likely to stick to the substrate in the background of a spot than to hybridize non-specifically on a spot containing some DNA. In this case, subtracting any value characterizing the target-substrate interaction may be an over-correction. A possibility is to use some control spots using exogenous DNA, and the intensities resulting from such non-specific hybridization, as more adequate determinants of background correction.

4.2.2 Log-transformation

The logarithmic transformation has been used to pre-process microarray data from the very beginning. There are several reasons for this: firstly, it provides values that are more easily interpretable and more meaningful from a biological point of view. Secondly, the log-transformation makes the distribution of the data more symmetrical and almost normal [26, 27].

4.2.3 Combining replicates and eliminating outliers

For reasons discussed above, microarray data generally involve large amounts of noise. Repeated measurements can help reduce such noise, and also facilitate comparison of the inter-experiment and within-experiment variations. The repeated measurements may be the different spots in cDNA arrays, or the same spot can be measured on different cDNA arrays. In many situations it is natural to combine the values of all replicates to obtain a single estimate that would be representative for the given gene/condition. Typically, the estimate is obtained by calculating a measure of central tendency, such as the mean, median or mode. Some measures of central tendency can be misleading, but there are also strong incentives to calculate a unique value for each gene representing its expression in a given condition. Such incentives may include the ability to compare various genes across different conditions or tissues, to store the values in expression databases for later retrieval, *etc*.

The value obtained after combining data from several replicates should include additional parameters of the distribution of the original values. Such parameters may include: the number of values, standard-deviation, *etc*. These additional parameters may be used to assess the confidence in a particular value (*e.g.*, mean) and to eliminate outliers.

4.2.4 Normalization

The driving force behind the extensive use of microarrays is the hope that eventually all meaningful comparisons between the gene expression levels in various conditions and/or tissues resulting from different experiments would be possible. A critical requirement before such comparisons are possible is to normalize the data in such a way that the data are independent of the condition and technology used. Research has been made to make data comparable both within and across technologies [28, 29]. Within a technology comparisons leads to the difficulties because of the variations in the intensities from different arrays used. This can be due to many causes, including different protocols, different amounts of mRNA, different settings of the scanner, differences between individual arrays or labelling kits, differences between the channels (dyes) used, etc. The goal of normalization is to make the values corresponding to individual genes comparable across arrays, by retaining the systematic effects resulting from the biological process of interest and by removing other systematic technical variations.

The need for normalization of expression data can be seen most clearly in self-self experiments. In such experiments, two identical mRNA samples are labelled with different dyes (Cy3 and Cy5) and hybridized onto the same slide. Since the experiment involves self-self hybridization, there should be no differential expression and, in an ideal situation, the intensities corresponding to the two dyes should be equal. However, it has been observed that the intensity corresponding to the red (R) dye is often lower than the intensity from the green (G) dye (red and green are commonly used colours for cDNA microarrays). Furthermore, the imbalance is usually not constant across the spots within and between arrays, but varies according to the location on the array, slide origin, or other variables.

Normalization procedures of varying complexity have been proposed to account for these problems, but it is quite difficult to predict which normalization procedure would be best suited to a particular data set. Sometimes, procedures might even introduce new sources of variation due to the uncertainty with which their parameters can be estimated. The normalization methods which have been presented in the literature can be divided into two groups: linear methods and non-linear methods. The linear methods generally involve either estimating one or more global constants for a microarray, or fitting a linear regression to the $\log(R)$ versus $\log(G)$ data [30, 31]. The non-linear methods that have been developed involve transforming the data onto the axes $((\log(R) + \log(G))/2$ versus $\log(R/G)$, and robustly fitting one or more robust lowess curves [32] to the data and computing the residuals from the curve fit [33, 34]. Rescaling of the points is done by dividing each final residual value by a robust estimate of the standard deviation of the residuals, the median absolute deviation. When a single curve is fitted, the method is referred to as slide normalization, and when a curve is fitted to the data for each individual array printer pin, the method is referred to as pin normalization. Spatial bias may still be present even after slide normalization. Other nonlinear normalization methods such as B-splines, wavelets, kernel smoothers and support vector regression have been discussed by Fujita *et al*. [35]. Figure 6 presents a flowchart from Park *et al*. [36] of the commonly used normalization methods. Many other model based techniques (both Bayesian and non-Bayesian) have also been proposed [37, 38, 39, 40] but not discussed in Figure 6.

