Statistical Challenges in Functional Genomics

Paola Sebastiani[*] and ***Marco Ramoni***[†]

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

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

Abstract

On February 12, 2001 the Human Genome Project announced that it had assembled 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 unparalleled discovery opportunities and it is reshaping biomedical sciences. One of the main aspects of this revolution is the introduction of heavily quantitative data-analytical 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.

Address: Paola Sebastiani, Department of Mathematics and Statistics, University of Massachusetts, Lederle Graduate Resarch Tower, Amherst, MA 01003. PHONE: (413) 545-0622, FAX: (413) 545-1801, EMAIL: sebas@math.umass.edu, URL: http://www.math.umass.edu.

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, for his insightful comments and helpful 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 all 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 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, putting it 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, leaders of the HGP consortium, Craig Venter of Celera Genomics, and U.S. President Clinton 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 [22], only twice the number of protein-coding genes in a worm or a fly. Because less than 50% of 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, 62, 64, 65] to have 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 knowledge discovery approach [56]. 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: what genes are differentially expressed in an abnormal/tumor cell compared to a normal cell? Which groups of genes are characteristic of a particular class of tumors? Is it is possible to identify genomic sub-classes of tumors to design more specific diagnostic tests and treatments? 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 of appropriate measures of uncertainty, make dubious distribution assumptions,

[1] Volume 409 of Nature, published February 15 2001 and available at `http://www.nature.com/genomics/human/`, reported 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/`, focused on the findings of the draft sequence reported by the privately funded company Celera Genomics

and are hardly portable across experimental platforms. Furthermore, little is 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 biology notions. 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 (Figure 1, right); 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 always pairs with T, and G always pairs with C.

Each strand of the double helix (Figure 1, left) consists of a sequence of nucleotides: the structural components, or building blocks, of DNA. A nucleotide is 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, 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) is the code for an amino acid. For example, the base sequence ATG codes for the amino acid methionine. Amino acids are the basic units of proteins, which perform most life functions, and the role of DNA is to provide instructions to the cells on when and how to produce a new protein.

The *genome* is an organism's complete DNA. Genomes vary widely in size across organisms, ranging from the 600,000 base pairs long genome of the bacterium *Escherichia coli* to the 3 billion base pairs of the human genome and, if unraveled, the whole DNA in a human cell would be six feet (two meters) long [13]. Except for mature red blood cells, all human cells, from fingernail cells to a neuron, contain the same DNA but, despite they carry the same set of instructions, these cells are actually different. These differences are due to the fact that segments of the DNA sequence become active by determining the creation of specific proteins in particular conditions and not in others. 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 messenger RNA was dis-

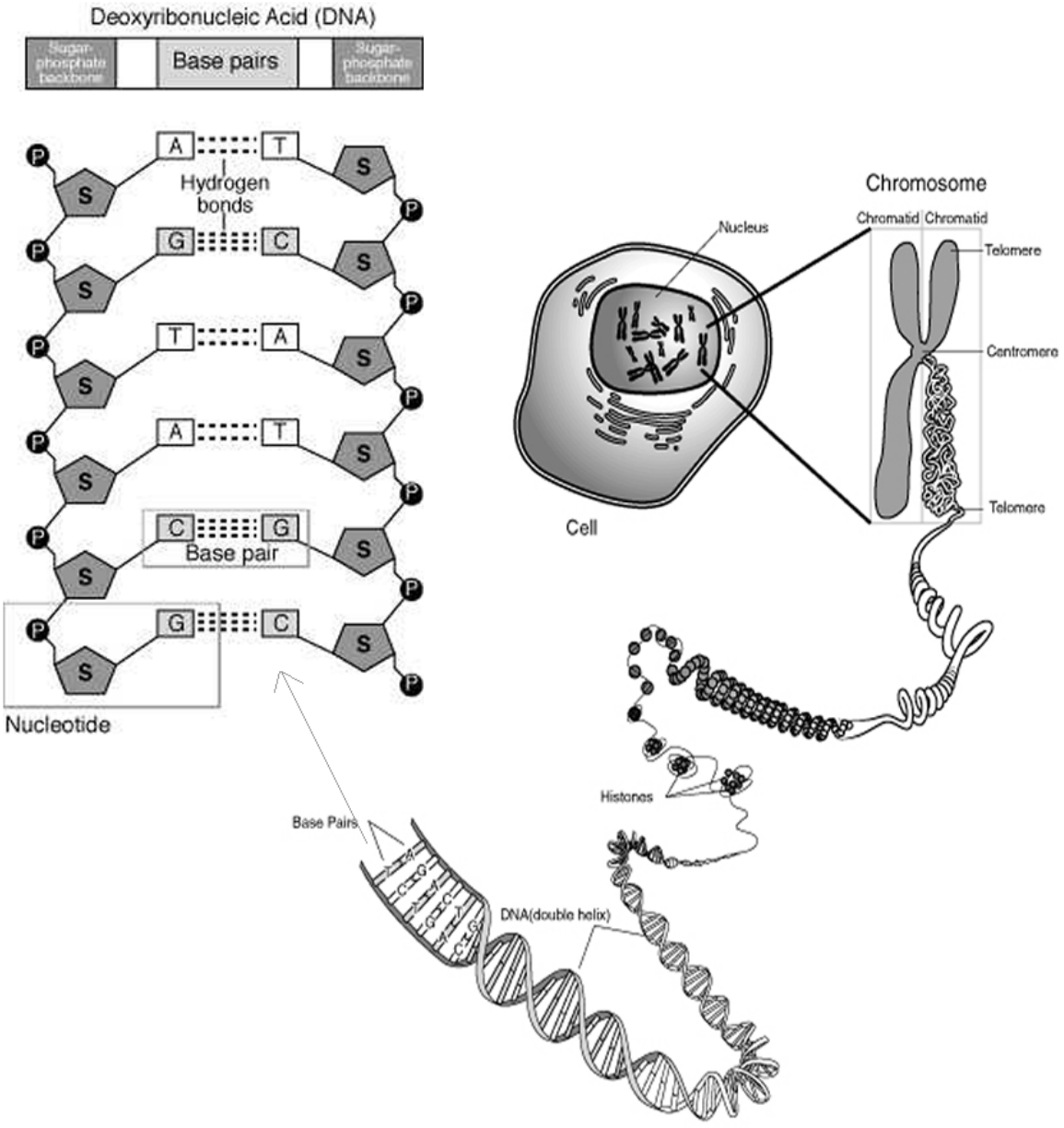


Figure 1: *The nucleus of the a cell contains the genetic code encoded in the DNA "packaged" in chromosomes. The DNA structure is a double helix where each strand consists of a sequence of nucleotides. Pictures taken from [43].*

