Statistical Challenges in Functional Genomics

Paola Sebastiani[*] ***Emanuela Gussoni***[†] ***Isaac Kohane***[‡] ***Marco Ramoni***[‡]

[*]*Department of Mathematics and Statistics*
University of Massachusetts, Amherst MA

[†]*Division of Genetics Children's Hospital Boston*
Harvard Medical School, Boston MA

[‡]*Children's Hospital Informatics Program*
Harvard Medical School, Boston MA

Abstract

On February 12, 2001 the Human Genome Project announced the completion of a draft physical map of the human genome - the genetic blueprint for a human being. Now the challenge is to annotate this map, by understanding the functions of genes and their interplay with proteins and the environment to create complex, dynamic living systems. This is the goal of *functional genomics*. Recent technological advances enable biomedical investigators to observe the genome of entire organisms in action by simultaneously measuring the level of activation of thousands of genes under the same experimental conditions. This technology, known as *microarrays*, provides today unparaleled discovery opportunities and is reshaping biomedical sciences. One of the main aspects of this revolution is the introduction of computation intensive, data analysis methods in biomedical research. This paper reviews the foundations of this technology and describes the statistical challenges posed by the analysis of microarray data.

Keywords: Bioinformatics, classification, clustering, differential analysis, gene expression, functional genomics, microarray.

Submitted to: Statistical Science. To appear as discussion paper (December 2, 2002).

Corresponding Author: Paola Sebastiani. Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA. PHONE: 413 545 0622, FAX: 413 545 1801, EMAIL: sebas@math.umass.edu, URL: http://www.math.umass.edu/∼sebas.

Contents

Acknowledgments

This research was supported by NSF grant ECS-0120309, by NHLBI grant HL-99-024, and by the Genomics Core of Beth Israel Deaconess Medical Center, Boston, MA. Authors are grateful to Stefano Monti, Whitehead Institute, and Joseph Jerry, Department od Veterinary and Animal Sciences, University of Massachusetts, for their insightful comments, and to the referees and the editor for their invaluable suggestions.

1. The Human Genome Project

The Human Genome Project (HGP) is a multi-year effort, coordinated by the Department of Energy and the National Institute of Health, to create a reference sequence of the entire DNA and to identify the estimated 30,000-40,000 genes of the human genome. Officially started in 1990, the HGP is expected to render its final results in 2005, but the staggering technological advances of the past few years will probably allow the completion of the project by April 2003. By then, the total cost of the project will be in excess of $3 billion, making the HGP one of the most funded single scientific endeavors in history, in the league of the Manhattan Project and the Apollo Space Program. The rationale behind such a herculean effort is that a panoramic view of the human genome would dramatically accelerate advances in biomedical sciences and develop new ways to treat, cure, or even prevent the thousands of diseases that afflict humankind. The HGP is also delivering a wealth of commercial opportunities: sales of DNA-based products and technologies are projected to exceed $45 billion by 2009 in the U.S alone.

In June 2000, Craig Venter of Celera Genomics, the U.S. President Clinton and the leaders of the HGP consortium announced the completion of a "working draft" DNA sequence of the human genome, whose details were published in February 2001 in two dedicated issues of Nature and Science[1]. The result of these efforts is a map of the human genes. This map consists of about 30,000-40,000 protein-coding genes [21], only twice the number of protein-coding genes in a worm or a fly. Because about 50% of these discovered genes have known functions, the challenge now is to annotate this map, by understanding the functions of genes, and their interplay with proteins and the environment to create complex, dynamic living systems. This is the goal of *functional genomics*.

Several projects around the world are currently under way to discover gene functions and to characterize the regulatory mechanisms of gene activation. One avenue of research focuses on gene expression level, and exploits the recent technology of microarrays [29, 65, 67, 68] to obtain a panoramic view of the activity of the genome of entire organisms. Microarray technology is reshaping traditional molecular biology by shifting its paradigm from a hypothesis driven to a data driven approach [60]. Traditional methods in molecular biology generally work on a "one gene in one experiment" basis, making the whole picture of gene functions hard to obtain. Microarray technology makes it possible to simultaneously observe thousands of genes in action and to dissect the functions, the regulatory mechanisms, and the interaction pathways of an entire genome.

A fundamental component of functional genomics is the development of computational methods able to integrate and understand the data generated by microarray experiments. Typical experimental questions investigated with microarray experiments are the detection of genes differentially expressed in an abnormal/tumor cell compared to a normal cell; the identification of groups of genes characterizing a particular class of tumors; the recognition, at molecular level, of novel sub-classes of tumors; the detection of gene regulatory mechanisms. Although the avalanche of genome data produced with microarrays grows daily, no consensus exists about the best quantitative methods to analyze them. Many methods lack appropriate measures of uncertainty, make dubious distributional assumptions, and are hardly portable across experimental platforms. Furthermore, little is

[1] Volume 409 of Nature, published February 15 2001 and available at http://www.nature.com/genomics/human/, reports the findings of the publicly sponsored HGP, while volume 291 of Science, published February 16 2001 and available at http://www.sciencemag.org/content/vol291/issue5507, focuses on the findings of the draft sequence reported by the privately funded company Celera Genomics

known about how to design informative experiments, how to assess whether an experiment has been successful, how to measure the quality of information conveyed by an experiment and, therefore, the reliability of the results obtained. The specific character of gene expression data opens unique statistical problems.

The aim of this paper is to offer an overview of these problems and the main approaches proposed to tackle them. To make the paper self-contained, the next section will review essential biological notions and we refer to [40] for more technical details. Section 3 describes the two most used microarray platforms: cDNA and synthetic oligonucleotide microarrays. Experimental design issues are described in Section 4, and Section 5 focuses on data quality issues. Section 6 describes techniques used for the analysis of gene expression data measured in comparative experiments, while Section 7 focuses on the supervised and unsupervised methods used to analyze gene expression data from experiments comparing several conditions. Section 8 lists some of the critical open problems and the challenges they pose to the statistical community.

2. The Biology of Gene Expression

Cells are the fundamental working units of every living system. The nucleus of each cell contains the chromosomes that carry the instructions needed to direct the cell activities in the production of proteins via the DNA (deoxyribonucleic acid). The structural arrangement of DNA looks like a ladder twisted into a helix, the sides of the ladder are formed by molecules of sugar and phosphate, while the rungs consist of pairs of nucleotide bases A (Adenine), T (Thymine), C (Cytosine) and G (Guanine) joined by hydrogen bonds. In base pairing, A pairs with T, and G pairs with C. Each strand of the double helix consists of a sequence of nucleotides that are made of one of the four bases A, T, G, C, a molecule of sugar and one of phosphate. The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence. The *genome* is an organism's complete DNA and encodes the *genetic code* required to create a particular organism with its own unique traits. The nucleotide bases A, T, C, and G are the letters that spell out these genetic instructions by producing a three-letter word code, where each specific sequence of three DNA bases (codons) encodes an amino acid. Amino acids are the basic units of proteins, which perform most life functions.

With few exceptions, all human cells contain the same DNA but, despite carrying the same set of instructions, cells are actually different. These differences are due to the fact that, stimulated by cell regulatory mechanisms or environmental factors, segments of DNA express the genetic code and provide instructions to the cells on when and in what quantity to produce specific proteins. These segments of DNA are the *genes* and the process by which they become active is called their *expression*.

The modern concept of gene expression dates back to 1961, when the theory of genetic regulation of protein synthesis was first described by Jacob and Monod [51]. The fundamental discovery was that differential gene expression, that is when and in what quantities a gene is expressed, determines differential protein abundance thus inducing different cell functions. The *gene expression level* is an integer valued or continuous measure that provides a quantitative description of the gene expression by measuring the number of intermediary molecules produced during this process. These molecules are the mRNA (messenger Ribonucleic acid) and the tRNA (transfer Ribonucleic acid),

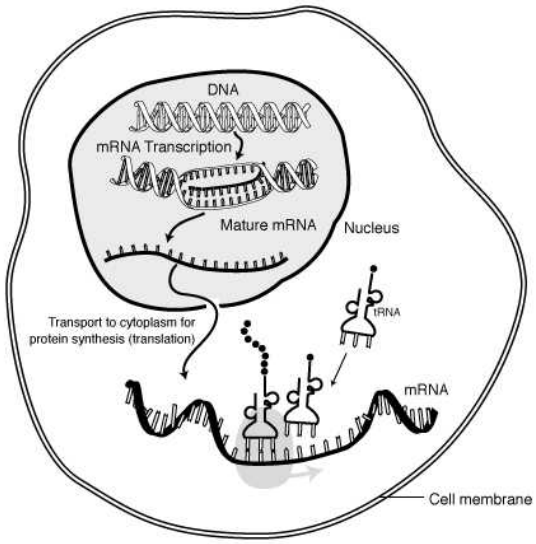

Figure 1: *During the expression process, a complementary copy of a gene code is transcribed into the mRNA. An appropriately modified copy migrates from the nucleus to the cytoplasm where it serves as a template for the protein synthesis. Picture taken from [47].*

and they are produced during the two steps of *transcription* and *translation* leading to the synthesis of a protein. This two-steps representation of the protein-synthesis process is depicted in Figure 1 and constitutes the *central dogma of molecular biology* [24]:

Transcription The first step of a gene expression is the creation of a complementary copy of the gene sequence stored in one of the two DNA complementary strands. The complementary copy of the gene DNA code transcribes U (Uracil) for A, A for T, G for C, and C for G into the mRNA.

Translation The mRNA transcript is moved from the nucleus to the cellular cytoplasm, where it serves as a template on which tRNA molecules, carrying amino acids, are lined up. The amino acids are then linked together to form a protein chain.

Because the gene expression consists of copying its DNA code into mRNA molecules, a measure of the gene expression level is the abundance of mRNA produced during this process [89]. This is the main intuition behind the large scale measurement of gene expression levels in microarrays that is described in the next section.

3. Large Scale Measurement of Gene Expression Levels

Quantitative methods to measure gene expression levels have been available to biologists for more than twenty years. Northern and southern blots (see [4, 106]) are techniques used to identify and locate mRNA and DNA sequences that are complementary to a segment of DNA. While these techniques are limited to examine a small number of genes at a time, a more recent technique, called Serial Analysis of Gene Expression (SAGE) [104], is able to measure the global gene expression from entire cells. SAGE technology was introduced in 1995 by a team of cancer researchers at Johns Hopkins to rapidly identify differences between cancer and normal cells. The main intuition behind this technology was that short but specific stretches of DNA are sufficient to uniquely identify the genes expressed in a particular cell. SAGE uses these *short sequence tags* to mark the transcripts of a gene and to identify the number of transcripts generated by each gene, thus providing a measure of the gene expression. This technology is useful for detecting and quantifying the absolute expression level of both known and unknown genes, but it is time-consuming as it involves multiple steps and extensive sequencing to identify the appropriate tags [66]. Microarray technology has rendered efficient this process by measuring, simultaneously, the relative expression level of a large number of genes and, in so doing, is reshaping the epistemological and methodological vision of molecular biology and biomedical sciences.

3.1 The Microarray Technology

The basic idea behind microarray technology is to simultaneously measure the relative expression level of thousands of genes within a particular cell population or tissue. Two key technical concepts behind this measurement process are *reverse transcription* and *hybridization*.

Reverse Transcription. The mRNA transcript of a gene can be experimentally isolated from a cell, and reversed-transcribed into a complementary DNA copy called cDNA. A collection of cDNAs transcribed from cellular mRNA constitutes the cDNA library of a cell. Similarly, double-stranded cDNA can be reversed transcribed into a complementary copy called cRNA. Technical details are described in [40, Ch 12].

Hybridization. Hybridization is the process of base pairing two single strands of DNA or RNA [64]. DNA molecules are double-stranded and these two strands melt apart at a characteristic melting temperature, usually above 65^oC. As the temperature is reduced and held below the melting temperature, single-stranded molecules bind back to their counterparts. The process of binding back is based again on the principle of base pairing, so that only two complementary strands can hybridize. In the same way, an mRNA molecule can hybridize to a melted cDNA molecule, when the mRNA contains the complementary code of the cDNA strands. When hybridization occurs, a single stranded DNA binds strongly to complementary RNA, and in a way that prevents the DNA strands from re-associating with each other [95].