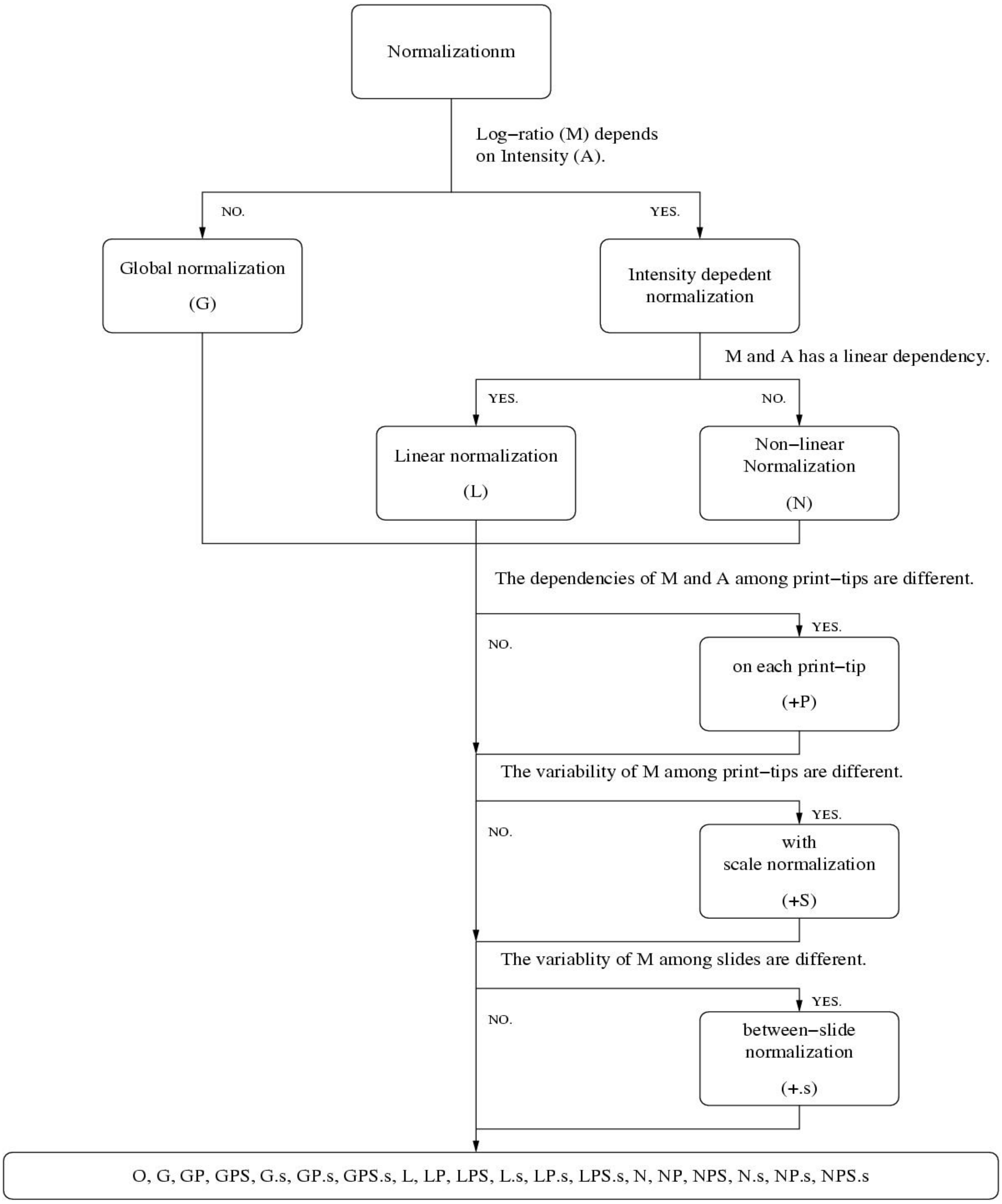
Normalizationm
Log−ratio (M) depends on Intensity (A).
NO.
YES.
Global normalization (G)
Intensity depedent normalization
M and A has a linear dependency.
YES.
NO.
Linear normalization (L)
Non−linear Normalization (N)
The dependencies of M and A among print−tips are different.
YES.
NO.
on each print−tip (+P)
The variability of M among print−tips are different.
YES.
NO.
with scale normalization (+S)
The variablity of M among slides are different.
YES.
NO.
between−slide normalization (+.s)
O, G, GP, GPS, G.s, GP.s, GPS.s, L, LP, LPS, L.s, LP.s, LPS.s, N, NP, NPS, N.s, NP.s, NPS.s

Figure 6: *Flowchart of normalization methods*

O: Original data
G: Global median normalization (GP, GPS, G.s, GP.s, GPS.s are its variants)
L: Intensity dependent linear normalization (LP, LPS, ls, LP.s, LPS.s and its variants)
N: Intensity dependent non-linear normalization (LOWESS) (NP, NPS, N.s, NP.s, NPS.s and its variants)
P: Print-tip normalization
S: Print-tip scale normalization
.s: Between-slide scale normalization

5. Selecting candidate genes and their functional annotation

In most cases, the purpose of a microarray experiment is to compare the gene expression levels in two different samples and to identify genes that are differentially expressed between these samples. Although this problem is simple in principle, it becomes more complex in reality because the measured intensity values are affected by numerous sources of fluctuations and noise [41, 42]. A few methods are discussed below that are used to find differentially expressed genes. These methods are independent of the technology used to generate the data (*e.g.*, cDNA or Affymetrix).

5.1 Commonly used methods to find differentially expressed genes

Fold change

Fold change is the simplest method for identifying differentially expressed genes. It evaluates the log ratio between two conditions (or the average of ratios when there are replicates) and considers all genes that differ by more than an arbitrary cut-off value to be differentially expressed [43, 44, 45]. This is not a statistical test as there is no associated value that can indicate the level of confidence in the designation of genes as differentially or not differentially expressed.

The most important drawback of this method is that, because the threshold is chosen arbitrarily, it may often be inappropriate. For instance, if we want to select genes with at least 2 fold-change to be differentially expressed and the condition under study does not affect any gene to the point of inducing a 2 fold change, no genes will be selected resulting in zero sensitivity. On the other hand, if the condition is such that many genes are changing dramatically, the method will select too many genes and will have a low specificity.

Another important disadvantage is related to the fact that the microarray technology tends to have a bad signal/noise ratio for genes with low expression levels. This leads to large variability at the low end and low variability at the high end for the log-transformed data. Since the fold change uses a constant threshold for all genes, it will introduce false positives at the low end while missing true positives at the high end [46, 47]. Intensity-specific thresholds have been proposed as a remedy for this problem [48].

The t- test

A better choice is to rank genes according to the absolute value of the t-statistics,

$$t = M_g^*/(S_g/\sqrt{n})$$

where M_g^* = the mean of the M_g-values ($M_g = \log_2 R_g/G_g$, R_g is the intensity for gene g on the red channel and G_g is the intensity for gene g on the green channel) for any particular gene across the replicate arrays, S_g = the standard deviation of the M_g values across the replicates for the gene, and n = the number of replicates. Any M_g-value that is an outlier will give rise to large standard deviations, which will usually prevent the gene in question from being spuriously identified as differentially expressed. However, genes with small sample variances have a good chance of having a large t-statistic even if they are not differentially expressed. This gene specific *t-test* is not affected by heterogeneity in variance across genes, because it only uses information from one gene at a time. It may, however, have low power because of the small sample size—the number of replicates. In addition, the variances that are estimated separately for each gene are not stable, and if the estimated variance is small, by chance, the corresponding t-value can be large even when the fold change is small.