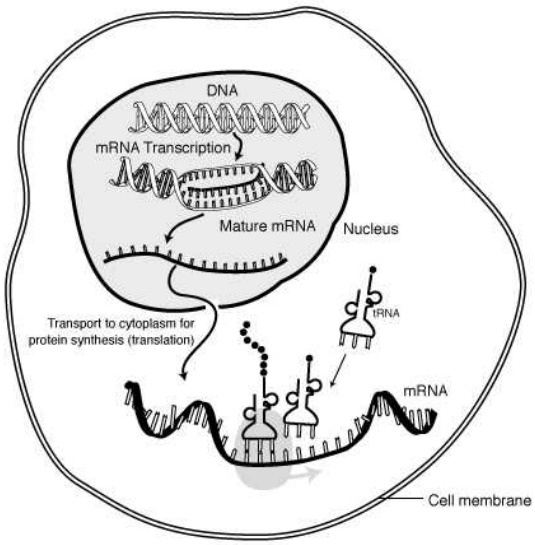


Figure 2: *During the expression process, a complementary copy of a gene code is transcribed into the mRNA. An appropriately modified copy migrates from the nucleus to the cytoplasm where it serves as a template for the protein synthesis. Picture taken from [43].*

covered and the theory of genetic regulation of protein synthesis was first described by Jacob and Monot [47]. The *gene expression level* is an integer valued or continuous measure that provides a quantitative description of the gene expression by measuring the amount of intermediary molecules produced during this process. These molecules are the mRNA (messenger Ribonucleic acid) and the tRNA (transfer Ribonucleic acid), and they are produced during the two steps of *transcription* and *translation* leading to the synthesis of a protein. This two-steps representation of protein-synthesis processes is depicted in Figure 2 and constitutes the *central dogma of molecular biology* [25].

Transcription The first step of a gene expression is the creation of a "complementary copy" of the code stored in one of its complementary strands. A feature of the nucleotide base pairing is that A always pairs with T, and G always pairs with C. The complementary copy of the gene DNA code transcribes T for the letter A, and G for the letter C (and viceversa) into the mRNA. The mRNA molecules look like a single DNA strand, except that the basis T is replaced by the nucleotide base U (Uracil).

Translation The mRNA 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.

The process of transcription initiates at the *promoter,* which is the part of a gene that contains the information to turn the gene on or off. The initial copy of mRNA (called mRNA *transcript*) contains the whole segment of DNA bases, which, in eukaryotes organisms, alternates *exons* — the region

of a gene that contains the code for producing the gene's protein — and *introns* — non-coding segments of the gene DNA. This initial copy is modified by removing the introns, and migrates out of the nucleus. During the translation process, cellular organelles, called *ribosomes*, function as "photocopy machines" by aligning several copies of mRNA into a "ribbon" that serves as template for the tRNA. As the whole gene expression consists of making a copy of its code into the mRNA, which is then lined up by the ribosomes to form the template for the tRNA, a measure of the gene expression level is the abundance of mRNA produced during this process [78]. This is the main intuition behind the parallel measurement of gene expression levels in microarrays that is described in the next section.

The definition of gene is currently moving away from its traditional *Mendelian definition* as *unit of heredity passed from parent to offspring* [43]. Defining a gene as a segment of DNA that contains the information for making a specific protein is a *functional definition*, and it is not completely satisfactory, as the same gene may determine more than one protein because of alternative splicing [15]. When its function is not even hypothesized, a gene is often described by an *Open Reading Frame* (ORF), which is a DNA sequence delimited by the sequence ATG, the starting *codon*, and one of the stop codons TAA, TAG and TGA.

3. Microarrays

Quantitative methods to measure gene expression levels have been available to biologists for more than twenty years. Northern and southern blots (see [3, 95]) 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) [93], is able to measure the global gene expression from entire cells or tissue. SAGE technology was introduced in 1995 by a team of cancer researchers at Johns Hopkins to rapidly identify differences between cancer cells 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 particular cell. SAGE uses these *short sequence tags* to mark the transcripts of a gene and identify the number of transcripts generated by a 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 [63]. Microarray technology has rendered efficient this process of measuring, simultaneously, the 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 Technology

The basic idea behind microarray technology is to simultaneously measure the expression level of thousands of genes by repeating the expression process backwards. Two key concepts behind this measurement process are *reverse transcription* and *hybridization*.

Reverse Transcription. The code transported by the mRNA can be experimentally isolated from a lump of cells and reversed-transcribed back into a copy, or clone, of DNA called cDNA. A collection of cDNAs transcribed from cellular mRNA constitutes the cDNA library of a cell.

Hybridization. Hybridization is the process of base pairing of two single strands of DNA or RNA [60]. 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, a mRNA molecule can hybridize to a melted cDNA molecule, when the mRNA contains the complementary code of the cDNA strands. When hybridization occurs, a singlestranded DNA binds strongly to complementary RNA, and in a way that prevents the strands from re-associating with each other [84].

Microarray technology is used to measure the level of expression of genes in a particular cell or tissue by hybridizing cellular mRNA extracted from the cell either to clones of an entire DNA sequence or to short specific segments known as *synthetic oligonucleotides* and called *oligos* in the bio-molecular jargon). The latter are short sequences of single-stranded DNA or RNA that bind readily to their complements. The tethered cDNA sequences or oligos are called *probes*, while the cellular mRNA extracted from the cell that contains the unknown expressed genes to be detected is called the *target* [72]. In both cases, the probe is taken to represent a gene of known identity. The target mRNA is labeled with florescent dye and, once on the microarray, genes expressed in the target will hybridize to their complementary probes. The more mRNA hybridize to a probe, the more intense the florescent dye will be on that probe. This mRNA abundance of a gene in the particular cell or tissue under analysis can be therefore measured by the emission intensity of the probe where the gene is located. The signal is filtered to remove noise generated by the microarray background and non-specific expression, i.e. spurious bindings of mRNA [29, 62].

3.2 cDNA Microarrays

cDNA technology was developed at Stanford University [78], although similar concepts can be traced back as far as the mid 80s [33]. A cDNA microarray consists of samples of cDNA strands that are fixed, in equal amount, to spots in a glass slide using a robot. Each strand of cDNA identifies uniquely, with its code, a gene so that each spot in the microarray corresponds to a gene. Investigators extract the total mRNA produced from two types of cells they are studying, for example healthy and tumor cells. They then label each of the two mRNA samples with a different fluorescent dye, one green (Cye3) and one red (Cye5). The pool of differentially labeled mRNA is allowed to bind to the complementary cDNA strands on the glass slides. During the hybridization, if segments of mRNA find their complementary portion among the samples of cDNA in the glass slide, they bind together. When the hybridization is complete, the glass slide is washed to remove the excess of the mRNA pool, 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 DNA spot and measures fluorescence for each dye separately, thus providing a measure of the mRNA abundance for each gene in the two cells. The intensity of the green spot measures the mRNA abundance of the gene in the cell whose mRNA was labeled with Cye3, while the intensity of the red spot measures the mRNA abundance of the gene in the cell whose mRNA was labeled with Cye5 and grey spots denote genes that were expressed in neither cell types.