Microarray technology is used to measure the relative level of expression of genes in a particular cell or tissue by hybridizing a labeled cDNA representation of the cellular mRNA to cDNA sequences (cDNA microarrays) or by hybridizing a labeled cRNA representation of the cellular mRNA to short specific segments known as *synthetic oligonucleotides* (synthetic oligonucleotide microarrays) [29, 65]. Synthetic oligonucleotides — also referred to as *oligos* in the bio-molecular jargon — are short sequences of single-stranded cDNA that bind readily to their complements. The tethered cDNA sequences or oligos are called *probes*, while the cDNA or cRNA representation of cellular mRNA extracted from the cell is called the *target* (this is the suggested common terminology of

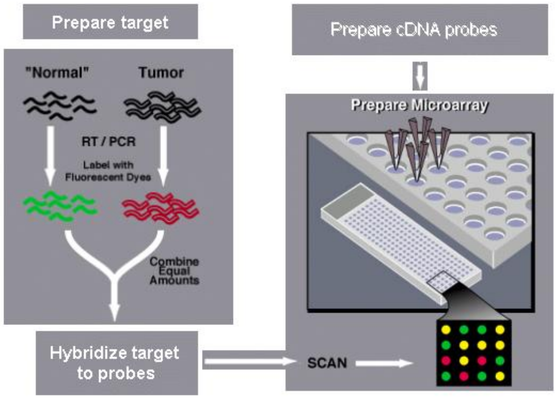

Figure 2: *A sketch of cDNA microarray technology. Selected probes are amplified by PCR, and the PCR product is printed to a glass slide using a high-speed robot. The targets are labeled representation of cellular mRNA obtained by reverse transcription of total RNA extracted from the test and reference cells, and the pooled target is allowed to hybridize with the cDNA spotted on the slides. Once the hybridization is completed, the slides are washed and scanned with a scanning laser microscope able to measure the brightness of each fluorescent spot; brightness reveals how much of a specific DNA fragment is present in the target.*

Phimister [80]). In both cases, the probes represent either genes of known identity or segments of functional DNA, also known as ESTs (expressed sequence tags). The target is labeled with fluorescent dye and hybridized to the probes. The higher the amount of cDNA or cRNA hybridized to a probe, the more intense the fluorescent dye signal will be on that probe. The relative mRNA abundance of a gene in a particular cell or tissue is therefore measured by the emission intensity of the probes. cDNA and synthetic oligonucleotide microarrays are the two most popular microarray technologies and are described in the next two sections.

3.2 cDNA Microarrays

cDNA technology was developed at Stanford University [89], although similar concepts can be traced back as far as the mid 80s [33]. The first step in the production of the microarray is the selection of the probes to be placed on the microarray, and the amplification of the corresponding cDNA clones by a technique known as polymerase chain reaction (PCR). PCR allows multiple rounds of amplification of a minimal amount of DNA to produce sufficient quantities of a sample. cDNA microarrays are produced by spotting PCR samples of cDNA strands in approximately equal amount on a glass slide using a high-speed robot. Each strand of cDNA identifies uniquely, with its code, a gene or an EST, so that each spot in the microarray corresponds to a gene or an EST.

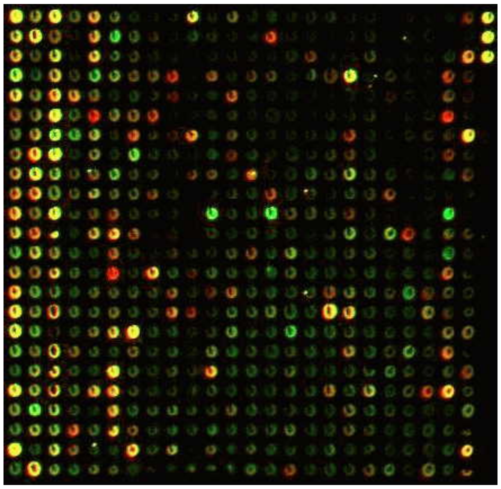

Figure 3: *A scanned image produced from a cDNA microarray experiment. Each spot represents a gene. Grey spots denote genes that were expressed in neither types of cells, colored spots identify genes that were expressed in one of the two cells or both. The color of the spot informs about the relative expression of the gene in the two cells.*

To prepare the target, investigators extract total RNA or mRNA produced from two types of cells, for example healthy and tumor cells or test and reference cells. Then, by using a single round of reverse transcription, the mRNA from the two samples is fluorescently labeled with Cy3 (green) and Cy5 (red) and the target mixture is hybridized to the probes on the glass slides. During the hybridization, if segments of the mRNA representation in the target find their complementary portion among the samples of cDNA in the glass slide, they will bind together. When the hybridization is complete, the glass slide is washed and laser excitement of the glass slide is used to yield a luminous emission that is then measured by a scanning microscope. Fluorescence measurements are made with a microscope that illuminates each spot and measures fluorescence for each dye separately, thus providing a measure of the relative mRNA abundance for each gene in the two cells. The intensity of the green spot measures the relative mRNA abundance of the gene in the cell whose reversed transcribed mRNA was labeled with Cy3, while the intensity of the red spot measures the relative mRNA abundance of the gene in the cell whose reverse transcribed mRNA was labeled with Cy5. Grey spots denote genes that were expressed in neither cell types.

These measurements provide information about the relative level of expression of each gene in the two cells. The monochrome images can be pseudo-colored to provide a quantitative measure of the relative expression of each gene in the two cells. This measure is adjusted to account for background noise caused by high salt and detergent concentrations during the hybridization or contamination of the target. Further details are discussed in Section 3.4. Figure 3 shows one of these images, in which spots are colored in red, green, yellow and grey. Each spot corresponds to a gene and the color of the spot informs about whether the gene is expressed (colored) or not, and about the relative level of expression in the two targets. Usually a measurement scale is provided to associate each color tone with a ratio between expression level in the two cells [11, 89].

Two limitations of cDNA technology are the risk of cross-hybridization and the large amount of total RNA (50-200 μg) required to prepare the target [29]. Cross-hybridization occurs when fragments of the reverse transcribed mRNA in the target hybridize to similar complementary probes, thus producing false detections. The large amount of mRNA for target preparation has implications on the range of detection, so that genes expressed at low level — less than 1 transcript per 100,000 — may fail to be detected. Several schemes to increase detection specificity are under development and a discussion can be found in [29].

3.3 Synthetic Oligonucleotide Microarrays

High-density synthetic oligonucleotide microarrays are fabricated by placing short cDNA sequences (*oligonucleotides*) on a small silicon chip by means of the same photolithographic techniques used in computer microprocessor fabrication. This proprietary technology, developed and commercialized by Affymetrix under the trademark of GeneChip®, allows the production of highly ordered matrices containing between 17,000 genes in the Affymetrix Murine Genome U74 set, and 33,000 genes in the Affymetrix Human Genome U133 set.

The rationale behind this technology is based on the concept of probe redundancy: a *set* of well-chosen small segments of cDNA is not only sufficient to uniquely identify a specific gene but also reduces the chances that fragments of the target will randomly hybridize to the probes, thus reducing the chances of cross-hybridization. Therefore, synthetic oligonucleotide microarrays represent each gene not by its cDNA but by a set of fixed-length independent segments unique to the DNA of the gene, as shown in Figure 4. On the GeneChip® platform, each oligonucleotide (probe) is 25-base long and each gene is represented by a number of *probe pairs* ranging from 11 in the new Human Genome U133 set, to 16 in the Murine Genome U74v2 set and the Human Genome U95v2. A probe pair consists of a perfect match (PM) probe, and a mismatch (MM) probe. Each PM probe is chosen on the basis of uniqueness criteria and proprietary, empirical rules designed to improve the odds that probes will hybridize with high specificity. The MM probe is identical to the corresponding PM probe except for the base in the central position, which is replaced with its complementary base as shown in Figure 4. The inversion of the central base makes the MM probe a further specificity control because, by design, hybridization of the MM probe can be attributed to either cross-hybridization or background signal caused by the hybridization of cell debris and salts to the probes [65, 67]. Each cell of an Affymetrix oligonucleotide microarray consists of millions of samples of a PM or MM probe, and probes tagging the same gene are scattered across the microarray to avoid systematic bias.

To prepare the target, investigators extract total RNA from a cell or tissue. mRNA is reversed transcribed into cDNA, which is made double stranded and then converted into cRNA using a transcription reaction that fluorescently labels the target. Once hybridization has occurred, the microarray is washed and scanned with a standard laser scanner. The scanner generates an image of the microarray that is gridded to identify the cells containing each probe, and analyzed to extract the signal intensity of each probe cell.

Although less flexible than cDNA microarrays because the experimenter cannot select the probes, synthetic oligonucleotide microarrays offer several advantages. Besides the decreased chance of cross-hybridization, synthetic oligonucleotide microarrays require a smaller amount of total RNA to prepare the target (5 μg); they have a wider dynamic range (the hybridization signal is linearly

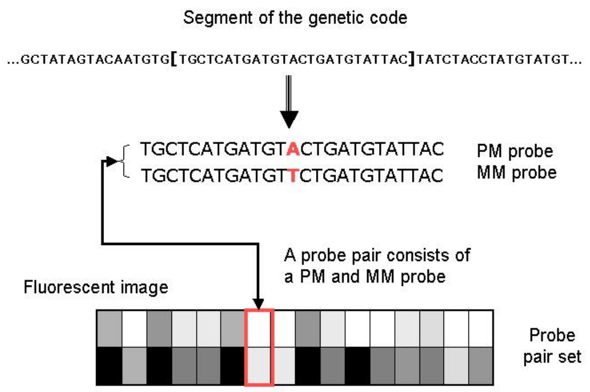
Segment of the genetic code
...GCTATAGTACAATGTG[TGCTCATGATGTACTGATGTATTAC]TATCTACCTATGTATGT...
TGCTCATGATGTACTGATGTATTAC
TGCTCATGATGTTCTGATGTATTAC
PM probe
MM probe
A probe pair consists of
a PM and MM probe
Fluorescent image
Probe
pair set

Figure 4: *An oligonucleotide microarray associates a gene with a set of probe pairs, in this case 20. Each probe pair consists of a perfect match probe (PM) and a mismatch probe (MM). Each PM probe is 25 bases long and is paired with the MM probe, in which the central base of the oligonucleotide is inverted. After hybridization of the target to the probes, the microarray is read with a laser scanner to produce an image, and the intensity of the MM probes is used to correct the intensity of the PM probes.*

related to up 500 fold mRNA abundance [65] compared to 10 fold in cDNA microarrays [84]); and a high detection specificity (mRNA transcript representations present in the target at the relative abundance of less than 1 in 10^6 can be detected [65]).

3.4 From Images to Data

In both cDNA and oligonucleotide microarrays, hybridization of the target to the probes determines a chemical reaction that is captured into a digital image by a scanning laser device. The next step is to translate the intensity of each hybridization signal into a table with numerical measures. The quality of image analysis process is crucial for the accurate interpretation of the data, and a variety of algorithms and software tools tailored to the different aspects of cDNA and oligonucleotide microarray images have been developed, see [9, 12].

The main steps of cDNA microarray image analysis are gridding, segmentation and intensity extraction, and have been recently reviewed in [93]. The gridding step recovers the position of the printed spots corresponding to the probes into the image. Because the position of the spots in the microarray is known, gridding is relatively straightforward although a series of parameters have to be estimated to account, for example, for shifts or rotations of the microarray into the image, or small translations of the spots. The segmentation consists of a classification of the image pixels into foreground and background, where foreground pixels correspond to spots of interest in the microarray, and background is noise resulting from high salt and detergent concentrations during the hybridization, or contamination of the target. Several segmentation methods have been proposed, which differ by the geometry of the spot they produce. For example, the method implemented in ScanAlyze[2] fixes a circle with constant diameter to all spots in the image, whereas the method implemented in the Axon software GenePix[3] estimates the diameter for each spot separately. The method developed by Chen *et al.* [16] and implemented in QuantArray[4] uses repeatedly the Mann-Whitney test to label groups of eight pixels at a time as background or foreground. The package spot[5] for the R software implements an adaptive shape segmentation developed by Yang *et al.* [110].

The intensity extraction step calculates the intensity of the red and green fluorescence of each spot, the background intensity, and some quality measures. The background intensities are used to correct the foreground intensities and, hence, the red and green intensities that become the primary data for the subsequent analysis. Background correction is motivated by the fact that intensity measured for each fluorescent channel includes a contribution that is not due to the hybridization of the target to the probes. Most packages calculate the foreground intensity as the mean or the median pixel values. To correct the intensity of the two channels, an estimate of the background intensity is usually subtracted from the foreground intensity. For example, ScanAlyze calculates the corrected intensity by the average number of foreground pixels for each channel minus the median number of background pixels. Corrected intensity values are calculated as the difference between median foreground pixels and background pixels in QuantArray. Spot computes the background intensity by a nonlinear filter called morphological opening, which works by creating a background image for the whole microarray and by sampling this background image at the nominal centers of the spots. Further details and empirical comparisons of different segmentation and background correction methods are in [110]. Note that background correction introduces negative values, when the foreground intensity is lower that the background intensity. Because background intensity larger than foreground intensity is considered an error, spots with negative corrected intensities are either disregarded or replaced by an arbitrary small positive number.