Modifications of t-tests
More stable estimates can be obtained to find differentially expressed genes but these are subject to bias when the assumption of homogenous variance between genes is violated. In such situations, modified versions of the *t-test* are both more powerful and less subject to bias.

The 'significance analysis of microarrays (SAM) is a modified version of the *t*-test (known as the *S*-test) [49]. In here, a small positive constant is added to the denominator of the gene-specific *t*-test (discussed above), *i.e,*.

$$S = M_g^{*}/[c + (S_g/\surd n)]$$

where the constant *c* can be taken to be the 90th percentile ($S_g/\surd n$) value). With this modification, genes with small fold changes are less likely to be selected as significant.

Another variant of the *t-test* is known as the *regularized t-test* [50]. This test combines information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator for a gene-specific *t-test.* Yet another variant of *t-test* is the *B-statistic* proposed by Lonnstedt and Speed [51]. It is a log posterior odds ratio of differential expression versus non-differential expression and allows for gene-specific variances, also combining information across many genes.

The *t-* and *B-tests* based on log ratios can be found in the Statistics for Microarray Analysis (SMA) package [52]; the *S-test* is available in the SAM software package [53]; and the *regularized t-test* is in the Cyber T package [54]. In addition, the Bioconductor package [55] has a collection of various analysis tools for microarray experiments. Additional modifications of the *t-test* are discussed by Pan [56].

Various model based approaches using the Bayesian framework have also been published to identify differentially expressed genes [57, 58].

5.2 Functional annotation

So far we have dealt with the analysis of expression data, eventually producing a list of genes that were significantly different in the samples considered. However, the ultimate aim of expression analysis is not to have the list of differentially expressed genes but to produce results that make sense biologically. Thus, there is a need to translate the list of differentially expressed genes into a functional profile that offers an insight into the cellular mechanism active in the given condition.

In this section, we look at tools/ontologies designed to map the tens, or sometimes hundreds, of differentially expressed genes to biological, molecular, and cellular functions.

Gene Ontology (GO)
Due to the complex and distributed nature of biological research, our current biological knowledge is spread over many databases maintained by many independent groups. Researchers usually need to visit many of these databases to integrate comprehensive annotation information for their genes. This search is further hampered by the wide variation in terminology in different databases.

Gene Ontology (GO) project addresses the need for consistent descriptions of gene products in different databases [59]. It includes three independent structured controlled vocabularies/ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species. A gene product might be associated with or located in one or more cellular components and may be active in one or more biological processes, during which it performs one or more molecular functions. The use of GO terms by collaborating databases facilitates uniform queries across them. The controlled vocabularies are structured so that they can be queried at different levels. For example, one can use GO to find all the gene products in the mouse genome that are involved in signal transduction.

Go itself is not populated with gene products of any organism, but rather GO terms are used as attributes of genes and gene products by related databases. Databases use these GO terms to annotate objects such as gene or gene products stored in their repositories.

DAVID (The Database for Annotation, Visualization and Integrated Discovery) Annotation Tool
DAVID knowledgebase [60] is designed to facilitate high throughput gene functional analysis. For a given gene list, it not only provides quick accessibility to a wide range of heterogeneous annotation data in a centralized location, but also enriches the level of biological information for an individual gene.

DAVID facilitates the analysis via four web-based analysis modules: 1) Annotation Tool - rapidly adds descriptive data from several public databases to lists of genes; 2) GoCharts - assigns genes to Gene Ontology functional categories based on user selected classifications and term specificity level; 3) KeggCharts - assigns genes to KEGG metabolic processes and enables users to view genes in the context of biochemical pathway maps; and 4) DomainCharts - groups genes according to PFAM conserved protein domains. The functionality provided by DAVID accelerates the analysis of genome-scale datasets by facilitating the transition from data collection to biological meaning.

FatiGO
FatiGO [61] takes two lists of genes and converts them into two lists of GO terms using the corresponding gene-GO association table. Then a Fisher's exact test for 2×2 contingency tables is used to check for significant over-representation of GO terms in one of the sets with respect to the other one. Multiple test correction to account for the multiple hypotheses tested is then applied.

OBO-Edit

OBO-Edit [62] is a graph-based tool with emphasis on the graph structure of an ontology and provides a user friendly interface. OBO-Edit is developed and maintained within the GO Consortium.

6. Hierarchical modeling and Baye's theorem

6.1 Basics

Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability.

A Bayesian approach to a problem starts with the formulation of a model that one hope is adequate to describe the situation of interest. One then formulate a prior distribution over the unknown parameters of the model, which captures ones beliefs about the situation before seeing the data. After observing some data, one applies Bayes' Rule to obtain a posterior distribution for these unknowns, which takes account of both the prior and the data. From this posterior distribution, one can compute predictive distributions for future observations.

In Bayesian analysis, both the model parameters and unobserved data (missing data) are treated as random variables. Let y = (y_1, y_2,…y_n) be the observed data and let $\theta = (\theta_1, \theta_2, \ldots.\theta_r)$ be a vector of unknown parameters. In order to make a probability statement about θ given y, one defines a joint distribution of θ and y. The joint probability density function can be written as a product of prior distribution $p(\theta)$ and the likelihood $p(\mathrm{y} \mid \theta)$:

$$p(\theta, \mathrm{y}) = p(\theta)\, p(\mathrm{y} \mid \theta) = p(\theta \mid \mathrm{y})\, \mathrm{p(y)} \qquad (1)$$

which leads to:

$$p(\theta \mid \mathrm{y}) = p(\theta, \mathrm{y}) \,/\, p(\mathrm{y}\,) = \ (p(\mathrm{y} \mid \theta)\, p(\theta))/\, p(\mathrm{y}) \qquad (2)$$

where, $p(\mathrm{y}) = \int p(\mathrm{y} \mid \theta)\, p(\theta)\, \mathrm{d}\theta$. An equivalent form of equation (2) is

$$p(\theta \mid \mathrm{y}) \propto p(\mathrm{y} \mid \theta)\, p(\theta)$$

where one omit the factor $p(\mathrm{y})$, which does not depend on θ and for a fixed value of y can be considered as a constant.