These measurements provide information about the level of expression of each gene in the two

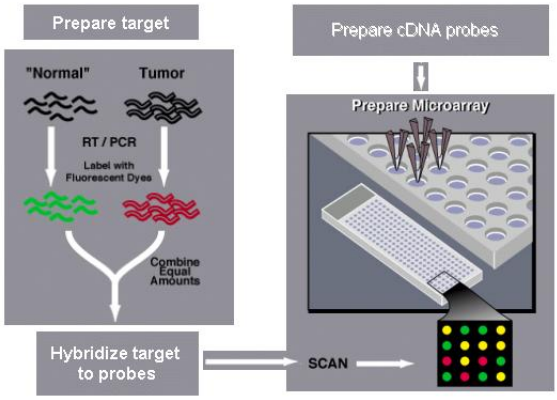


Figure 3: *A sketch of cDNA microarray technology. The method uses a robot to precisely apply tiny droplets containing clones of functional DNA (cDNA) to glass slides (probes). Researchers then attach fluorescent labels to the mRNA extracted from the cell they are studying. The labeled target is allowed to hybridize with the complementary DNA strands on the slides. Once the hybridization is completed, the slides are put into a scanning microscope able to measure the brightness of each fluorescent dot; brightness reveals how much of a specific DNA fragment is present in the target.*

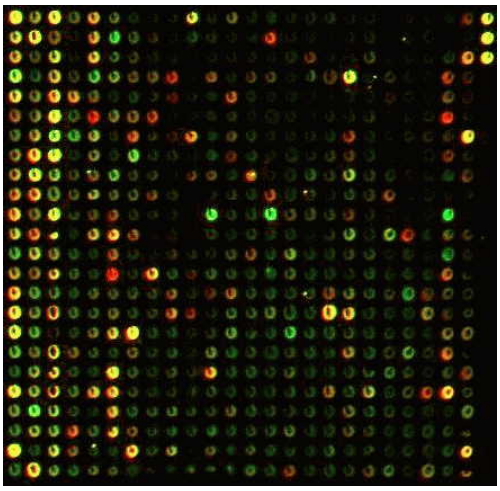


Figure 4: *A scanned image produced from a cDNA microarray experiment. Each spot denotes a gene. Grey spots denote genes that were expressed in neither types of cells, colored spots identify genes that were expressed in one of the two cells or both. The color of the spot informs about the relative expression of the gene in the two cells.*

cells. The monochrome images from the scanner can be imported into software and pseudo-colored to provide a quantitative measure of the relative expression of each gene in the two cells, which is adjusted to account for background noise. Figure 4 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 the gene 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. Examples of the application of this technology are discussed in [78] and, more recently, in [10].

3.3 Synthetic Oligonucleotide Microarrays

High-density synthetic oligonucleotide microarrays are fabricated by placing short DNA sequences (*oligonucleotides*) on a small silicon chip by means of photolithographic techniques used in computer microprocessor fabrication. This proprietary technology, developed and commercialized by Affymetrix of Santa Clara, CA, under the trademark of GeneChip®, allows the production of highly ordered matrices containing almost 20,000 genes (Affymetrix Human Genome U133 Set).

The rationale behind this technology is the concept of "probe redundancy", that is, a set of well-chosen small segments of DNA is not only sufficient to uniquely identify a specific gene but it will also reduce the chances that fragments of an unrelated mRNA will randomly hybridize to the probe. Therefore, synthetic oligonucleotide microarrays represent a gene not by the whole copy of its functional DNA but rather by a set of fixed-length independent segments unique to the functional DNA of the gene, as shown in Figure 5. On the GeneChip® platform, each oligonucleotide (probe) is 25-base long and each gene is represented by 16-20 *probe pairs*. 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, so that each such probe should hybridize only with the complementary portion of mRNA produced by the whole functional DNA sequence when the gene is expressed. The MM probe is identical to the corresponding PM probe except for the fact that the base in the central position is inverted. The inversion of the central base makes the MM probe a specificity control because, by design, hybridization of the MM probe can be attributed to background noise, or non gene specificity of the PM probe, and it is used to remove background and non-specific hybridization [64, 62]. Each cell of an Affymetrix oligonucleotide microarray consists of millions of samples of a PM or MM probe, and each gene is represented by a *probe set* of 16-20 probe pairs. Probes are scattered across the array to avoid systematic bias.

One microarray can be used only to detect the genes that are expressed and their specific expression level in a particular experimental condition. Investigators extract the total mRNA produced by a lump of cells, and label it with a fluorescent dye. They then let the labeled mRNA hybridize with the probe pairs in the microarray. Each segment of the target hybridizing to a probe will render it fluorescent, thus increasing the emission of the probe. Once hybridization has occurred, the microarray is washed to remove the excess of mRNA, and it is scanned with a standard laser scanner. The scanner generates an image of the microarray and the image is gridded to identify the cells containing each probe. The intensity of each cell in the image is then taken as a proxy of the expression level of the corresponding probe, and the expression level of each gene is computed as a robust average of the specific hybridization of the probe pair set. This specific hybridization of a probe-pair is computed as either the difference between the emission intensity of each PM and its related

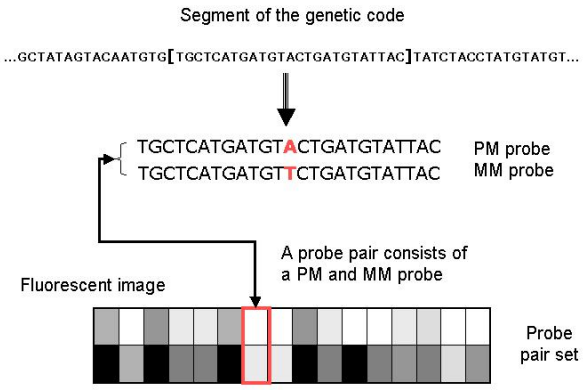


Figure 5: *An oligonucleotide microarray associates a gene with a set of probe pairs, in this case 20. Each probe pair consists of a perfect match probe (PM) and a mismatch probe (MM). Each PM probe contains millions of oligonucleotides 25 bases long and it is paired with the MM probe, in which the central base of the oligonucleotides is inverted. After mRNA hybridization, the microarray is read with a laser scanner to produce an image, where the intensity of the MM probes is used to correct the intensity of the PM probes to provide a measure of the specific hybridization.*

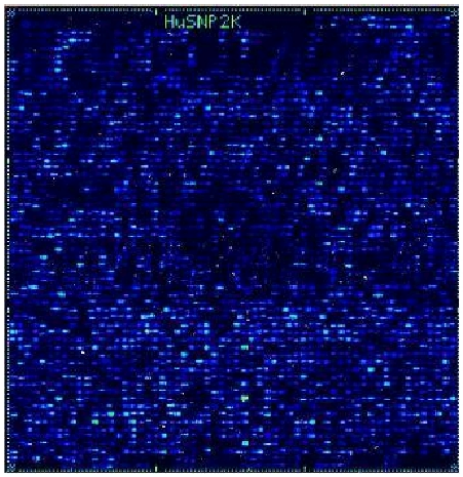


Figure 6: *Scanned image of a synthetic oligonucleotide microarray. Grid cells represent* probes *and the intensity of each matrix cell measures the quantity of hybridized oligonucleotides in a probe.*

MM, or as the logarithm of the intensity ratio PM/MM. When averaged over the entire probe pair set, the first measure determines the *average difference*, whereas the second measure determines the *log-average*. The analytical software package provided by the platform manufacturer includes a decision procedure to assess whether a given probe set has been hybridized at all. The decision procedure, based on Wilcoxons Signed Rank test, assesses whether the hybridization of a gene probe set has occurred (P for present), has not occurred (A for absent), or has been only marginal (M). These three labels are called *absolute calls* and it is suggested to consider only P-labeled genes in the analysis of the experimental data [11]. An alternative method [61], measures the expression level of each gene as the intensity difference of each probe pair, rather than their average. In an effort to increase the reliability and reproducibility of measurements obtained by oligonucleotide microarrays, researchers have also suggested to use only a selected subset of all probe pairs set [82], but no sound principle has been introduced yet to guide this selection. Still, the vast majority of published research and high-level statistical analysis relies on the measures provided by the manufacturer's software.

