Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Dottorato di Ricerca in Fisica

Settore scientifico disciplinare: Fisica applicata (FIS/07)

XX CICLO

NEW APPROACHES TO OPEN PROBLEMS IN GENE EXPRESSION MICROARRAY DATA

Tesi di Dottorato
di
Daniela Marconi

Coordinatore:
Prof. Fabio Ortolani

Supervisore:
Prof. Renato Campanini

Correlatore:
Dott. Valter Gattei

Bologna, Marzo 2008

"Any good poet, in our age at least, must begin with the scientific view of the world; and any scientist worth listening to must be something of a poet, must possess the ability to communicate to the rest of us his sense of love and wonder at what his work discovers."

(E. Abbey, The Journey Home)

ABSTRACT

In the past decade, the advent of efficient genome sequencing tools and high-throughput experimental biotechnology has lead to enormous progress in the life science. Among the most important innovations is the microarray technology. It allows to quantify the expression for thousands of genes simultaneously by measuring the hybridization from a tissue of interest to probes on a small glass or plastic slide. The characteristics of these data include a fair amount of random noise, a predictor dimension in the thousand, and a sample noise in the dozens.

One of the most exciting areas to which microarray technology has been applied is the challenge of deciphering complex disease such as cancer. In these studies, samples are taken from two or more groups of individuals with heterogeneous phenotypes, pathologies, or clinical outcomes. these samples are hybridized to microarrays in an effort to find a small number of genes which are strongly correlated with the group of individuals. Even though today methods to analyze the data are well developed and close to reach a standard organization (through the effort of International project like Microarray Gene Expression Data -MGED- Society [5]) it is not infrequent to stumble in a clinician's question that do not have a compelling statistical method that could permit to answer it.The contribution of this dissertation in deciphering disease regards the development of new approaches aiming at handle open problems posed by clinicians in handle specific experimental designs.

In Chapter 1 starting from a biological necessary introduction, we revise the microarray technologies and all the important steps that involve an experiment from the production of the array, to the quality controls ending with preprocessing steps that will be used into the data analysis in the rest of the dissertation. While in Chapter 2 a critical review of standard analysis methods are provided stressing most of problems that

In Chapter 3 is introduced a method to address the issue of unbalanced design of miacroarray experiments. In microarray experiments, experimental design is a crucial starting-point for obtaining reasonable results. In a two-class problem, an equal or similar number of samples it should be collected between the two classes. However in some cases, e.g. rare pathologies, the approach to be taken is less evident. We propose to address this issue by applying a modified version of SAM [17]. MultiSAM consists in

a reiterated application of a SAM analysis, comparing the less populated class (LPC) with 1,000 random samplings of the same size from the more populated class (MPC) A list of the differentially expressed genes is generated for each SAM application. After 1,000 reiterations, each single probe given a "score" ranging from 0 to 1,000 based on its recurrence in the 1,000 lists as differentially expressed. The performance of MultiSAM was compared to the performance of SAM and LIMMA [73] over two simulated data sets via beta and exponential distribution. The results of all three algorithms over low-noise data sets seems acceptable However, on a real unbalanced two-channel data set regarding Chronic Lymphocytic Leukemia, LIMMA finds no significant probe, SAM finds 23 significantly changed probes but cannot separate the two classes, while MultiSAM finds 122 probes with score >300 and separates the data into two clusters by hierarchical clustering. We also report extra-assay validation in terms of differentially expressed genes Although standard algorithms perform well over low-noise simulated data sets, MultiSAM seems to be the only one able to reveal subtle differences in gene expression profiles on real unbalanced data.

In Chapter 4 a method to address similarities evaluation in a three-class problem by means of Relevance Vector Machine [82] is described. in fact, looking at microarray data in a prognostic and diagnostic clinical framework, not only differences could have a crucial role. In some cases similarities can give useful and, sometimes even more, important information. The goal, given three classes, could be to establish, with a certain level of confidence, if the third one is similar to the first or the second one. In this work we show that Relevance Vector Machine (RVM) [2] could be a possible solutions to the limitation of standard supervised classification. In fact, RVM offers many advantages compared, for example, with his well-known precursor (Support Vector Machine - SVM [3]). Among these advantages, the estimate of posterior probability of class membership represents a key feature to address the similarity issue. This is a highly important, but often overlooked, option of any practical pattern recognition system. We focused on Tumor-Grade-three-class problem, so we have 67 samples of grade I (G1), 54 samples of grade 3 (G3) and 100 samples of grade 2 (G2). The goal is to find a model able to separate G1 from G3, then evaluate the third class G2 as test-set to obtain the probability for samples of G2 to be member of class G1 or class G3. The analysis showed that breast cancer samples of grade II have a molecular profile more similar to breast cancer samples of grade I. Looking at the literature this result have been guessed, but no measure of significance was provided before.

Acknowledgments

First, I'd like to acknowledge Renato Campanini, whom has served has the adviser for the research described in this dissertation and the Lab-mates of MIG group.

Thanks to Valter Gattei, head of Clinical and Experimental Hematology Research Unit in Aviano, has been a pleasure to work with. The long brainstormings in these years enhanced my wonder about the secrets of biology.

I would like to thank my parents for supporting my education throughout my life and above all in the last weeks, to give me a warm haven over the hill where I polished the draft. Also a big thank you to my sister who always roots for me, even if sometimes she think that doing a PhD it's a way to avoid to "act my age". Thanks to my little sweet niece Caterina because, having her smile on my desktop make any reboot more appealing.

A special thanks goes to all my friends Roberta, Letizia, Simona, Serena, Francesca,Vittorio, Antonio and many others... despite my withdraw of these last months they never let me short of love and support.

Finally, I must express my most heartfelt gratitude to Giulio who, besides to be the best "peer reviewer" of my life, is above all my most fervent well-wisher.

Contents

Chapter 1

Microarray: Surveying Genetic Information

1.1 Post-genomic Era

The 14th of April of 2003 [65]the International Sequencing Consortium announced that The Human Genome Project (HGP) had been completed, 99% of the human genetic code was sequenced. We are now moving from the *pre-genomic era* characterized by the effort to sequence the human genome, to a *post-genomic era* that concentrates on harvesting the fruits hidden in the genomic text.

An overarching challenge in this post genomic era is the management and analysis of enormous quantities of sequence data. In this context the Human Genome Project is best understood as the 20th century's version of the discovery and consolidation of the periodic table[54]. The Human Genome Project aims to produce biology's periodic table; not a rectangle reflecting electron valences, but a tree structure depicting ancestral and functional affinities among the human genes. The biological periodic table will make it possible to define unique "signature" for each building block. Molecular biology has tended, in the last decades, to examine genes individually. The reasons could be found in the limits imposed by the technologies like Northern Blot or Southern Blot, prior to the advent of high-throughput technologies like microarray [69, 60]. Recently the advent of microarray technology has made it possible to monitor the expression levels of thousands of genes in parallel. Arrays offer the first promising tool for addressing the challenges of the post-genomic era, by providing a systematic way to survey variation in DNA and RNA. This technology have been widespread applied during the last several years, and it seems likely to become a standard tool both of research in molecular biology and of clinical diagnostic. It is now time to gain a global prospective on the cell by asking

genome-wide questions.

1.2 Biological Basis of Genetic Information

1.2.1 Building Blocks of Life

Cells are considered to be the most basic units of life (both for prokariotes or eukaryotes). This is the fundamental finding which was stated in the first half of the 19th century by Matthias Schleiden and can be regarded as the cornerstone of the cell theory. The theory was later enhanced by the discovery that all necessary genetic information for a whole organism is present in the DNA of each individual cell in the organism.

DNA is most commonly recognized as two paired chains of chemical basis, spiraled into what is commonly known as the double helix. DNA is a large polymer with a linear backbone of alternating sugar and phosphate residues. The sugar molecule contains five carbon atoms, labeled 1' through 5', The backbone is created by a series of bonds between the 3' carbon of one unit's sugar molecule, the phosphate residue, and 5' carbon of the next unit. DNA strands have an orientation determined by the numbering of the carbon atoms, which by convention starts at the 5' end and finishes at the 3' end. A single-stranded DNA sequence is therefore always written in this canonical 5'→ 3' direction , unless otherwise stated.

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), the two main information carrying molecules in the cells, have a similar structure as both are built up from nucleotide monomers. A nucleotide consists of a phosphate group, a pentose (five-carbon sugar) and an organic base. In DNA, the pentose is deoxyribose and the organic bases adenine (A), guanine (G), cytosine (C) and thymine (T). In RNA, the pentose is ribose and thymine is replaced by uracil (U). Nucleotides can bind to each other to form pairs: adenine pairs with thymine or uracil, cytosine pairs with guanine (see Figure 1.1 on page 3).

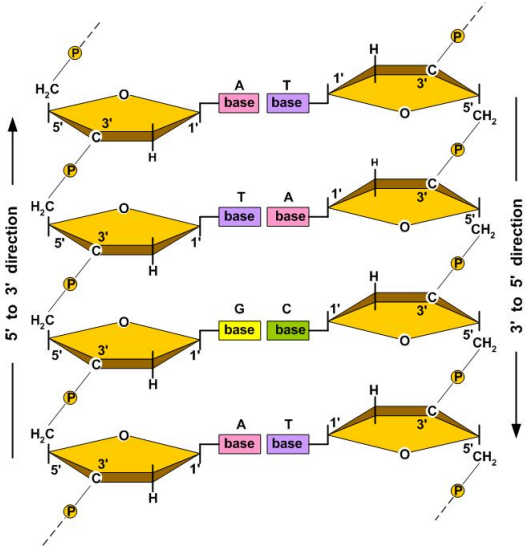

Figure 1.1: The backbone and double-helix structure of DNA

This relation is called base-pair complementary and is the basic process behind th functionality of the cell and the subsequent construction of proteins. All living organism are composed largely of proteins, and their importance was well stated by the distinguished scientist Russel Doolittle, who wrote "we are our proteins".

1.2.2 The Central Dogma of Molecular Biology

A major finding of molecular biology is that DNA specifies RNA and RNA determines proteins. Specifically, the flow of information from DNA to RNA to proteins is called the central dogma of molecular biology [27]. DNA is the primary source of information for the building and functioning of a cell. This information is encoded in nucleotide triplets called codons. DNA is transcribed into RNA, which is then translated into proteins. The overall process of transcription and subsequent translation is referred to as gene expression. Although the information flows strictly from nucleic acids to proteins, the relation between DNA, RNA and protein is circular as proteins carry out, or at least support, the synthesis of DNA and RNA (seeFigure 1.2 on page 4).

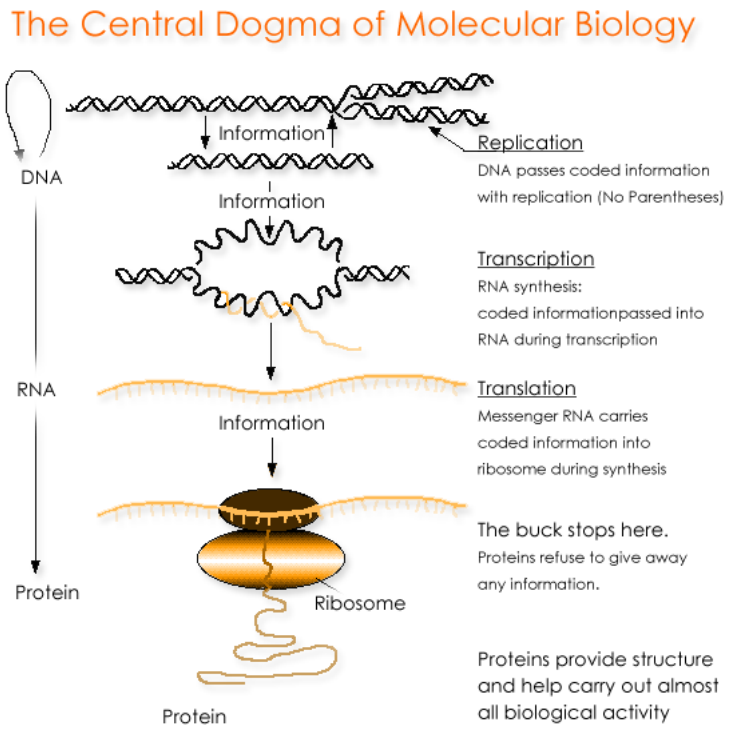

Figure 1.2: The Central Dogma of Molecular Biology

1.2.3 Genes and Genetics Code

Each cell of an organism has one or more DNA molecules. Each DNA molecule forms a chromosome. The complete set of chromosomes inside a cell is called a genome. The number of chromosomes in a genome is characteristic of a particular species. For example, every cell in Homo sapiens has 46 chromosomes.

A DNA molecule contains certain contiguous stretches which encode information for building proteins. However, some portions of DNA molecule do not contain encoded information but rather are termed "junk DNA". This may actually be a misnomer, as it has been suggested that junk DNA may indeed perform unrecognized and valuable functions. A gene is a contiguous stretch of DNA that contains the information necessary to build a protein or an RNA molecule. Gene lengths vary, but human genes normally have 10,000 base pair. The starting and the ending points of genes can be recognizes by specific cell mechanism. A protein is composed of a chain of amino acids. The mechanism by which genes specify the sequence of amino acids in a protein is called the *genetic code*. To be specific, a triplet of nucleotides is used to specify each amino acid. Such a triplet is called a *codon*. Given the four bases types, the total number of possible combinations within nucleotide triplets is 64. However, these 64 combinations can only refer to the

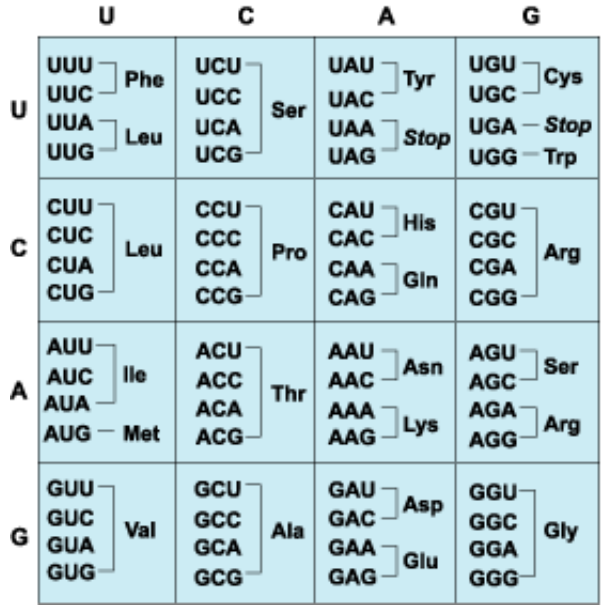

Figure 1.3: Genetic code mapping codons to amino acids

twenty amino acids which actually occur (see Figure 1.3 on page 5).

There is therefore redundancy in coding, and several different triplets will correspond to the same amino acid. Moreover, three of the possible codons (UGA, UAG and UAA) do not code for any amino acid and are used instead to signal the end of a gene. Such redundancy is actually a valuable feature of the genetic code, rendering it more robust in the event pf small errors in the transcription process.

1.2.4 Transcription and Gene Expression

Transcription is the process of synthesizing RNA using genes as a templates. A gene is *expressed* when, through the transcription process, its coding is transferred to an RNA molecule. To initiate a transcription process, the DNA double helix is "unzipped", starting at the promoter site of a gene. The *promoter site* is a region on the 5' side of the DNA strand which indicates that a gene is forthcoming. Once the DNA double helix has been opened at this starting point, one DNA strands serves as a template strand. An RNA molecule is constituted by binding together ribonucleotides complementary to the template strand until the STOP codon is met. This resulting RNA is called *messenger RNA*, or, briefly, mRNA. After the transcription process, the mRNA will be transported to cellular structures called ribosomes to guide the manufacture of proteins. For eukaryotes, many genes are composed of alternating parts called introns and exons. After transcription, the introns are spliced out from the mRNA. This means that only the exons will partecipate in protein synthesis. Alternative splicing occurs when the

same genomic DNA can give rise to two or more different mRNA molecules on the basis of alternative selection of introns and exon, generally resulting in the production of different proteins. Because of the changes which result through the splicing of introns and exons, the entire gene as found in the chromosome is usually called the genomic DNA, and the spliced sequence consisting of exons only is called the *complementary DNA* or *cDNA*. The cDNA can be obtained by a *reverse transcription process* which transforms mRNA back into DNA and which is one of the basics biological step in the flowchart of a microarray experiment. Scientist involved in gene expression research usually find it easier to work with expressed sequence tags (ESTs) instead of the entire gene. An EST is a unique short subsequnce (only few hundred base pairs in length), generated from the DNA sequence of a gene. Scientific com unity also identify with ESTs sequences of the genome which are expressed but not well annotated . Is important to stress that not all the *transcribed genetic* information will be really *translated* into proteins, for this reason *genomics* and *proteomics* are nowadays two different branches of post-genomic era. Research and technologies have to concentrate both on proteomic and genomic level to gain the correct view of what is going on into the cell.

1.2.5 Genotype and Phenotype

Genomes belonging to the same species vary slightly from organism to organism in a phenomenon known as *genome variation* (or *genetic variation*). It is subtle variability in genomes that is responsible for the evolution and diversity of organisms. Some genomes variations are unique to an organism, while others are passed on through generations via reproductive cells. Most genomes variations involve only a few bases. Some common variations include the replacement of one base by another (*substitution*), the excision of a base (*deletion*), the addition of a base (*insertion*), and the removal of a small subsequene of bases and their reinsertion in the opposite order (inversion) or in another location (translocation). Such genome variations are due to mutations and polymorphisms. A polymorphism is a genome variation in which every possible sequence is present in at least one percent of a population, whereas a mutation refers to a genome variations that is present in less than one percent of a population.

In each organism, DNA resides in chromosomes. A cell may contain a single set of chromosomes (the haploid state) or two chromosome sets (the diploid state); in the latter case, each chromosome is represented by two copies. An exception is the pair of sex chromosomes, which draws one copy from the father and one from the mother. The two members of a pair of chromosomes are called homologous chromosomes. The existence of genome variation means that some genes may differ slightly from individual to individual. When this happens, each alternate version of a gene is called an allell. In fact, every diploid cell carries two alleles of each gene, one in each of a pair of homologous chromosomes. When both alleles are the same, the organism is said to be homozygous for that gene. In the latter case, only one of the alleles, the dominant allele, may be

expressed, while the other is not expressed.

The common and properly-functioning version of a gene is referred to as wild-type allele. As with the redundancy present in genetic coding, the presence of two versions of each gene is another protective mechanism provided by nature. If one copy should happen to be defective, the other copy is available to compensate. Due to this protective mechanism, many genome variations do not produce any noticeable. However, the effects of a small percentage of genome variations are noticeable, with both beneficial and deleterious results. Most of current research focus on elucidating genotype-phenotype relationships, as the relationships between disease and their genomic basis are termed. Genotype refers to the genetic, makeup of an individual, while the outward characteristics of the individual are its phenotype. They are, naturally, connected as the phenotype is shaped, develops, and functions on the information provided by and encoded in the genotype.

1.3 Microarrays

1.3.1 Introduction

The recent development of high-density DNA microarray technology enables researchers to capture the snapshot of cells on a genome-wide scale at the transcriptional level. The expression levels of thousands, or even tens of thousands, of genes can be monitored using a single microarray chip. Microarrays measure the presence of mRNA. The mRNA can be extracted from cells, tissues, etc. By analyzing extracted mRNA, one obtains a quantitative assessment of the genetic activity of the location from which the mRNA was extracted. Microarrays derive an expression level for each gene, a scalar value corresponding to the amount of mRNA which in turn corresponds to the gene in question. The molecular principle of microarray is the inherent ability of nucleic acids to bind to complementary sequences (hybridization). This allows simultaneous probing of a complex mixture of nucleic acids using an array of complementary sequences which are spatially ordered on a glass surface. The biological sample to be analyzed is deposited on the array where sequences of the sample hybridize to arrayed sequences. In figure 1.4 a simple scheme of hybridization process is showed.