Often the prior on θ depends in turn on other parameters φ that are not mentioned in the likelihood. So, the prior $p(\theta)$ must be replaced by a prior $p(\theta \mid \varphi)$, and a prior $p(\varphi)$ on the newly introduced parameters φ is required, resulting in a posterior probability

$$p(\theta, \varphi \mid \mathrm{y}) \propto p(\mathrm{y} \mid \theta)\, p(\theta \mid \varphi)\, p(\varphi)$$

This is the simplest example of a Bayesian hierarchical model. The process may be repeated; for example, the parameters φ may depend in turn on additional parameter ϕ, which will require its own prior. Eventually the process must terminate, with priors that do not depend on additional model parameters. The parameters that are not of interest, so-called nuisance parameters, are integrated out from the full posterior.

A computational challenge in applying Bayesian methods is that the integration required for inference is generally not tractable in closed form, and thus must be approximated numerically. Intractable integrations are quite common in case of nuisance parameters (typically unknown variances). Markov chain Monte Carlo (MCMC) integration methods, such as the Metropolis-Hastings algorithm [63, 64] and the Gibbs sampler [65, 66], provide often a feasible approximate numerical solution to the above mentioned problem. MCMC methods work by sampling from the probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps acts as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps.

The convergence of the Markov chain to the correct stationary distribution can be guaranteed for an enormously broad class of posteriors. This convergence is also the source of difficulty in actually implementing MCMC procedures for two reasons. First, it forces us to make decision about when to stop the sampling algorithm and summarize its output. Second, determination of the quality of the estimates produced may be problematic, as the samples are not i.i.d draws from the posterior but correlated samples.

6.2 Modeling with WinBUGS

All the Bayesian models described in this thesis have been implemented using WinBUGS (the MS Windows operating system version of BUGS: Bayesian Analysis Using Gibbs Sampling) [67]. WinBUGS is a versatile package that has been designed to carry out Markov chain Monte Carlo (MCMC) computations for a wide variety of Bayesian models. The software is currently distributed electronically from the BUGS Project website (http://www.mrc-bsu.cam.ac.uk/bugs/overview/contents.shtml). WinBUGS implements various MCMC algorithms to generate simulated observations from the posterior distribution of the unknown quantities (parameters) in the statistical model. The idea is that with sufficiently many simulated observations, it is possible to get an accurate picture of the distribution. Convergence diagnostics, model checks comparisons, and other plots are also available.

6.3 Hierarchical models for expression data

In this thesis, we have attempted to build hierarchical models for (1) solving signal saturation using spot data (Publication III), (2) solving signal saturation using pixel data (Publication IV), and (3) finding differentially expressed genes between two experimental conditions that include simultaneous correction for signal correction, array effects, and dye effects (Publication V).

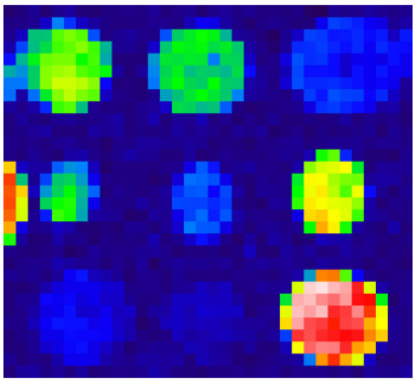

Figure 7: *A small portion of a scanned image from the Cy3 channel. It displays nine spots using "rainbow" colour map, where the blue end of the spectrum represents low pixel values and the red end of the spectrum represents high pixel values.*

Figure 7 displays nine spots on the hybridized array using "rainbow" colour map. The spots are composed of pixels where the blue end of the spectrum represents low pixel values and the red end of the spectrum represents high pixel values. The red pixels in the figure are saturated. The aim behind Publications III and IV was to estimate the true underlying signal (T) of a spot by combining information from multiple (*e.g.*, three) scans made at varying scanner sensitivities. Multiple scans ensure that the intensity level of the weakly expressed genes exceeds the intrinsic noise level of the scanner and saturation of the highly expressed genes is avoided. The underlying logic in both these publications is quite similar but they deal with different types of input data. Figure 8 demonstrates a pictorial representation of the hierarchical model used in Publication III using spot intensity data.