The rationale behind the use of paired PM and MM probes is that the specific hybridization, represented by PM probes, should be always stronger than the non-specific hybridization, represented by 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 [64, 49]. However, a large proportion — often as high as 25% — of expression levels measured in GeneChip® microarrays are negative numbers so that, for these genes, the average hybridization of the MM probes is larger than the average hybridization of the PM probes, thus raising the question of whether there are design errors in the probe choice. This kind of design errors, even on a massive scale, are not unusual: almost 60% of one of the three microarrays containing the entire murine genome was found to contain to be non-specific [4]. The new statistical software provided by Affymetrix[2] replaces negative expression values by imputed numbers, so that the expression measures are non negative. However, treating measurement errors as missing values and replacing them with imputed data may hide important information about the data variability and the procedure is still under scrutiny of the scientific community.

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. We will term the set of expression levels measured for a gene across different conditions its *expression profile*, whereas we will use the term *genomic landscape* of a sample to denote the expression level of the genes measured in that sample in a particular condition. From the experimental design perspective, the main difference between cDNA and oligonucleotide microarrays is that one cDNA microarray is sufficient to compare the expression level of genes in two different experimental conditions, while oligonucleotide platforms require one microarray per condition. Both types of microarrays are nowadays used to answer the same broad classes of questions.

[2] `http://www.affymetrix.com/products/statistical_algorithms_reference_guide.html`

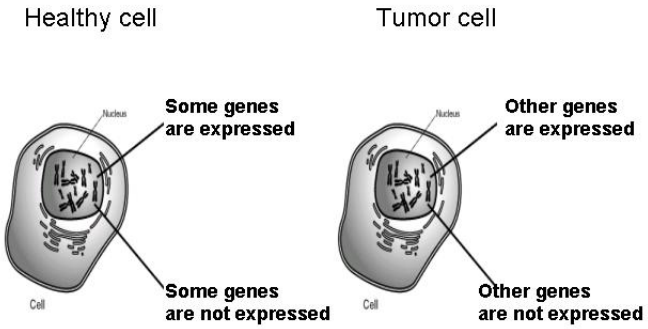


Figure 7: *Microarray technology enable investigators to identify genomic differences between two samples.*

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 genomic landscape 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 7, 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 [56], 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 genomic landscaping of several conditions at the time to characterize, for example, the genomic fingerprint of different types of cancer, [1], or the effect of changing several experimental factors simultaneously [19]. In both cases, each sample consists of the gene expression levels measured in cell lines grown or observed in a particular condition, and different samples can be assumed to be stochastically independent. An different class of experimental questions involve the study of the temporal evolution of gene expression profiles, so that different samples may be stochastically dependent. Studies to this class try to understand, for instance, the process that turns a locally growing tumor into a metastatic killer [20], the yeast sporulation cycle [85], or the response of human fibroblasts to serum [45]. Although the dependency structure among samples requires a different analysis, the common feature of these experiments is to compare the genomic landscape of cells in different conditions.

More advanced experiments try to gain understanding into the biological mechanism associ-

ated with the genomic differences of cell types observed in different experimental conditions, when the experimenters wish to discover the dependency structure between genes within the same experimental condition, or the dependencies among genes between different experimental conditions. Examples of these studies are the discovery of the genes that can be used to aid a more precise diagnosis of particular types of cancer [55], or the discovery of temporal dependencies among gene expression in some biological processes [73].

4.2 Experimental Design

The design of microarray experiments is a critical, albeit still neglected, issue of modern functional genomics. Besides technical issues of probe/microarray choice and design, the most fundamental design issue is the choice of the number of replications required to stake a statistically sound claim. Although microarray technology has rendered gene expression measurement blazing 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 [27]. Arguments have been made to show that a single replication of a comparative experiment is not sufficient to achieve reproducible results [59] but, despite the increasing awareness that data generated by even the most accurate microarray are very noisy, many discoveries reported in main stream journals are often based on experiments with three replications [97].

The main difficulty of this experimental design aspect is caused by the parallel nature of experiments conducted with microarrays: the best 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. An orthogonal problem is the selection of the time points to sample in order to study the temporal evolution of a biological system. 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.

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 [56], 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. The potential effect of the exogenous sources of variability should be accounted for during the experimental design, by taking into consideration variance components.

A further experimental design issue arises from the common problem of mRNA paucity. It is often the case that a single cell line is unable to produce enough mRNA in the desired condition. In this situation, common practice is to *pool* together the mRNA extracted from different animals. While obvious reasons of variability control suggest to use the same pooled sample for each experimental condition, the determination of the number of units to pool together is still an open issue.

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 [70, 53, 51, 19], 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 [24]. 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 testing hypotheses, the experimental design questions are far from being answered.

5. Data Preprocessing

A common strategy to reduce data variability and data dimensionality is to perform two preprocessing operations before undertaking any analysis of the data: *normalization* and *filtering*. The goal of the first operation is to remove systematic distortions across microarrays in order to render comparable 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 decrease the dimensionality of the data by removing genes that are not sufficiently differentiated.

5.1 Normalization of Microarray Data

One well known problem of cDNA technology is the consistent unbalance of the fluorescent intensities of the two dyes Cye3 (green) and Cye5 (red), as Cy3 is systematically less intense than Cy5 [74, 98], and normalization techniques were originally introduced to render the gene expression levels measured by the two different dyes comparable [29]. Although synthetic oligonucleotide microarrays do not suffer from a known systematic distortion similar to the dye fluorescence unbalance of cDNA microarrays, comparative experiments conducted on this platform require to hybridize two targets with two separate microarrays, and a variety of causes, including variations of the amount of mRNA in the two targets or the quantity of dye used to label the two targets, may introduce errors. Normalization techniques are therefore used to "remove" these experimental errors.

For each gene g, we denote by (y_{g1}, y_{g2}) the pair of expression levels measured in the two conditions. 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 $(\sum_g y_{g1})/(\sum_g y_{g2})$, [74]. The rationale behind this choice is that the total quantity of mRNA 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, the technique

is also called total mean normalization. Variants of this procedure estimate β by the ratio of the median, or by the ratio of the trimmed means.

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 conditions. 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 have been proposed to handle possibly nonlinear transformations [7, 44, 98].

One problem of normalization with calibration applied to cDNA data is that, when $\alpha > 0$, small values of the systematically larger intensity are replaced by negative numbers. As the normalized expression levels are often compared by computing their ratio in log-scale, negative numbers are then disregarded. To avoid this bias, other normalization techniques try to calibrate the ratios y_{g2}/y_{g1} [17] or the log-ratios $\log(y_{g2}/y_{g1})$, [98].