The analysis of oligonucleotide microarray images exploits the fact that the image produced by the scanning laser device describes the probes by squares of an approximately known number of pixels organized in a lattice. Furthermore, the image contains some alignment features recognizable as the checker-board patterns at the corners of the image in Figure 5. Because the approximate physical dimension of each probe in the image is known, once the positions of the alignment features are determined, a basic grid is created to determine the pixels describing each probe cell by using some form of linear interpolation. To extract the intensity of each probe, the original proprietary algorithm employed by Affymetrix software used the 75th percentile of the pixel intensities, after removing the boundary pixels whose intensity could be distorted.

Awareness of potential misalignment of the basic gridding algorithm, with consequent failure to extract the correct signal intensity, has led researchers to develop adaptive pixel selection algorithms, see [88, 114]. The adaptive pixel selection algorithm of Schadt *et al.* [88] begins by removing pixels of extreme intensity and then, iteratively, adjusts the edges by removing those pixels that contribute most significantly to the coefficient of variation. Some constraints are also imposed to avoid bias

[2] http://rana.lbl.gov/EisenSoftware.htm

[3] http://www.axon.com/GN_GenePixSoftware.html

[4] http://www.packardbioscience.com/products/521.asp

[5] http://experimental.act.cmis.csiro.au/Spot/index.php

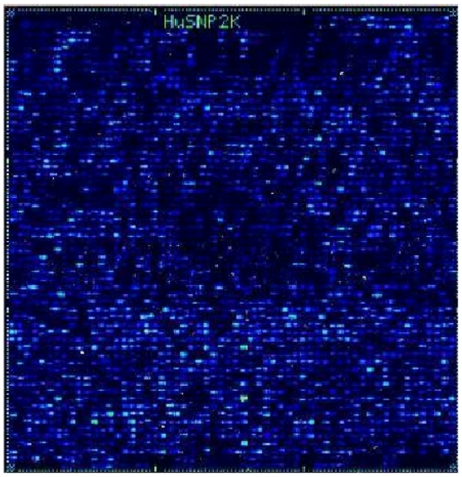

Figure 5: *Scanned image of a synthetic oligonucleotide microarray. Grid cells represent* probes *and the intensity of each matrix cell measures the quantity of hybridized oligonucleotides in a probe. The checker-board patterns at the corners of the image are the alignment features used to grid the image.*

of boundary pixels. The algorithm in [114] corrects for misalignment resulting in an improved selection of pixels attributed to individual probe cells, and a substantial reduction in the variance of pixel intensities. The main motivation of this method is the fact that probe cells are often not equally spaced, so that gridding by linear interpolation can cause misalignment by as many as three pixels. To accommodate for this deformation, the algorithm uses an iterative procedure that translates the initial location of probe cells by maintaining the lattice structure of neighbor cells. The most recent Affymetrix software for image analysis has a new algorithm to compute the background intensity that accounts for potential spatial effects. Essentially, the image is split into k (default value 16) rectangular zones, and background intensity is computed as the lowest 2% intensity of the cells in each rectangular zone. The background intensity for each cell is calculated as a weighted average of the background intensities in each rectangular zone, with weights that account for the distance of the cell from the centers of each rectangular zone. This estimate of the background intensity is then subtracted to the probe cell intensity. Negative values resulting from background adjustments are set equal to a user defined value (the default value is 0.5).

Because the relative mRNA abundance is represented by the intensity of a probe pair set that consists of a number of probe pairs, the intensities of the probe cells are summarized to yield a relative measure of the gene expression level. The latest statistical algorithm produced by Affymetrix (MAS 5.0) generates, for each probe set, three measures: a detection call, a detection p-value, and a signal value. The detection calls assess the quality of the hybridization, whereas the detection p-values represent the confidence in this assessment. The signal is a proxy for the relative expression level of the gene represented by the probe set. Full details are described in [1] and we summarize them briefly.

Detection calls and p-values are generated by first calculating a discriminant score R_i for each probe pair PM and MM given by

$$R_i = \frac{I(PM_i) - I(MM_i)}{I(PM_i) + I(MM_i)}$$

where $I(PM_i)$ and $I(MM_i)$ are the extracted intensities for the ith perfect match probe and mismatch probe. The score R_i is bounded above by 1, and measures the ability of the ith probe pair to detect its intended target. A positive value implies that the perfect match intensity $I(PM_i)$ is larger than $I(MM_i)$, and the strength of detection ability of the ith probe pair increases with R_i. A negative value implies that the mismatch intensity $I(MM_i)$ exceeds $I(PM_i)$ and highlights a poor detection ability of the ith probe pair. To avoid bias, saturated cells (defined as mismatch probes cells with intensity above a fixed threshold) as well as probe cells in which $I(PM_i) \leq I(MM_i) + 0.015$ are disregarded.

To determine the detection p-value, the scores R_i computed for the probe pairs in a probe set are compared with a user defined threshold τ (typically $\tau = 0.015$), and the null hypothesis of no difference between the median discrimination score and τ is tested by the one sided Wilcoxon's Signed Rank test. The detection p-value is simply the p-value computed by assuming an asymptotic normal distribution for the Wilcoxon signed rank statistic when more than 12 probe pairs are used, whereas exact calculations are carried out when the retained number of probe pairs is less than 12. Detection calls describe whether the hybridization of a probe set has occurred (P for present), has not occurred (A for absent), or has been only marginal (M), and are assigned on the basis of a significance range for the detection p-value. Suggested settings are to call the hybridization present if the detection p-value is smaller than 0.04, marginal if the detection p-value is between 0.04 and 0.06, and absent otherwise. Lastly, the signal that measures the relative expression level of a probe set is computed by the "One-Step Tukeys Biweight Estimate", which essentially produces a robust average of the differences between $I(PM_i)$ and $I(MM_i)$, with weights that take into account the distance between $I(PM_i) - I(MM_i)$ and the median intensity difference.

The rationale behind the use of paired PM and MM probes is that the specific hybridization, represented by the intensity of the PM probes, should be stronger than the non-specific hybridization represented by the intensity of the MM probes, and such a consistent pattern across the probe set is unlikely to occur by chance. Several studies have been produced to support this claim, for example [53, 67]. However, mismatch values $I(MM_i)$ can be higher than perfect match values $I(PM_i)$ for a number of reasons, such as cross-hybridization occurring when the probe sequence has high homology with another unknown sequence, or errors in the probe sequences that cause low specificity. Therefore, a weighted average of the difference between $I(PM_i)$ and $I(MM_i)$ can produce negative intensity values. In fact, the previous Affymetrix software MAS 4.0 used to return probe set intensity values called "Average difference" that could be negative. A series of rules are employed by the latest Affymetrix software MAS 5.0 to avoid the calculation of negative signal values. Particularly, if the mismatch value $I(MM_i)$ is higher than the perfect match $I(PM_i)$, then the mismatch is assumed to provide no additional information about the estimate of the signal and it is replaced by an imputed value called idealized mismatch (IM). This idealized mismatch is either a value smaller than $I(PM_i)$, or an estimate based on the average ratio between perfect match and mismatch values.

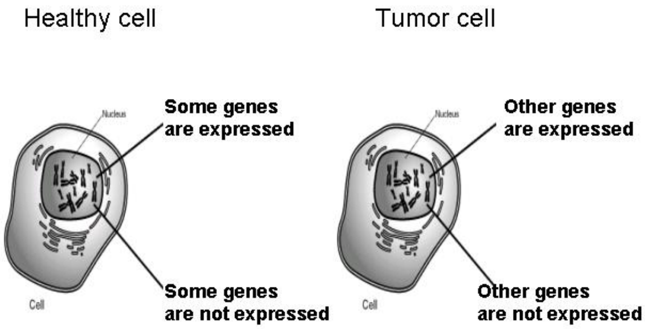

Figure 6: *Microarray technology enables investigators to detect the genes differentially expressed in two samples.*

4. Experimental Questions and Experimental Design

Both cDNA microarrays and oligonucleotide microarrays provide a panoramic view of the activity of genes under particular experimental conditions, and are nowadays used to answer the same broad classes of questions. In the following, we will term the set of expression levels measured for a gene across different conditions its *expression profile*, whereas we will use the term *sample molecular profile* to denote the expression level of the genes measured in a sample in a particular condition.

4.1 Experimental Questions

By providing a measure of expression of a gene in terms of its mRNA abundance, microarray technology lets the experimenters observe the molecular profile of a cell, or cell line — distinct families of cells grown in culture — in a particular condition. The simplest experiment we can devise using this technology is a *comparative* experiment, illustrated in Figure 6, to identify the genes differentially expressed in two conditions. An example of this experimental setting is the comparison of metastatic versus non-metastatic derivatives of a tumor cell line [60], in which samples of cells from the two conditions are extracted from several patients. The experimental conditions can be specific levels of controllable environmental factors, such as extreme temperatures or starvation, or the modification (*knock-in*) or the removal (*knock-out*) of a specific portion of the genome.

More complex experimental questions involve the molecular profiling of several conditions at a time to characterize, for example, the genomic fingerprint of different types of cancer [2], or the effect of changing several experimental factors simultaneously [18]. In both cases, each sample consists of the gene expression levels measured in cells grown or observed in a particular condition, and different samples can be assumed to be stochastically independent. A different class of experimental questions involves the study of the temporal evolution of gene expression profiles, so that different samples may be stochastically dependent. Studies in this class try to understand, for instance, the process that turns a locally growing tumor into a metastatic killer [19], the yeast sporulation cycle [96], or the response of human fibroblasts to serum [49]. Although the dependency structure among samples requires a different analysis, the common feature of these experiments is to compare the molecular profiles of cells in different conditions.

Advanced experiments investigate the regulatory mechanism of cells observed in different experimental conditions. When functioning normally, regulatory pathways in cells modulate the level and duration of gene expression, thus ensuring that cells respond to physiological and extracellular stimuli in an appropriate manner. However, a broad range of diseases can result when the activation of a gene regulation pathway triggers either an under- or over-production of certain proteins. Fundamental problems are the discovery of new gene regulatory pathways and of causal dependencies among gene expression [81].

4.2 Experimental Design

The design of microarray experiments is a critical, albeit still neglected, issue of modern functional genomics. One important difference between cDNA microarrays and synthetic oligonucleotide microarrays is that the former are designed for the so-called competitive hybridization: two targets can be simultaneously hybridized to the probes on one microarray. The first key issue in designing a comparative cDNA experiment is to choose between direct and indirect comparisons. In the first case, the target mixture is hybridized to the same microarray whereas, in the second case, the mRNA representations from the two treated cells are mixed with the reverse transcribed mRNA from a reference cell, and the two mixtures are hybridized to two different microarrays. A discussion of the pros and cons of direct versus indirect comparisons is in [111].

Besides technical issues of probe/microarray choice and design, the most fundamental design issue common to both cDNA and synthetic oligonucleotide microarrays is the choice of the number of replications required to stake a statistically sound claim. Although microarray technology has rendered gene expression measurement blazingly fast, the cost of a single experiment — up to $1200 for a single high resolution synthetic oligonucleotide microarray — is still a significant factor in the experimental choices of biomedical investigators. Comparative experiments reported in main stream biomedical journals were originally limited to one replication of an experiment [26]. Arguments have been made to show that a single replication of a comparative experiment is not sufficient to achieve reproducible results [63] but, despite the increasing awareness that data generated by even the most accurate microarray are very noisy, many discoveries reported in mainstream journals are often based on experiments with three replications [108].

The main problem of this experimental design aspect is caused by the parallel nature of experiments conducted with microarrays: the number of replicates necessary to obtain an accurate measure of the expression level of a gene g may not be the same number needed for a different gene. Furthermore, responses to the micro-environment conditions, such as the time of the day or washing conditions appear to have a significant impact on gene expression. Eric Lander [60], leader of one of the largest genomic centers in the world, reports that “It is well known among *aficionados* that comparison of the same experiment performed a few weeks apart reveals considerably wider variation than seen when a single sample is tested by repeated hybridization.” Therefore, while replicated experiments should increase the amount of information needed to carry out a statistical analysis, they may also increase variability among replicates.

An additional experimental design issue arises from the common problem of mRNA paucity. It is often the case that a single cell is unable to produce detectable mRNA in the desired condition. In this situation, common practice is to either *pool* together the mRNA extracted from different samples, or to amplify the cellular RNA. While obvious reasons of variability control suggest using the same pooled sample for each experimental condition, the determination of the number of units to pool together is still an open issue. An interesting discussion of this experimental design issue is in [111].

When the objective of the experiment is the study of the temporal evolution of a biological system, the researchers need also to choose the time points to sample. These experiments are usually performed by sampling the gene expression profile using a microarray at predefined temporal intervals and then mounting these snapshots of the genome activity into “movies” that capture the dynamics of the process. The specificity of each gene becomes, here, even more important: the optimal sample points to observe the evolution of a gene during a process may be not the same for another gene on the same microarray.