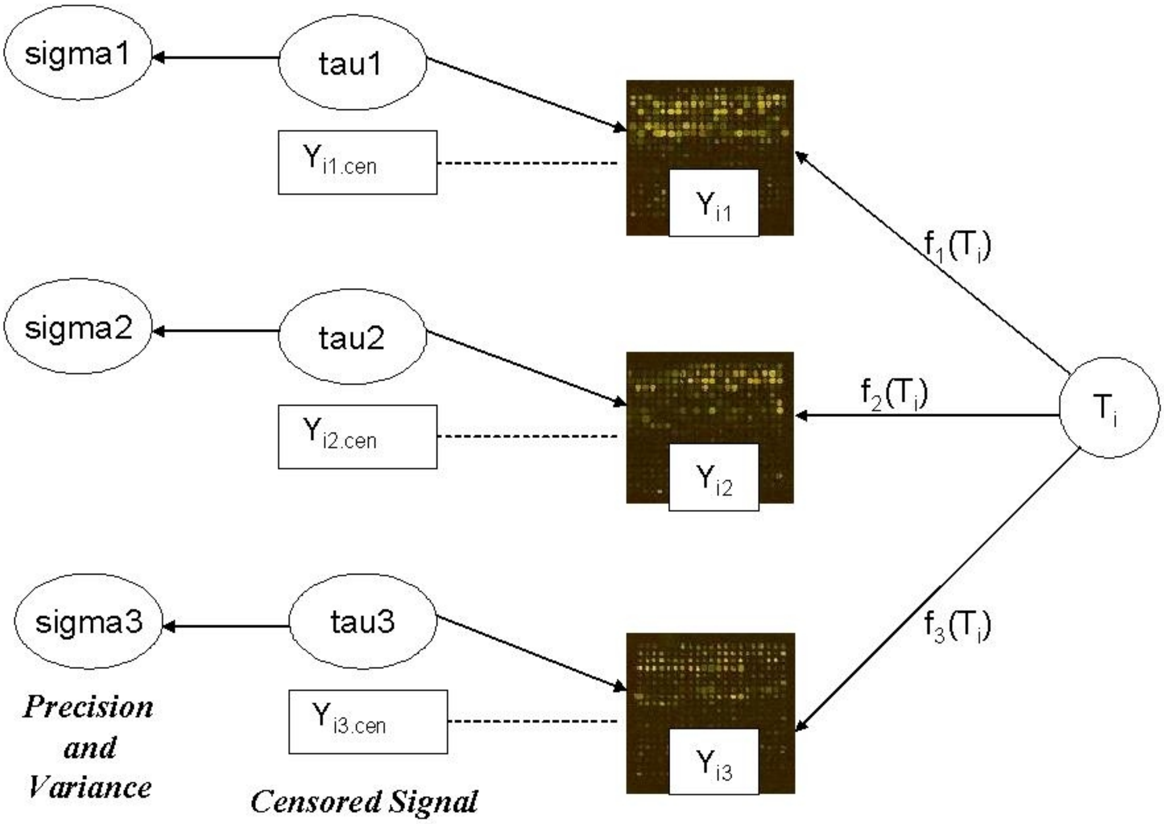

Figure 8: *A pictorial representation of the hierarchical model used in Publication III, where Y_{is} denotes a measured signal from spot i of scan s, $Y_{is\cdot cen}$ denotes a censored observation, T_i denotes the underlying signal that needs to be estimated, f_i's are functions used for calibrating the three scans, tau's are precision parameters and sigma's are the square roots of variances.*

Publication IV models signal saturation, using the measured signal from the pixels that forms the spots. Saturation happens for a pixel when the signal from the pixel exceeds the scanner's upper threshold of detection; therefore, modelling the signal from pixels gives a more truthful description of the saturation phenomenon than modelling the signal from the spots (spot signal is the summary of signal from the pixels comprising the spot). However, modelling the signal from pixels is computationally more demanding as each spot consists of 80–100 pixels. Figure 9 demonstrates a schematic diagram showing the connection between Publication III and Publication IV.

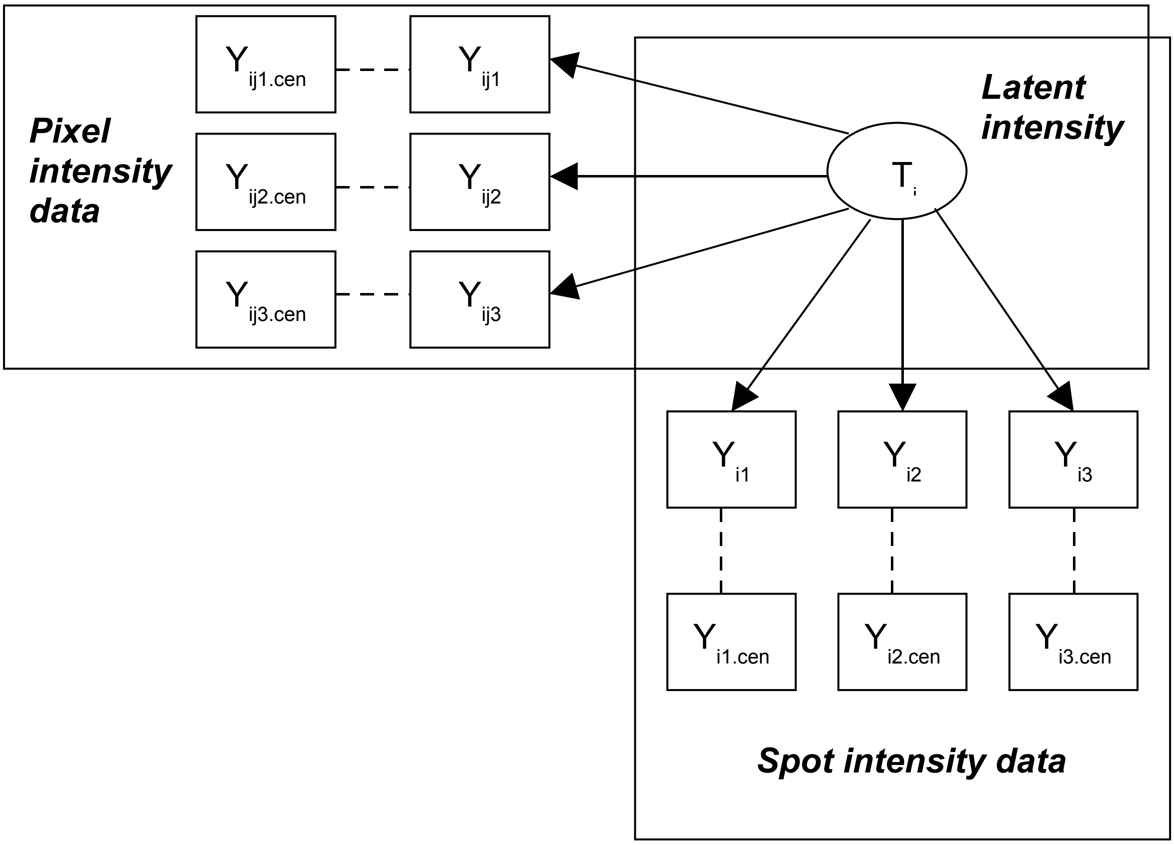

Figure 9: *A schematic diagram showing the connection between Publication III (using spot intensity data) and Publication IV (using pixel intensity data). Y_{is} denotes the observed intensity for spot i under scan s, $Y_{is.cen}$ denotes the corresponding censored observation, Y_{ijs} denotes the observed intensity for pixel j of spot i under scan s, $Y_{ijs.cen}$ denotes the corresponding censored observation, T_i denotes the underlying signal that needs to be estimated.*

Publication V was an extension of Publication III and models differential expression by correcting for signal saturation, array effects, and dye effects. A pictorial representation of the hierarchical model of Publication V is presented in Figure 10.