In the context of microarray technology, the nucleotide sequences attached to the array surface are frequently called *probes*[1], while sequences of the sample are termed *targets*.

McLachlan briefly reviewed the history of the microarray technology. In the 1980, a group led by R.P. ekins in the Department of Molecular Endocrinology at the University College, London was the first to use simple microspotting techniques to manufacture

[1]In this dissertation the nomenclature proposed by Duggan et al. [35]will be used, and the term *probes* refers to the DNA on the array while target refers to the labeled DNA in solution.

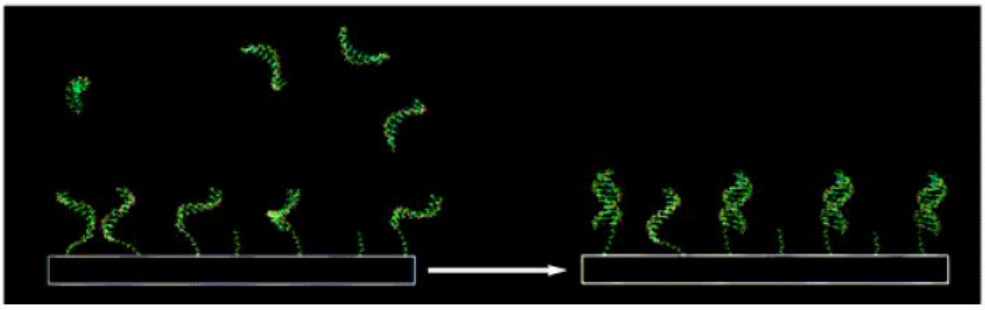

Figure 1.4: Target samples hybridizes to complementary probe samples.

arrays for high-sensitivity immunoassay studies. Numerous groups of researchers have furthered the technology introduced by Ekins and his colleagues. In the United States, notable research has been accomplished by Stephen P.A. Fodor and his colleagues at Affymetrix, Inc. (Santa Clara, California)[60], as well as groups at Stanford University, particularly Patrick O. Brow, in the Department of Biochemestry and Biophysics[69]. Brown and his colleagues at Stanford are credited with engineering the first DNA microarray chip, while Fodor and colleagues at Affymetrix, Inc., created the first patented DNA microarray wafer chip, the GeneChip. Numerous commercial entities and academic groups have since contributed to advancements in DNA microarray technology, and PubMed has registered an incredible growth in number of publications registered since the 1995 (see 1.5) for the topic "microarray" and also for "microarray" related to "cancer" investigation.

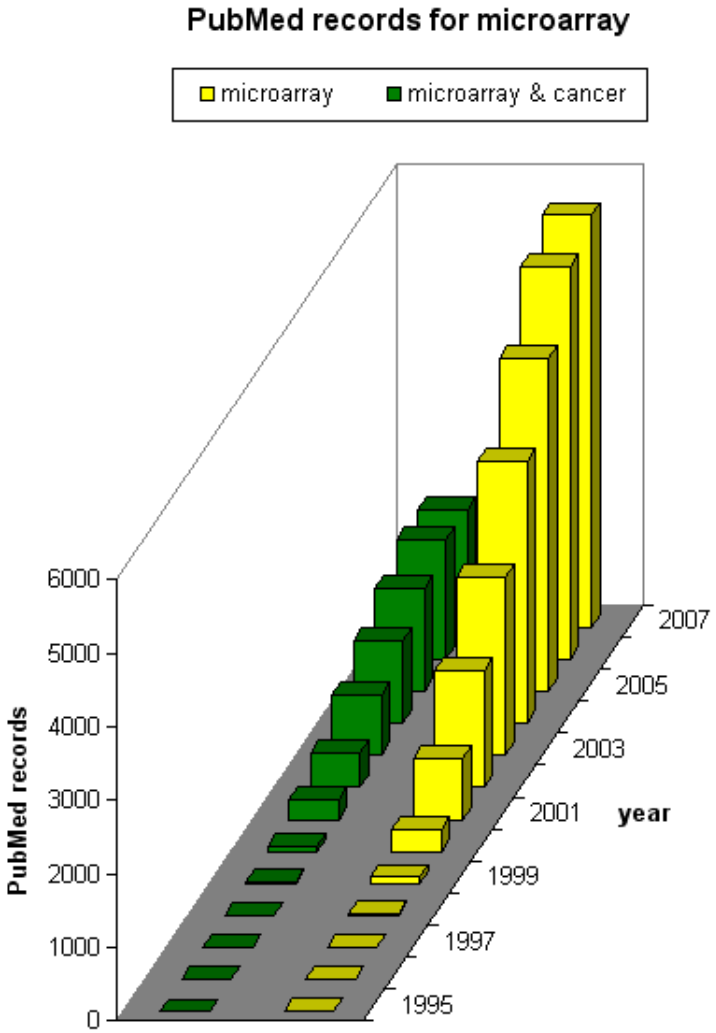

Figure 1.5: PubMed records for microarray

In the most general form a DNA array is a chip made of nylon membrane, glass or plastic. Usually the chip is arranged in a regular grid-like pattern and segments of DNA strands are either deposited or synthesized within individual grids. Once the array is prepared, a microarray experiment involves some basics steps like sample preparation and labeling, sample hybridization and washing, and microarray image scanning and processing.

A growing appreciation of the potential value of microarray results beyond the summarized description found in most papers (including supplementary data) has led to the creation of public repositories for microarray data—for example, ArrayExpress [63] (`http://www.ebi.ac.uk/arrayexpress`) of the European Bioinformatics Institute (EBI)

and the Gene Expression Omnibus (GEO;http://ncbi.nlm.nih.gov/geo)of the National Center for Biotechnology Information (NCBI) at the National Institutes of Health. Depositing data in these repositories has become a condition for publication in several journals and is likely to become one for publication in most[80]. These databases rapresent for biostatisticians, biophysicists and bioinformaticians a unique resources by which they can pose many different questions to the transcriptome without further money or experimental investment. Even in this dissertation repository-data set from GEO will be used to evaluate the feasibility of the methods.

In this context, as systems are needed for the management and storage of microarray data, standards and the use of ontologies[2] are crucial to managing and sharing these data. The brainchild of Alvis Brazma and Alan Robinson of the EBI, the Microarray Gene Expression Data (MGED) Society was initially formed as an international grassroots organization to develop standards for databases. There are several complementary projects in the MGED:

- The MIAME (Minimum Information About a Microarray Experiment) project aims to define the information that should always be included in databases and also provides guidelines to the authors and reviewers of manuscripts that describe microarray experiments. Philosophically, MIAME defines the information that is required to permit another researcher to understand the experiment and the data; practically speaking, it is a checklist of what should be supplied for publication. A complication is that each study has different types of associated information that are relevant, and judgments must be made about what is relevant. As of September 2007, two commercial and two academic databases are listed at the MGED website as being compliant with MIAME guidelines, although there are certainly others.

- The MAGE (MicroArray and Gene Expression) project aims to provide a standard for the representation of microarray expression data that would facilitate the exchange of microarray information between different data systems. The original MGED project provided a standard XML format, called the MicroArray Mark-up Language format (MAGE-ML), for reporting microarray data and its associated information.

- The MGED-OWG (Ontology Working Group) aims to provide standard terms for the annotation of microarray experiments. These terms will enable structure

[2]"An ontology is an explicit specification of some topic. For our purposes, it is a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other. Ontologies therefore provide a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary."(From Stanford Knowledge Systems Lab)

queries of elements of the experiments. Furthermore, the terms will also enable unambiguous descriptions of how the experiment was performed. The terms will be provided in the form of an ontology which means that the terms will be organized into classes with properties and will be defined. A standard ontology format will be used. For descriptions of biological material (biomaterial) and certain treatments used in the experiment, terms may come from external resources that are specified in the Ontology. Software programs utilizing the Ontology are expected to generate forms for annotation, populate databases directly, or generate files in the established MAGE-ML format. Thus, the Ontology will be used directly by investigators annotating their microarray experiments as well as by software and database developers and therefore will be developed with these very practical applications in mind.

All these projects are contributing to fix standards for microarray databases.

1.3.2 Microarray Chip Manufacture

There are two main approaches to manufacture of microarray chips: deposition of DNA fragments by robotic spotting and *in situ* synthesis[55]. Manufacture by robotic deposition may proceed through the deposition of PCR-amplified cDNA clones or the printing of already-synthesized oligonucleotides. In situ fabrication can be divided into photolithography, ink jet printing, and electrochemical synthesis[88].

In this dissertation the Operon deposition-based platform will be used in chapter 3 while two different platforms based on in situ synthesis will be used in chapter 4 (Affymetrix HGU133 and Whole Genome Agilent 4 x 44K).

1.3.2.1 Deposition-Based Manufacture

The manufacture of deposition-based arrays involves the consideration of three issues: the selection of DNA probes, preparation of the probes and the printing process. To probes that are to be printed on the array are chosen directly from databases including GeneBank, dbEST, and UniGene. Additionally, full length cDNAs, collections of partially sequenced cDNAs (or ESTs) or randomly chosen cDNAs from any library of interest can be used.

In the process of deposition-based manufacture, the DNA probes are prepared away from the chip. Process can be either be polymerase chain reaction (PCR) products or oligonucleotides. The PCR technique was developed in 1983 through the work of Kary B. Mullis. This techniques creates billions of copies of specific fragments of DNA from a single DNA molecule. After the amplification, the PCR products are partially purified by precipitation and/or gel-filtration to remove unwanted salts, detergents, PCR primers and proteins present in the PCR cocktail. Alternatively, DNA probes can be prepared by pre-synthesizing DNA oligonucleotides for use on the array.

Once the DNA probes are determined and prepared, a typical printing process follows five steps:

1. Robots dip thin pins into the wells of solutions to collect the first batch of DNA
2. The pins touch the surface of the arrays to spot the DNA is spotted onto a number of different arrays, depending on the number of arrays to be made and the amount of liquids the pins can hold.
3. The pins are washed to remove any residual solution and ensure no contamination of the next sample.
4. The pins are dipped into the next set of wells.
5. Return to step 2 and repeat until the array is complete [31].

The deposition-based manufacture are based exclusively on *dual labeling*[3] scheme.

1.3.2.2 In situ Synthesis

Arrays synthesized in situ are fundamentally different from spotted arrays in the following aspects[88]:

- **Selection of probes**. Probe selection is performed based on sequence information alone. Therefore, every probe synthesized on the array is known. In contrast, with cDNA arrays, which deal with expressed sequence tags, the function of the corresponding sequence is often unknown. Additionally, since this selection methods avoids duplicating sequences among gene family members, this approach can distinguish and quantitatively monitor closely-related genes.
- **Preparation of probes**. The probes are printed (Agilent Sureprint technology) or photochemically synthesized (light direct synthesis for Affymetrix) base-by-base on the surface of the array. In the preparation of the probes there is no cloning and no PCR process involved.

As both the Agilent and the Affymetrix data will be used in this dissertation, below some more details will be given in order to have in mind their specificity in subsequent analysis.

[3]Platform dependent labeling process is necessary to allow the detection of which probes are bound to the microarray. In the most common dual labeling experiments two samples are hybridized to arrays,each labeled with Cy3 or Cy5 dyes, which are excited by green and red lase, respectively. For single labeling platform, biotin-labeled cDNA is constructed, with a carefully defined protocols to ensure reproducibility.

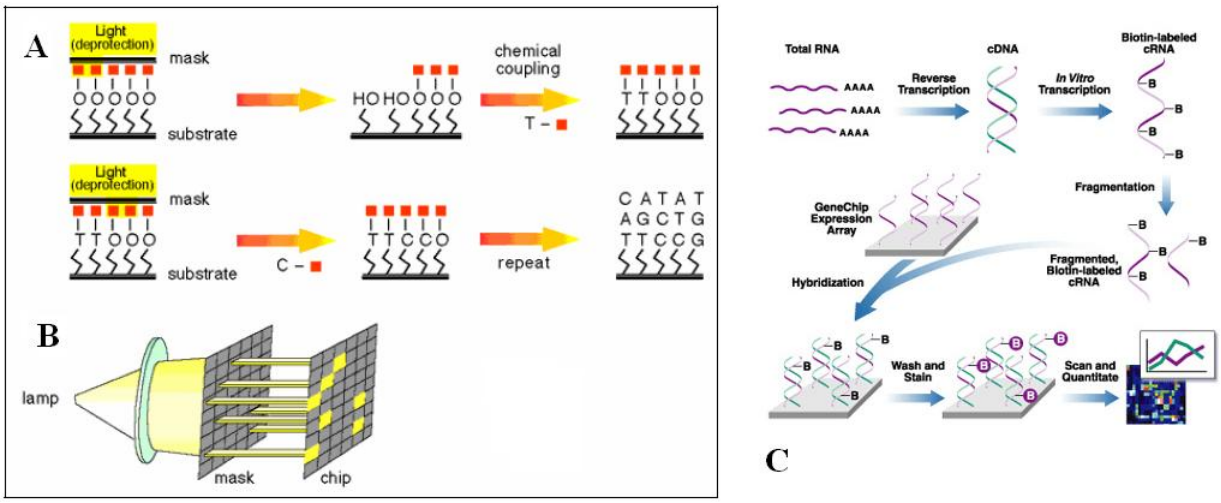

Figure 1.6: A)Light directed oligonucleotide synthesis. A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group. Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites. The process is repeated, activating different sets of sites and coupling different bases allowing arbitrary DNA probes to be constructed at each site; B)Schematic representation of the lamp, mask and array; C)Steps of wet-lab GeneChip experiment.

1.3.2.3 The Affymetrix GeneChip platform

To control the oligonucleotide synthesis, Affymetrix uses photolithographic masking (see figure 1.6). They attach synthetic linkers modified with photochemically removable protecting groups to a glass substrate and direct light through a photolithographic mask to specific areas on the surface to produce localized photodeprotection. The first of a series of chemical building blocks, hydroxyl-protected deoxynucleosides, is incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. Next, light is directed to different regions of the substrate by a new mask, and the chemical cycle is repeated. Highly efficient strategies can be used to synthesize arbitrary polynucleotide at specified locations on the array in a minimum number of chemical steps. Thus, given a reference sequence, a DNA probe array can be designed that consists of a highly dense collection of complementary probes with virtually no constraints on design parameters. The amount of nucleic acid information encoded on the array in the form of different probes is limited only by the physical size of the array and the achievable lithographic resolution. Compared to cDNAs, the gene expression is quantified by non-competitive hybridization, meaning that only one biological sample (the sample of interest) is fluorescently labeled and hybridized to the microarray. The expression of each gene is measured by comparing the hybridization to a set of 20 probe pairs (*probeset*), each of which is 25 base pairs long. The first type of probe in each pair is the perfect match (PM) which exactly corresponds to the gene sequence, whereas the second is the mismatch (MM), created by changing the middle (the 13^{th}) base of the original sequence (see). The idea of this construction is to provide a control mechanism for random variation and cross-hybridization. Of note, the use of the MM probes in the summarization of the values for each gene is nowadays controversial [47, 12] and many non commercial algorithms tends today to not use the MM values in the summarization process.

1.3.2.4 The Agilent oligonucleotide platform

There are seven main steps in the creation of Agilent microarrays. First, glass wafers are coated with a surface that will make a strong bond with both the glass and the nucleic acids that are to be printed. Then, the reagents for oligo synthesis are inspected for quality and purity, while the cDNA for deposition microarrays is prepared and rigorously qualified for printing. Next, the prepared glass and the nucleic acids come together through the careful orchestration of inkjet printing, which involves multiple real-time quality control feedback mechanisms to monitor the presence, size, shape and position of every feature. Following printing, the microarrays undergo a process to permanently bind the printed DNA to the surface of the microarray, and deactivate the surface around the features. This minimizes the surface's ability to bind non-discriminantly with sample that could lead to high background signal. The process of printing oligonucleotide

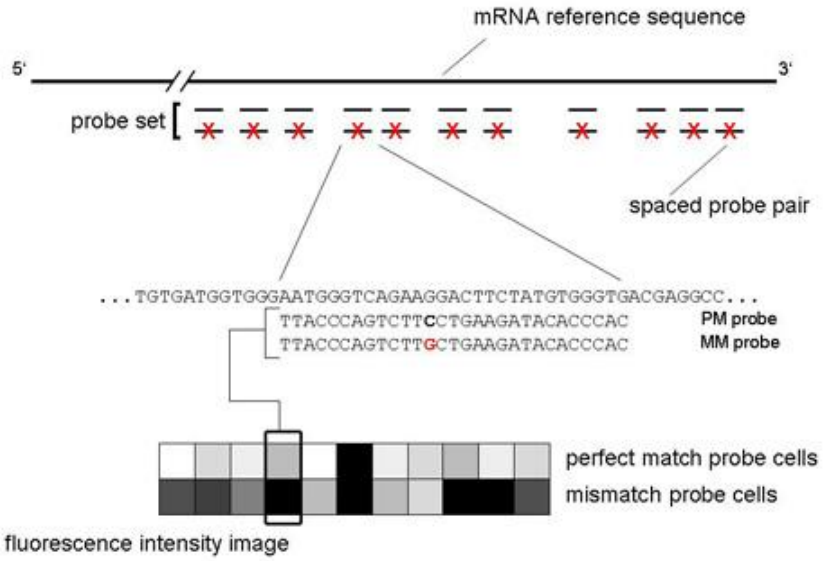

Figure 1.7: Probe pairs set. Gene sequences of MM and PM are shown for a specific probe pair.

microarrays is nearly identical to the process described for cDNA microarrays, however, instead of printing the oligos fully prepared onto the microarray surface, they are actually synthesized base-by-base in repetitive print layers using standard phosphoramidite chemistry (see figure). After the inkjet head and reservoirs are washed and thoroughly dried, they are connected to bottles containing the four different phosphoramidite nucleotides that make up the building blocks of in situ nucleic acid synthesis. This ensures a constant supply of reagents flowing to the inkjet head during printing.The oligo print run commences with the firing of a test pattern to select the best nozzles for printing. Then the iterative oligonucleotide synthesis loop begins when the first nucleotide of each oligo is printed onto the activated glass surface of the microarrays. In phosphoramidite synthesis reactions, the reactive sites on the nucleotides are blocked with chemical groups that can be removed selectively. This allows the bases to be added to the oligo chain one base at a time in a very controlled manner. After the first base is printed, the trityl group that protects the 5' hydroxyl group on the nucleotide is removed and oxidized to activate it, enabling it to react with the 3' group on the next nucleotide. In between each step, the excess reagents are washed away so that they won't randomly react later in the synthesis. An advantage of Agilent platform is that can use both dual or single labeling scheme.