In more complex experiments conducted to study the effect of different experimental factors, the choice of the number of replications is paired with the choice of the experimental treatments to test. Some recent research has addressed the issue of the experimental design for microarray data [18, 55, 57, 77, 111], by proposing classical factorial experimental designs, but we believe the choice of the experimental design is very much an open problem. The theory of statistical experimental design seeks experimental plans that allow a specific statistical analysis to be carried out to test particular hypotheses [23]. Because to date no agreement exists about the appropriate statistical analysis of gene expression data produced with microarrays, and because many experiments with microarrays are conducted to generate rather than test hypotheses, critical experimental design issues are still far from being solved.

5. Data Preprocessing

To answer the experimental questions, the quantitative measurements of gene expression data produced by microarray experiments are analyzed using statistical and machine learning methods. A common strategy to reduce data variability and dimensionality is to perform two preprocessing operations known as *normalization* and *filtering* on either the raw or transformed data, before undertaking any data analysis. The goal of the normalization operation is to remove systematic distortions across microarrays to render comparable the experiments conducted under different conditions. The aim of the filtering operation is two-fold: to reduce variability by removing those genes whose measurements are not sufficiently accurate, and to reduce the dimensionality of the data by removing genes that are not sufficiently differentiated. The transformation of the raw data, advocated by several authors, should even out the intensity values that are usually extremely skewed.

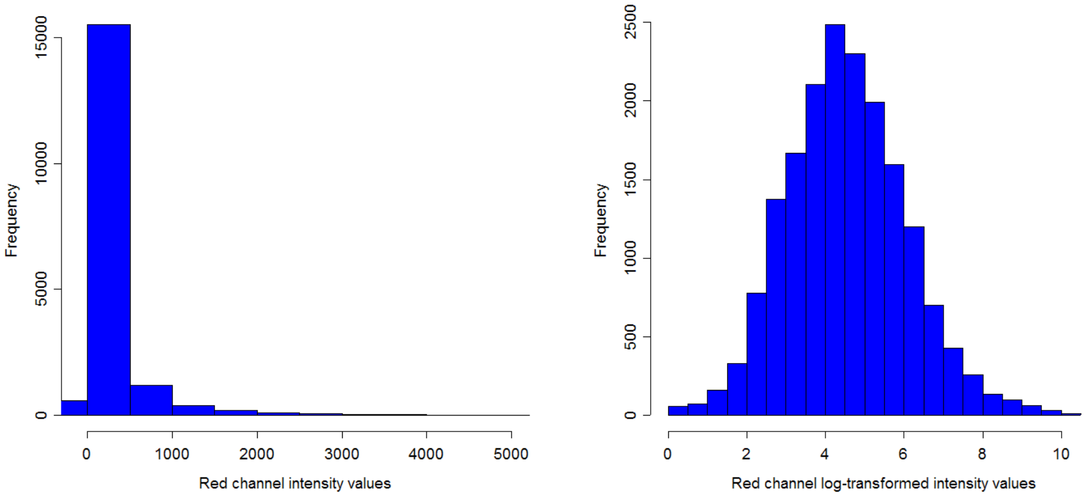

Figure 7: *Histogram of the corrected intensity values for the red channels (left), and histogram of the corrected intensity values after negative intensities were removed and positive intensities were log-transformed (right).*

5.1 To Log or not to Log Transform?

Suppose the microarray experiment was conducted to compare the expression level of G genes in two cells. For each gene g, denote by (y_{g1}, y_{g2}) the pair of relative expression levels measured in the two conditions. If the experiment was conducted by a direct comparison with one cDNA microarray, (y_{g1}, y_{g2}) will denote the corrected intensity values for the red and green channel in the spot corresponding to the gene g. When the experiment is conducted with two synthetic oligonucleotide microarrays, (y_{g1}, y_{g2}) will denote the signal values for the probe set describing the gene g.

Because the corrected intensity values are highly skewed, log-transforming the raw data (y_{g1}, y_{g2}) produced by cDNA microarray experiments is strongly recommended by several authors to even out intensity values, see for example [110]. A fairly common assumption is that the logarithmic transformation produces normally distributed data [73]. As an example, the histogram in the left plot of Figure 7 describes the distribution of corrected intensity values for the red channel in an experiment conducted to compare gene expression levels in normal and malignant lymphocytes [2]. The background correction was done by subtracting the median background intensity from the average foreground intensity for each spot, and the foreground and background intensities were computed using SCANALYZE. Note that the background correction introduces a small proportion of negative values and, typically, spots with negative corrected intensities are either disregarded or the negative intensity is replaced by an arbitrary small number. The distribution of positive intensity values is extremely skewed, and the histogram in the right plot of Figure 7 describes the distribution of the log-transformed intensities, after removal of the negative values. The logarithmic transformation removes most of the original asymmetry but some left skewness is still visible. Other examples are reported in the Speed group microarray page[6]. This residual lack of symmetry after the logarithmic transformation is typical of Gamma distributed data [71], so that rather than the logarithmic transformation, some power or variance transformations would be more suitable. Examples of such transformations are discussed in [86].

Although the choice for the best data transformation of the red and green corrected intensities is still an open problem, it is generally acknowledged that the corrected intensity values measured with competitive hybridization on cDNA microarrays should be transformed. No such similar consensus exists when the data (y_{g1}, y_{g2}) are produced by synthetic oligonucleotide microarrays, possibly because the original Affymetrix statistical software MAS 4.0 used to return a fairly large proportion of negative intensity values, often 25% of the data. Authors have suggested using the cubic root transformation [101], or the log-transformation of appropriately truncated data [46], but well accepted data analysis protocols [39] do not use any data transformations. Empirical evidence that the bulk of positive values produced by synthetic oligonucleotide microarrays analyzed with MAS 4.0 follows a log-normal distribution was recently provided by [45], whereas questions about the correct transformation to use, if any, of intensity values calculated by the Affymetrix statistical software MAS 5.0 are still open.

5.2 Normalization of Microarray Data

A well known problem with cDNA technology is the consistent imbalance of the fluorescent intensities of the two dyes Cy3 (green) and Cy5 (red) — Cy3 is systematically less intense than Cy5 [82, 110]. Although a simple dye-swap experiment in which each hybridization is repeated twice with reversed dye assignment to the two targets would be the best way to remove this systematic bias [111], normalization techniques are commonly used to render the gene expression levels measured by the two different dyes comparable [29]. Synthetic oligonucleotide microarrays do not suffer from a known systematic distortion similar to the dye fluorescence imbalance of cDNA microarrays, but a comparative experiment conducted on this platform requires hybridization of each target to a different microarray. A variety of experimental errors, including variations of the amount of mRNA used to create the target hybridized to each microarray, or the quantity of dye used to fluorescently label each target, may introduce errors. Normalization techniques are therefore used in an attempt to "remove" the experimental errors.

Assuming that the amount or type of dye used to label the two targets as well as variations of the quantity of cellular mRNA used in the two targets induce contaminations, the observed expression level y_{g2} masks the correct expression level $\tilde{y}_{g2}$ one would observe if the second experiment were conducted in exactly the same conditions of the first experiment. Formally, we can write

$$y_{g2} = f(\tilde{y}_{g2})$$

and normalization techniques consist of estimating the function $f(\cdot)$ to recover $\tilde{y}_{g2} = f^{-1}(y_{g2})$. *Total Intensity Normalization* approximates $f(\cdot)$ with the zero-intercept regression line $y_{g2} = \beta\tilde{y}_{g2}$, and estimates β by $(\sum_g y_{g1})/(\sum_g y_{g2})$, [82]. The rationale behind this choice is that the total quantity of mRNA representation hybridizing from each target should be the same. When β is estimated by the ratio $(\bar{y}_1/\bar{y}_2)$, where $\bar{y}_1$ and $\bar{y}_2$ are the average expression levels in the two targets,

[6] http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html

the technique is also called *Total Mean Normalization*. Variants of this procedure estimate β by the ratio of the medians or by the ratio of the trimmed means.

An alternative technique, known as *Normalization with Calibration*, relies on the assumption that only a very small proportion of genes in a microarray should have substantially different levels of expression across the two cells. Following this principle, the function $f(\cdot)$ is approximated with the regression line $y_{g2} = (\tilde{y}_{g2} - \alpha)/\beta$, and the parameters α and β are estimated from the data (y_{g1}, y_{g2}) by fitting the linear regression $y_1 = \alpha + \beta y_2$. In so doing, the regression line for y_1 versus $\tilde{y}_2$ will have zero intercept — thus removing systematic deviations — and unitary slope — thus capturing the intuition that the majority of gene expression levels across the two experimental conditions should remain unchanged. Normalization with calibration can be adjusted to account for specific non-linear effects, and nonparametric regression techniques, such as lowess regression, have been proposed to handle possibly nonlinear transformations [7, 110] or spatial effects [48, 109].

All these normalization techniques can be used either globally or locally. Global normalization uses all genes in the microarray to identify a transformation of the expression data to calibrate the measures in the two samples. Local normalization uses only those genes known to remain constantly expressed across the two particular experimental conditions, or *housekeeping genes*, a library of genes believed to have nearly constant expression level in a variety of experimental conditions. Well accepted protocols [7, 20, 39] use the subset of genes detected as hybridized by the Affymetrix software.

One problem of normalization with calibration applied to intensity data is that, when $\alpha > 0$, small values of the systematically larger intensity are replaced by negative numbers. To avoid this bias, other normalization techniques try to calibrate the ratios y_{g2}/y_{g1} [16], or the log-ratios $\log(y_{g2}/y_{g1})$, [109, 110]. Clearly, these techniques are applicable to microarray data that can be paired, as for example data generated by direct comparisons with cDNA microarrays.

Extending normalization techniques to repeated experiments is not straightforward. Yang *et al* [110] provide a comprehensive overview of normalization techniques for repeated experiments with cDNA microarrays. For oligonucleotide microarrays, a common approach to normalization of multiple experiments is to choose one replication as baseline and to apply normalization with calibration, or total intensity normalization, to the other replications [39]. Because the results will differ according to the chosen baseline, authors have suggested computing the baseline as the average expression profile across all microarray samples [101]. An open question remains whether normalization of replicated experiments with oligonucleotide microarrays is needed at all. In replicated experiments, in which more than one microarray is hybridized to a replication of the same target, changes in the amount of cellular mRNA used to prepare the target, or changes in the amount of fluorescent dye should be considered part of the experimental error. If no systematic errors are introduced, one can assume that the measurement observed for gene g in the replicate k of the experimental condition i is

$$y_{gik} = \mu_{gi} + \epsilon_{gik}$$

where ϵ_{gik} is the error in replicate k, and μ_{gi} is the correct expression level of gene g in condition i. The assumption that the experiment is reproducible would require that, on average, the experimental errors compensate, so that normalization is not necessary. This is for example the approach adopted in [76]. The error variance can be modeled to account for the different sources of variability. An approach along this line is presented in [52, 107] for the analysis of repeated cDNA-based expression levels, transformed in log scale.

The issue of normalization of repeated comparative experiments differs from the normalization needed when more than two experimental conditions — either different targets or the same target tested at different time steps — are analyzed. For example, when the objective of the whole experiment is to examine the temporal behavior of a genomic system during a cell cycle, it is common practice to take only one replication of the gene expression data at each time point [32, 49, 96], and standard normalization techniques are used to make the expression levels measured at different time points comparable. Although a preferable solution would be to take a few replicates of each measurement, cost constraints often make this solution impractical.

5.3 Filtering

Several techniques are available to reduce data dimensionality and variability by removing some gene measurements. It is surprising to realize that *ad hoc* rules are commonly used, and that the choice of the genes to be removed can differ substantially according to the microarray platform and the technique chosen to analyze the data.

For expression data measured with a cDNA microarray, it is common practice to disregard those genes with negative or small expression levels (before or after normalization). Typically, all those spots in which the foreground intensity does not exceed the background intensity by more than 1.4 fold are disregarded, or replaced by an arbitrary small number. The Affymetrix statistical software MAS 5.0 assigns a detection call to each probe set to assess the amount and quality of hybridization, and it is suggested to discard all genes whose expression level is labeled as A (absent) or M (marginal) in all samples. This procedure is justified by the empirical evidence that expression levels smaller than 10 are actually measurement errors [1]. However, a large proportion of genes would often be discarded by this procedure, and investigators tend to adopt less stringent criteria to select a subset of the genes to be further analyzed. A common strategy is to retain only those genes whose minimum fold-change exceeds a particular threshold d in a preset number of experiments c, for example $d = 3$ and $c = 1$ in [14]. The choice $c = 2$ was originally suggested by DeRisi *et al.* [26] to analyze expression levels measured with cDNA microarrays, and an insightful analysis of the empirical success of this rule is described in [87]. Golub *et al.* [39] suggest to further score genes by their standard deviation, so that to limit the analysis to those genes that vary most across experiments, and a similar approach is proposed in [30]. Other authors remove “spiked” genes, that is, those genes with one abnormally large or abnormally small measurement [99]. The recent book [12] contains a comprehensive description of other filtering techniques most commonly used.