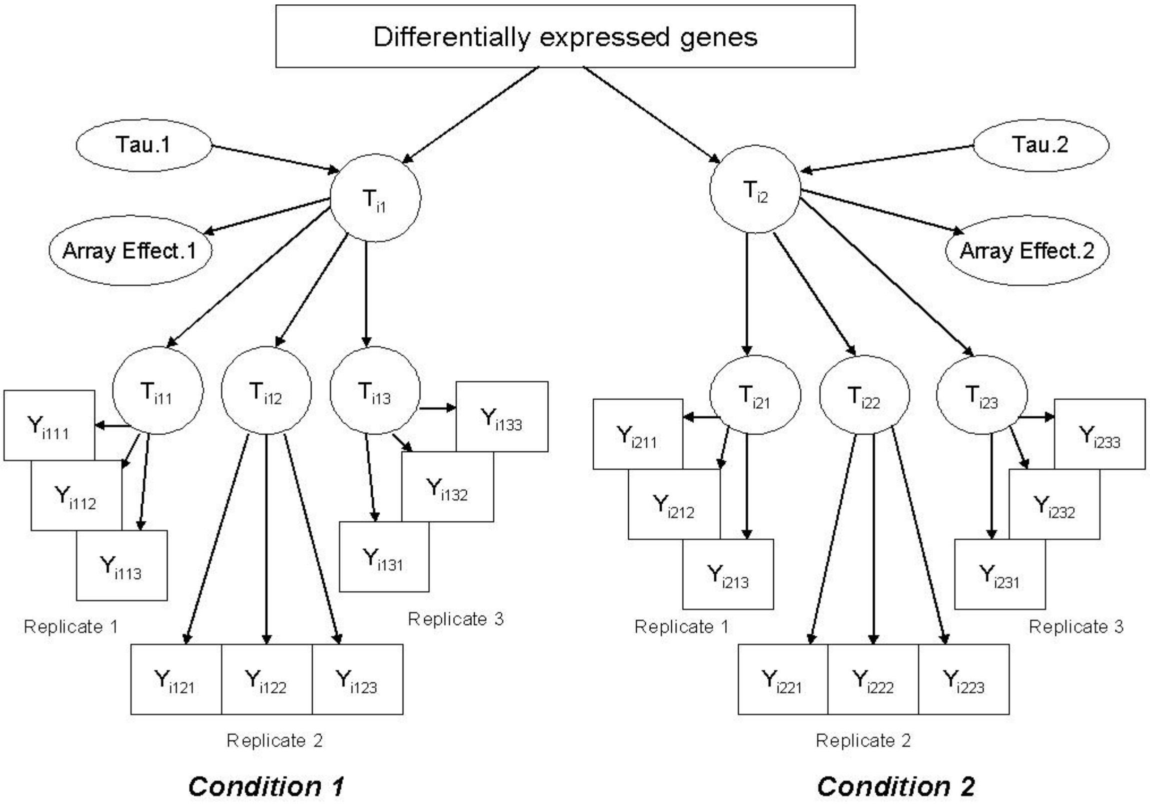

Figure 10: *Pictorial representation of the hierarchical model of Publication V, where Y_{icrs} denotes the observed intensity for spot i under condition c and scan s of replicate r, T_{ic} denotes the true latent intensity of the gene i under condition c, T_{icr} denotes the true latent intensity of the gene i under condition c of replicate r, and tau's are precision parameters.*

7. Conclusions

Microarrays are high-throughput biological assays that allow measuring the expression profiles of a large number of distinct genes. As a result, immense amounts of noisy data, which are corrupted by systematic and random noise occurring from various sources, are produced in such experiments. These data are then used for an improved understanding of the function of genes, including knowing when, where, and to what extent a gene is expressed. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms, broader cellular functions, and biochemical pathways. Inference based on these noisy data can be misleading, and therefore a major effort of research is currently directed towards developing methods for extracting improved gene signals and towards sharper methods of data analysis. The main focus of this thesis is to describe approaches for improving gene signal, and propose optimal designs for conducting microarray experiments. In addition, this thesis also describes methods for analyzing data using Bayesian framework.

7.1 Summary of publications

Microarray experiment consists of large number of steps and errors can be introduced at any of these steps while performing the experiment. As many as possible of these errors should be taken into consideration while designing the layout of the experiment. A careful planning of the experiment before the actual execution would eventually minimize the effect of unwanted variations and maximize the precision of the estimates of the parameters of interest. Publication II of this thesis describes approaches for planning and designing microarray experiments optimally for any number of dyes, arrays and conditions, considering both technical and biological replicates.

Once a careful design has been laid out and the experiment conducted, there is a need to measure the expression of genes from the hybridized arrays as precisely as possible. Publication I, Publication III and Publication IV aim at achieving this goal, but their focus is at different stages of the analysis. Publication I aim at improving image segmentation by using an additional dye, SYBR green RNA II. A strong signal emitted by SYBR green labelled probes and a low signal from the background allow clear distinction of foreground and background signals for the spots. This was used to learn about the spot quality and to flag spots which are not reliably hybridized and corrupted by noise. It was tested that the segmentation and quantification results obtained using this approach performed better than those produced by the commercial image analysis software, GenePix.

Publication III proposes a Bayesian hierarchical method for improving the quality of signal from DNA microarrays by analysing spot intensity data collected from several scans at varying scanner sensitivities. The method improves the accuracy at which expression can be measured at all ranges and extends the dynamic range of measured gene expression at the high end.