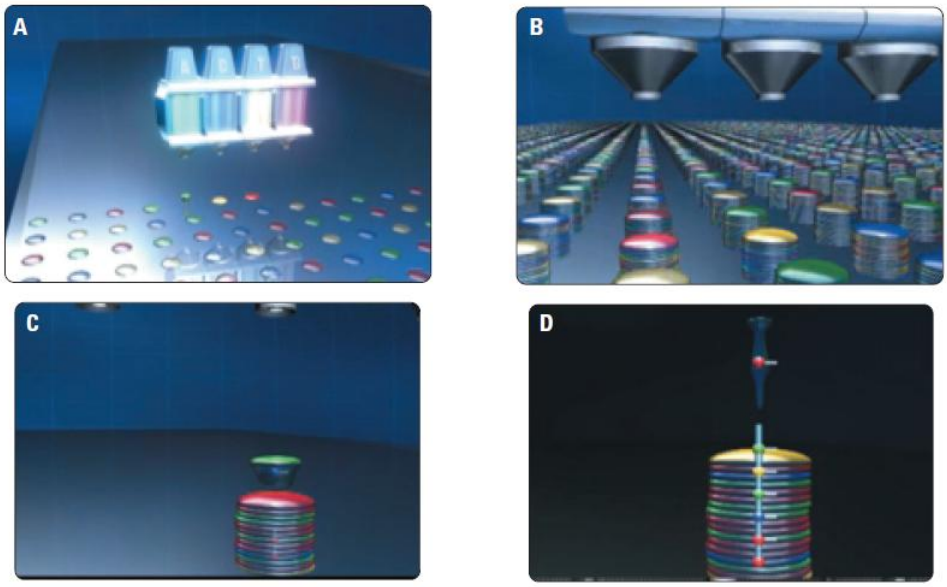

Figure 1.8: These four images communicate the general mechanism for oligo synthesis via inkjet printing. A shows the first layer of nucleotides being deposited on the activated microarray surface. B shows the growth of the oligos after multiple layers of nucleotides have been precisely printed. C is a close-up of one oligo as a new base is being added to the chain, which is shown in figure D.

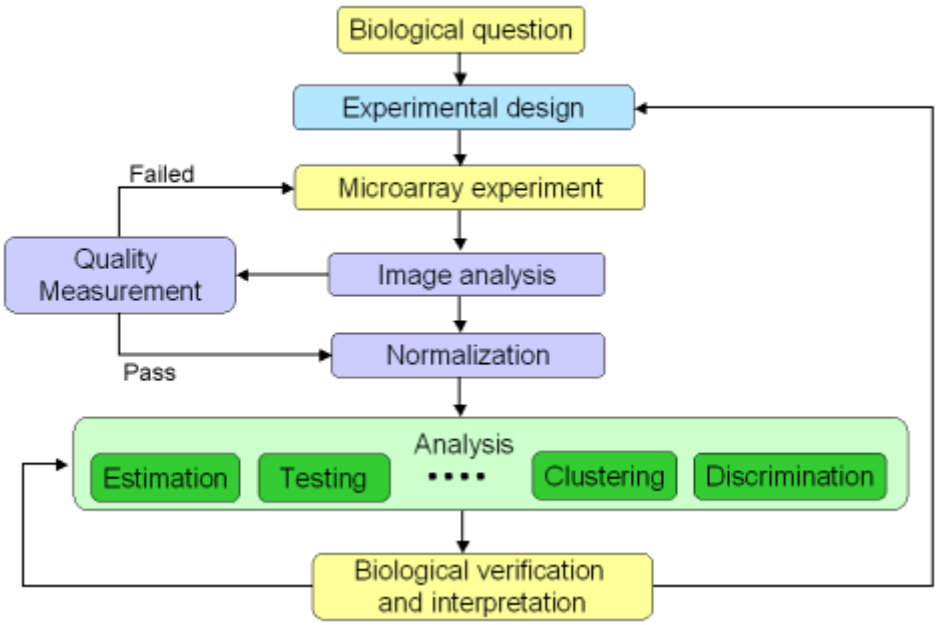

Figure 1.9: Experimental steps in microarray experiments

1.3.3 About Microarray Experiments

Steps of a typical microarray experiment are described here with a particular focus on the critical issues that regards specifically some of the work curated during the analyses done for this thesis project. Figure 1.9 shortly summarize the general flowchart of a microarray experiment.

For what concerns exclusively the wet-lab part of the experiment is instead summarized here: firstly, RNA is extracted from a tissue sample. The RNA is reverse transcribed to cDNA that is labeled by incorporating modified nucleotides. Dyes are either already linked to these nucleotides (direct labelling) or coupled to them by a further chemical reaction (indirect labelling). The labelled cDNA is hybridized to the microarray probes. Each cDNA binds only to specific oligonucleotides on the array because of the base-pair complementary. While Affymetrix arrays are hybridized with a single sample, cDNA arrays demand the co-hybridization of two samples, one of which is a reference sample. To prevent non-specific hybridization, blocking agents such as Cot-1 DNA or salmon sperm might be added. Subsequently, the dye is excited by a laser, so that the amount of annealed cDNA can be quantified by measuring the fluorescence intensities.

After the wet lab the experiments have necessary to take an interdisciplinary course that involves a continuous exchange of ideas and competences between biologists (or clinicians) and data analysts.

1.3.3.1 Experimental Design

In every experiment involving data collection, in order for the experiment to provide the data necessary for the analysis, the experiment needs to be designed. The design of the experiment is a crucial but often neglected phase in microarray experiments. If the experiments are not designed properly, no analysis method will be able to obtain valid conclusions. During the design phase, the first problem to face with is to find an equilibrium between the expensive cost of each array and the necessity to have enough replicates to control the noise at different levels.

In a strict linguistic sense, to replicate means duplicate, repeat, or perform the same task more than once. Replication allows the experimenter to obtain an estimate of the experimental error. This estimate of error can become the basis for drawing conclusions whether the observed differences in data are significant. Replication is a wide misunderstood term in the microarray field. Often, the misunderstanding is related to the definition of the task to be performed. Thus, if the purpose is to understand and control the noise introduced by the location of the spot on the slide, one can replicate spots by printing exactly the same conditions. Finally, if the purpose is to control the biological variability, different mRNA samples can be collected from similar specimens and the microrarray should be used in exactly the same conditions from all other points of view. The common misunderstanding is related to the fact that often researchers refer to replicates without specifying which one factor was varied while keeping everything else constant.

In the context of experimental design is important to make some distinctions and punctualization:For probe and spot replicates, an empirical approach would just choose a reasonable number knowing that any number of replicates is better than not having replicates at all. If this approach is chosen, a good minimum for the number of spot/probe replicates is 3. Is important to stress, however, that most of the not custom platforms do not allow the researcher to choose the number of probe replicates. In particular, Affymetrix HGU133 and Agilent 4x44K have decided to introduce more than one replicate only for probes that are "well known", i.e. to say those probes that are known to play a central role in biological process or molecular function. This inevitably introduce a *literature bias*, that can not be avoided in standard experimental designs.

Once the design of the array is fixed, the distinction between an array that is a technical replicate or a biological replicate is often a confounding question. To make technical replicates means to extract the mRNA from a single target and then use it into different arrays, while to have biological ones means two extract the mRNA from similar targets and then use an array for each target mRNA. The technical replicate aims to control the noise generated from experimental random and systematic errors. In most of the experimental designs technical replicates are often neglected in order to collect more biological replicates.

How many microarrays is enough? Statisticians do not like simple answers to this

question; it depends on the goals of the study, the resources, and the reliability of the technology – specifically how accurate the chips are, and how often a hybridization fails. However the following guidelines apply to most situations. If an exploratory study aims to find large (more than two-fold) differences between two conditions, then a design with three samples per condition is usually adequate. If the aim is to find smaller differences, or almost all of the large differences, then five samples per group are necessary to obtain sufficiently reliable enough estimates of variation among samples within conditions, in order to distinguish true differences between conditions. This applies to both treatment and control conditions. Six samples per condition allows meaningful permutation tests, which can give more accurate, and less conservative, estimates of p-values and false discovery rates. If there are more than two conditions, and the treatments do not drastically alter the cell physiology, then the number of samples within any one condition can be somewhat less; with four or more conditions, one can obtain reasonable estimates of within-condition variation with only four samples per condition. All of these suggestions assume that there are no outlying samples, which should be discarded; it is wise to do one or two more per condition in clinical situations, where outliers occur commonly, and it is safer to do one more for animal experiments, where sometimes one animal in a condition appears very different than all the others. The question of how many replicates to do depends on how small the differences are that you want to detect, and the noise level in your system. Different systems have different noise levels, and the only way to estimate the noise is to do three or four replicate hybridizations. For Affymetrix systems with the best analysis at NCBI find that 3 to 5 chips per group gives useful information. Usually many more cDNA chips are needed for comparable levels of accuracy. To estimate replicability of a two-color chip, hybridize three pairs of replicate dye-swaps (6 chips) using the same two (different) RNA samples. To do meaningful clustering requires at least 20 samples, and generally many more.

The most common design for two color (competitively hybridized spotted) arrays is the 'reference design': each experimental sample is hybridized against a common reference sample. Although this effectively means that only one sample of interest is hybridized per chip, the reference design has several practical advantages over more efficient designs:

- it extends easily to other experiments, if the common reference is preserved;
- is robust to multiple chip failures; and
- reduces incidence of laboratory mistakes, because each sample is handled the same way.

The reference sample is used in many chips, therefore the reference mRNA needs to be abundant. When comparing treatment versus control samples the most natural reference is the wild type or the biological controls, which are often the most abundant. However if

the study aims to compare each of several samples against all others, there is no natural control. A reliable alternative is a common reference obtained by pooling all samples. This enables samples to be compared with each other indirectly. A pooled reference sample reduces the number of extreme gene ratios (which have large errors) on each chip. Some labs take this further and create a "universal reference": a pool of mRNA derived from several standard cell lines, which they use most often in their experiments. Using a universal reference enables them to compare results for all their experiments.

One complication in two-color arrays is that the two dyes don't get taken up equally well, so that the amount of label per amount of RNA differs (dye bias). An early proposal to compensate for dye bias was to make duplicate hybridizations with the same samples using the opposite labeling scheme.The intent was to compensate dye bias by averaging ratios from dye-swapped hybridizations. However dye bias is not consistent, and in practice the ratios in dye-swap experiments do not precisely compensate each other. Normalization methods such as lowess give more consistent results, although dye-swapping makes it easier to compensate for dye-bias. However the dye-swap is the basis for most other efficient designs: the general principles of a good two-color design are that it should be balanced (i.e. every sample appears equally often in red and green) and the samples whose ratios are most interesting should appear on the same chips most often. As explained later in this chapter some of the normalization methods are able to efficiently compensate for the dye-bias.

The last question about experimental design, that will be extensively addressed in chapter 3, regards the balancing of data collection between two different conditions. In fact, to have an equal number of samples for both the conditions would be crucial in order to respect some of the assumption made by standard microarray data analysis methods. However, above all in disease related investigation, this is not always possible for clinical or biological specific reasons.

1.3.3.2 Image Scanning and Processing

After the completion of hybridization, the surface of the hybridized array is scanned to produce a microarray image. As previously mentioned, samples are labeled with biotin fluorescent dyes that emit detectable light when stimulated by laser. The emitted light is captured by the photo-multiplier tube (PMT) in a scanner, and the intensity is recorded. Most scanners contains one or more lasers that are focused onto the array (for two-channel microarrays, the scanner uses at least two lasers).

One-channel microarray, such as an oligonucleotide array, yields one image per array, whereas a two-channel microarray yields two images per array, one image per channel. The scanner reads a microarray by dividing it up into a very large number of pixels and recording the intensity level of fluorescence at each pixel. The resulting rectangular array of pixels and their associated intensities constitutes the image of the microarray. The image must be converted into spot intensities for analysis. The purpose of this

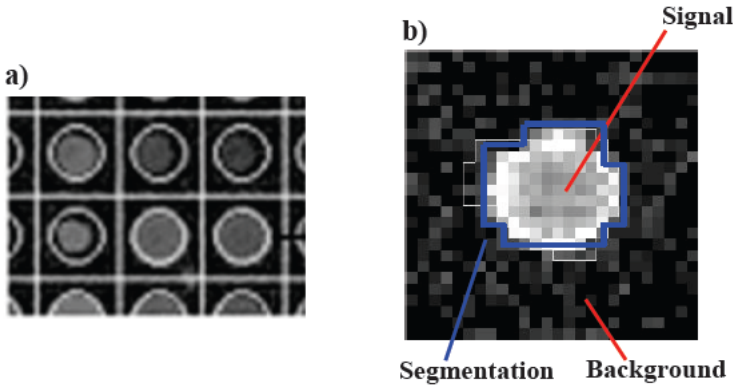

Figure 1.10: a) Gridding of microarray image; b) spot's foreground and background intensities.

conversion is to assign to every DNA sequence that was spotted on the microarray an intensity measure, called the spot intensity, reflecting the amount of labeled sample that hybridized to it. The task of quantifying a scanned image is often carried out in three steps. First, the location of each spot in the array is defined by assigning coordinates to the center of each spot: this is called gridding. Second, the signal (the set of pixels that correspond to labeled cDNA hybridizing to its complementary sequence spotted on the microarray), is separated from the background (the set of pixels that correspond to labeled cDNA hybridizing non-specifically to the microarray): this is called segmentation. Finally, each spot is assigned two intensity values (see 1.10):

- Foreground Intensity: denotes the average intensity of the pixels designated as signal. The average used is often the median.
- Background Intensity: this is the average intensity of the pixels around the spot that were designated as background. Again the average used is the median.

In principle, the intensities of those pixels not corresponding to spots should be zero. However, this never happens. Instead, because of various reasons such as non-specific binding of the labeled sample to the array substrate and substrate fluorescence, these pixels emit a low, but not insubstantial, level of fluorescence that may vary with location. The concern is that the spot intensities may also contain a certain amount of this non-specific fluorescence, called the background fluorescence. It is customary therefore to estimate a background intensity from data, eventually deciding to subtract the background from the foreground intensity. However actually there is much evidence that in presence of good quality arrays is not useful to subtract the background.

1.3.3.3 Noise

Due to their nature, microarrays tend to be very noisy. Even if an experiment is performed twice with exactly the same materials and preparations in exactly the same conditions, it is likely that after the scanning and image processing steps, many genes will probably be characterized by different quantification values. In reality noise is introduced at each step of various procedures:mRNA preparation (tissue, kits and procedures vary), transcription (inherent variation in the reaction, enzymes), labelling (type and age of label), amplification, pin type, surface chemistry, humidity, target volume, slide inhomogeneities, target fixation, hybridization parameters (time, temperature, buffering, etc), unspecific hybridization (labelled cDNA hybridized on areas which do not contain perfectly complementary sequences), non-specific background hybridization, artifacts (dust), scanning (gain settings, dynamic range limitations), etc.

The challenge appears when comparing different tissues or different experiments. Is the variation of a particular gene due to the noise or is it a genuine difference between the condition tested? Furthermore, when looking at a specific gene, how much of the measured variance is due to the gene regulation and how much to noise? The noise is an inescapable phenomenon and the only weapon that the researcher seems to have against it is replication.

1.3.3.4 Microarray Data Cleaning and Preprocessing

1.3.3.4.1 Data Transformation It is common practice to transform DNA microarray data from the normalized row intensities into log intensities, before proceeding with *high-level analysis*. There are several objectives of this transformation:

- There should be a reasonable even spread of features across the intensity range.
- Variability should be constant at all intensity levels.
- The distribution of experimental errors should be approximately zero.
- The distribution of intensities should be approximately bell-shaped.

Figure 1.11 shows the histogram of intensities of a typical microarray of a typical microarray data set before and after log transformation.

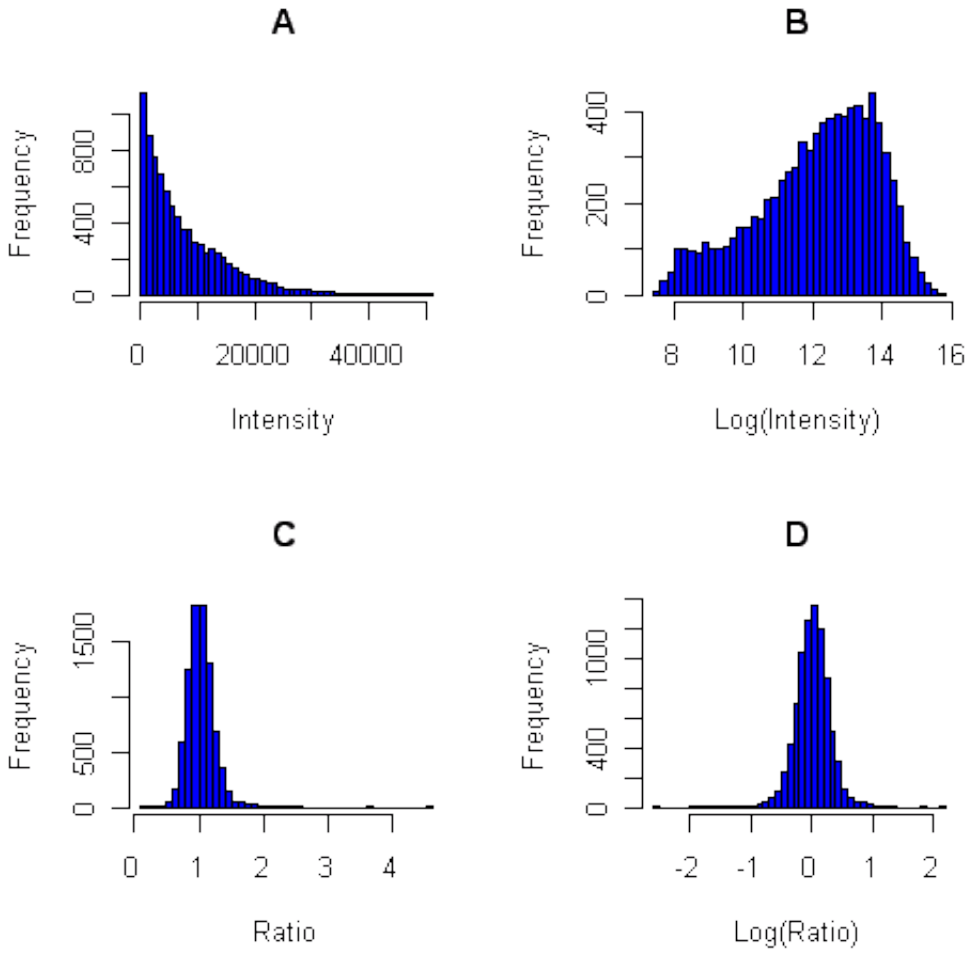

Figure 1.11: Histogram of the intensities (A) before and (B) after the log transformation of an example data set. Histogram of the ratios of intensity (C) before and (D) after the log transformation of an example data set.

We can see that the raw data is very heavily clumped together at low intensities and sparsely distributed at high levels. By contrast, the data is more evenly spread over the intensity range after the log transformation. The transformation greatly reduces the skewness of the distribution and simplifies visual examination. Microarray data analysis typically uses logarithms to base 2. In processing, the ratio of the raw Cy5 and Cy3 intensities is transformed into the difference between the logs of intensities of the Cy5 and Cy3 channels. Therefore, 2-fold up-regulated genes correspond to a log ratio of +1, and 2-fold down-regulated genes correspond to a log ratio of -1. Genes that are not differentially expressed have a log ratio of 0. These log ratios have a

natural symmetry which reflect the biological structure and is not present in the row fold differences. Figure 1.11 illustrates histograms of ratios of the intensity data set before and after log transformation.

1.3.3.5 Missing Value Estimation

DNA microarray experiments can generate data sets with multiple flagged spots. Flagged values occur for diverse reasons, including insufficient resolution, image corruption, or slide contamination by dust or scratches. This flag is often treated as missing values. Missing data may also occur systematically as a result of the robotic methods employed in generating the microarrays. Unfortunately, many algorithms for gene expression analysis require a complete data set as input. Therefore, methods for estimating missing data are sometimes used before these algorithms can be applied.