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, whereas local normalization uses only a subset of the genes on the microarray, which are either 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 [38, 21, 7] use the subset of genes detected as hybridized by the Affymetrix software.

Extending normalization techniques to repeated experiments is not straightforward. Yang *et al* [98] give 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 [38]. To avoid the lack of symmetry of this procedure, the baseline is often computed as the average expression profile [90]. 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 with a replication of the same target, changes in the amount of cellular mRNA 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 measurements observed for gene g in the replicate k of the experimental condition i are

$$y_{gik} = \tilde{y}_{gi} + \epsilon_{gik}$$

where ϵ_{gik} is the error in replicate k, and $\tilde{y}_{gi}$ is the correct expression level of gene g without the experimental error. The assumption that the experiment is reproducible would require that, on the average, the experimental errors compensate, so that each ϵ_{gik} is generated from a normal distribution with zero expected value, while the error variance can be modeled to account for the different sources of variability. An approach along this line is presented by [96, 48] for the analysis

of repeated cDNA-based expression levels. To further account for the variability across repeated samples of the same target, one may assume that

$$y_{gik} = \tilde{y}_{gik} + \epsilon_{gik}$$

where $\tilde{y}_{gik}$ is the realization of the gene g expression level in the k replicate of the experimental condition i. By random modeling the gene expression level, one can also take into account the variability between expression levels of the same gene in the replication of the same experimental condition. This approach is used in [80] to develop an integrated Bayesian differential analysis of gene expression data that overcomes the need for arbitrary normalization techniques.

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 [45, 85, 32], and normalization techniques are typically used to make the expression levels comparable. Although the correct solution would be to take few replicates of each measurement, cost constraints often make this solution impossible. More study is here needed to ascertain whether statistical modeling techniques can be used to account for measurement errors.

5.2 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 differs substantially according to the microarray platform and the statistical analysis to be performed.

For expression data measured with cDNA microarray, it is common practice to disregard those genes with negative or small expression level (after normalization). The software developed by Affymetrix deploys a decision procedure to assess the amount of hybridization of each gene, and it is suggested to discard all genes whose expression level is labeled as A (absent) or M (marginal). This procedure is justified by the intuition that expression level smaller than values ranging between 10 and 100 are actually measurement errors [11]. However, the genes discarded by this procedure would often amount to the vast majority, and investigators tend to adopt less stringent criteria to select a subset of the genes to be further analyzed. A common strategy retains only those genes with a change in the normalized expression level exceeding 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 [27] to analyze expression levels measured with cDNA microarrays, and an insightful analysis of the empirical success of this rule is described in [77]. Golub *et al.* [38] 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 by [30]. Other authors remove "spiked" genes, that is, those gene with one abnormally large or abnormally small measurement [88].

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 that 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 (possibly repeated) comparative experiments. The objective of these analysis is to identify the genes with 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 (*fold change*), whereas methods in the second category use the data to estimate the difference in expression of each gene in the two conditions.

6.1 Ratio-based Analysis

Early comparative experiments based on cDNA microarrays technology measured differences of gene expression across two conditions in terms of *fold-change*, computed as the ratio of the expression levels [78, 79, 27]. Particularly, only genes with a fold-change exceeding 2 were usually described as differentially expressed. The need to define a threshold to assess significant differentially expressed genes in two conditions is the motivation of a series of articles focused on statistical fold-analysis.

We let μ_{gi} denote the true expression level of gene g in condition i, so that $\rho_g = \mu_{g1}/\mu_{g2}$ denotes the unobservable "true" expression level ratio 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 [17], the authors use the naive ratio estimator $r_g = y_{g1}/y_{g2}$, where y_{gi} is the expression level measured in condition i, $i = 1, 2$. Assuming that the measurements from the two different channels (corresponding to the Cye3 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 $(1 - \alpha)\%$ confidence interval for the ratio ρ_g. The assumption of a constant coefficient of variation c in the two conditions let 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 by [69], this approach disregards ancillary information during the computation of the distribution of the ratio statistic. Despite the fact that expression levels should be positive numbers, the measurements of the two channels are supposed to follow normal distributions. The inappropriate distribution assumptions is corrected in [69], 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 correct. In fact, the same spot of cDNA in the microarray is simultaneously hybridized with the pool of mRNA 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 described by Lee *et al.* in [59] uses a mixture model to describe the joint distribution of the log-ratio of the measurements from the two channels as follows:

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

where 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, one can estimate the components of the mixture by using for example the EM algorithm [26]. The estimates can then be 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 of the paired measurements for each gene g by

$$\log(y_{g1k}/y_{g2k}) = \mu + \alpha_g + \beta_k + (\alpha\beta)gk + \epsilon_{gk} \quad g = 1, \ldots, G, \quad k = 1, \ldots, n \tag{1}$$

where G is the total number of genes in the microarray, and n is the total number of replicates of the experiment. The parameter α_g represents the "gene-effect", described as the specific ratio of expression level when either the gene is expressed or unexpressed in each replication of the experiment. The parameters β_k captures the "microarray-effect" and the interaction terms $(\alpha\beta)gk$ account for possible variation of each gene expression ratio in each replication of the experiment. The errors ϵ_{gk} are assumed to have zero mean. Despite acknowledging that all the effects in model (1) should be treated as random effects, the authors propose to estimate the parameters α_k using the standard two-way Anova estimator

$$\hat{\alpha}_g = \frac{1}{n}\sum_k \frac{y_{g1k}}{y_{g2k}} - \frac{1}{nG}\sum_{gk}\frac{y_{g1k}}{y_{g2k}} \quad g = 1, \ldots, G.$$

The estimates $\hat{\alpha}_g$ are then used as proxy of l_g to estimate the posterior probability that the genes are differentially expressed. Several authors have improved the whole procedure, by relaxing the parametric assumption on the mixture model [31, 71], or using a larger number of fixed effects [51], or random effects [96].

The scope of this stream of works is limited to gene expression data measured by cDNA microarrays, where the expression measurements across the two experimental conditions are paired by

design. When the expression data are measured with 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 [41, 97]. Typically, this threshold varies between 2 and 3 [37, 66, 46, 76], but it can be as low as 1.7 [58], and no published work addresses the problem of the extent of false positive and false negative rates produced by this "naive" fold analysis.