Publication IV solves the problem of improving the data quality and signal saturation but differs from Publication III in the type of data modelled. Since saturation occurs for the pixels comprising the spot and not for the spot, signals from pixels were used to model saturation phenomenon. Modelling pixel intensity data gives a more truthful description of the saturation phenomenon as opposed to spot intensity data, because spot summary data do not have a sharp threshold value beyond which saturation would have an effect.

Publication V focuses on the analysis aspect of the microarrays. It proposes a Bayesian hierarchical model for finding differentially expressed genes between two experimental conditions using an integrated statistical approach where signal correction, systematic array effects, dye effects, as well as differential expression, are all modelled jointly.

7.2 Future directions

This thesis suggests several directions for further research. An obvious extension would be to apply the integrated model and method of Publication V on the data obtained by image analysis using SYBR green RNA II, as proposed in Publication I. This extension would help improve the signal further, as intensities occurring from spots corrupted by noise would be avoided. Another possibility for further research would be to perform image analysis using SYBR green RNA II jointly with the other processing steps of microarrays analysis.

Another immediate extension of the model in Publication V would be to extend it for identifying genes that show differential expression over a time course. The current model in Publication V is successfully implemented using the WinBUGS software. WinBUGS gives the user the possibility to easily handle and modify the code. This ease is balanced against the long running time when dealing with genomic data. A possibility would be to implement the existing models in C or C++ for a realistic run time of the models.

Overall, the methods and techniques developed here could be used for the processing of data from other high-throughput techniques.

References

[1] Genome Project Statistic, NCBI Friday, 19 August, 2008.

[2] Lennon GG, and Lehrach H, “Hybridization analyses of arrayed cDNA libraries”, Trends Genet., 7, 314–317, 1991.

[3] Kafatos FC, Jones CW, and Efstratiadis A, “Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure”, Nucleic Acids Res., 7, 1541–1552, 1979.

[4] Gillespie D, and Spiegelman S, “A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane”, J. Mol. Biol., 12, 829–842, 1965.

[5] Southern EM, Case-Green SC, Eider JK, Johnson M, Mir KU, Wang L, and Williams JC, “Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids”, Nucleic Acids Res., 22, 1368–1373, 1994.

[6] Zhao N, Hashida H, Takahashi N, Misumi Y, and Sakaki Y, “High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression”, Gene, 156, 207–213, 1995.

[7] Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, and Jordan BR, “Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones”, Genomics, 29, 207–216, 1995.

[8] Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, and Solas D, “Light-directed, spatially addressable parallel chemical synthesis”, Science, 251, 767–773, 1991.

[9] Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, and Adams CL, “Multiplexed biochemical assays with biological chips”, Nature, 364, 555–556, 1993.

[10] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, and Fodor SP, “Light-generated oligonucleotide arrays for rapid DNA sequence analysis”, Proc. Natl. Acad. Sci., 91, 5022–5026, 1994.

[11] Schena M, Shalon D, Davis RW, and Brown PO, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”, Science, 270, 467–470, 1995.

[12] Shalon D, Smith SJ, and Brown PO, “A DNA microarray system for analyzing complex DNA samples using two-colour fluorescent probe hybridization”, Genome Res., 6, 639–645, 1996.

[13] DeRisi JL, Iyer VR, and Brown PO, “Exploring the metabolic and genetic control of gene expression on a genomic scale”, Science, 278, 680–686, 1997.

[14] Liang P, and Pardee AB, “Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction”, Science, 257, 967-971, 1992.

[15] Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW, “Serial analysis of gene expression”, Science, 270, 484-487, 1995.

[16] Albert B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P, “Molecular Biology of the Cell”, Garland Science, 4th edition, 2002.

[17] Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, and Davis RW, “Microarrays: biotechnology's discovery platform for functional genomics”, Trends in Biotechnology, 16, 301-306, 1998.

[18] Fisher W, and Zhang M, “A Biochip Microarray Fabrication System Using Inkjet Technology”, IEEE Transactions on Automation Science and Engineering, Volume 4, Issue 4, 488 – 500, 2007.

[19] Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, and Eberwine JH, “Amplified RNA synthesized from limited quantities of hetrogenous cDNA”, Proc. Natl. Acad. Sci., 87, 1663-1667, 1990.

[20] Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, and Coleman P, “Analysis of gene expression in single live neurons”, Proc. Natl. Acad. Sci., 89, 3010-3014, 1992.

[21] Manduchi E, Scearce LM, Brestelli JE, Grant GR, Kaestner KH, and Stoeckert CJ, “Comparison of different labelling methods for two-channel high-density microarray experiments”, Physiol. Genomics, 10, 169-179, 2002.

[22] Badiee A, Eiken HG, Steen WM, and Løvlie R, “Evaluation of five different cDNA labelling methods for microarrays using spike controls”, BMC Biotechnology, 3, 23, 2003.

[23] Yang Y, Buckley M, Dudoit S, and Speed T, “Comparison of methods for image analysis on cDNA microarray data”, J. Comput. Graph. Stat., 11, 108-136, 2001.

[24] Schena M, “DNA Microarrays: A Practical Approach”, Oxford University Press, 1999.

[25] Schena M, “Microarray Biochip Technology”, Eaton, 2000.

[26] Long A, Mangalam H, Chan B, Tolleri L, Hatfielf GW, and Baldi P, “Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework”, J. Biol. Chem., 276(23), 19937-19944, 2001.

[27] Speed TP, “Hints and prejudices – always log spot intensities and ratios”, Technical report, University of California, Berkley, 2000. http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html

[28] Elo LL, Lahti L, Skottman H, Kylaniemi M, Lahesmaa R, and Aittokallio T, “Integrating probe-level expression changes across generations of Affymetrix arrays”, Nucleic Acids Research, Vol. 33 (22): e193, 2005.