Suppose a microarray data set is represented by a matrix where each rows corresponds to one gene and each column rapresents an experimental condition. A simple approach is to replace a missing entry with the average expression over the rows (Row Average Method). This method is not optimal since it does not take into account the correlation structure of the entire data set. Troyanskaya et al. propose two more complex algorithms based on K-nearest neighbors (KNNimpute) and singular value decomposition (SVDimpute) that are the most used in most of the microarray literature. In this dissertation not so much details will be given about imputation methods because the choice (analyzing the target data sets) is to not use imputation at all. In fact, accordingly with Troyanskaya , it is a good rule of thumb exercising caution when drawing critical biological conclusion from data that is partially imputed. So the choice will be to simply flag the data and then using some cautions about biological conclusions if the differentially expressed profile contains flagged genes.

1.3.3.6 Background Subtraction

The scanning of arrays results in optical or background noise affecting pixel intensities. Images obtained from spotted arrays contain specific information on this background noise from the pixels not associated with spotted regions. High density-oligonucleotide have minimal space between the segments of the array where probes are attached, known as cells; therefore, background information iis difficult to obtain and not commonly used.

Tipically, image processing software will produce an absolute expression measure X and a background measurement B for each spot or cell. If, as is likely, X is the result of signal and additional background noise, then it is a biased estimate of the true hybridization that we intend to measure (that is, it is likely to be systematically too high). To obtain an unbiased measure of expression, conventional wisdom is to subtract the background considering X-B. If both X and B are unbiased and background adds to the signal, then X-B is unbiased. Even in these circumstances, however, there are important

trade-offs to be evaluated in deciding whether and how to subtract background noise. Because both X and B are estimates, the variability of X-B is larger then the variability of X alone; thus, subtracting background adds variance. This especially problematic in the low-intensity range, where the variance of B can be of the same order of magnitude as X.

Generally, the assumptions of unbiasedness and additivity are far too optimistic. Also some researchers have found that the background estimates produced by some of the most popular image -process algorithms are not sufficiently reliable.

One alternative is to avoid background subtraction altogether and only use X to estimate the expression level. This avoids introducing the additional variance from inaccurately estimate background and is generally conservative in making declarations of differential expression and practical.

To see this, say that the true expressions in two samples being compared are e_1 and e_2. The observed values are $X_1 = e_1 + B_1 + \epsilon_1$ and $X_2 = e_2 + B_2 + \epsilon_2$ where ϵ_1 and ϵ_2 are errors in measurements of the true signal. Because both B_1 and B_2 are positive, the log ratio of the non background-corrected raw expression values $\frac{X_1}{X_2}$ is likely to be closer to 1 then the true ratio $\frac{e_1}{e_2}$. this bias toward one is stronger for low intensity genes. In summary, not subtracting background can be an attractive alternative, as it does not rely on potentially problematic background estimates and loses sensitivity mostly for low intensity genes; the exception are experiments with major spatial artifacts affecting only one channel.

In practice, decisions about background subtraction need to be made based on careful visualization of the data. The following rules of thumb are helpful for cDNA arrays:

- Inspect images of background alone (see 1.3.3.10 for methods and tools) and focus on major spatial artifacts affecting only one channel. If those are present, then background subtraction is critical.

- Avoid generating negative values and zeros after background subtraction; these indicate that error in the measurements of background are greater then the signal in the spot. The spot does not necessarily need to be discarded. Also avoid infinitesimal values. The resulting estimated ratio is both unreliable and likely to generate extreme ratios.

- Inspect MAplots discussed in 1.3.3.10. Major "fishtail effects" as in figure 1.3.3.6 at low intensities often indicate that spot-level background cannot be carried out reliably. Consider alternatives such as global background subtraction, subtraction of the faintest background, or no background subtraction.

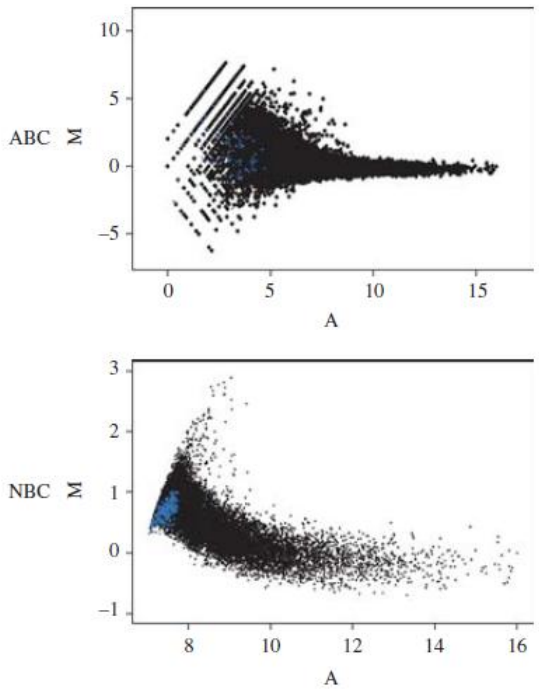

Figure 1.12: MAplot of the same array: (ABC) with "fishtail effects" caused by Adding Background Correction and (NBC) with No Background Correction.

1.3.3.7 Probe-level Analysis of GeneChip Arrays

GeneChip arrays pose the challenge of summarizing data from a probeset into a single measure, which estimates the level of expression of the gene of interest. Affymetrix software provides default approaches for this step. Two important issues suggest that the probe-level data should be considered an integral part of GeneChip data analysis. First, visualization of probe-level data can help identify artifacts. Second there is evidence[47, 48] that alternative summarizations to the defaults currently implemented by Affymetrix may provide improved ability to detect biological signal.

Typically, Affymetrix GeneChip microarrays have hundreds of thousands of probes. These probes are grouped together into probesets. Within a probeset each probe interrogates different parts of the sequence for a particular gene. Summarization is the process of combining the multiple probe intensities for each probeset to produce an expression value.

Various measures of expression have been proposed (for example see [47, 58]). The first version of Affymetrix's analysis software used an average over probe pairs of the differences $(PM_{ij} - MM_{ij})$, $j = 1, ..., J$, for each array i(where J is the number of probe pairs for a given probeset). Specifically, for each probeset on each array i, $AvDiff$ is defined by:

$$AvDiff = \frac{1}{\#A} \sum_{j \epsilon A} (PM_j - MM_j) \tag{1.1}$$

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 standard deviations from the average of the $d_{(2)}, ..., d_{(J-1)}$ the $j-th$ smallest difference. $\#A$ represents the cardinality of A.

Summary statistics, such as AvDiff, are motivated by the underlying statistical model:

$$PM_{ij} - MM_{ij} = \theta_{ij} + \varepsilon_{ij},\ j = 1, ..., J \tag{1.2}$$

The expression quantity on array I is represented with the parameter θ_i. $AvDiff$ is an appropriate estimate of θ_i if the error term ε_{ij}has equal variance for $j = 1, ..., J$. However the equal variance assumption does not hold for GeneChip probe level data, since probes with larger mean intensities have larger variances [47].

In the latest version of their software, Affymetrix uses a log transformation that is successful at reducing the dependence of the variance on the mean. Specifically, the MAS 5.0 signal is defined as the anti-log of a robust average (*Tukey biweight*) of the values $log\,(PM_{ij} - IM_{ij})$, where IM is the Ideal Mismatch (for details see [1]). A model for MAS 5.0 is:

$$log\,(PM_{ij} - IM) = log\,(\theta_i) + \varepsilon_{ij},\ j = 1, ..., J \tag{1.3}$$

Li and Wong [58] reported that variation of a specific probe across multiple arrays could be considerably smaller than the variance across probes within a probeset. In the scale, the between-array standard deviation is in general five times smaller than the within-probeset standard deviation [47, 58] . To account for this strong probe affinity effect, they proposed a multiplicative model

$$PM_{ij} - MM_{ij} = \theta_i \cdot \phi_j + \varepsilon_{ij},\ i = 1, \ldots, I,\ i = 1, \ldots, J \tag{1.4}$$

The probe affinity effect is represented by ϕ_j.

Using data from a spike-in experiment, Irizarry et. al. [47] found that appropriately removing background and normalizing probe level data across arrays results in an improved expression measures motivated by a log scale linear additive model. The model can be written as

$$T\,(PM_{ij}) = e_i + a_j + \varepsilon_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J \tag{1.5}$$

where T represents the transformation that background corrects, normalizes and logs the PM intensities, e_i represents the log_2 scale expression value found on arrays $i = 1, \ldots, I$, a_i represents the log_2 scale affinity effects for probes $j = 1, \ldots, J$, and represents error as above. A robust linear fitting procedure, such as median polish, was

used to estimate the log scale expression values e_i. The resulting summary statistic is referred to as RMA (Robust Multi-array Analysis).

Irizarry et. al. [47] demonstrated, using a spike-in study, that RMA has many advantages as compared to MAS 5.0. They showed that:

- RMA has a better precision; in particular, for lower expression values they found that RMA provides a greater than 5-fold reduction of the with-in replicate variance.
- RMA provided more consistent estimates of fold change.
- RMA provided higher specificity and sensitivity when using fold change analysis to detect differential expression. This greater sensitivity and specificity of RMA in detection of differential expression provides a useful improvement for researchers using the Affymetrix GeneChip technology.

1.3.3.8 Within-array Normalization

The complexity of the microarray experimentation process often introduces systematic bias into intensity measurements. Among other sources of variability, systematic bias can be caused by the concentration and amount of DNA pooled on the microarrays, wear to arraying equipment such as spotting pins, the quantities of mRNA extracted from samples, reverse transcription bias, lack of spatial homogeneity of the slides, scanner settings, saturation effects, background fluorescence, linearity of detection response and ambient conditions (ozone concentration) [2].The purpose of within-array normalization is to remove the effects of any systematic source of variation at array level.

For Affymetrix microarrays there are two levels at which within array level normalization can occur: probe-level, and probeset-level. The topic of probe-level normalization is considered extensively in . At this level, it is raw probe intensities, possibly after a background correction, that are normalized. Probeset-level normalization occurs when all the probes in a probeset are normalized together as a group. For instance, we could compute the mean (or median) value of a probeset, normalize these summaries and then adjust individual probes based on the adjustment to the summary.

For multichannel data, in addition, dye bias is present in almost all multichannel experiments. Generally, Cy5 (red) intensities tend to be higher than Cy3 (green) intensities, but the magnitude of the difference generally depends on the overall intensity [2]. The reasons for the imbalance between the channels are as follows [79]:

- The Cy3 and Cy5 labels may be differentially incorporated into DNA samples with varying frequencies of occurrence.
- The Cy3 and Cy5 dyes may have different emission responses to the excitation laser at different frequencies of occurrence.

- The Cy3 and the Cy5 emissions may be differentially measured by the photomultiplier tube at different intensities.
- The Cy3 and the Cy5 intensities measure at various areas on the array may differ due to a tilt in the array which results in variation in focus.

Several reports have indicated that $log_2(ratio)$ values Can have a systematic dependence on the intensity . This most commonly appears as a deviation from zero for low-intensity spot.

In this context, for multichannel microarray, normalization can be applied to adjust the bias among multiple channels, several approach are described below.

1.3.3.8.1 Standardization or Z-score Normalization Data sets are standardized to ensure that the mean and the standard deviation of each data set are equal. The method is simple; from each measurement on the array, subtract the mean measurement of the array and divide by the standard deviation. After this transformation, the mean of the measurements on each array will be zero, and the standard deviation will be one. An alternative to using the mean and standard deviation is to use the median and the median absolute deviation from the median (MAD). This have the advantage of being more robust to outliers than simply using the mean and standard deviation.

1.3.3.8.2 LOWESS: Locally Weighted Linear Regression The LOWESS transformation, also known as LOESS, stands for Locally Weighted polynomial regression [22, 23]. In essence, this approach divides the data into a number of overlapping intervals and fits a polynomial of the form:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots \tag{1.6}$$

Polynomials are very nice mathematical objects in the sense that they can approximate a large category of functions. However, the polynomial approximation has few general problems. Firstly, the approximation is good only in a small neighborhood of the chosen point and the quality of the approximation gets worse very quickly as one gets further away from the point of approximation. Secondly, polynomial approximation is very prone to over-fitting if higher degree polynomials are used. The approach used by LOWESS/LOESS deals with both issues in an elegant way. Firstly, the degrees of the polynomials used are limited to 1 (in LOWESS) or 2 (in LOESS) in order to avoid the over-fitting and the excessive twisting and tuning. Secondly, since the polynomials approximation is good only for narrow intervals around the chosen point, LOWESS will divide the data domain into such narrow intervals using a sliding window approach. The sliding window approach starts at the left extremity of the data interval with a window of a given width w. The data points that fall into this intervals will be used to fit the first

polynomial in a weighted manner. The points near the point of estimation will weigh more than the points further away. The procedure continues by sliding the window to the right, discarding some data points from the left but capturing some new data points from the right. A new polynomial will be fitted with this local dataset and the process will continue sliding the window until the entire data range has been processed. The result is a smooth curve that provides a model for the data. The smoothness of the curve is directly proportional to the number of points considered for each local polynomial, i.e., proportional with the size of the sliding window. If there are n data points and a polynomial of degree d is used, one can define a smoothing parameter f as a user-chosen parameter between $\frac{d+1}{n}$ and 1.The LOWESS will use $n \cdot f$ (rounded up to the nearest integer) points in each local fitting. Large values of f produce smooth curves that wiggle the least in response to variations in the data. Smaller values of q produce more responsive curves that follow the data more closely but are less smooth. The effects of a LOWESS normalization for a sample of the Lung dataset are showed in figure 1.13.

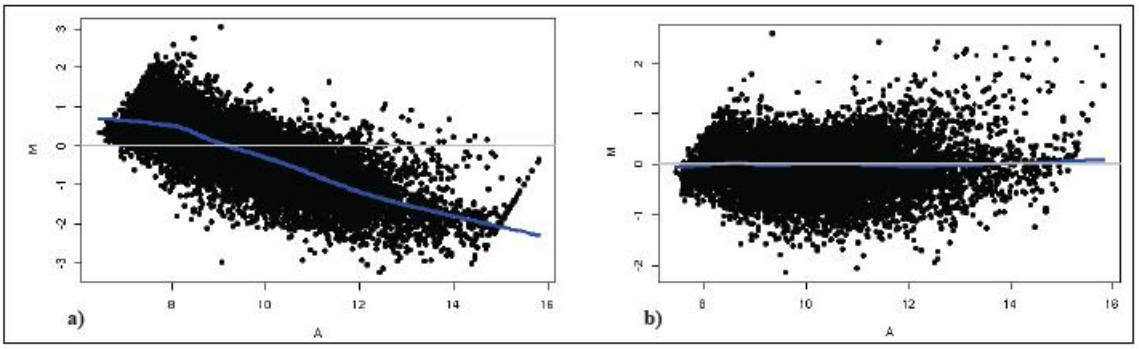

Figure 1.13: a) MA-plots for a sample Agilent array. The blue line is the LOWESS fit. An intensity-dependent effect is noticeable on the plot. b) LOWESS normalization could remove the intensity-dependent bias.

The biggest advantage of LOWESS is that there is no need to specify a particular type of function to be used as a model. The only parameter that need to be specified by the user are the degree of the polynomials d and the smoothing factor f. Disadvantages of LOWESS include the fact that it does not produce a regression function, or model, that can be easily represented by a mathematical formula. In particular, the dye bias distortion model found on a particular dataset cannot be transferred directly to another dataset or group of researchers. LOWESS need to be applied every time, on every data set and will produce a slightly different model in each case.

1.3.3.8.3 Within-print-tip-group normalization This would be the optimal methods if the platform under analysis is designed to have different prin-tip groups. In this

case every grid in the array is printed using the same print-tip. Different experiments may be done using different printing set-ups depending on the layout of the tips in the print-head of the arrayer (e.g. 4 by 4 or 2 by 2 print heads). Some systematic differences may exist between the print-tips, such as slight differences in the length or in the opening of the tips, and deformation after many hours printing. Alternatively, print-tip groups are proxies for spatial effects on the slide. Within-print-tip-group normalization is simply a "*print-tip* + A" dependent normalization, that is:

$$log_2\frac{R}{G} \rightarrow log_2\frac{R}{G} - c_i(A) \tag{1.7}$$

where $c_i(A)$ is the LOWESS fit to the MA plot for the $i-th$ grid only, $i = 1, \ldots, I$, where I represents the number of print-tips. Within-print-tip-group normalization is equivalent to LOWESS normalization, for each print-tip group independently (see bottom panel of figure 1.14 for an example of print-tip-group LOWESS fit).

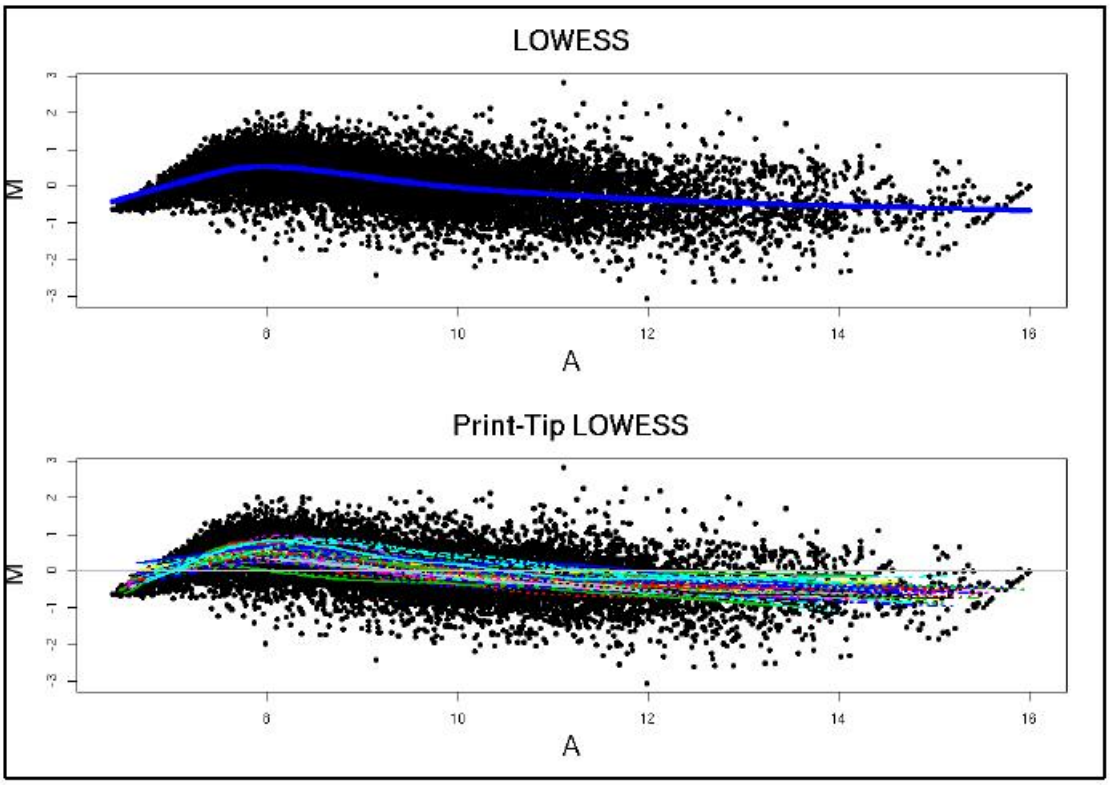

Figure 1.14: Top panel shows a LOWESS fit (blue line) of a MA-plot for a two color array. Bottom panel shows print-tip LOWESS fit: in this case different lines represent LOWESS fits that were done separately for genes belonging to different print-tip groups.