All these filters depend on arbitrary thresholds used to decide when a value is abnormally large or small, or when the variability of the measurements is too high. The impact of normalization and filtering strategies is unclear and few systematic studies are available to provide investigators with a description of the properties of these preprocessing techniques and guidance on choosing the one most appropriate for their particular problem.

6. Analysis of Comparative Experiments

This section describes the most popular techniques for the analysis of gene expression data in repeated comparative experiments. The objective of the analysis is to identify the genes with a significant expression change across two conditions. The approaches to this problem can be classified in two broad categories. Methods in the first category, known as fold analysis, estimate the ratio between the expression levels of each gene in the two conditions, whereas methods in the second category use the data to estimate the expected difference in expression of each gene in the two conditions.

6.1 Fold Analysis

Early comparative experiments based on cDNA microarray technology measured differences of gene expression across two conditions in terms of the fold-change: the ratio of the expression levels [26, 89, 90]. Particularly, genes showing a negative or positive fold-change of at least two were deemed as differentially expressed across the two conditions. The need to choose a threshold to identify significant differentially expressed genes in two conditions is the motivation of a series of articles focused on statistical fold-analysis.

We let $\rho_g = \mu_{g1}/\mu_{g2}$ denote the unobservable "true" fold-change for gene g in the two conditions. When $\rho_g = 1$, the expression level of the gene g has not changed, while $\rho_g < 1$ and $\rho_g > 1$ indicate differential expression of the gene g in the two conditions. Particularly, $\rho_g < 1$ means that the gene is *down-regulated* by condition 1, whereas $\rho_g > 1$ means that the gene is *up-regulated* by condition 1. Statistical approaches to ratio-based differential analysis estimate the ratio ρ_g with some statistic r_g, and decide whether deviations of the estimate r_g from 1 can be attributed to a real difference of the gene expressions in the two conditions, rather than sampling variability. In the first published work following this approach [16], the authors use the naive ratio estimator $r_g = y_{g1}/y_{g2}$. Assuming that the measurements from the two different channels (corresponding to the Cy3 and Cy5 fluorescent dyes) are independent and normally distributed, and that they have constant coefficient of variation for all genes in both conditions, the authors derive an approximate distribution of the ratio statistic r_g that can be used to find a $(1-\alpha)\%$ confidence interval for the ratio ρ_g. The assumption of a constant coefficient of variation c in the two conditions lets the distribution of r_g depend on c, which is estimated by Maximum Likelihood. The authors also propose an iterative procedure to simultaneously estimate c and the normalization factor to render comparable the measurements from the two channels.

As noted in [74], this approach disregards ancillary information during the computation of the distribution of the ratio statistic, because the product $y_{g1} \times y_{g2}$ contains information about the variability of r_g. Furthermore, despite the fact that expression levels should be positive numbers, the measurements of the two channels are assumed to follow normal distributions. This inappropriate distributional assumption is corrected in [74], by assuming that the measurements of the two channels follow Gamma distributions, and a Bayesian method is proposed to estimate the fold-change of each gene to account for the "between microarrays" variability. Although this second approach is based on sounder distributional assumptions about gene expression measurements, it relies on the unconventional assumption that the experimental error across microarrays also follows a Gamma distribution.

Distributional assumptions aside, both approaches treat the pair of measurements of each gene in the cDNA microarray as independent, but this choice does not seem to be always correct. In direct comparisons, the same spot of cDNA in the microarray is simultaneously hybridized to the pool of mRNA representation in the target mixture. In other words, the two targets compete for hybridization to the probes so that, by design, each pair of measurements should be treated as a matched pair. Alternative approaches that model directly the ratio $r_g = y_{g1}/y_{g2}$, or its logarithm $l_g = \log(r_g)$, overcome this difficulty. The method introduced by Lee *et al.* in [63] uses a mixture model to describe the joint distribution of the log-ratio of the measurements from the two channels as follows:

$$f(l_g) = pf_E(l_g) + (1-p)f_U(l_g).$$

The parameter p is the unknown proportion of genes that are differentially expressed; $f_E(l_g)$ is the density function of l_g, when the gene g is differentially expressed, and $f_U(l_g)$ is the density function of l_g, when the gene g is not differentially expressed. By assuming a normal distribution for l_g, for each g, the mixture components are estimated by using the EM algorithm [25]. The estimates are then used to compute the posterior probability

$$pf_E(l_g)/f(l_g)$$

that each gene g is differentially expressed in the two experiments. When more than one replication is available, this procedure is applied to a "polished" summary of the original expression ratios that is computed as follows. By taking into account the sources of variability of each gene measurement, the authors model the log-ratio $l_{gk} = \log(y_{g1k}/y_{g2k})$ of the paired measurements for each gene g by

$$l_{gk} = \mu + \alpha_g + \beta_k + (\alpha\beta)_{gk} + \epsilon_{gk} \quad g = 1, \ldots, G, \quad k = 1, \ldots, n \tag{1}$$

where G is the total number of genes in the microarray, and n is the total number of replicates of the experiment. The parameters α_g represents the "gene-effect", described as the correct fold-change of gene g across all replications of the experiment. The parameter β_k captures the "microarray-effect", due for example to between microarray differences in the fluorescent dye, or the amount of mRNA used to prepare the target. The interaction term $(\alpha\beta)_{gk}$ accounts for possible variations of each gene fold-change in each replication of the experiment. The errors ϵ_{gk} are assumed to have zero mean. Although the authors acknowledge that all the effects in model (1) should be treated as random effects, they propose to estimate the parameters α_g using the standard two-way Anova estimator

$$\hat{\alpha}_g = \frac{1}{n}\sum_k l_{gk} - \frac{1}{nG}\sum_{gk} l_{gk} \quad g = 1, \ldots, G$$

which does not require any assumptions about the error distribution. Each estimate $\hat{\alpha}_g$ is then used as proxy of l_g to estimate the posterior probability that the gene g is differentially expressed. Note that, in the absence of pure replications, model (1) is over parameterized because the gene-array interaction $(\alpha\beta)_{gk}$ and the random error ϵ_{gk} are not distinguishable. In fact, pure replications are rarely conducted, and the microarray effect should be treated as a random block effect. Several authors have modified this approach by relaxing the parametric assumption on the mixture model

[31, 78], by using a larger number of fixed effects to model dye and spot effects [55], or by using random effects [107].

The scope of this stream of work is limited to direct comparisons of gene expression data with cDNA microarrays, where two targets are hybridized to the probes on the same microarray. In this case, the expression measurements from each microarray are paired by design. When cDNA are used for indirect comparisons, or the expression data are measured with synthetic oligonucleotide microarrays, there is no unique pairing of the data. To conduct the fold analysis on repeated experiments, researchers compute the average of the normalized expression levels in the two experimental conditions, and impose an arbitrary threshold on the ratio (or log-ratio) of the two averages. Unfortunately, no consensus exists about this threshold, even across different studies on the same organism by the same investigator [43, 108]. Typically, this threshold varies between 2 and 3 [38, 50, 70, 85], but it can be as low as 1.7 [62], and no published work addresses the problem of the extent of false positive and false negative rates produced by this "naive" fold analysis. A very elegant solution is presented in [46], which describes a Bayesian method for the fold analysis under the assumption that appropriately truncated gene expression data follow a log-normal distribution.

6.2 Differential Analysis

We now describe the hypothesis that a gene g is not differentially expressed in two experimental conditions by $H_0 : \mu_{g1} = \mu_{g2}$, while differential expression occurs under the alternative hypothesis $H_a : \mu_{g1} \neq \mu_{g2}$. To identify the set of genes that are differentially expressed, one needs to test, for each gene, the null hypothesis and then selects the set of genes for which the null hypothesis is rejected. We continue to denote by y_{gik_i}, $g = 1, \dots, G$, $i = 1, 2$, and $k_i = 1, \dots, n_i$ the expression level data generated by a comparative experiment. When the expression levels are measured with cDNA microarrays by direct comparisons, the replications of each condition are equal, say $n_1 = n_2 = n$, while there is no need to impose this restriction for data measured with oligonucleotide microarrays, or indirect comparisons with cDNA microarrays. The standard statistic used to test the null hypothesis is the t statistic

$$t = \frac{|\bar{y}_{g1} - \bar{y}_{g2}|}{\sqrt{s_g^2}}$$

where $\bar{y}_{g1}$ and $\bar{y}_{g2}$ are the average expression levels of gene g in the two conditions, and s_g^2 is an estimate of the variance σ_g^2 of the sample mean difference. Large values of the t statistic would offer evidence in favor of differential expression. The two main problems are the choice of the estimate s_g^2, and the identification of a threshold to reject the null hypothesis.

When the two samples are not independent — as for data collected with cDNA microarrays in direct comparisons — an appropriate estimate of σ_g^2 appears to be

$$s_{Dg}^2 = \frac{\sum_k [(y_{g1k_i} - y_{g2k_i}) - (\bar{y}_{g1} - \bar{y}_{g2})]^2}{n(n-1)} \equiv \frac{s_{g1}^2}{n} + \frac{s_{g2}^2}{n} - 2\frac{s_{g12}}{n}. \tag{2}$$

where the term $s_{g12} = \sum_i (y_{g1k} - \bar{y}_{g1})(y_{g2k} - \bar{y}_{g1})/(n-1)$ is an estimate of the covariance of the two sample means. When the two samples are independent — for example when data are collected with oligonucleotide microarrays or indirect comparisons with cDNA microarrays— σ_g^2 can be estimated by

$$s_{Ig1}^2 = \frac{\sum_{k_i}(y_{g1k_i} - \bar{y}_{g1})^2}{n_1(n_1-1)} + \frac{\sum_{k_i}(y_{g2k_i} - \bar{y}_{g2})^2}{n_2(n_2-1)} := \frac{s_{g1}^2}{n_1} + \frac{s_{g2}^2}{n_2}, \tag{3}$$

or by

$$s_{Ig2}^2 = \frac{\sum_i \sum_{k_i}(y_{gik_i} - \bar{y}_g)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) := s_{gp}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \tag{4}$$

where $\bar{y}_g$ is the average expression across the two experiments. The estimate in Equation (3) is appropriate when the variances of the gene expression data in the two conditions are different, and its use is suggested in [28, 69]. The estimate in Equation (4) uses the typical pooled estimate of the common variance, and it is used less often, see for example in [76]. Because of the large variability of gene expression data measured with microarrays, authors have suggested some forms of penalization for the denominator of the t-statistic. For example, Golub *et al.* [39] suggest estimating σ_g by the quantity

$$s_{S2Ng} = \frac{s_{g1}}{\sqrt{n_1}} + \frac{s_{g2}}{\sqrt{n_2}},$$

and refer to the ratio $|\bar{y}_{g1} - \bar{y}_{g2}|/s_{S2Ng}$ by *signal-to-noise ratio*. Because $s_{S2Ng} > s_{Ig1}$ unless either $s_{g1} = 0$ or $s_{g2} = 0$, the signal-to-noise ratio statistic penalizes those genes that have large variances in both conditions compared to those genes that have a large variance in one class and a low variance in the other. The justification for this choice is that when a gene is differentially expressed in the two conditions, it is biological reasonable to expect expression data distributed with very different variances. One objection to this justification is that one is interested in the distribution of the t-statistic under the null hypothesis of no differential expression.

Other forms of penalization are justified by the fact that because of the wide range of measurements, the estimate s_{Ig1} may be very small for some gene g and may produce an inflated value of the t-statistic. Therefore, authors have suggested to estimate σ_g by $a + s_{Ig1}$, and the constant a is chosen to minimize the coefficient of variation of the t-statistic in [101], while Efron *et al.* [30] suggest to replace a by the 90th percentile of the standard error of all the genes.

The most popular approach to choose a threshold is distribution free. The main idea is to compute the value of the t-statistic from the data in which the sample labels that represent the experimental conditions are randomly reshuffled. By repeating this process several times, it is possible to construct the empirical distribution of the t-statistic under the null hypothesis of no differential expression. From the empirical distribution function, one can select a gene specific threshold to reject the null hypothesis with a particular significance. This method is implemented in the popular program GENECLUSTER 2.0b[7] that conducts the differential analysis based on the signal-to-noise ratio statistic, or the standard t statistic in which $s_g = s_{Ig1}$. The program SAM[8] implements a distribution free differential analysis using the t-statistic with denominator $a + s_{Ig1}$. Because of the

[7] http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html

[8] http://www-stat.stanford.edu/̃tibs/SAM/index.html

large number of genes, authors have also developed algorithms for multiple comparison adjusted p-values, see for example [28].

Distribution free methods tend to be widely used in practice although few authors have suggested making distribution assumptions on the gene expression data. For example, Baldi and Long [5] introduce a Bayesian parametric version of the analysis based on the t-statistic, in which expression data transformed in logarithmic scale are assumed to follow a normal distribution. Usually, the t statistic is applied to data that are normalized using one of the methods described in Section 5.2. A model-based approach to simultaneously normalize and estimate the difference of gene expression between two experimental conditions is presented in [99]. Although their integrated modelling approach is appealing, a limitation is the large sample approximate distribution for the t-statistic.