6.2 Differential Analysis

Suppose that an experiment, comparing two conditions 1 and 2, produces expression level data y_{gik_i}, $g = 1, \ldots, G$, $i = 1, 2$, and $k_i = 1, \ldots, n_i$. When the expression levels are measured with cDNA microarrays, the replications of each condition are equal $n_1 = n_2 = n$, while there is no need to impose this restriction for data measured with oligonucleotide microarrays. For each gene, we let μ_{gi} denote the true expression level in condition i. The hypothesis that the gene g is not differentially expressed in the two conditions is equivalent to setting $H_0 : \mu_{g1} = \mu_{g2}$, 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 $H_0 : \mu_{g1} = \mu_{g2}$, and to select the set of genes for which the null hypothesis is false. The standard statistic used for testing the null hypothesis is

$$t = \frac{|\bar{y}_{g1} - \bar{y}_{g2}|}{\sigma_g}$$

where $\bar{y}_{g1}$ and $\bar{y}_{g2}$ are the (normalized) average expression level of gene g in the two conditions, and σ_g is the standard error of the sample mean difference. When the two samples are independent — as for data collected with oligonucleotide microarrays — σ_g can be computed as

$$\sigma^2_{Ig} = \frac{\sum_{k_i}(y_{g1k_i} - \bar{y}_{g1})^2}{n_1(n_1-1)} + \frac{\sum_{k_i}(y_{g2k_i} - \bar{y}_{g2})^2}{n_2(n_2-1)} \equiv \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}, \tag{2}$$

and, when the two samples are not independent — as for data collected with cDNA microarrays — a more appropriate calculation of σ_g is

$$\sigma^2_{Dg} = \frac{\sum_k[(y_{g1k_i} - y_{g2k_i}) - (\bar{y}_{g1} - \bar{y}_{g2})]^2}{n(n-1)} \equiv \frac{S_1^2}{n} + \frac{S_2^2}{n} - 2\frac{S_{12}}{n}. \tag{3}$$

where the term $S_{12} = \sum_i(y_{g1k} - \bar{y}_{g1})(y_{g2k} - \bar{y}_{g1})/(n-1)$ is an estimate of the covariance of the two sample means. Notwithstanding some inconsistencies in the calculation of σ_g, the use of this t-statistic was first introduced in the differential analysis of oligonucleotide microarray experiments by [38] under the name of *signal-to-noise ratio*. Adopting the same t-statistic, Tusher *et al.* [90] use resampling techniques to identify a gene-specific threshold and to define an acceptance and rejection region for the null hypothesis. The procedure — called Significance Analysis of Microarrays SAM — computes a gene-specific threshold by taking into consideration the false positive rate, in this case the number of genes mistakenly detected as differentially expressed. The same t-statistic is applied in [98] in a similar manner to the differential analysis of gene expression data measured

with cDNA microarrays, although σ_g is not corrected to model the within pairs dependency. A Bayesian parametric version of the analysis based on the t-statistic is described in [5], whereas [30] propose an empirical Bayes procedure.

A model-based approach to estimate the difference of gene expression between two experimental conditions is presented by [88]. The idea is to model the expression level y_{gi} of gene g in the experiment i by

$$y_{gi} = \delta_i + \lambda_i(a_g + b_g x_i) + \epsilon_{gi}$$

where ϵ_{gi} is the experimental error, δ_i, λ_i are parameters that model chip-specific variability, x_i is a dummy variable that denotes the experimental condition, and a_g, b_g are gene specific parameters. Particularly, by setting $x_i = 0, 1$, b_g represents the difference of expression levels of gene g across the two conditions. The weighted-least squares estimates of the parameters δ_i, λ_i are used to adjust the expression level y_{gi} into $(y_{gi} - \hat{\delta}_i)/\hat{\lambda}_i$ that are themselves used to estimate b_g. Although the modeling approach is appealing, because the authors attempt to take into account different sources of variability that can affect the expression level of each gene, the final inference is based on large sample approximations. Given the fact that often comparative experiments are based on three or four replications, relying on large sample approximations may be a serious limitation for the applicability of this method.

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 [38] or non-Hodgkin's lymphoma [1] — or controlled experimental factors as sex and age [48]. The different experimental conditions can also be time points, when the experimenter wishes to analyze the evolution of a physiological response [45] or to identify genomic features of a cell cycle [73], or to track down the genetic mechanisms that switch a locally growing tumor into a metastatic killer [20]. 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 $y_g = (y_{g1}, \ldots, y_{gn})$ collects the expression level y_{gj} for gene g measured in the n samples, while each column $e_j = (y_{1j}, \ldots, y_{Gj})$ 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 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 the type of analysis described here, although its goal is mainly to "describe" the genomic differences of two conditions. In cancer genomic experiments, for example, the goal may be the development of new diagnostic tools, based on the genome landscape of tumor cells, to obtain reliable prognosis to inform the therapeutic strategy. 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 [38] or breast cancer [94] — and uses each patient sample as an instance of the genomic landscape of the specific type of tumor. The goal of the analysis would be to determine the *genomic fingerprint* of each type of tumor, to make it possible a genome-based diagnosis of a specific tumor [55].

Class Discovery. Multiple microarray experiments can also be used to help investigators create new classifications by discovering new classes characterized by a specific genomic fingerprint. 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 [1]. For example, in the analysis of gene expression data collected from tissues of breast cancer patients affected, the goal may be the identification of new molecular taxonomies of breast cancers characterized by particular genomic fingerprints. 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 genomic landscape 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 label new cases in a *test set*. Suppose the $G \times n$ data matrix Y contains the expression profiles of G genes measured in c different conditions, and that each condition i is measured n_i times for a total of $n = \sum_i n_i$. 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 genomic landscapes 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 genomic landscape. The classification can be based on a decision rule that lets a class be chosen by minimizing the expected loss. We call this approach model-based versus a model-free approach in which the space of gene expression data is partitioned in such a way 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 [67] and naive

Bayes classifiers [39]. 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

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

where F is the cumulative distribution function of the logistic distribution or of the standard normal distribution, y_{ghk} is the expression level of gene g in the replication k of condition h, 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 probabilities $p(i|y_1, \ldots, y_G)$ 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 sparsity of the microarray data, which often consists of thousands of genes and few observations for each gene. A Bayesian method for fitting probit regression and tacking the "small n large p" problem has been recently proposed by [94] 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

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

where $p(y_g|i)$ is the density function of the expression level of gene g in class i. 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 observation available for each gene [50].

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 observation for each pair (y_g, i) is at least 25 [67]. Similar feature selection methods are available for naive Bayes classifier [68]. 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 model, 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.

Examples of model-free approaches to classification are methods for discriminant analysis as Fisher linear discriminant analysis, nearest neighbor classification trees, [39], or support vector machines [92]. A comprehensive review of classical statistical methods for discriminant analysis applied to gene expression based tumor discrimination is presented in [28], with a critical assessment of pros and cons of each method. Support vector machines are another supervised classification technique. Support vector machines use a training set in which genes known to belong to the same

functional class are assigned the same class label, and genes known not to be members of that class are assigned the same different class label. The two-labeled data constitute the training set that is used to learn to distinguish between members and non-members of the functional class on the basis of their expression data. Having learned the expression features of the class, the support vector machine can be used to recognize and classify the genes in the data set on the basis of their expression [9].

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 to classification are currently more popular. 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 [86], or selecting genes that are highly correlated with a dummy pattern associated with each class [38].

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 genomic landscape, or both simultaneously. 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 Eisen plot.

The same approach can be applied to the columns of the data matrix to identify samples with a similar gene expression landscape. 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 "visually", by searching for large contiguous patches of color representing groups of genes sharing similar expression patterns or groups of samples sharing similar gene expression landscape. The identification of these patches allows the extraction of subgroups of genes to be used to re-cluster the samples, 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 genomic-based subclasses of non-Hodgkin lymphoma [1], cutaneous malignant melanoma [8], breast cancer [83], and lung cancer [7].