After within-print-tip-group normalization, all normalized log-ratios from different print-tip groups will be centered around zero. However, it is possible that the log-ratios from the various print-tip groups have different spreads and some scale adjustment is required. One approach is to assume that all log-ratios from the i-th print-tip group follow a normal distribution with mean zero and variance $a_i^2\sigma_i^2$, where 2σ is the variance of the true log-ratios and a_i is the scale factor for the $i-th$ print-tip group. In order to perform scale normalization, the scale factors a_i for the different print-tip groups must be estimated and then eliminated. Enforcing the natural constraints $\sum_{i=1}^{I} \log a_i^2 = 0$ with I denoting the total number of print-tips on the array, the maximum likelihood estimate for a_i is

$$\widehat{a}_i^2 = \frac{\sum_{j=1}^{n_i} M_i^2}{\sqrt[I]{\prod_{k=1}^{I} \sum_{j=1}^{n_i} M_{kj}^2}} \tag{1.8}$$

where M_{ij} denotes the $j-th$ log-ratio in the $i-th$ print-tip group, $j = 1, \ldots, n_i$.

1.3.3.8.4 Distribution Normalization (Quantile Normalization) While the purpose of LOWESS is to correct the mean of the data sets, the objective of distribution normalization is to make the distribution of the transformed spot intensities as similar as possible across the arrays.

In , Bolstad et al. propose a method for distribution normalization. Processing involves the following steps [88]:

1. Standardize the data

2. For each array D_i, order the standardize measurements from lowest to highest. Let D_{i1} be the smallest measurement in the D_i, and D_{in} be the greatest measurement, where n is the number of measurements in D_i.

3. Compute a new distribution D' whose lowest values of all the arrays being normalized, i.e., $D'_1 = avg\{D_{11}, \ldots, D_{m1}\}$, where m is the number of arrays; whose second-lowest value is the average of the second-lowest values from each of the arrays, i.e., $D'_2 = avg\{D_{12}, \ldots, D_{m2}\}$; and so on until the highest value is the average value of the arrays, i.e., $D'_n = avg\{D_{1n}, \ldots, D_{mn}\}$.

4. Replace each measurement on each array with the corresponding average in the new distribution according with the corresponding average in the new distribution according to its rank. For example, if a particular measurement of array D_i is the 100th smallest value in the array, replace it with the 100th smallest value D'_{100} in the new distribution.

Distribution normalization is an alternative to LOWESS normalization. It useful where the different arrays have different distributions of values. The assumption behind this

method is that given a series of arrays, a small number of genes may be differentially expressed, however, the overall distribution of spot intensities should not vary too much.

1.3.3.9 Between-slide Normalization

After within-slide normalization was done, all normalized log ratios should be centered around zero. However, in many experiments expression levels must be compared across different slides. It is important to note that individual slides in a multiple slide comparisons may need to be adjusted for scale when the different slides have substantially different spreads in their intensity log-ratios. Failing to perform a scaling normalization could lead to one or more slides having undue weight when averaging log-ratios across slides. Same principles used for within-slide print-tip groups scaling (see 1.3.3.8) can be used for multiple slide scale adjustment (see also). In practice, the need for scale normalization between slides will be determined empirically. In general there is a trade-off between the gains achieved by scale normalization and the possible increase in variability introduced by this additional step. In case where the scale differences are fairly small it may be thus be preferable to avoid this step.

1.3.3.10 Quality Control

It is useful if data analysis is not seen as the last step in a linear process of microarray exploration but rather as a step that completes a loop and provides the feedback necessary to fine-tune the laboratory procedures that produced the microarray. Thus array quality assessment is as aspect that should be included among the goals of the data analysis.

Microarray data are affected by systematic errors, such as intensity-dependent bias or spatial-dependent bias. Performing a "quality control analysis", referred to microarray data, means to identify these errors to remove them before further data analysis is conducted. In the worst case, when is not possible to adjust for these biases, a chip could be marked as outlier, and removed from subsequent analysis. Quality control is a fundamental step in microarray data analysis, and should always be considered before doing *high-level analysis*.

1.3.3.10.1 Two-color Arrays Quality Control The simplest approach to quality control analysis of two-color arrays is to look at "diagnostic plots" of spot statistics, such as red and green background intensities. Although visual inspection lacks the rigorous basis of statistical analysis, it provides an important tool to detect artifacts. Visual inspection of the data is usually supported by spatial false-color representations of intensities. Is important however to join a quantitative evaluation of quality based on some standard score, just for example signal to noise ratio, variance for replicated spots etc.

The ArrayQuality Bioconductor package (`http://www.bioconductor.org/repository/release1.5/package/html/arrayQuality.html`) is a complete and powerful tools for general two-color arrays quality control because it enables to obtain at the same time qualitative (Diagnostic plot) and quantitative (Quality-score comparison plot) information about the slides.

The arrayQuality report for Diagnostic plot, an example is showed in figure 1.15, consists of eight different diagnostic plots (most of them widely accepted in the microarray scientific community) described below :

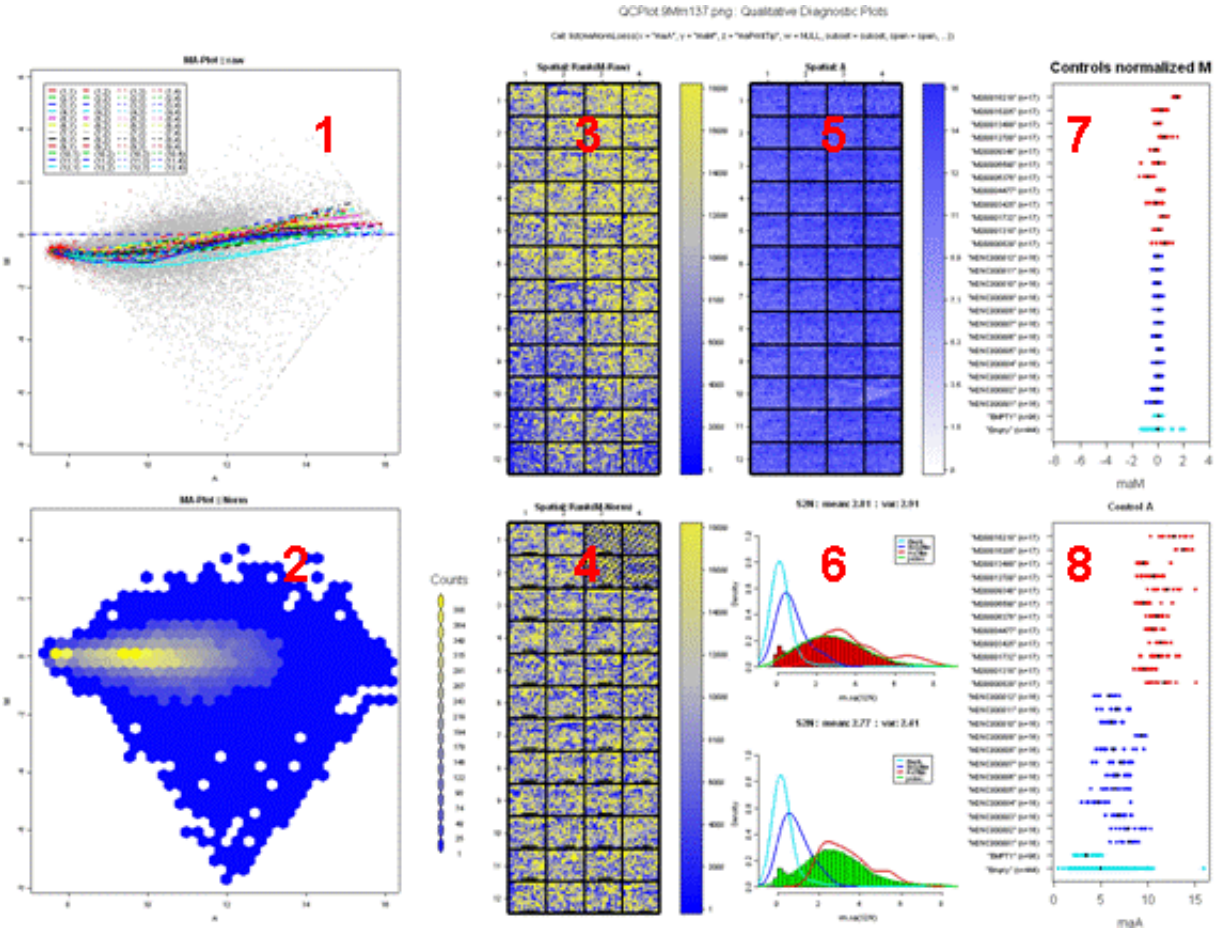

Figure 1.15: arrayQuality diagnostic plot for 2 color arrays..1)MA-plots of raw data; 2)MA-plot of normalized raw data; 3)Spatial plot of rank of raw M values; 4)Spatial plot of normalized M values ranks; 5)Spatial plot of raw A values; 6)Histogram of the signal-to-noise log-ratio (SNR) for Cy5 and Cy3 channels; 7)Dot plot of controls normalized M values; 8)Dot plot of controls A values, without background subtraction.

1. MA-plot of raw M. A brief introduction to the ideas behind MA-plot. In early

efforts in microarray quality control data were typically displayed by plotting the log intensity $\log_2 R$ in the red channel versus the log intensity $\log_2 G$ in the green channel. Such plots tend to give an unrealistic sense of concordance between the red and the green intensities and can mask interesting features of the data. Is preferable thus to plot the intensity log ratio $M = \log_2 \frac{R}{G}$ versus the mean log intensity $A = \log_2 \sqrt{RG}$. An MA-plot amounts to a 45° counterclockwise rotation of the $(\log_2 G, \log_2 R)$-coordinate system followed by scaling of the coordinates. It is thus another representation of the (R,G) data in terms of the log ratios M, which directly measure differences between the red and the green channels and are quantities of interest to most investigators. MA-plots are more revealing than their $\log_2 R$ versus $\log_2 G$ counterparts in terms of identifying spot artifacts and for normalization purposes. The first plot of figure 1.15 is an MA-plot of raw M with no background subtraction. The colored lines represent the loess curves for each print-tip group. The red dots highlight any spot with corresponding weighted value less than 0. Users can create their own weighting scheme or function. Things to look for in a MA-plot are saturation of spots and the trend of loess curves, which is an indicator of the amount of normalization to be performed.

2. MA-plot of normalized data density. By default, print-tip loess normalization is used. Instead of the typical MA-plot, a density MA-plot is used to highlight density of dots on the plot. A light yellow color indicates a high density of dots, whereas blue color represents a lower density. This plot gives you information on the bulk of your data intensity (low/high signal).

3. Spatial plot of rank of raw M values (no background subtraction): Each spot is ranked according to its M value. We use a blue to yellow color scale,where blue represents the higher rank (1), and yellow represents the lower one. Missing spots are represented as white squares. This is a quick way to visually detect uneven hybridization and missing spots.

4. Spatial plot of normalized M values ranks. By default, print-tip loess normalization is used. Each spot is ranked according to its M value. We use a blue to yellow color scale,where blue represents the higher rank (1), and yellow represents the lower one. Missing spots are represented as white squares. In addition, flagged spots are highlighted by a black square. This type of graphical representation helps verify that normalization removed any spatial effects.

5. Spatial plot of raw A values. The color indicates the strength of the signal intensity, i.e. the darker the color, the stronger the signal. Missing spots are represented in white.

6. Histogram of the signal-to-noise log-ratio (SNR) for Cy5 and Cy3 channels. The

mean and the variance of the signal are printed on top of the histogram. In addition, overlay density of SNR stratified by different control types (status) are highlighted. Their color schemes are provided in the legend. The SNR is a good indicator for dye problems. The negative and empty controls density lines should be closer, almost superimposed.

7. Dot plot of controls normalized M values. Controls with more than 3 replicates are represented on the Y-axis. Controls M values should be tight. and close to 0.

8. Dot plot of controls A values, without background subtraction. Controls with more than 3 replicates are represented on the Y-axis. Intensity of positive controls should be in the high-intensity region, negative and empty controls should be in the lower intensity region. Positive controls range and negative/empty controls range should be separated.

In figure 1.16 an example of quality-score comparison plot generated via arrayQuality package is reported. This is a more quantitative comparison of slide quality.Through the package is possible to extract some statistical measures from the test slide and to compare them against results obtained for a collection of slides of "good quality" to assess the quality of the hybridization. Is possible to choose a wide range of measures to quantify the quality of a typical hybridization: single channel measures (range of foreground signal, MAD of background, signal to noise ratio, etc.), two channel measures (median A values for each type of controls, amount of normalization needed, etc.), percentage of flagged spots. Some measures have been negated such that the quality scale had an increasing trend from problematic to good quality.

For more details about measure selected for comparison refer to `http://ugrad.stat.ubc.ca/R/library/arrayQuality/doc/guide.html`. For each measure, the following is represented on the graph:

- Boxplot of the reference slides values.
- 1st and 3rd quantiles before scaling for each boxplot.
- Y-axis on the right : for each measure, we have printed 2 values. The first one is the percentage of reference slides measures under your slide's result. The second one is your slide value for this measure before scaling.
- All the results are scaled in order to be able to compare them on the same graph.
- The red dots are the test slide scaled values

The quality check arranged by arrayQuality package are becoming a standard in two colors microarray data, for this reasons the Agilent FeatureExtraction software provide

a similar report for the each array. In this dissertation for the two dual-label platforms analyzed in this data set we use both the quality control tools

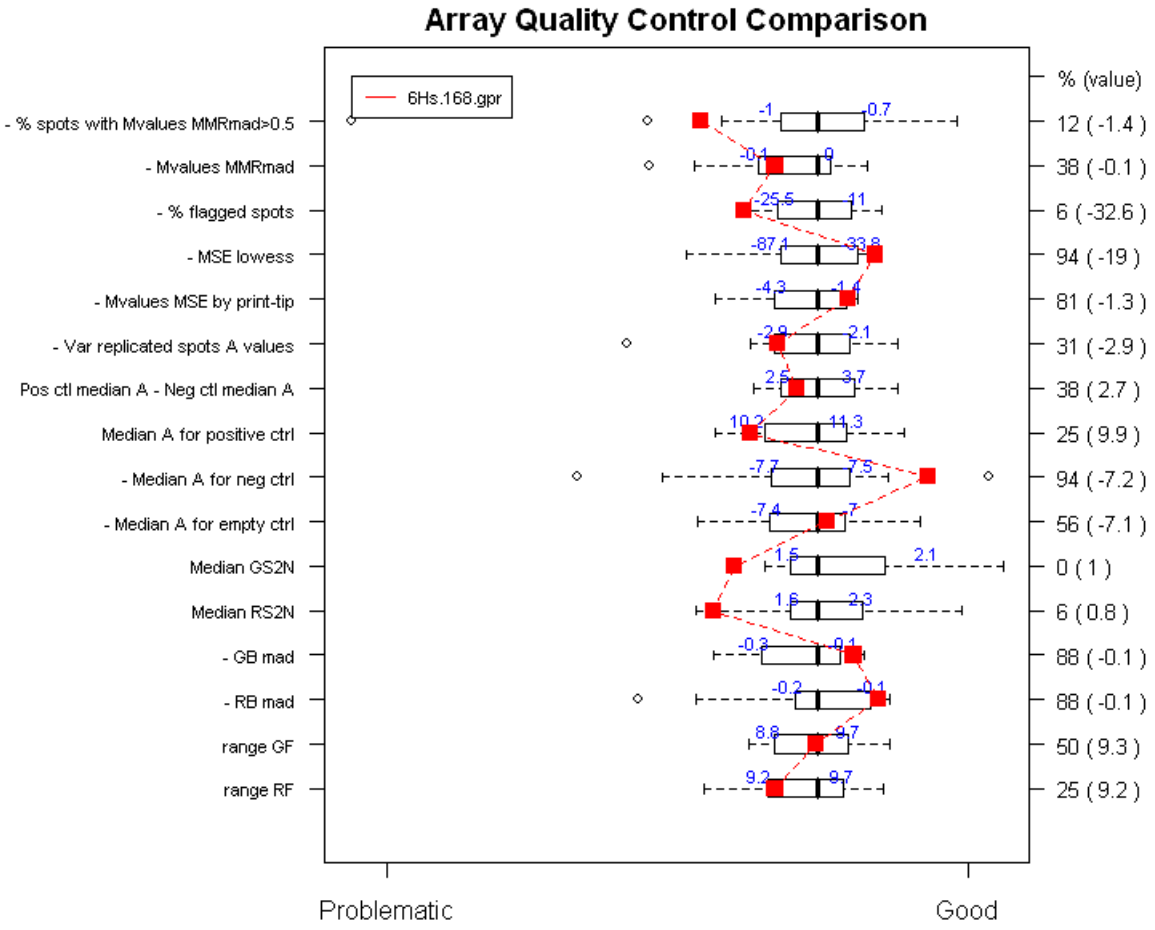

Figure 1.16: Quality-score comparison plot

1.3.3.10.2 GeneChip Array Quality Control In 1.3.3.10.1 we stressed the importance of quality control check in two-color arrays data analysis. Some of diagnostic tools we described, with few modifications, can be also used with Affymetrix chips. Lets briefly review some of them.

MA-plots for single channel microarray platforms are computed from the means and differences of log-expression values from two chips. In this case we need at least two arrays to make an MA-plot. While it would be possible to look at MA-plots for every possible pair of arrays, this will be an immense number of plots for any dataset consisting of many arrays. To reduce the number of possible comparisons a probewise median array

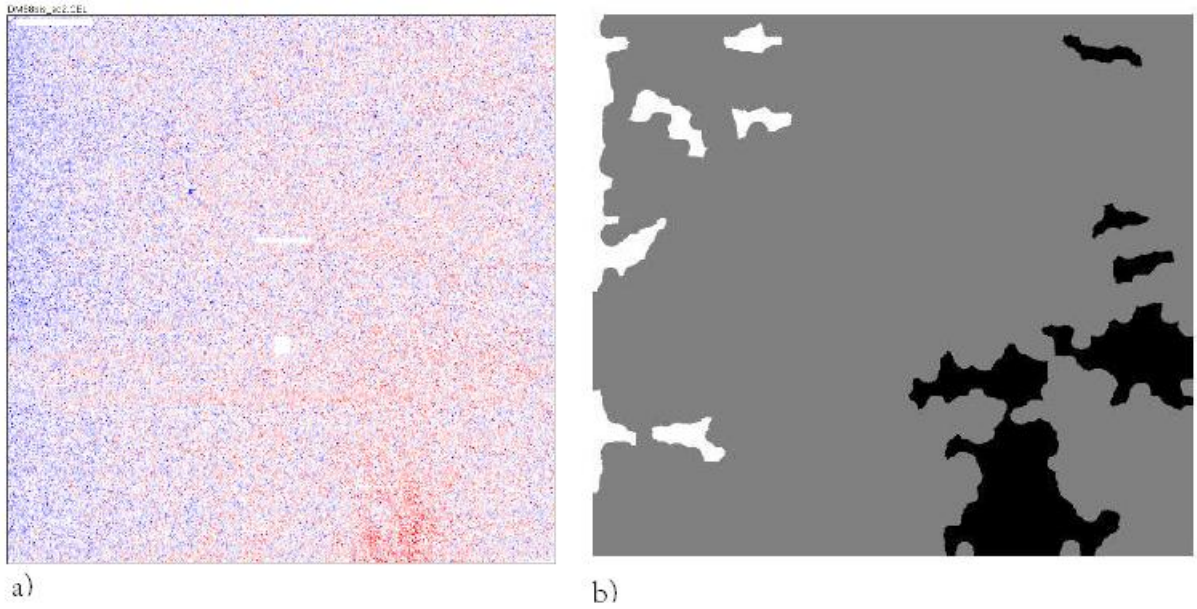

Figure 1.17: a)Spatial defects revealed by PLM; b) Spatial defects revealed by Harshlight

can be created and then each array compared to this pseudo-array. MA-plots are a useful tool in assessing intensity-dependent bias.