7. Analysis of Multiple Conditions

Some of the most interesting applications of microarray technology are based on data collected under multiple experimental conditions. These conditions can be, for example, different known classes of the same tumor — such as acute leukemia [39] or non-Hodgkin's lymphoma [2] — or controlled experimental factors such as sex and age [52]. The different experimental conditions can also be time points, when the experimenter wishes to analyze the evolution of a physiological response [49] or to identify genomic features of a cell cycle [81], or to track down the genetic mechanisms that switch a locally growing tumor into a metastatic killer [19]. These different experiments are designed to answer different questions and they require different data analysis tools.

7.1 Main Objectives

Data are typically collected in a $G \times n$ array Y, where G is the number of genes whose expression level is measured in each of the n samples. Each row $y_g = (y_{g1}, \dots, y_{gn})$ collects the expression level y_{gj} for gene g measured in the n samples, while each column $e_j = (y_{1j}, \dots, y_{Gj})$ collects the expression level of the G genes in sample j. The expression levels can be either absolute or relative with respect to a common reference sample. The n samples are typically collected from $c \leq n$ conditions. We will continue to denote by n_i the number of samples taken in each condition i, so that $n = \sum_{i=1}^{c} n_i$. The main experimental goals of multiple microarray experiments fall neatly into two broad classes:

Class Prediction. The experimenter chooses c conditions and measures repeatedly the expression level of the same set of genes in each condition. Each condition is regarded as a class label, and the goal of the analysis is to detect the genes that are differentially expressed in at least two conditions, or that are good predictors of the class. The analysis described in Section 6 is a particular example of this type of analysis, although its goal is mainly to "describe" the molecular differences of two conditions. In cancer genomic experiments, for example, the goal may be the development of new diagnostic tools, based on the molecular profiles of tumor cells. To do this, the experimenter may collect samples from patients known to be affected by different types of the same tumor class — such as different types of leukemia [39] or breast cancer [105] — and uses each patient sample as an instance of the molecular profile of the specific type of tumor. The goal of the analysis would be to determine the molecular profile of each type of tumor, to make possible a molecular-based diagnosis of a specific tumor [59].

Class Discovery. Multiple microarray experiments can also be used to help investigators create new classifications by discovering new classes characterized by a specific molecular profile. There is little doubt that the current taxonomy of cancer lumps together molecularly distinct diseases with distinct clinical phenotypes, with the consequence that patients receiving the same diagnosis can have different clinical courses and treatment responses [2]. For example, in the analysis of gene expression data collected from tissues of breast cancer patients, the goal may be the identification of new molecular taxonomies of breast cancers characterized by particular profiles. Again, the advantage of such discovery could be to aid the diagnosis, as well as to tailor treatments to more specific diagnoses. Sometimes, the distinction among different classes is observable only through the dissection of the dynamics of the genomic system. In these cases, the different conditions are represented by time points and the goal is to identify groups of genes behaving in a similar way.

The solution to class prediction problems requires the development of classification rules able to label the molecular profile of a sample, whereas the goal of class discovery studies is to create new classes from the available data. Formally, the distinction between the two tasks is that the former relies on a labeled data set, while the latter relies on an unlabeled data set. Supervised and unsupervised machine learning methods are currently used to tackle both tasks.

7.2 Supervised Classification

Supervised classification techniques are used to learn a classification rule from a set of labeled cases (called the *training set*) to classify new unlabeled cases in a *test set*. Each condition i is regarded as a class label, and the columns of the data matrix Y are the labeled cases used to learn mappings of molecular profiles to class labels. This mapping can be constructed in two ways. One approach models the dependency of the class labels on the gene expression, and this dependency is used to compute the probability of each class label, given its molecular profile. The classification can be based on a decision rule that selects a class by minimizing the expected loss. We call this approach model-based in contrast to a model-free approach that partitions the space of gene expression data so that each element of the partition corresponds to one and only one class label. Well known model-based classification methods are multinomial logistic or probit regression [71] and naive Bayes classifiers [41]. In multinomial logistic/probit regression, the probability distribution of the class labels $p(i|y_1, \ldots, y_G)$, $i = 1, \ldots, c$, is modeled as

$$p(i|y_1, \ldots, y_G) = F^{-1}(\beta_0 + \sum_g \beta_g y_g)$$

where F is the cumulative distribution function of the logistic distribution or of the standard normal distribution, and β_g are regression parameters. The probabilities are estimated directly from the training set and, to classify a case with known gene expression data, say $y_1, \ldots, y_G$, it is sufficient to compute the probability $p(i|y_1, \ldots, y_G)$ for all i, and to select the class with maximum probability. The classification rule can be adjusted to account for misclassification costs. A difficulty with this approach, known as “small n large p” problem, is the typical sparseness of microarray data, which often consists of thousands of genes (large p) and few observations for each gene (small n). A Bayesian method for fitting probit regression and tacking the “small n large p” problem has been recently proposed by West *et al.* [105], for the classification of different types of breast cancers.

Naive Bayes classifiers rely on the assumption that expression measurements within a microarray are conditionally independent given the class membership, so that the stochastic dependency between class labels and gene expression values can be modeled as

$$p(i, y_1, \dots, y_G) = p(i) \prod_g p(y_g|i).$$

where $p(y_g|i)$ is the density function of the expression level of gene g in class i, and $p(i)$ is the marginal probability of the ith class. Once the terms $p(i)$ and $p(y_g|i)$ are estimated from the training data, it is possible to predict the class of a new unlabeled case by computing the posterior distribution of the class labels, given the gene expression values observed in the new case. The conditional independence assumption of the classifier simplifies the dependency structure of the class labels on the gene expression data, and the classification rule can be learned efficiently and accurately, despite the small number of observations available for each gene [54].

The classification accuracy of both regression and naive Bayes classifiers can be improved by selecting the subset of genes with highest predictive accuracy. In logistic regression, for example, the selection of genes can be done by using standard large sample model selection techniques, which are reliable when the number of observations for each pair (y_g, i) is at least 25 [71]. Similar feature selection methods are available for the naive Bayes classifier [72]. However, the staggering cardinality of the model space requires the adoption of heuristic search strategies. For example, if one limits attention to the set of all additive logistic regression models, the cardinality of the model space would be 2^G, where G can be as large as 12,625, in the case of experiments carried out with the Affymetrix Human Genome U95A chip.

Although model-based approaches provide a quantification of the uncertainty of the predictive model and a principled way to select a subset of the most predictive genes, model-free approaches are currently the most popular. Examples of model-free approaches to classification are methods for discriminant analysis such as Fisher linear discriminant analysis, nearest neighbor classification trees, [41], or support vector machines [103]. A comprehensive review of classical statistical methods for discriminant analysis applied to gene expression based tumor discrimination is presented in [27], with a critical assessment of pros and cons of each method. The selection of genes with predictive properties is often based on heuristic rules, such as filtering out genes with a fold-change not exceeding a particular threshold [97], or selecting genes that are highly correlated with a dummy pattern of zeros and ones mirroring the class partition [39].

Support vector machines are a supervised classification technique of increasing popularity that uses the training data Y in which genes known to belong to the same functional class are assigned the same class label, say $i = 1$, and genes known not to be members of that class are assigned the same different class label, say $i = -1$. The two-labeled data constitute the training set for the support vector machine that is used to learn to distinguish between members and non-members of the functional class on the basis of their expression data. Formally, a support vector machine maps the binary labeled training data $y_1, \dots, y_G$ into a high dimensional feature space F, where

$f_g = \phi(y_g)$. In the feature space F, the two classes of data are separated by a hyperplane (w, b) with maximum margin γ. The optimal solution is known to be $w = \sum_g \alpha_g i_g \phi(y_g)$, where i_g is the label assigned to the gene g, and the parameters α_g are positive real numbers chosen to maximize the function

$$\sum_g \alpha_g - \sum_{gh} \alpha_g \alpha_h i_g i_h < \phi(y_g), \phi(y_h) >$$

$$\text{subject to} \quad \sum_g \alpha_g i_g = 0$$

where $< \phi(y_g), \phi(y_h) >$ is the dot product in the feature space. The real number b is found by maximizing the hyperplane margin

$$\gamma = \min_g i_g \{< w, \phi(y_g) > -b\}.$$

Having learned the expression features of the two classes, the support vector machine can be used to recognize and classify the genes in the data set on the basis of their expression [10]. The classification is based on the decision function

$$d(y) = \text{sign}(< w, \phi(y) > -b) = \text{sign}(\sum_g \alpha_g i_g < \phi(y_g), \phi(y) > -b)$$

so that if the decision function for the new gene with expression profile y is $d(y) > 0$, the gene is assigned to the same functional class of the genes labeled by $i = 1$ in the training set. Note that the parameter α_g associated with the profile y_g expresses the weight that this point has on the decision function. Particularly, only a subset of the initial training point will have non-zero weights α_g. These points are called the support vectors. Because both the learning algorithm and the decision function depend on the dot product $< \phi(y_g), \phi(y_h) >$, the specification of the map $\phi(\cdot)$ can be done indirectly via the kernel function $K(x, y) =< \phi(x), \phi(y) >$. Typical kernel functions are the dot product, when $\phi(\cdot)$ is the identity, and some power or exponential function of the dot product.

7.3 Unsupervised Classification and Clustering

Unsupervised classification techniques, such as clustering or multidimensional scaling, can be used to group either genes with a similar expression profile or samples (e.g. patients) with a similar molecular profile, or both. The average-linkage hierarchical clustering proposed by Eisen *et al.* [32] is today one of the most popular analytical methods to cluster gene expression data. Given a set of n expression values measured for G genes, this approach recursively clusters genes, or samples, according to some similarity measure of their measurements. When applied to gene expression profiles, the method treats each row of the $G \times n$ data matrix Y as an n-dimensional vector, and iteratively merges genes into a single cluster. Relationships among the genes are represented by a tree (*dendrogram*), whose branch lengths reflect the degree of similarity between the genes. The similarity measure commonly used is the correlation between pairs of gene expression data, but other measures have been used, such as Euclidean distance or information-theoretic metrics. The resulting tree sorts the genes in the original data array Y, so that genes or groups of genes with similar expression patterns will be adjacent. The ordered table can be displayed graphically, together with the dendrogram, for the investigators' visual inspection. Figure 8 provides an example of such graphical display known as an Eisen plot. Software for the cluster analysis and visualization is available from the Eisen's Lab web page.[9]

The same approach can be applied to the columns of the data matrix to identify samples with a similar molecular profile. Hierarchical clustering applied to the rows and columns of the data array Y will return a sorted image of the original data. The image of the sorted data is typically used to support the operation of partitioning genes or samples into separated groups with common patterns. This operation is done by visual inspection, by searching for large contiguous patches of color that represent groups of genes sharing similar expression patterns or groups of samples sharing similar molecular profiles. The identification of these patches allows the extraction of subgroups of genes to be used to re-cluster the samples and, conversely, the extraction of subgroups of experiments to be used to re-cluster gene expression patterns. Although the choice of the subsets is arbitrary and the final result heavily depends on the genes or samples selected at each step of the procedure, this method has been successfully applied to identify, for example, new molecular classes of non-Hodgkin lymphoma [2], of cutaneous malignant melanoma [8], of breast cancer [94], and of lung cancer [7].

Notwithstanding these interesting results, this approach is not without problems. The subjective nature of partitioning by visual inspection may lead one to disregard some important information or to include irrelevant information. Decades of cognitive science research have shown that the human eye tends to overfit observations, to selectively discount variance and "see" patterns in randomness [102, 37]. Permutation tests are sometimes used to validate the partitions found by this procedure [32], and a bootstrap-based validation technique is presented in [56]. The Gap statistics of Tibshirani *et al.* [100] can also be used to find the optimal number of groups in the data. A second problem of this approach is the dilution of distance measures in average-linkage hierarchical clustering. When genes are assigned to the same subtree, the similarity measure between subtrees, or between single genes and subtrees, is computed by using a subtree profile calculated as the average of the subtree member profiles. As the subtree grows, this average profile becomes a less adequate representation of the subtree members. A solution to this problem can be the adoption of single-linkage clustering or complete-linkage clustering [82].

Relevance networks [13] are a non hierarchical clustering method which does not suffer from this dilution problem. For each pair of genes, the method computes a similarity between their expression measures, such as correlation or mutual information on appropriately discretized expression measures, and assigns genes whose similarity measure is above a preset threshold to the same cluster. This method can be regarded as a graphical representation of the matrix of all pairwise distances between gene expression profiles, since genes assigned to the same cluster are linked by an edge whose thickness is proportional to the similarity between the two elements. Although this method does not rely on visual inspection, the division into clusters is entrusted to an arbitrary threshold.