Notwithstanding these interesting results, this approach is not without problems. The subjective

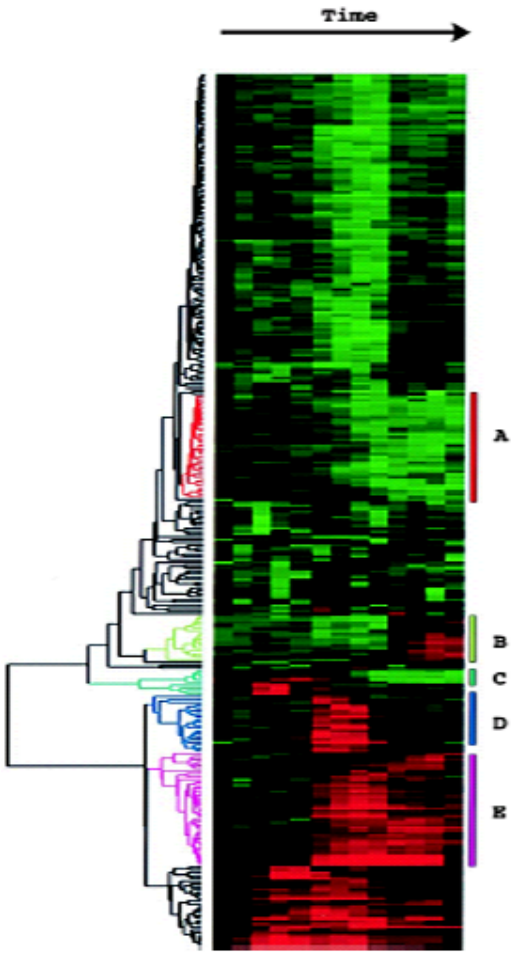


Figure 8: *Example of Eisen plot applied to 517 gene expression data measured in 13 experiments displaced along time. The image is a graphical display of the data array Y with rows sorted using the average-linkage hierarchical clustering procedure. Each row of the image represents a gene, and each column represents an experiment. Each cell (g, j) of the image represents, graphically, the log-fold ratio of gene g expression in experiment j and the same gene expression in experiment 1. Cells with log ratios of 0 are colored black, increasingly positive log ratios with reds of increasing intensity, and increasingly negative log ratios with greens of increasing intensity. A representation of the dendrogram is appended to the image. Contiguous patches of color, labeled 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].*

nature of partitioning by visual inspection may lead 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 [91, 36]. Permutation tests are sometimes used to validate the partitions found by this procedure [32], and a bootstrap-based validation technique is presented in [52]. The Gap statistics of Tibshirani *et al.* [89] 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 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 [74].

Relevance networks [12] are a non hierarchical clustering method which does not suffer of 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 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 an 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 [87] to identify groups of genes with similar patterns across different experimental conditions. Similar to k-means clustering are the self-organizing maps of Kohonen [54]. 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 [86], and they have been combined with nearest neighbor classification method to discriminate between two types of acute leukemia [38].

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" [40], "biclustering" [18], "coupled two way clustering" [35], or the "plaid model" [57]. 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 sam-

ples. Biclustering is a method for clustering simultaneously genes and samples by using a similarity measures 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, 16] 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 labeling the cells of the array Y. In this way, the problem of simultaneously clustering 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) ($r = 1, \ldots, R$, $c = 1, \ldots, C$) to similar cells of Y, then the likelihood function of the matrix Y, conditional on a known labeling of rows and column, can be represented as

$$p(Y|H) = \prod_{r=1}^{R} \prod_{c=1}^{C} \prod_{g(r)} \prod_{j(c)} p(y_{g(r)j(c)}|\theta_{r,c}\theta_h)$$

where $g(r)$ are the genes assigned the same label r, $j(c)$ are samples assigned the same label c, and $p(y_{g(r)j(c)}|\theta_{r,c}\theta_h)$ is the density function of gene-samples assigned the same label pair. When some knowledge about R and C is available, one can use the EM algorithm to estimate the unknown parameters for a specification of the density function $p(y_{g(r)j(c)}|\theta_{r,c}\theta_h)$ using a mixture model approach. Alternatively, if some initial labeling of the experiments is available, one can use some agglomerative clustering procedure to iteratively relabel rows and columns. Some relevant work in this area is presented in [99] for one dimensional clustering, and in [7]. Although model-based clustering relies on distributional assumptions on gene expression profiles and samples, the validity of these assumptions can be assessed using statistical validation techniques. One of the main advantage 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.