In 1.3.3.10.1 we have shown methods to detect spatial artefacts in cDNA arrays. Suárez-Fariñas et al. proposed an extremely simple method, namely "harslight", able to find localized artifacts, like speck of dust on the face of the chip, on Affymetrix arrays. This methods produces an Error Image (E) for each chip, which indicates the deviation of this chip's log intensities from the other chips in the experiment. Formally, E is calculated as $E^{(i)} = L^{(i)} - median_i\left(L^{(i)}\right)$ where $L^{(i)}$is the log-intensity matrix of the chip i. Given that chip intensity of each cell is highly determined by the sequence of the probe , this deviation should be near zero except for the probes belonging to the probe sets related to the genes that are differentially expressed. Since probes belonging to a probe set are (more or less) randomly distributed over the chip, probes of related genes are rarely located next to one other (www.affymetrix.com/support/technical/technotes), so that no obvious pattern should be discernible. Suárez-Fariñas et. al [81], to automatically detect spatial defects, developed an R-Package which spots faulty patterns on E using a battery of diagnostic tests based on both imaging processing and statistical approaches (see [81] for more details). Figure 1.17 b) is an example of the visual output result produced by harshlight on an Affymetrix chip. Large black and white areas indicates positive and negative systematic variation for E.

Based on the model of equation 1.5, Bolstadt observed that many departures from quality standards attributable to processing failures will be reflected by inflated residuals

from the fits to the model. Summarizing the residuals on the chip can therefore be expected to provide good discrimination among chips producing data of varying quality. Quantities related to residuals can be imaged, highlighting pattern of residuals that deviate substantially from an overall estimated scale (for more details about the method see). Using the PLM R's package, figure 1.17 a) gives an image of residuals for the same sample used to generate figure 1.17 b) . Marked red and blue areas indicates large absolute residuals, suggesting that fitted model is less than optimal. Detected spatial artifacts in a) and b) of figure 1.17agree each others very well.

1.3.4 Deciphering Disease with Microarray

One of the most exciting areas to which microarray technology has been applied is the challenge of deciphering complex disease such as cancer. Also in this dissertation the new approaches proposed in chapters 3 and 4 move in this direction trying to give answers to some open questions posed by collaborators (clinicians and biologists). Since most tumors exhibit unique expression patterns, gene expression data are often referred to as "signatures" or "portraits" [21]. The simultaneous monitoring of large-scale gene expression levels enables the identification of cell types which share these common expression patterns. As summarized by Chung et al. [21], previous studies have shown that DNA microarrays can help investigators to develop expression-based classifications from many types of cancer, including breast [46], leukemia[53] that are targets of the data sets analyzed during this research project.

In these studies, samples are taken from two or more groups of individuals with heterogeneous phenotypes, pathologies, or clinical outcomes. these samples are hybridized to microarrays in an effort to find a small number of genes which are strongly correlated with the group of individuals. These genes are often called *informative genes* [43], since they may help biomedical researchers to understand disease mechanisms. They can also be used to resolve levels of heterogeneity among cells that are apparent by eye and to provide a more accurate prognosis and prediction of response therapy [21]. Figure 1.18 illustrates the mechanism of using gene expression profiles to distinguish individuals with different phenotypes [88].

This approach is very promising since tumors of different types (such as malignant and benign tumors or good and poor prognosis tumors) can be very difficult to distinguish by conventional morphological, histological, clinical or pathological means. An examination of their expression patterns or signatures, offer much better potential for accurate discrimination. After Looking at this signatures from a functional point of view, with the help of functional genomics approaches like GeneOntology[3] or pathway analysis, can also give more insights in order to understand mechanism involved in the development and characterization of each pathology.

The contribution of this dissertation in deciphering disease regards the development of new approaches aiming at handle open problems posed by clinicians in handle specific

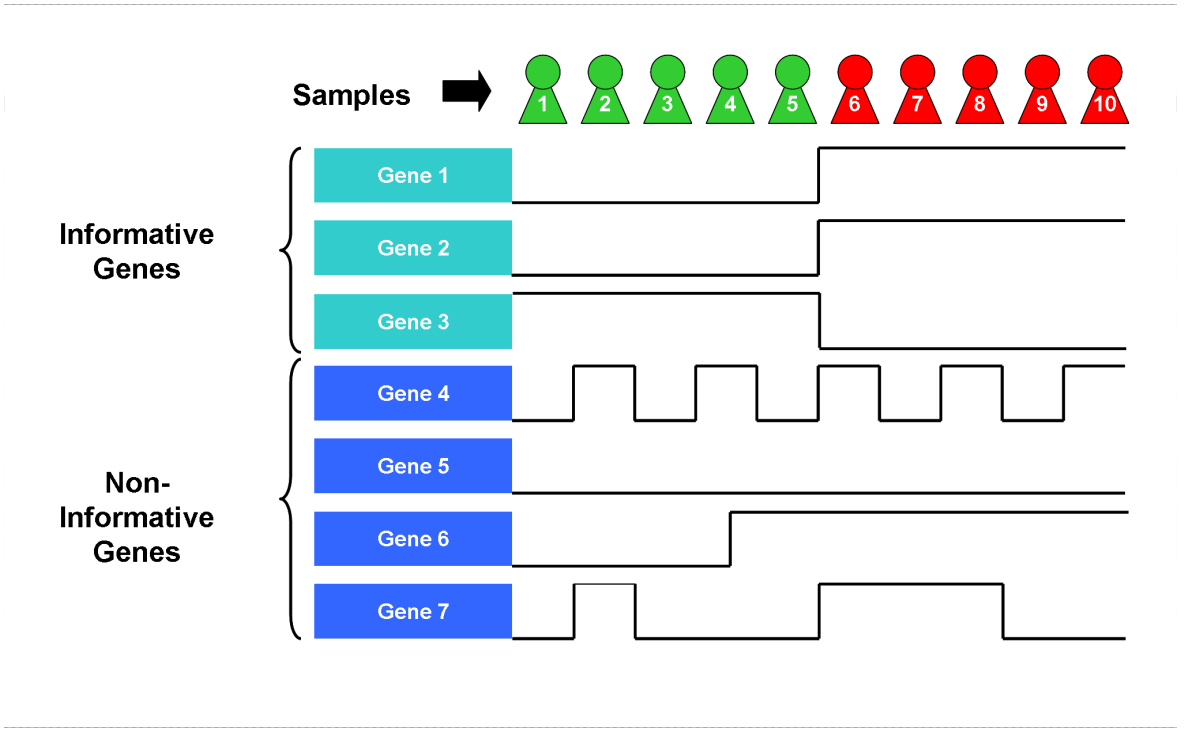

Figure 1.18: A simplified illustration of the mechanism of "expression signature". The polylines show the expressions profiles in an example gene expression data set. The first three genes are "informative genes" which exhibit expression profiles strongly correlated to the phenotype structure.

experimental designs.

Chapter 2

Standard Sample-Based Analysis Approaches

2.1 Introduction

In section 1.3.4 is declared the intention to face some open problems in micro array sample-based gene expression analysis in order to investigate disease related data set. In this context, take as witnesses Newton and his famous phrase *"If I have seen further it is by standing on the shoulders of Giants"* (Letter to Robert Hooke 1676), it is crucial to review what are the basis of these new approaches. As stressed before (see figure 1.5) we have seen an incredible growth of publications in this field and obviously this review will not be exhaustive. The aim of this chapter is to give a general idea about trends in gene expression microarray data analysis focusing the attention at two different levels:

- problems faced or not in this context
- algorithms of interest for the approaches developed in the next chapters

2.2 Selection of Informative Genes

Microarray data are often extremely asymmetric in dimensionality. A specific characteristic of of this data is that the number of samples in a microarray experiment is typically by far smaller than that of the genes; this is known as "large p, small n" problem in statistics. At one extreme, a microarray data set usually contains thousands or even ten of thousands of genes. At the other extreme, the number of samples is usually no more than few hundreds. Such extreme asymmetry between the dimensionality of genes and samples presents several challenges to conventional supervised and unsupervised methods (for more details about supervised and unsupervised method (see next sections),

which were generally designed to process a large number of data object with relatively few attributes. Some of the challenges these method encounter in processing asymmetric microarray data sets are listed below:

- Proximity measure. Most methods for sample-based analysis rely on some proximity function to measure the distance or similarity between a pair of samples. However, when the number of attributes is very high, some proximity measures, such as the Euclidean distance, may become meaningless. That is, the distance of an object to its nearest neighbor approaches the distance to the farthest neighbor.

- Overfitting. The problem of "overfitting" occurs when an algorithm adapts to the training samples too exactly, loosing sufficient ability to generalize in the prediction of new samples. In consequence, while the classification of the training examples may be perfect, the accuracy of prediction with test samples drops dramatically. The large number of features characteristics of microarray data sets may render the predictive algorithm prone to overfitting to the limited number of training samples.

- Multiplicity. A parallel examination of a large volume of genes may incorrectly identify some as being differentially expressed between different samples when, in fact, these differences may be due to random variation [2, 79]. Multiple testing will be discussed in section

- Course of dimensionality. The term course of dimensionality was coined by R. Bellman. In the context of sample based analysis, it refers to the exponential growth of the hypothesis space with respect to the number of features. In general, a clustering or classification method needs to search the feature space to find a solution. Given the large number of genes and the exponential growth of the search space the efficiency of learning will drop dramatically.

These difficulties suggest the appropriateness of reducing the data dimensionality to improve the effectiveness and efficiency of the sample-based algorithms. In fact, many genes in a microarray experiment are irrelevant to the problem under study and need not to be considered in the clustering or classification process.For example, a study of a typical biological process, such as a determination of the differences between tumor samples, seldom involves more than a few dozen of genes [88].

Determining which genes to be used in the clustering or classification procedure is essential for the success of sample-based analysis. Gene selection is necessary not only to reduce dimensionality but also to identify those genes that are closely related to the cell types. In general the approach to gene selection can be categorized as supervised and unsupervised, depending on whether the class labels of the samples are given a priori. In the following subsections, these methods will be discussed in more details.

2.3 Selecting Differentially Regulated Genes

In many cases, the purpose of the microarray experiment is to compare the gene expression levels into different specimens. In such comparative studies, a very important problem is to determine those genes that are differentially expressed in the two class compared [31]

Although simple in principle, this problem becomes more complex in reality because the measured intensity values are affected by numerous sources of fluctuation and noise [31]. In this context, distinguish between genes that are truly differentially regulated and genes that are simply affected by noise becomes a real challenge. All methods discussed here are independent of the technology used to obtain the data. The main difference between the different types of data is the pre-processing as we have seen in chapter 1.

2.3.1 Criteria

The performance of a gene selection method can be calculated in terms of positive predicted value (PPV), negative predicted value (NPV), specificity and sensitivity. In general for any diagnosis or classification method, one can compare the truth with the results reported by the method. In a binary decision situation such as changed/unchanged, the results can always be divided into 4 categories:truly changed that are reported as changed (True Positives - TP), unchanged that are reported as changed (False Positives - FP), truly changed that are reported as unchanged (False Negatives - FN) and truly unchanged that are reported as such (True Negatives - TN). Based on these is possible to define

$$PPV = \frac{TP}{TP + FP} \tag{2.1}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.5}$$

All this quantities range from 0 to 1. A perfect method would yield no false positives and no false negatives. However in disease related experiment clinicians often ask to optimize the selection (being more conservative) in order to avoid useless extra-assay biological validation on genes that are not truly differentially regulated.

2.3.2 Fold Change

The simplest and most intuitive approach to finding the genes that are differentially regulated is to consider their fold change. Typically an arbitrary threshold such as 2 or 3 fold is chosen and the difference are considered as significant if it is larger than the threshold. In a screening experiment involving many genes, most genes will not change. Thus the experiment/control ratio of most genes will be grouped around 0 which means that their logs will be grouped around 0, this is quite evident in the histogram of ratios presented in figure 1.11 on page 23. The horizontal axis of such a plot represents the log ratio values. In consequence, selecting differentially regulated genes can be simply done by setting thresholds on this axis and selecting the genes outside such thresholds. This thresholding methods can resemble a classical hypothesis testing situation, see subsection 2.3.3 for more details. The difference is that in a hypothesis testing situation the threshold are chosen very precisely in order to control the probability of the Type I error (calling a gene differentially regulated by mistake) while in the fold change method, the threshold are chosen arbitrarily.

The latter is one of the most important drawback. Another important disadvantage is that fold change constant thresholding can introduce false positives at the low end (thus reducing the specificity), while missing true positives at the high end (thus reducing the sensitivity). This is a quite evident consequence of the funnel shaped aspect of an MA-plot (see for example 1.3.3.6 on page 26) that shows a bad signal/noise ratio for low expression levels.

2.3.3 Hypothesis Testing

Another possible approach to gene selection is to use univariate statistical tests to select differentially expressed genes.

Scientific theory can generally never be verified, but only disproved. In statistics, this leads to the procedure of setting up a research (or alternative) hypothesis and a contradictory null hypothesis. The research hypothesis is supported if we can show that there is evidence against the null hypothesis. Hypothesis testing involve several important steps. The first step is to clearly define the problem. Such a problem might be stated as follows [31]: "The expression level c of a gene is measured in a given condition. It is known from literature that the mean expression level of the given gene in normal condition is μ. We expect the gene to be up-regulated in the condition under study and we would like to test whether the data support this assumption".

The second step is to generate two hypotheses. These are statistical hypotheses and, unlike biological hypotheses, they have to take a certain, very rigid form. In particular, the two hypotheses must be mutually exclusive and all inclusive. Mutually exclusive means that the two hypotheses cannot be both true at the same time. All inclusive means that their union has to cover all possibilities. In other words, no matter what

happens, the outcome has to be included in one or the other hypothesis. One hypothesis will be the *null hypothesis*. The other hypothesis will be the *alternative* or *research hypothesis*.

The null hypothesis, traditionally denoted by H_0 is the claim that is initially assumed to be true (the "prior belief" claim). The alternative hypothesis, denoted by H_a, is the assertion that is contradictory to H_0. The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, is possible to believe in the truth of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then reject H_0 or fail to reject H_0. The alternative hypothesis has to reflect our expectations. If is believable believe that the gene should be up-regulated, the research hypothesis will be:$H_a : c > \mu$. The null hypothesis has to be mutually exclusive and also has to include all the other possibilities. Therefore, the null hypothesis will be $H_0 : c \leq \mu$.

The third step is to calculate an appropriate test statistic based on data and to choose a significance level, or a rejection region. The key to a good test is a good test statistic. The test statistic is generally a sample statistic that reflects how far the observed data is from the situation described by the null hypothesis. Many test statistics, T, have the form $T = \frac{r}{s}$. Here r could be an estimate of the size of the biological effect being tested: the further the data is from the null hypothesis (i.e., the more likely that the null hypothesis is false), the larger the value of r. The denominator, s, is a standard error that measures the variability of r. Thus T measures how large the biological effect r is relative to its variability. It is no accident that T has the form of a signal-to-noise ratio with r as the "signal" and s as the "noise".

The probability distribution of the test statistic under the null hypothesis is called its *null distribution*. Based on the null distribution, we can calculate the *p-value*, the probability of observing a value as extreme as that observed if the null hypothesis was true. Clearly, the smaller the p-value, the greater is the weight of evidence against the null hypothesis. A typical decision rule for a test states that the null hypothesis is rejected if and only if the p-value is less than a specified value called the *significance level* of the test. In other words, the p-value is the probability of drawing the wrong conclusion by rejecting a true null hypothesis. Choosing a significance level means choosing a maximum acceptable level for this probability. The significance level is the amount of uncertainty we are prepared to accept in our studies. For instance, when we choose to work at a significance level of 10% we accept that 1 in 10 cases our conclusion can be wrong. Usual significance level are 1%, 5%. The final step in the hypothesis testing procedure is to compare the value of the calculated test statistic with the rejection region, that is the set of all test statistics values for which will be rejected; in this way is possible either to reject or not reject the null hypothesis, with the chosen significance level.

The basic for choosing a particular rejection region lies in an understanding of the errors that one might be faced with in drawing a conclusion. Let us consider that the

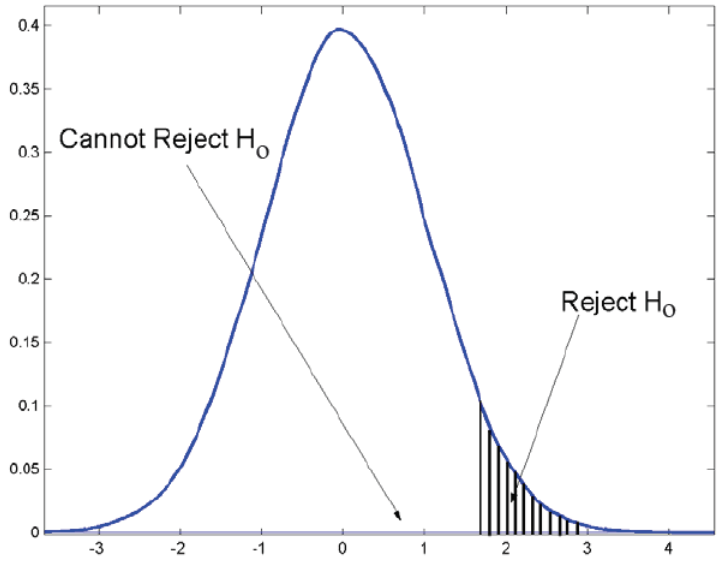

Figure 2.1: If computed statistic takes a value inside the dashed area, the Null Hypothesis is rejected

true situation is known; let "assume that H_0 is actually true and H_a is false". In this case:

- If we accept H_0, we have drawn the correct conclusion. We will call the instances in this category true negative: they are negative because they go against our research hypothesis and they are true because H_a is indeed true and our H_a is false. If H_0 is "the gene is not regulated", and H_a is "the gene is either up or down regulated", true negatives will be those genes which are not regulated and are reported as such by our hypothesis testing algorithm.

- If we reject H_0, we have drawn an incorrect conclusion. We will call the instances in this category false positive. They are positive because they go with our research hypothesis and they are false because they are reported as such by our algorithm while, in fact, they are not. Non-regulated genes reported as regulated would be false positive.

Rejecting a null hypothesis when it is in fact true is called a *Type I error*. The probability of a Type I error is usually denoted by α. Let us consider a normal distribution such as the one presented in figure 2.1 .

Using a threshold such as the one shown in the figure 2.1 will classify as up-regulated any genes expressed at a level higher than the threshold. However, this means all the genes in the shaded area of the graph will be false positives since H_0will be rejected for

them. Therefore, the probability of a Type I error corresponds directly to the significance level chosen.

Let now "assume that H_0 is false and H_a is true". In this case:

- If we accept H_0, we have drawn an incorrect conclusion. The instances in this category will be false negative (genes that are in fact regulated but are not reported as such by our algorithm).
- If we reject H_0, we have drawn the correct conclusion. The instance in this category are true positive.