When some prior knowledge about the number of groups in the data is available, k-means clustering can be used as an alternative to hierarchical clustering to provide an optimal grouping of rows and/or columns of the data array Y into a preset number of clusters. K-means clustering starts

[9] http://rana.lbl.gov/EisenSoftware.htm

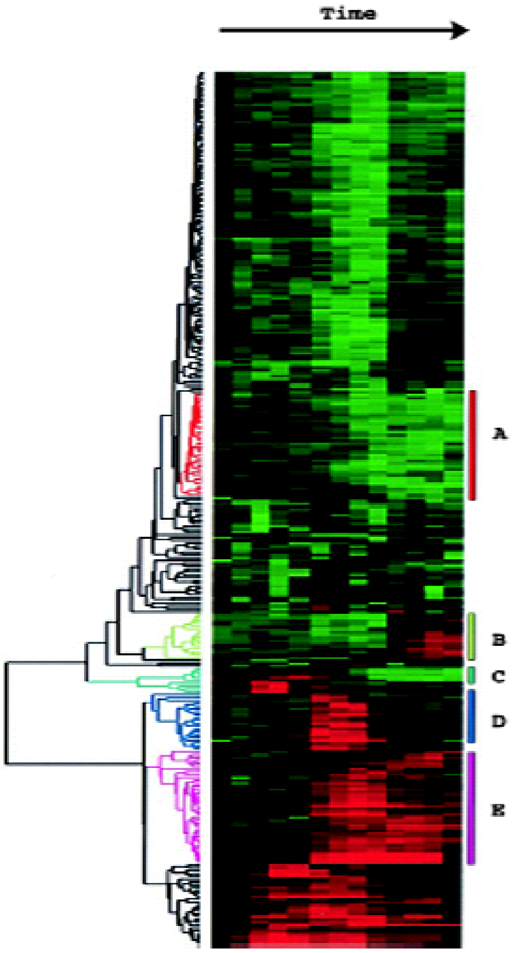

Figure 8: *Example of Eisen plot applied to 517 gene expression data measured in 13 experiments displaced along time. The image is a graphical display of the data array Y with rows sorted using the average-linkage hierarchical clustering procedure. Each row of the image represents a gene, and each column represents an experiment. Each cell (g, j) of the image represents the fold-change of gene g, relative to the first time point expression value, in logarithmic scale. Cells with log fold-change equal to 0 are colored black, increasingly positive log fold-changes with reds of increasing intensity, and increasingly negative log fold-changes with greens of increasing intensity. A representation of the dendrogram is appended to the image. Contiguous patches of color, labeled by the investigators with the letters A, B, C, D and E, are taken to indicate groups of genes that share similar expression patterns. The image is reproduced from [32].*

with a random assignment of the rows (columns) of the data matrix into k disjoint groups, and the rows (columns) are iteratively moved among the clusters until a partition with optimal properties is found. Typically, the criterion to find the optimal partition is minimizing the within-cluster variability while maximizing the between-cluster variability. The within-cluster variability is measured by the average distance between cluster members and the cluster profile, while the between-cluster variability is a measure of the distance of each cluster member from the other cluster profiles. K-means clustering is used by Tavazoie *et al.* [98] to identify groups of genes with similar patterns across different experimental conditions. Similar to k-means clustering are the self-organizing maps of Kohonen [58]. A self-organizing map uses a 2- or 3- dimensional projection of each cluster profile and provides a straightforward graphical representation of the result. Self-organizing maps have been used to identify classes of genes with similar functions in the Yeast cell cycle [97], and they have been combined with the nearest neighbor classification method to discriminate between two types of acute leukemia [39]. An implementation of the method is in GENECLUSTER 2.0b.

One potential danger of searching an optimal sorting of the data array Y by independently looking for an optimal arrangement of rows and columns is to overlook the association between gene expression data and samples. Clustering methods that address the issue of sorting simultaneously rows and columns of the matrix Y have recently been proposed, such as "gene shaving" [42], "biclustering" [17], "coupled two way clustering" [36], or the "plaid model" [61]. Gene shaving is a block clustering technique to cluster genes and samples simultaneously. The algorithm uses an iterative procedure to identify subsets of highly correlated genes that vary greatly between samples. Biclustering is a method for clustering simultaneously genes and samples by using a similarity measure of genes and samples. The idea of coupled two way clustering is to cluster pairs of small subsets of genes and samples. The rationale of this approach is that only a small subset of the genes is expected to participate in any cellular processes, which by themselves are supposed to take place only in a subset of the samples. Therefore, the algorithm looks for pairs of a relatively small subset of genes and samples yielding stable and significant partitions. The plaid model is a block clustering technique that produces overlapping clusters.

All these clustering methods are model-free: they do not rely on any assumptions about the distribution of genes or samples. In contrast, model-based procedures [6, 15] regard clustering as the task of merging together the observations generated by the same probability distribution. Cast in this framework, the simultaneous clustering of genes and samples can be regarded as the task of identifying a hidden variable labelling the cells of the array Y. In this way, the problem of simultaneously grouping rows and columns could be solved by estimating the hidden variable and, subsequently, by finding the genes and the samples that share the same label. If we let H be the hidden variable that assigns the same label (r, c) to the similar cells of Y, then the likelihood function of the data matrix Y, conditional on a known labelling h of rows and column, can be represented as

$$p(Y|h,\theta) = \prod_r \prod_c \prod_{g(r)} \prod_{j(c)} p(y_{g(r)j(c)}|\theta_{r,c})$$

where $\theta = \{\theta_{r,c}\}$. The index $g(r)$ specifies the genes assigned the same label r, whereas the index $j(c)$ specifies the samples assigned the same label c, and $p(y_{g(r)j(c)}|\theta_{r,c})$ is the density function of the genes and samples assigned the same label pair (r, c). The overall likelihood can then be written as $\sum p(Y|h,\theta)p(h|\eta)$, where $p(h|\eta)$ is the probability that $H = h$ that depends on parameters η. The EM algorithm can be used to estimate the unknown parameters for a specification of the density function $p(y_{g(r)j(c)}|\theta_{r,c})$ and the probabilities $p(h|\eta)$. Alternatively, if some initial labelling of the experiments is available, an agglomerative clustering procedure can be used to iteratively relabel rows and columns. Some relevant work in this area is presented in [112] for one dimensional clustering, and in [7]. Although model-based clustering relies on distributional assumptions of gene expression profiles and samples, the validity of these assumptions can be assessed using statistical validation techniques. One of the main advantages of a model-based approach is the possibility of using sounds statistical methods to assess the significance of the similarity between genes or samples and to identify the best number of clusters consistent with the data [34].

7.4 Time Series Analysis

Several applications of genome-wide clustering methods focus on the temporal profiling of gene expression. The intuition behind this analytical approach is that genes showing a similar expression profile over time are acting together, because they belong to the same or, at least similar, functional categories. Temporal profiling offers the possibility of observing the regulatory mechanisms in action and tries to break down the genome into sets of genes involved in the same, or at least related, processes. However, the clustering methods described in the previous section rest on the assumption that the set of observations for each gene are exchangeable over time: pairwise similarity measures, such as correlation or Euclidean distance, are invariant with respect to the order of the observations and, if the temporal order of a pair of series is permuted, these distance measures will not change. While this assumption holds when expression measures are taken from independent biological samples, it may be no longer valid when the observations are a time series.

Although the functional genomic literature is becoming increasingly aware of the specificity of temporal profiles of gene expression data, as well as of their fundamental importance in unravelling the functional relationships between genes [18, 19, 20], traditional clustering methods are still used to group genes on the basis of their similarity. For example, Holter *et al.* [44] describe a method to characterize the time evolution of gene expression levels by using a time translational matrix to predict future expression levels of genes based on their expression levels at some initial time, thus capturing the inherent dependency of observations in time-series. This approach relies on the clustering model obtained using a timeless method, such as singular value decomposition [3], and then infers a linear time translational matrix for the characteristic modes of these clusters. The advantage of this approach is that it provides, via the translational matrix, a stochastic characterization of a clustering model that takes into account the dynamic nature of temporal gene expression profiles. However, the clustering model which this method relies upon is still obtained by disregarding the dynamic nature of the observations, while we expect that different assumptions on the correlation between temporal observations will affect the way in which gene profiles are clustered together.

When the goal is to cluster gene expression patterns measured at different time points, the observations for each gene are serially correlated and clustering methods should take into account this dependency. The method of [83] is a Bayesian model-based approach to cluster temporal gene expression patterns that accounts for the temporal dependencies using autoregressive models. The method represents gene expression dynamics as autoregressive equations and uses an agglomerative procedure to search for the most probable set of clusters, conditional on the available data.

Features of this method are the ability to take into account the dynamic nature of gene expression time series during clustering, and a principled way to identify the number of distinct clusters. As the number of possible clustering models grows exponentially with the number of observed time series, a distance-based heuristic search procedure is used to render the search process feasible. In this way, the method retains the important visualization capability of hierarchical clustering but acquires an independent measure to decide when two series are different enough to belong to different clusters. Furthermore, the reliance of this method on an explicit statistical model of gene expression dynamics makes it possible to use standard statistical techniques to assess the goodness of fit of the resulting model and validate the underlying assumptions. When the autoregressive order is equal to zero, this method subsumes, as a special case, model-based clustering of atemporal (i.e. independent) observations. The method is implemented in the program CAGED[10] described in [91].

8. Open Challenges

Microarray technology makes it possible the simultaneous execution of thousands of experiments to measure gene expression levels in a variety of conditions. This article has reviewed the biology of gene expression, the technology of microarrays, and several statistical issues involved in the analysis of gene expression data, including experimental design, data quality, data analysis and validation. Although a massive effort is under way to improve methods and technology, several issues are still open and are particularly relevant to the statistical community.

Experimental design The design of a microarray experiment is an unprecedented challenge. The main character of microarray technology is to make possible the parallel execution of thousands of experiments that are not independent of each other. For example, the measurements of the gene expression data are subjected to common experimental errors, such as those due to the amount of fluorescent dye used to label the target in each experimental replicate, or the amount of mRNA in each sample target. The challenge is the design of parallel and dependent experiments that can exploit the full power of this technology. Because no agreement exists about the appropriate statistical analysis of gene expression data produced with microarrays, and because many experiments with microarrays are conducted to generate rather than test hypotheses, critical experimental design questions are still far from being answered.

Quality assessment and normalization A very important issue when analyzing gene expression data is the ability to assess whether the execution of an experiment was successful, or to evaluate the quality of the experimental data. By this we mean the ability to decide whether the effects of random components such as variations in the amount of dye, or variations of the mRNA samples, are not large enough to irremediably mask the signal in the data. The normalization and gene filtering techniques discussed in Section 5 seem to be *ad hoc* bias-correction procedures, but their effect is unclear and their use is questionable in many applications. Some initial efforts in this direction are presented in [45]. The authors investigate whether probability distributions such as the Benford's law of the first significant digit, or the Zipf's law, can provide reference distributions to be used as gold standard in data quality assessment. Although very preliminary, their results are suggestive and open the way to a general probabilistic way to measure the reliability and quality of microarray

[10] http://www.genomethods.org/caged

data.

Differential analysis The last two years have witnessed an increasing number of research articles proposing methods for the differential analysis of gene expression data measured in comparative experiments. Many of these methods use one of the t-statistic described in Section 6 with an *ad hoc* chosen denominator, and the most disconcerting fact is the lack of empirical and theoretical studies to help choose the best method. The consequence seems to be that the choice of the differential analysis method is driven by the availability of software rather than the quality and appropriateness of the method. Furthermore, many of the original problems associated with gene expression data measured by the Affymetrix software MAS 4.0 have been overcome by the latest statistical software MAS 5.0 with a consequent change in research priorities. Particularly, the improved quality of the data produced by the new software opens the way to the development of full parametric methods.

Survival analysis While extensive work has been conducted to develop methods for the differential analysis of gene expression data measured in two conditions, very little is known about the analysis of gene expression data in which the training signal is a continuous variable. Particularly important to cancer genomics applications is the development of methods for the selection of genes that are predictive of the survival time of patients treated with a particular therapy. Some preliminary work is this area is in [75, 79].

Metric selection An open issue in the analysis of gene expression data is the selection of the metrics most suitable to answer specific biological questions. As an example, popular clustering methods use correlation, Euclidean distance, Kulback-Liebler information distance and, typically, different distances sort gene expression profiles in different ways. Similarly, the classification induced by support vector machines depends on the specification of the kernel function. An important contribution would be the development of a formal way for determining which metrics are most relevant, or robust, for different questions.

Does clustering provide the right answer? Clustering techniques are extremely popular tools for the comparative analysis of gene expression data collected in a variety of conditions. The main reason for using clustering methods is the intuition that co-regulated genes have similar patterns, or similar levels of expression [32]. However, clustering techniques by themselves cannot discover the dependency structure between genes. Popular knowledge representation formalisms such as Bayesian networks [22] and dynamic Bayesian networks seem to be the ideal modeling tool for capturing the dependency structure among genes. The big challenge is whether the data structure available — large number of parameters for few observations — makes Bayesian networks induced from gene expression data reliable. The wealth of genomic information grows daily and one may imagine that full Bayesian methods could be used to integrate the data with prior knowledge in a coherent way. Some initial attempts are in [35, 92, 113].