[29] Elo LL, Katajamaa M, Lund R, Oresic M, Lahesmaa R, and Aittokallio T, “Improving identification of differentially expressed genes by integrative analysis of Affymetrix and Illumina arrays”, Omics, 10(3):369-80, 2006.

[30] Finkelstein D, Ewing R, Gollub J, Sterky F, Cherry M, and Somerville S, “Microarray data quality analysis: lessons from the AFGC project”, Plant Mol. Biol., 48, 119-131, 2002.

[31] Richmond T, and Somerville S, “Chasing the dream: plant EST microarrays”, Curr. Opin. Plant Biol., 3, 108-116, 2000.

[32] Cleveland WS, “Robust locally weighted regression and smoothing scatterplots”, J. Am. Stat. Assoc., 74, 829–836, 1979.

[33] Dudoit S, Yang YH, Callow MJ, and Speed TP, “Statistical methods for identifying expressed genes in replicated cDNA microarray experiments”, Statistica Sinica, 12, 111-139, 2002.

[34] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP, “Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation”, Nucleic Acids. Res., 30, e15, 2002.

[35] Fujita A, Sato JR, de Oliveira Rodrigues L, Ferreira CE, and Sogayar MC, “Evaluating different methods of microarray data normalization”, BMC Bioinformatics, 7, 469, 2006.

[36] Park T, Yi SG, Kang SH, Lee SY, Lee YS, and Simon R, “Evaluation of normalization methods for microarray data”, BMC Bioinformatics, 4, 33, 2003.

[37] Kerr MK, Martin M, and Churchill GA, “Analysis of variance for gene expression microarray data”, Journal of Computational Biology, 7, 819-837, 2000.

[38] Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, and Paules RS, “Assessing gene significance from cDNA microarray expression data via mixed models”, Journal of Computational Biology, 8(6), 625-637, 2001.

[39] Bhattacharjee M, Pritchard CC, Nelson PS, and Arjas E, “Bayesian integrated functional analysis of microarray data”, Bioinformatics, vol. 20, no. 17, pp. 2943–2953, 2004.

[40] Lewin A, Richardson S, Marshall C, Glazier A, and Aitman T, “Bayesian modeling of differential gene expression”, Biometrics, vol. 62, no. 1, pp. 10–18, 2006.

[41] Draghici S, Kuklin A, Hoff B, and Shams S, “Experimental design, analysis of variance and slide quality assessment in gene expression arrays”, Current Opinion in Drug Discovery and Development, 4(3), 332-337, 2001.

[42] Schuchhardt J, Beule D, Wolski E, and Eickhoff H, “Normalization strategies for cDNA microarrays”, Nucleic Acid Research, 28(10), e47i-e47v, 2000.

[43]. Schena M, Shalon D, Heller R, Chai A, Brown PO, and Davis RW, “Parallel human genome analysis: microarray-based expression monitoring of 1000 genes”, Proc. Natl. Acad. Sci., 93, 10614-10619, 1996.

[44]. DeRisi JL, Iyer VR, and Brown PO, “Exploring the metabolic and genetic control of gene expression on a genomic scale”, Science, 278, 680-686, 1997.

[45]. Draghici S, “Statistical intelligence: effective analysis of high density microarray data”, Drug Discov. Today, 7, S55-S63, 2002.

[46] Rocke DM, and Durbin BA, “Model for measurement error for gene expression arrays”, J Comput Biol., 8, 557–569, 2001.

[47] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, and Tsui KW, “On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data”, J Comput. Biol., 8, 37–52, 2001.

[48] Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, and Quackenbush J, “Within the fold: assessing differential expression measures and reproducibility in microarray assays”, Genome Biol., 3(11), research0062, 2002.

[49] Tusher VG, Tibshirani R, and Chu G, “Significance analysis of microarrays applied to the ionizing radiation response”, Proc. Natl. Acad. Sci., 98, 5116–5121, 2001.

[50] Baldi P, and Long AD, “A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes”, Bioinformatics, 17, 509–519, 2001.

[51] Lonnstedt I, and Speed T, “Replicated microarray data”, Statistica Sinica., 12, 31-46, 2002.

[52] R package: statistics for microarray analysis.
http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html

[53] SAM: Significance Analysis of Microarray. http://www-stat.stanford.edu/~tibs/SAM

[54] Cyber T. http://www.igb.uci.edu/servers/cybert/

[55] Bioconductor. http://www.bioconductor.org

[56] Pan W, “A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments”, Bioinformatics, 18, 546–554, 2002.

[57] Baldi P, and Long AD, “A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes”, Bioinformatics, vol. 17, no. 6, pp. 509–519, 2001.

[58] Ramoni MF, and Sebastiani P, “Bayesian methods for microarray data analysis”, in Proceedings of the IMA Workshop 1: Statistical Methods for Gene Expression: Microarrays and Proteomics, Minneapolis, Minn, USA, September-October 2003.

[59] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G, “Gene ontology: tool for the unification of biology”, Nat Genet., 25(1), 25-29, 2000.

[60] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, and Lempicki RA, “DAVID: Database for Annotation, Visualization, and Integrated Discovery”, Genome Biology, 4(5), P3, 2003.

[61] Al-Shahrour F, Díaz-Uriarte R, and Dopazo J, “FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes”, Bioinformatics, 20(4), 578-580, 2004.

[62] OBO-Edit. http://oboedit.org

[63] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E, “Equation of state calculations by fast computing machines”, Journal of Chemical Physics 21, 1087-1092, 1953.

[64] Hastings WK, “Monte Carlo sampling methods using Markov chains and their applications”, Biometrika, 57, 97-109, 1970.

[65] Geman S, and Geman D, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”, IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721-741, 1984.

[66] Gelfand AE, and Smith AFM, “Sampling-based approaches to calculating marginal densities”, Journal of the American Statistical Association 85, 398-409, 1990.

[67] Spiegelhalter DJ, Thomas A, and Best NG, “WinBUGS”, Version 1.2. User Manual, MRC Biostatistics Unit, 1999.