The second type of mistake is called a *Type II error*. The probability of a Type II error is denoted by β. The probability of avoiding a Type II error corresponds to correctly picking the instances that do not belong to the distribution reference. In our case, this corresponds to finding differentially regulated genes (which do not belong to the distribution representing the normal variation of expression levels). This is exactly the purpose of an hypothesis testing algorithm. The higher this probability, the better will the algorithm be at finding such genes. This probability is called power of the test and calculated as $1 - \beta$. A natural tendency is to try to minimize the probability of making an error. The probability of making a Type II error is not directly controllable by the user. However, the probability of making a Type I error, α, is exactly the significance level. Therefore, one might be tempted to use very high standards and choose very low values for this probability. Unfortunately, this is not a good idea since α and β are closely related and using a lower value for α has the immediate consequence of reducing the power of the test.

The classical statistical methods to performing an hypothesis test on data of the type control vs treatment is the t-test. Two version of this test, the paired t-test and unpaired t-test, are applicable to data sets containing two groups of observations. Is important to stress that t-statistics require some restrictive assumption like:

- Data are assumed to be normally distributed.
- Data are assumed to have equal variance.

However, if this is not the case, the value of this statistics may not represent the true degree of differential expression. As result, using p-values obtained from the t-distribution as a test of gene expression may be meaningless in these instances. In fact, there are many sources of variability in a microarray experiment, and outliers are frequent. Thus, the distribution of intensities of many genes are often not normal in real data set. In this case , non-parametric methods which do not place any assumption on the observed data. These non parametric method do not rely on the estimation of parameters (such

as the mean and the standard deviation) in describing the distribution of the variable of interest in the population for example and details refer to the work of Deng et al [28].

Comparing gene expression across two condition for a single gene is an instance of one of the most statistical questions. Estimation and testing tools in this case are very well developed, as mentioned before. In genomics applications, however, there is increasing consensus[64]that is not efficient to consider each gene in isolation, and gains can be made by considering the ensemble of gene expression measures at once. This occurs for at least two reasons: first, genes measured on the same array type in the same laboratory are al affected by a number of common sources of noise; second, changes in expression are all part of the same biological mechanism, and their magnitudes, although different, are not completely unrelated.

Joint estimation of many related quantities is a time-honored problem in statistics dating back at least to the pioneering work of Stein and colleagues and continuing with empirical Bayes approaches. In the case of microaray, examples of implementation are provided in the work of Smith (for details see[73]) who use empirical Bayes approaches joined with linear models to asses the differential expression.The method is implemented in a R-based Bioconductor package called LIMMA [74]. Another approach of a al genes-based modified t-statistics can be found in the SAM (Significance Analysis of microarray Data) algorithm developed by Tusher et al.[84] that will be described in more details in chapter 3. In the microarray scientific community, these two are the most used algorithm for select differentially regulated genes, even though there are numerous different methods in literature.

2.3.4 Multiple Comparisons

In analyses aimed at selecting differentially expressed genes , there are several approaches for reporting the degree of reliability of results. Conventional approaches based on gene-specific p-values are generally criticized on the grounds of the multiplicity of comparison involved. In fact, analyzing microarray data involves performing a very large number of statistical tests, as a test is being run on each and every gene.

One important drawback of doing this is that if we perform multiple tests in parallel, the level of significance for the whole set of tests does not equal the level of significance for the single tests. Let us study this phenomenon. The significance level α was defined as the acceptable probability of a Type I error. This corresponds to a situation in which the null hypothesis is rejected when it is in fact true. The genes that are called differentially regulated when in fact they are not, will be false positives. In terms of hypothesis testing, when the test statistic T for a gene is more extreme than the threshold T_α, we will call this gene differentially regulated. However, the gene may be so just due to random effects. This will happen with probability α. If we do not a mistake, we will be drawing the correct conclusion for that given gene. This will happen with probability:

$$Prob(correct) = 1 - \alpha \tag{2.6}$$

Now we have to take in consideration the fact that there are many such genes. Let us consider that there are G such genes. For each of them we will follow the same reasoning. However, at the end, we would like to draw the correct conclusion from all of them. The probability of such an event is easily computed by:

$$Prob(globally\, correct) = (1 - \alpha) \cdot (1 - \alpha) \ldots (1 - \alpha) = (1 - \alpha)^G \tag{2.7}$$

Then the probability of being wrong somewhere would be 1 minus the probability of being correct in all experiments:

$$Prob\,(wrong\ somewhere) = 1 - Prob\,(globally\ correct) = 1 - (1 - \alpha)^G \tag{2.8}$$

In this situation being wrong means drawing the wrong conclusion for at least one gene. Equation 2.8 is very close to unity for large G and the expected number of false positives is $\alpha \cdot G$, which is very large for very large G. Thus the number of false positives can be so high as to overwhelm and totally obscure any actual effects.

In the multiple hypothesis literature equation 2.8 is termed, Family Wise Error Rate (FWER), and is defined as the probability of one or more false positives occurring among all significant hypothesis. Most conventional multiplicity adjustment attempt to control the FWER. The most popular approach is to report the false discovery rate (FDR)[8], for a group of genes or for a specified cutoff value of a statistic of choice. Assuming that the population of genes truly divides into two groups, the altered and the unaltered genes, and that a statistical approach selects a set of "significant genes", the the FDR is an estimate of the fraction of truly altered genes among the the genes declared significant. This approach often reflect appropriately the fact that array experiments are performed to guide future validation work on individual genes, which is usually expensive and time-consuming. It is also directly interpretable as the probability that a gene in the list of selected genes if one takes the set of genes on the arrays as the population of reference for the calculation of the probability. Additional discussion and comments are summarized in [34]. Application of FDR within SAM is instead treated in chapter 3.

A drawback of the methods based on hypothesis testing is that they tend to be a bit conservative. Not being able to reject a null hypothesis and call a gene differentially regulated does not necessarily mean that the gene is not so. In many cases, it is just that insufficient data do not provide sufficient statistical proofs to reject the null hypothesis. However, those gene that are found to be differentially regulated using such methods will most likely be so.

Another point concern the experimental design, using hypothesis testing with few number of samples for each class is really risky in order to obtain reliable answers to the

posed biological question. The method proposed in [73] claim to be able to work with a reduced number of samples but the suggestion is to work at least with a number of samples per class greater then six. But is well known for people who works in this field that sometimes for economical and clinical problems laboratory try to plan experiment with just 3 samples per class. In this case all the hypothesis testing assumptions (requested by parametric methods) could be not considered valid anymore. Another point is that using this method in presence of a particularly unbalanced design is not recommended, because such a design affects evidently the equality of variances.

2.3.5 Beyond Two Groups

Methods for identifying genes that are differentially expressed across two experimental conditions can be extended to more general settings in which multiple conditions. Extension include time course experiments where conditions correspond to multiple time points, factorial designs, in which the effects of multiple factors and their interaction are explored simultaneously and so forth. The empirical Bayes method described in [73] approach this kind of problems with a model-based approach. In general, multilevel analysis can proceed by first specifying a statistical model for expression as a function of conditions. In this context a particularly interesting approach to micro array data analysis and selecting differentially regulated genes is the Analysis Of Variance (ANOVA) [14]. This method will be briefly revised here. The idea behind ANOVA is to build an explicit model about the sources of variance that affect the measurements and use the data to estimate the variance of each individual variable in the model.

For instance, Kerr and Churchill [49, 50] proposed the following model to account for the multiple sources of variation in a microarray experiment:

$$\log\left(y_{ijkg}\right) = \mu + A_i + D_j + G_g + \left(AD\right)_{ij} + \left(AG\right)_{ig} + \left(VG\right)_{kg} + \left(DG\right)_{jg} + \epsilon_{ijkg} \qquad (2.9)$$

In this equation μ is the overall mean signal of the array, A_i is the effect of the i^{th} array, D_j represents the effect of the j^{th} day, G_g is the variation of g^{th} gene, $\left(AD\right)_{ij}$ is the effect of the array -dye interaction, $\left(AG\right)_{ig}$ is the effect of a particular spot on a given array (array-gene interaction), $\left(VG\right)_{kg}$ represent the interaction between the k^{th} variety and the g^{th} gene, $\left(DG\right)_{ig}$ is the effect of dye-gene interaction and ϵ_{ijkg} represents the error. Finally $\log\left(y_{ijkg}\right)$ is the measure log-ratio for gene g of variety k measured on array i using dye j.

Sum of squares are calculated for each of the factors above. The mean squares will be obtained by dividing each sum of squares by its degree of freedom. This will produce a variance-like quantity for which an expected value can be calculated if the null hypothesis is true. Essentially, each individual test asks the question whether a certain component

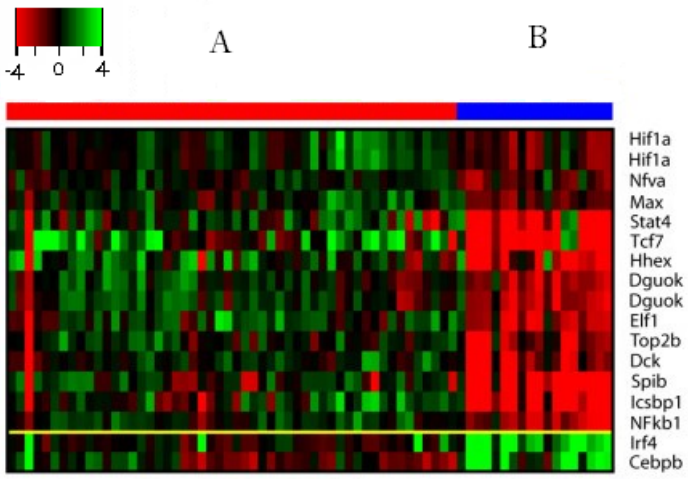

Figure 2.2: Example of heatmap with samples ordered by class membership in the classical green (up-regulation) and red (down-regulation) scale.

has the variance significantly different from the variance of the noise. the differentially regulated genes will be the genes for which the $(VG)_{kg}$ factor is significant..

The advantage of ANOVA is that each source of variance is accounted for. The caveat is that ANOVA requires a very carefully experimental design. Another limits of ANOVA (and in general of model based approaches) is that, in a problem with more than two classes, is possible to knows just which are the genes involved [73]. In fact, this kind of analysis cannot reveal intrinsic correlation between specimens or direction (up/down) of these regulations.

2.4 Visualization and Unsupervised Analysis

As with quality control and signal extraction, multivariate analysis of microarrays relies substantially on visualization. The most commonly used tool is a color map (*heatmap*) of either relative or absolute hybridization intensities after proper normalization. This type of map was introduced by Eisen et al. [36]. An expression map is arranged as a matrix in which each row corresponds to a gene and each column to an array, and the color represent the expression level. Rows and columns are often sorted in a way that facilitates visualization; for example genes may be grouped by functional classes, sample may be ordered by time or class membership. In figure 2.2 there is an example of a a heatmpa order by class membership.

However the most common way to order samples and genes is, like in [36], using cluster algorithms, in particular Hierarchical Clustering (treated more extensively later in this section).

Clustering algorithms divide a set of objects (genes or samples) into groups so that gene expression pattern within a group are more alike than pattern across groups [64]. There are two broad approaches to clustering techniques. Hierarchical techniques provide a series of successively nested clusters, and their result resembles a phylogenic classification. Nonhierarchical techniques generally find a single partition, with no nesting. Both are used extensively in microarray analysis for two main goals. The first is representing distances among high-dimensional expression profiles in a concise visually effective way, such as are tree or dendrogram. The second is to identify candidate subgroups in complex data. The two tasks can be closely tied as a concise lower-dimensional representation of the objects studied often simplifies manual identification of subgroups. Clustering techniques have been successful in supporting visualization and as method for generating hypotheses about the existence of groups of genes or samples with similar behavior. An example of the latter is the identification of novel subtypes of cancer. Oftentimes, even when phenotype information is available, unsupervised cluster analysis are used to explore gene expression data and form gene groups that are then correlated to phenotype,

Quackenbush [67] provides an excellent reviews of the use of cluster analysis in microarrays. Some general comments apply to the use of clustering techniques in genomic analysis. First, these techniques are exploratory: their strength is in providing rough maps and suggesting directions for further study. In good studies, context and meaning for groups found by automated algorithms is provided by substantial additional work either in the lab or on databases. The outcome of a clustering procedure is therefore the beginning, rather then the end, of a genome biometry analysis.

Second clustering results are sensitive to a variety of user-specified inputs. The clustering of a large and complex set of objects, such as a genome, is akin to arrange books in a library[31]: it can be done sensibly in many different ways, depending on the goals. From this perspective, good clustering tools are responsive to user's choices, not insensitive to them, and sensitivity to input is a necessity of cluster analysis rather than a weakness. This also means, however, that use of a clustering without a through understanding of its workings, the meaning of inputs, and their relationship to the biological questions of interest is likely to yield misleading results.

Third, clustering results are generally sensitive, sometimes extremely so, to small variations in the samples and the genes chosen and to outlying observations. This means that a number of the data-analytic decisions made during normalization, filtering, data transformations, and so forth will have an effect on clustering. In this context, it is challenging, but all more important, to provide accurate assessments of the uncertainty that should be associated with clusters found. Uncertainty from sampling and outliers can be addressed using resampling techniques [49].

2.4.0.1 Hierarchical Clustering

Hierarchical clustering has almost become the de facto standard for gene expression data analysis, probably because of its intuitive presentation of the clustering results. The whole clustering process is presented as a tree called a dendrogram; the original data are often reorganized in a heat map demonstrating the relationships between genes or conditions.

In hierarchical (agglomerative) clustering [31], each expression profile is initially assigned to one cluster; at each step, the distance (through Euclidean distance or Pearson's correlation, and others) between every pair of clusters is calculated and the pair of clusters with the minimum distance is merged; the procedure is carried on iteratively until a single cluster is assembled.

After the full tree is obtained, the determination of the final clusters is achieved by cutting the tree at a certain level or height, which is equivalent to putting a threshold on the pairwise distance between clusters. Note that the final cluster positions is thus rather arbitrary.

As we mentioned, in every step of agglomerative clustering, the two clusters that are closest to each other will be merged. Here comes the problem of how we define the distance between two clusters. There are four common options:

1. Single linkage. The distance between two clusters is the distance between the two closest data points in these clusters (each point taken from a different cluster).

2. Complete linkage. The distance between two clusters is the distance between the two furthest data points in these clusters.

3. Average linkage. Both single linkage and complete linkage are sensitive to outliers [32]. Average linkage provides an improvement by defining the distance between two clusters as the average of the distances between all pairs of points in the two clusters.

4. Ward's method. At each step of agglomerative clustering, instead of merging the two clusters that minimize the pairwise distance between clusters, Ward's method merges the two clusters that minimize the 'information loss' for the step. The 'information loss' is measured by the change in the sum of squared error of the clusters before and after the merge. In this way, Ward's method assesses the quality of the merged cluster at each step of the agglomerative procedure.

These methods yield similar results if the data consist of compact and well separated clusters. However, if some of the clusters are close to each other or if the data have a dispersed nature, the results can be quite different [32].Ward's method, although less well known, often produces the most satisfactory results.

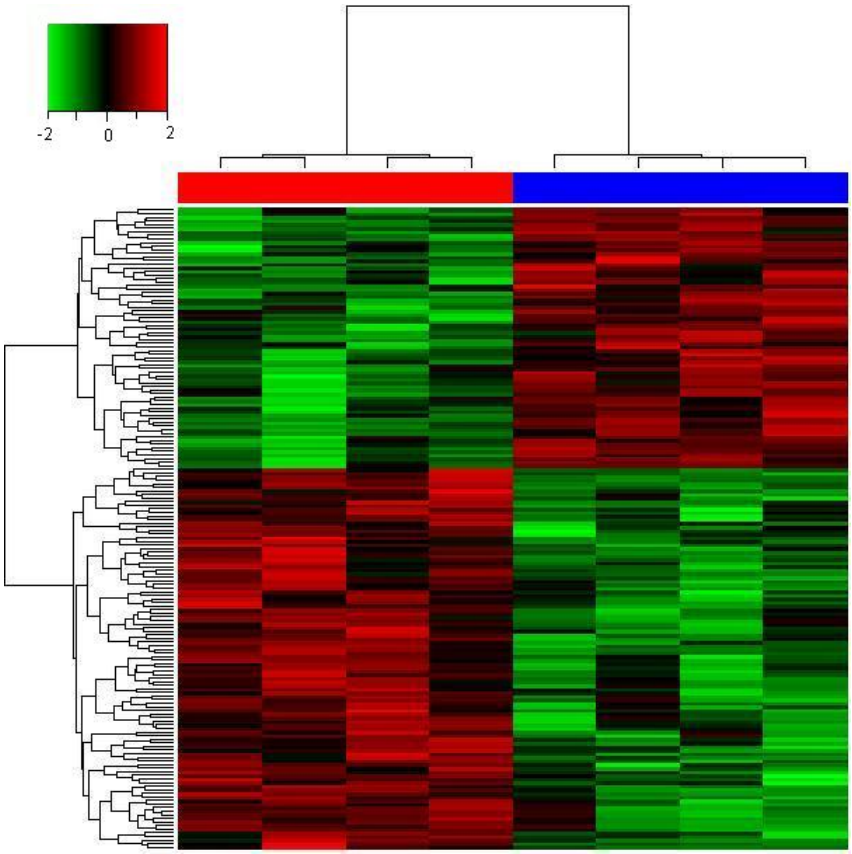

Figure 2.3: Example of a heatmap reordered on the basis of a hierarchical clustering algorithm.

A heatmap presenting the gene expression data, with a dendrogram to its side indicating the relationship between genes (or experimental conditions), is the standard way to visualize the result of hierarchical cluster analysis on microrray data. The length of a branch in the dendrogram is proportional to the pairwise distance between the clusters. Usuallly this method in most of publication in microarray scientific community is used in supervised approach. In figure 2.3 an example of hierarchical clustering-guide heatmap.

2.5 Class Prediction

Class prediction or prognostic prediction tries to predict the class membership (or survival or protein expression or any prognostic variable) of a set of subjects given their gene expression data. Although related to the selection of differentially expressed genes (see section 2.3), these are different objectives that answer different biological questions and require different methods (unfortunately, this difference is not always recognized in em-

pirical work) [30]. Ranking genes often precedes trying to use genes for class prediction, but genes that show large expression differences are not necessarily good predictors.

The goal is to predict the clinically relevant characteristic of a subject (be it class membership, survival, prognosis or any other variable of interest) given the genetic profile of this subject. This is also an area of extremely active research, where the disciplines of statistics and machine learning have contributed much.

The complexity of gene expression array analysis is stimulating the development of novel specific statistical modeling tools for this purpose. However, the existing body of pattern recognition and prediction algorithms developed in computer science and statistics can provide an excellent starting point for prediction in molecular problems for the immediate future. Dudoit et al. [33] offer a practical comparison of discriminant methods for the classification of tumors using gene expression. Relevant methods from the traditional modeling tradition include: linear discriminant analysis [40], tree-based algorithms [90, 16], Support Vector Machine (SVM) [86, 45], Relevance Vector Machine (evolution of SVM) [82, 59], Nearest-Neighbor classifier[26] are just some examples of the available methods successfully applied in the gene expression class prediction.

Available reviews[33, 71, 68] show that relatively simple and well known methods such as k-nearest neighbor (KNN) and diagonal linear discriminant analysis (DLDA), together with support vector machines (SVMs), perform very well in most classification tasks in microarray data. Because of their performance and DLDA, KNN and SVM should probably be used routinely as benchmarks when proposing new methods.

It is hard to say which could be the best algorithm, because until now the benchmark data sets used to evaluate the performance are regarded as "easy data sets". Despite of which algorithm is used, is important to stress some issues that probably deserve more attention. First, for the user it quickly becomes evident that many methods yield non-unique solutions or, in other words, can return different solutions of very similar quality (e.g., prediction error rate), which itself leads to the question of how to choose among solutions. A direct way of approaching this problem is via model combination and model averaging.