Validation Validation of cluster analysis is a very important issue that deserves further attention. Because clusters of similar genes/experiments are often identified by visual inspection or by imposing arbitrary thresholds, an independent quantitative validation of the results is required to assess whether the clusters are indeed capturing the signal in the data. Permutation tests as in [7] or bootstrapping the results [56] are often used to show that clustering applied to data in which the signal has been removed does not identify meaningful groups of genes/experiments. However, it is important to stress that these tests do not prove the functional validity of the groups identified in the data. An increasing number of studies use an independent biological validation of the identified groups [2, 39], but on such a small number of cases (for example 40 patients in [2]) this validation does not seem to provide much support. Some authors show the validity of their results by using different clustering techniques [7, 8]. The development of sound validation tests ranks among the top priorities in the field.

Eric Lander [60] wrote that developing experimental designs able to take advantage of the full power of microarray technology is the challenge for biologists of this century but he also acknowledges that the greatest challenges are fundamentally analytical. The newly born functional genomic community is in great need of tools for data analysis and visual display of the results, and the statistical community could offer an invaluable contribution toward an efficient collection and use of functional genomic data.

References

[1] Affymetrix Inc. *Statistical Algorithms Description Document*, 2002. Available from http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. S. T. Tran, X. Yu, J. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warmke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106, 2000.

[4] J. C. Alwin, D. J. Kemp, and G. R. Stark. Methods for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA*, 74:5350–5354, 1977.

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

[6] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[7] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas

by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA*, 98:13790–13795, 2001.

[8] M. Bittner, P. Meltze, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhinik, A. Ben-Dork, N. Sampask, E. Dougherty, E. WangI, F. MarincolaI, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[9] D. D. L. Bowtell. Options available — from start to finish — for obtaining expression data by microarray. *Nature Genetics*, 21:25–32, 1999.

[10] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97:262–267, 2000.

[11] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetic Supplement*, 21:33–37, 1999.

[12] A. J. Butte, A. Kho, and I. S. Kohane. *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA, 2002.

[13] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA*, 97:12182–12186, 2000.

[14] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell*, 12:323–337, 2001.

[15] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. MIT Press, Cambridge, MA, 1996.

[16] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, 2:364–374, 1997.

[17] Y. Cheng and G. M. Church. Biclustering of expresssion data. In *Proceedings of the 8th International Conference on Intelligent Systems and Molecular Biology*, pages 93–103. AAAI Press, 2000.

[18] G. A. Churchill and B. Oliver. Sex, flies and microarrays. *Nature Genetics*, 29:355–356, 2001.

[19] E. A. Clark, T. R. Golub, E. S. Lander, and R. O. Hynes. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature*, 406:532–535, 2000.

[20] H. A. Coller, C. Grandori, P. Tamayo, T. Colbert, E. S. Lander, R. N. Eisenman, and T. R. Golub. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. USA*, 97:3260–3265, 2000.

[21] The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[22] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, NY, 1999.

[23] D. R. Cox and N. Reid. *The Theory of the Design of Experiments*. Chapman and Hall/CRC, Boca Raton, FL, 2000.

[24] F. H. C. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

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

[26] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[27] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(576):77–87, 2002.

[28] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying genes with differential expression in replicated cDNA microarrays experiments. *Statistica Sinica*, 12:111–139, 2001.

[29] J. D. Duggan, M. Bittner, Y. Chen, P Meltzer, and J. M. Trent. Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, 21:10–14, 1999.

[30] B. Efron, J. D. Storey, and R. Tibshirani. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[31] B. Efron, J. D. Storey, and R. Tibshirani. Microarrays, empirical Bayes methods, and false discovery rate. Technical report, Department of Statistics and Division of Biostatistics, University of Stanford, 2001.

[32] M. B. Eisen, P. T Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[33] R. Ekins and F. W. Chu. Microarrays: Their origins and applications. *Trends in Biotechnology*, 17:217–218, 1999.

[34] C. Fraley and A. E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[35] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[36] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, 97:12079–12084, 2000.

[37] T. Gilovich, R. Vallone, and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314, 1985.

[38] R. Glynne, S. Akkaraju, J. I. Healy, J. Rayner, C. C. Goodnow, and D. H. Mack. How self-tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature*, 403:672–676, 2000.

[39] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 15:531–537, 1999.

[40] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis*. W. H. Freeman and Company, New York, 7th edition, 1999. Available from http://www.ncbi.nlm.nih.gov/books/.

[41] D. J. Hand. *Construction and Assessment of Classification Rules*. Wiley, New York, NY, 1997.

[42] T. Hastie, R.Tibshirani, M. B. Eisen, A. A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. O. Brown. ’Gene Shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1:1–21, 2000.

[43] F. C. P. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. S. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728, 1998.

[44] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. Banavar. Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. USA*, 98:1693–1698, 2001.

[45] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass. Making sense of microarray data distributions. *Bioinformatics*, 18:576–584, 2002.

[46] J. G. Ibrahim, M. H. Chen, and R. J. Gray. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97:88–99, 2002.

[47] National Human Genome Research Institute. A glossary of genetic terms. Available from http://www.nhgri.nih.gov/DIR/VIP/Glossary, 2001.

[48] R. A. Irizarry, G. Parmigiani, M. Guo, T. Dracheva, and J. Jen. A statistical analysis of radiolabeled gene expression data. In *Proceedings of the 33rd Symposium on the Interface: Computing Science and Statistics*, Fairfax Station, VA, 2001. Interface Foundation of North America.

[49] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson Jr., M. B. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 97:8409–8414, 1999.

[50] L. Jackson-Grusby, C. Beard, R. Possemato, M. Tudor, D. Fambrough, G. Csankovszki, J. Dausman, P. Lee, C. Wilson, E. S. Lander, and R. Jaenisch. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genetics*, 27:31–39, 2001.

[51] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[52] W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson. The contribution of sex, genotype and age to transcription variance in Drosophila melanogaster. *Nature Genetics*, 29:389–395, 2001.

[53] M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*, 28:4552–4557, 2000.

[54] A. D. Keller, M. Schummer, L. Hood, and W. L. Ruzzo. Bayesian classification of DNA array expression data. Technical Report UW-CSE-2000-08-01, Department of Computer Science and Engineering, Seattle, WA, 2000.

[55] K. M. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarrays. *Genetical Research*, 77:123–128, 2001.

[56] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA*, 98:8961–8965, 2001.

[57] M. K. Kerr and G. A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2:183–201, 2001.

[58] T. Kohonen. *Self Organizing Maps*. Springer, Berlin, DE, 1997.

[59] S. R. Lakhani and A. Ashworth. Microarray and histopathological analysis of tumours: The future and the past? *Nature Reviews Cancer*, 1:151–157, 2001.

[60] E. S. Lander. Array of hope. *Nature Genetics Supplement*, 21:3–4, 1999.

[61] L. Lazzeroni and A. B. Owen. Plaid models for gene expression data. Stanford Biostatistics Series 211, Department of Health Research and Policy, Stanford University, Stanford, CA 94305-5405, 2000.

[62] C. K. Lee, R. Weindruch, and T. A. Prolla. Gene-expression profile of the ageing brain in mice. *Nature Genetics*, 25:294–297, 2000.

[63] M. T. Lee, F. C. Kuo, G. A. Whitmorei, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA*, 18:9834–9839, 2000.

[64] G. G. Lennon and H. Lehrach. Hybridization analyses of arrayed cDNA libraries. *Trends Genet*, 7:314–317, 1991.

[65] R. J. Lipshutz, S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics Supplement*, 21:20–24, 1999.

[66] D. J. Lockhart and C. Barlow. Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews*, 2:63–68, 2001.

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

[68] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.

[69] I. Lönnstedt and T. P. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

[70] D. H. Ly, D. J. Lockhart, R. A. Lerner, and P. G. Schultz. Mitotic misregulation and human aging. *Science*, 287:2486–2492, 2000.

[71] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[72] T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.

[73] R. Nadon and J. Shoemaker. Statistical issues with microarrays: Processing and analysis. *Trends in Genetics*, 18:265–271, 2002.

[74] M. N. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52, 2001.

[75] D. V. Nguyen and D. M. Rocke. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, 2002. In press.

[76] A. B. Olshen and A. N. Jain. Deriving quantitative conclusions from microarray expression data. *Bioinformatics*, 18:961–970, 2002.

[77] W. Pan, J. Lin, and C. T. Le. How many replicates of arrays are required to detect gene expression changes in microarrays experiments? A mixture model approach. Technical report, Division of Biostatistics, School of Publih Health, University of Minnesota, 2001.

[78] W. Pan, J. Lin, and C. T. Le. A mixture model approach to detect differentially expressed genes with microarray data. Technical report, Division of Biostatistics, School of Public Health, University of Minnesota, 2001.

[79] P. J. Park, L. Tian, and I. S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 2002. In press.

[80] B. Phimister. Going global. *Nature Genetics Supplement*, 21:1, 1999.

[81] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29:153–159, 2001.

[82] J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427, 2001.

[83] M. Ramoni, P. Sebastiani, and I. S. Kohane. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, 99:9121–9126, 2002.

[84] A. Relogio, C. Schwager, A. Richter, W. Ansorge, and J. Valcarcel. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acid Research*, 30, 2002. In press.

[85] C. J. Roberts, B. Nelson, M. J. Marton, R Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 287:873–880, 2000.

[86] D. M. Rocke and B. Durbin. A model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8:557–569, 2001.

[87] C. Sabatti, S. L. Karsteny, and D. Geschwindy. Thresholding rules for recovering a sparse signal from microarray experiments. Technical report, Departments of Human Genetics and Statistics, UCLA., 695 Charles Young Drive South, Los Angeles, CA 90095-7088., 2001.

[88] E. E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 80:192–202, 2000.

[89] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.

[90] M. Schena, D. Shaloni, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA*, 93:10614–10619, 1996.

[91] P. Sebastiani, M. Ramoni, and I. Kohane. Bayesian model-based clustering of gene expression dynamics. In Parmigiani G., R. Irizarry, and S. L. Zeger, editors, *The Analysis of Microarray Data: Methods and Software*. Springer, New York, 2003. In press.

[92] E. Segal, B. Taskar, A. Gasch, N. Friedman, and Daphne Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 1:1–9, 2001.

[93] G. K. Smyth, Y. H. Yang, and T. P. Speed. Statistical issues in cDNA microarray data analysis. Technical report, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, May 2002.

[94] T. Sorlie, C. M. Perou, R. Tibshirani R, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, 98:10869–10874, 2001.

[95] E. Southern, K. Mir, and M. Shchepinov. Molecular interactions on microarrays. *Nature Genetics Supplement*, 21:5–9, 1999.

[96] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–297, 1998.

[97] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhui, S. Kitareewani, E. Dmitrovsky, E. S. Lander, and T. R Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.

[98] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

[99] J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11:1227–1236, 2001.

[100] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society, B*, 2001. In press.

[101] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121, 2000.

[102] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[103] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

[104] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[105] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson Jr, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA*, 98:11462–11467, 2001.

[106] B. White. Southerns, Northerns, Westerns, and Cloning: “molecular searching” techniques. In *MIT Biology Hypertextbook*. Massachusetts Institute of Technology, 1995. Available at http://esg-www.mit.edu:8001/esgbio/rdna/rdna.html.

[107] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, P. C. Afshari, and R. S Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8:625–637, 2001.

[108] J. J. Wyrick, F. C. P. Holstege, E. G. Jennings, H. C. Causton, D. Shore, M. Grunstein, E. S. Lander, and R. A. Young. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*, 402:418–421, 1999.

[109] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: A robust composite method addressing sinlge and multiple slide systematic variation. *Nucleic Acids Research*, 30, 2002. In press.

[110] Y. H. Yang, S. Dudoit, P. Luu, and T. S. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, pages 141–152. SPIE, 2001.

[111] Y. H. Yang and T. P. Speed. Design issues for cDNA microarray experiments. *Nature Genetics Reviews*, 3:579–588, 2002.

[112] K. Y. Yeung, C. F. A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. Technical Report UW-CSE-2001-04-02, Department of Computer Science and Engineering, University of Washington, Seattle, WA, 2001.

[113] C. Yoo, V. Thorsson, and G.F. Cooper. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In *Proceedings of the Pacific Symposium on Biocomputing*, 2002. Available from http://psb.stanford.edu.

[114] H. Zuzan, C. Blanchette, H. Dressman, E. Huang, S. Ishida, J. R. Marks, J. R. Nevins, R. Spang, M. West, and V. E. Johnson. Estimation of probe cell locations in high-density synthetic-oligonucleotide DNA microarrays. Technical report, 1Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, October 2001.