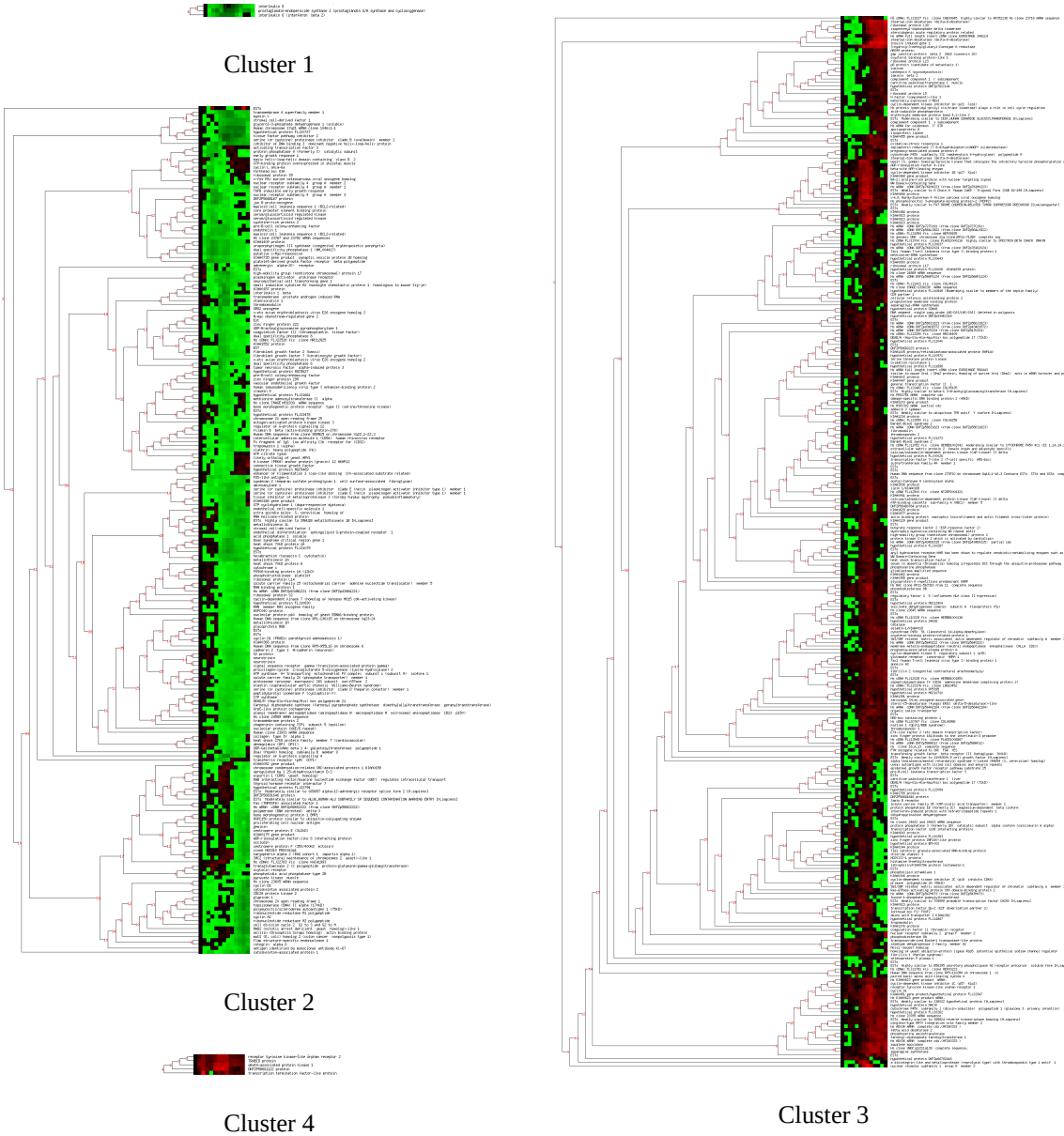
7.4 Time Series Analysis

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 should be acting together in the process under consideration because they belong to 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 realizations of 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 unraveling the functional relationships between genes [18, 19, 20], traditional clustering methods are used to group genes on the basis of their similarity. For example, Holter *et al.* [42] 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 timeless method, such as singular value decomposition [2], 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, which 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. CAGED (Cluster Analysis of Gene Expression Dynamics) [75] is a model-based approach to cluster temporal gene expression patterns able to account for 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. CAGED features 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, the method uses a distance-based heuristic search procedure able to render the search process feasible. In this way, CAGED retains the important visualization capability of hierarchical clustering and 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. Ramoni *et al.* [75] use CAGED to cluster a set of 517 gene expression patterns observed during the temporal deployment of the transcriptional program underlying the response of human fibroblasts to serum [45]. By using simple first order autoregressive equations, the algorithm groups the gene temporal patterns into four clusters, while Iyer *et al.* identify, by visual inspection of the data array produced by average-linkage hierarchical clustering, eight subgroups of genes. Interestingly, these eight subgroups are merged into two of the clusters found by CAGED, thus supporting the claim that the human eye tend to overfit. Figure 9 shows the graphical display of the four clusters found with CAGED on the same data used in Figure 8. When the autoregressive order is equal to zero, this method subsumes, as a special case, model-based clustering of atemporal (i.e. independent) observations.



Cluster 1

Cluster 2

Cluster 4

Cluster 3

Figure 9: *The four clusters of temporal gene expression data found with* CAGED *on the data of Figure 8.* CAGED *splits the 517 genes into four clusters, and each cluster merges, in a sorted manner, genes that are generated by the same stochastic process. Note that the images display contiguous patches of color that are eventually merged into four distinct clusters. As in the Eisen's plot in Figure 8, a representation of the dendrogram is appended to each cluster, and attached to each node branch is the Bayes factor of the model in which the subtrees are merged versus the model in which the subtrees are left disjoint.*

8. Open Challenges

Microarrays 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 microarray, 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 both methods and technology, several challenges are still open and are particulary relevant to the statistical community.

Experimental design The design of a microarray experiment is an unprecedent challenge. The main character of microarray technology is to make it 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 testing hypotheses, the experimental design questions are far from being answered.

Quality assessment and normalization A very important issue when analyzing gene expression data is to be able 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 so large to mask irremediably the signal of 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. Although the experimenter could be less preoccupied with this issue by repeating each experiment a certain number of times, costs and time constraints do not usually make taking many replicates a realistic options.

Differential analysis The last two years have witnessed an increasing number of research articles proposing methods for the differential analysis of gene expression data. Particular attention has been given to the differential analysis of gene expression data measured with cDNA microarrays, but very little work has addressed the issue of comparing gene expression data collected with oligonucleotide microarrays. The naive fold analysis described in Section 6 does not appear to give reproducible results, but it is very much the only simple option available to biologists. Analytical methods are needed for the differential analysis of gene expression data that address the specific aims of microarray experiments of not providing scientific discoveries but formulating scientific hypotheses.

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. Very popular machine learning tools such as Bayesian networks [23] 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 [100, 81, 34].

Validation Validation of cluster analysis is a very important issue that deserves further attention. As often clusters of similar genes/experiments are detected by visual inspection, or by imposing arbitrary thresholds, it is necessary to independently validate the results to make sure that clusters are indeed capturing the signal in the data. Permutation tests, or bootstrapping the results, 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, these tests do not prove that the groups found in the data are meaningful. Some independent, biological validation of subgroups of genes/experiments found by clustering is carried out in [1, 38], although on such a small number of cases (for example 40 patients in [1]) the validation does not seem to provide much support. Some authors show the validity of their results by using different clustering techniques [8, 7]. The development of sound validation tests ranks among the top priorities in the field.

Eric Lander [56] 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] 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.

[2] 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.

[3] 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.

[4] A. Ananthaswamy. Chip chop. *New Scientist*, March 14th, 2001. Available at `http://www.newscientist.com/news/news.jsp?id=ns9999512`.

[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] 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.

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

[11] A. Butte, A. Kho, and I. S. Kohane. *Microarrays, Informatics and Functional Genomics*. MIT Press, Cambridge, MA, 2002.

[12] 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.

[13] D. Casey. *Primer on Molecular Genetics*. U.S. Department of Energy, Human Genome Management Information System, Oak Ridge National Laboratory, 1st edition, 1992. Available at `http://www.ornl.gov/hgmis/publicat/primer/intro.html`.

[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] B. Chabot. Directing alternative splicing: Cast and scenarios. *Trends Genet*, 12(11):472–478, 1996.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

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

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

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

[26] 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.

[27] 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.

[28] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, Department of Statistics University of California, Berkeley, Berkeley, CA, 2000.

[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] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. Banavar. Dynamic modeling of gene expression data. *PNAS*, 98:1693–1698, 2001.

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

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

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

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

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

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

[57] 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.

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

[59] 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.

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

[61] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, 98:31–36, 2001.

[62] 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.

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

[64] 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.

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

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

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

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

[69] 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.

[70] 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.

[71] 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.

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

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

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

[75] M. F. Ramoni, P. Sebastiani, and I. S. Kohane. Cluster analysis of gene expression dynamics. Technical report, Childrens Hospital Informatics Program, Harvard Medical School, Boston, MA, 2001.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] P. Sebastiani, M. Ramoni, and I. Kohane. Bayesian differential analysis of gene expression data. Technical report, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, 01003, 2001.

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

[82] D. Selinger, K. Cheung, R. Mei, E. M. Johanson, C. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature Biotechnology*, 18:1262–1267, 2000.

[83] 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.

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

[85] 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.

[86] 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.

[87] 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.

[88] 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 Res*, 11:1227–1236, 2001.

[89] 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.

[90] 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.

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

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

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

[94] 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.

[95] 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`.

[96] 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*, 2001. In press.

[97] 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.

[98] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. *Statistica Sinica*, 2001. To appear.

[99] 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.

[100] 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`.