Regardless of which model(s) are used, two general problems can affect all models/algorithms. First, most of the available methods assume additive effects of genes. Non-additive relationships or interactions, also called synergistic (or antagonistic) effects, are present when the outcome (e.g., being of class A) depends not just on the sum of the independent contributions of X and Y, but on their combined effects. Non-additive relationships are likely both between genes and between genes and other factors. Second, the predictive capacity of many models can be hampered by unrecognized heterogeneity within classes that are regarded as homogeneous. Little work has been done in this area.

A final set of problems involves the biological interpretation of class prediction models (together with making sense of information for potentially tens of thousands of coefficients). Most methods for building predictors tend not to return models that allow for

easy biological interpretation of why and how those predictors are used, and how the genes in the predictors affect and relate to the class prediction.

To evaluate the performance of a predictor, it is common to provide the error rate of the predictions. However, many papers, including 'high-profile' ones, report error rates that are severely biased, leading to overoptimistic claims about the performance of different methods. This is a most unfortunate situation because lack of appropriate rigor in the application and adherence to appropriate rules of evidence undermines trust in the promises of these technologies. These severe problems were addressed in the bioinformatics literature by Simon et al. [71]. In spite of the seriousness of the problem, the practice of reporting severely biased error rates is still common.

One possible problem is reporting the 'resubstitution rate', the error rate computed from the very same observations as were used to build the classifier, because the resubstitution error rate is severely biased down due to overfitting: if we fit a classifier to a data set, we can expect it to 'adapt to' some peculiarities of the data, which will make it work well with those data, but might lead it to work poorly with data not yet seen by the classifier or learner. This problem is even more serious with microarray data, where there are thousands of genes that can be part of a predictor. With so many variables, and so few samples, it is very easy to find a predictor that works perfectly in a completely random data set [71]. To solve this problem either cross-validation or bootstrap have been used; both methods build the predictor using a subset of the data, and then predict the values for the remaining data, thus insuring that the predictions are from data not used for the training.

A second common problem is to carry out the cross-validation after the gene selection: all samples are used for gene selection, and the cross-validation process does not include gene selection. This leads to very optimistic estimates of the error rate [71].

As final note about class prediction regards the stability of the results[30].Suppose a predictor has been built that includes 20 genes. How far can we take biological interpretation on the relevance of these genes? A paper by Somorjai et al. [75] suggests that often not very far; the problem is the instability or non-uniqueness of results, a phenomenon called the 'Rashomon effect' by L. Breiman [15]. It is very common that, if we re-run a given procedure with only minor changes or using bootstrap samples, we end up with very different sets of models, suggesting that there are many different 'optimal' subsets of genes (because there are many different descriptions that give approximately the same minimum error rate; Somorjai et al. show how this can arise because of small sample sizes and an extremely small sample per feature ratio (i.e., very small number of arrays relative to the number of genes). They suggest using a variety of classifiers or predictors and finding whether the same features are selected; if the same set of genes is repeatedly selected, we would be more confident that the set is reasonably robust. Of course, this way of examining robustness to selection methods cannot be used if feature selection is carried out using the same filter method for different classifiers (e.g., find-

ing the 200 genes with smallest adjusted p-value, and then using those 200 genes with DLDA, KNN and SVM). Additionally, the bootstrap can be used to examine variation in solutions achieved. The multiplicity problem deserves much more careful attention and prompts for cautious interpretation of results.

2.6 Philosophical Issues in Microarray Data

This brief review focus on general trends in microarray data analysis and stress some of the problems that concern each specific approach. However, below is reported some considerations that are more general but that has to attract the attention of gene expression microarray researcher community. Sometimes, in fact, the "bioinformatics"[1] community are concentrated on "reinventing the wheel" [30]. Many publications[2], in fact, focus on reinventing solutions for problems (sometimes too much specific) that has already been successfully solved by others. The interest, instead have to concentrate to general issues hided in these kind of data, but often neglected.

2.6.1 Microarray: "experiments" or observational studies?

Although microarray studies are often referred to as 'experiments', they are frequently observational studies [4]. The differences between observational and experimental studies are well known in statistics and epidemiology, and affect both analyses and interpretation of results. Observational studies present several potential problems, particularly the following.

- Background differences between groups and presence of potential confounding variables; confounding is a pervasive problem. Potter (2003) illustrates it with examples of the relation between vegetable consumption and cancer being confounded by differences in smoking associated with vegetable consumption (smokers also tend to eat fewer vegetables) and differences in expression profiles between cancer types being related to the unmeasured confounding of age and sex. A related problem is interaction, such as when the degree of association between an exposure factor (e.g., expression of gene A) and the disease is different for different levels of the confounding variable, such as sex; there is evidence that this might be the case in lung cancer (Patel, Bach and Kris, 2004). The problems of confounding and interaction are discussed in more detail below.

[1] This is the keyword that, today, identifies people working on handling, analyzing bio-data with computers

[2] This comment refer above all to those publications that grows in a pure bioinformatics context, without any connection with biologist's (or clinician's) interest

- Biases arising from handling of units (e.g., case samples are frozen several hours after collection whereas control samples are frozen immediately;) or from biases during the selection of subjects for the study or from informative patterns of missingness.

- Samples too small to allow for generalizations to the populations of interest, and problems of reproducibility.

These issues are well known in epidemiology, which studies patterns of disease and possible factors that affect these patterns of disease by using mainly observational data. However concerns related to microarrays being often observational studies are mostly absent from standard papers and textbooks on microarray design and analysis[31, 88].

2.6.2 Rule of Thumb: Round-table of different expertises

Successful use of microarrays to answer biologically relevant questions will require close collaboration between biologists and statisticians during the complete process of the study[30]. The need for statisticians' advice during the experimental design has been discussed before (see section 1.3.3.1). However, it should be remembered that full details of the experimental set-up are necessary for the use of appropriate statistical methods.

Statisticians need to realize that there are often many subtleties in the interpretation of microarray results that preclude simple mappings from RNA expression data to phenotypes . At the same time, statistical help is needed to insure that the statistical model and test being used is addressing the biological questions of interest. What in any case is unrealistic is to expect that if the biologist sends a file with 15 000 rows by 200 columns (genes by subject) to the statistician, the statistician will return to the biologist the list of, say, 30 genes that are the answer to the biological question. But that is, in fact, what some users often expect from software tools or statistical consulting, and what some statisticians might believe is possible/desirable. This also means that the questions asked are sometimes reformulated to accommodate the available software.

The problem of these expectations and procedures is that they lack key ingredients often needed to provide an answer to the underlying biological question. Diàz-Uriarte [30] lists some typical questions that a statistician might ask that are reported below

- Are genes grouped in families, and are we interested in the overall responses of groups of genes, or should we look at individual genes?

- Are certain genes or spots in the array more relevant biologically, maybe because they are easier to measure reliably with other assays?

- Is there additional information on which genes are likely to be differentially expressed?

- Do you really need the best possible predictor that statistical computing will get you, or do you want a small list of genes very likely to be differentially expressed?
- In what stage of the scientific discovery process is this study, and how tight control do you require over the type I error rate?
- What other information and variables about the patients, besides the microarray data, do you have available?
- What population do you expect the results of these studies to be relevant for?
- Are these the original, complete data, and are these the original biological questions, or have the data and questions gone through an already long run of analyses which has already filtered data and reoriented hypotheses?
- What is the next stage of this study, or what do you want to do with these results?
- What additional studies could be done to confirm the results from these analyses?

Only after these (and other) questions have been answered is it time to search for the appropriate strategy. This research project starts exactly in this way. Each of the approaches explained in the next chapters are the attempts to give answers to questions emerged after long "round tables" with the clinicians and the biologists of CRO[3].

2.6.3 Availability of source code

Many new method papers are published every month, and biologists and applied statisticians do not have the time to implement each and every idea that is published, nor to deal with the complications associated with patented algorithms. Sometimes, however, when researchers ask for software from authors of method papers they face with a great deal of practical or not explained difficulties.

Some of the reasons for making source code available in bio informatics and micro array research (summarized by Dugout, Gentleman and Quackish) are reported here:

- full access to the algorithms and their implementation, which allows users to understand what they are doing when they run a particular analysis
- the ability to fix bugs and extend and improve the supplied software
- encouraging good scientific computing and statistical practice by providing appropriate tools,instruction and documentation

[3]National Oncology Research Center in Aviano (Italy)

- providing a workbench of tools that allow researchers to explore and expand the methods used to analyze biological data
- ensuring that the international scientific community is the owner of the software tools needed to carry out research
- promoting reproducible research by providing open and accessible tools with which to carry out that research (reproducible research as distinct from independent verification)

The approaches explained in chapter 4 relies in part on the lack of source code that could have given an answer to open question posed by the biologists at CRO.

Chapter 3

MultiSAM: Facing with Unbalanced Design

3.1 The Open Problem: Uneven Sample Size

As seen in section 1.3.3.1, in microarray experiments for gene expression profile, the experimental design is often a crucial start-point to obtain reasonable results. In particular, in a two-class problem (to compare two different conditions), it should be quite obvious to collect an equal, or at least similar, number of samples between the two classes.

In fact, this should be a basic rule of thumb in order to can use standard methods both in hypothesis testing and in class prediction framework.

In hypothesis testing , choosing to use robust modified t-statics (see 2.3.3 on page 49) that have resulted to be more effective in the problem posed by micrarray data, we have to consider that they make stringent assumptions (the same requests by classical t-test), i.e. a) data are normally distributed and 2) homoscedasticity (i.e. equal variance). Above all for the second condition, collect an equal number of samples in both class is a way to control the homogeneity of variance and the applicability of Central Limit Theorem, as well explained by Devore in [29].

In class prediction, as well, some is well known the problems of uneven-sample-size bias. As reported by Wood et al. [87] for a classification algorithm when the training data sets with uneven class sizes were used, the model was undesirably biased towards the class with the large training size. That is to say, the larger the training sample size for one class is, the smaller its corresponding classification error rate is, while the smaller the sample size, the larger the classification error rate.

So the best solution would be to design a balanced experiment. However, this is not always possible. A complete and balanced design is an ideal that is not always achieved in practice. The cost of conducting a microarray experiment, in terms of labor, time and money, sometimes makes it impractical to use a balanced design. Also, some studies

set out with a balanced design but become unbalanced during implementation because of technical and administrative problems. Another important issue rises when clinicians decide to investigate disease subset with rare and peculiar clinical and biological features, that come to the attention for their severe prognosis. During the last three years of our collaboration, the research team at CRO[1] runs into the latter situation and pose to us the necessity to handle a challenging unbalanced microarray data set.

Looking at the literature, most of the publications about gene expression profiles do not face this kind of problem, for two different reasons:

- often the experimental design is balanced in it self (see just for example [38, 10])
- sometimes, despite of an unbalanced experimental design (that as we have seen affect variances and error rates), standard method are used[2] (see for example [52, 9].

Just some papers approach this problem in different ways. Concerning hypothesis testing Lee et al.[56] tackle the issue using ANOVA; but in there the unbalancing of the design regards a time course[3] experiment so the situation cannot exactly resemble the standard two-class gene profiling problem. Moreover a up-to-date study [85] shows that ANOVA fails in case of unbalanced design. Even if in CRO's data set the application of class prediction algorithm was prohibitive for the size of the smaller class, there is some papers that face (without a definitive solutions) the problem of uneven samples size[20, 87]. Considering such a state of the art, the choice is to find an alternative solution.

3.2 The IGHV3-21 B-cell Chronic Lymphocytic Leukemia data set

To clear up the circumstances that have lead to an unbalanced design regarding the gene expression profiling of a particular sub set of B-Cell chronic Lymphocytic Leukemia (B-CLL) is important to contextualize this disease and its specifically spread and feature in Italy[13].

The clinical course of B-cell chronic lymphocytic leukemia (CLL) can be in part foreseen by the presence of mutated (M) or unmutated (UM) immunoglobulin variable (IGV) genes[70]. Analysis of IGV heavy chain (IGHV) gene usage has revealed expression of a biased IGHV repertoire in CLL compared with normal B cells [83].Such molecular peculiarities occur both in poor-prognosis UM CLL and in highly stable/indolent M CLLs [83].

[1]National Oncology Research Center (Aviano-Italy)

[2]Standard methods can have however a margin of success on those two-class problems that are "easy" to separate (normal vs. tumor for example) and with a low level analysis.

[3]Monitoring of gene expression at different time point

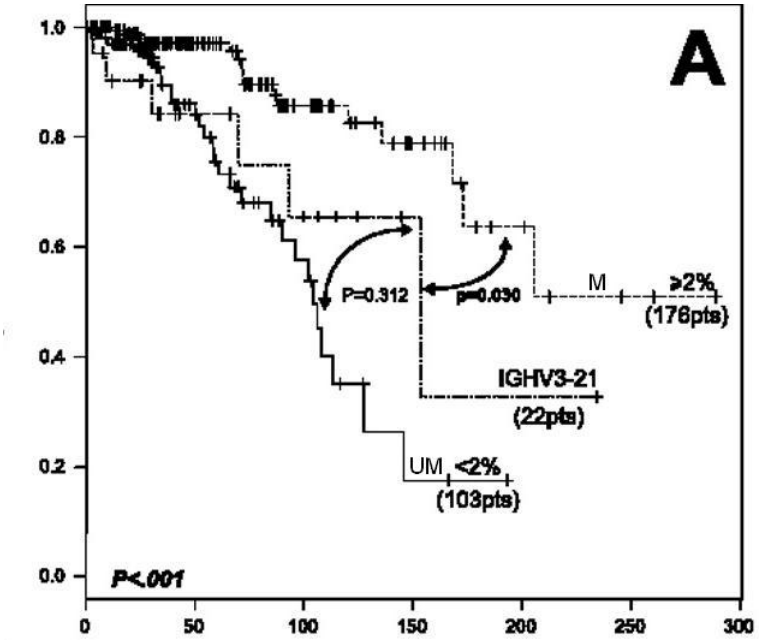

Figure 3.1: **Kaplan-Meier survival curve analysis in IGHV3-21 CLL patients.** (A) Comparison of survival probabilities in IGHV3-21 patients (22 patients), patients with M IGHV gene configuration (at least 2%, 176 patients), and patients with UM IGHV gene configuration (less than 2%, 103 patients).

Usage of the IGHV3-21 gene by CLL has been initially reported in about 10% of Scandinavian patients12,13 and subsequently confirmed in small CLL series from other Northern European countries[62]. Although IGHV3-21 CLL usually displays M IGHV genes, patients experience overall survivals similar to those of UM CLL[42]. Recent studies reported a low incidence of IGHV3-21 CLL in mediterranean countries, thus suggesting that the frequency of IGHV3-21 CLL may be related to geographic, ethnic, or environmental background[42, 78]. Italy may provide a valuable model to analyze IGHV3-21 distribution among CLL because the country includes both continental and mediterranean areas. To conclusively dissect the issue of a

In [13], survival data[4] were analyzed in 301 patients, 279 with a non–IGHV3-21 CLL (176 with M and 103 with UM IGHV genes) and 22 with IGHV3-21 CLL (14 M and 8 UM). In the context of non–IGHV3-21 CLL, a significantly longer survival was documented for M cases compared with UM CLLs. The survival curve of IGHV3-21 CLL was similar to that of UM non–IGHV3-21 cases (p-value 0.312) and significantly different from M CLLs (p-value 0.030; see figure 3.1).

[4]For detailed introduction to survival data analysis please refer to [25]. For the survival analysis performed at CRO on IGVH3-21 please refer to [13].

From the above discussion is possble to summarise as follows the principal characteristic of this rare and peculiar subset of CLL that are of interest for the planning of the experimental design:

- IGHV3-21 CLL is a rare subset of CLL that has low incidence in general (and in particular in mediterranean countries);
- Eventhough IGHV3-21 CLL usually displays M IGHV genes, patients experience overall survivals similar to those of UM CLL;
- IGHV3-21 CLLs are known to have a clinical and features charactersitic that are not macroscopically far from non-IGHV3-21 CLLs; in this context, identify a gene expression profile of IGHV3-21 CLLs with respect to non-IGHV3-21 CLLs will be a challenging experience.

Collecting patients from different Italian oncological centers the study end up with only 13 IGVH3-21 CLL (11 M and 2 UM) patients and with 52 non-IGVH3-21 (35 M and 17 UM) patients with a sufficient high quality mRNA material to land on a microarray. Details on platform, experimental design, pre-processesing and filtering choices are extensively described in Appendix A.

Conscious that the main characteristic of this microarray experiment is to be a challenging unbalanced data set next section will focus on the construction and validation of resample-based methods to find a significant gene expression profile.

To give an idea of how the entire data set behave in term of class membership in figure 3.2is reported a dendrogram realized with the entire data matrix (65 samples x 28513 genes) with a hierarchical clustering algorithm (euclidean distance joined with ward algorithm). The procedure was repeated (data not shown) with different hierarchical methods and different distances. The generated hierarchic clustering clearly demonstrated that all 13 IGHV3-21 CLLs were intermingled among non–IGHV3-21 cases, thus indicating that the 2 subgroups shared the large part of the features investigated.

3.3 From SAM to MultiSAM

The first way to approach an open problem like unbalanced data sets is to look a what was already successful in the state of the art, trying to improve or correct the pitfalls that rise with this kind of problem. Also Lee et al.[56] looked at linear models and ANOVA as a possible choice. In the case of IGHV3-21 data set this is not an optimal solution because as stressed in [85], linear models and ANOVA do not perform well on really noisy data (when the sources of noise are not identifiable as factors in the models). So in our quite long experience with Operon microarray data sets, we notice an high level of noise not ever controlled by pre-processing and normalization steps.

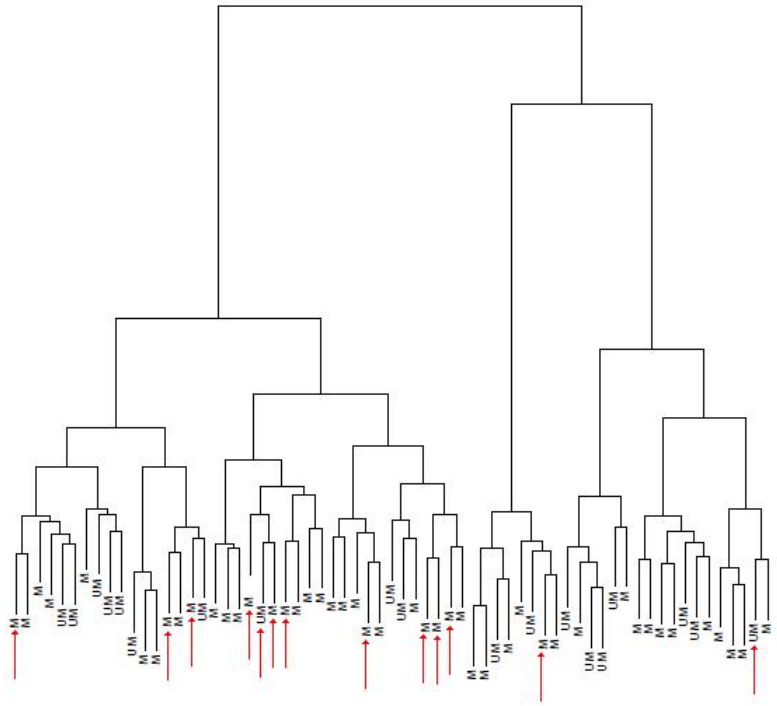

Figure 3.2: Global hierarchical clustering dendrogram based on the entire data matrix (65 samples x 28513 genes) of IGHV3-21 data set. Red arrows represent the 13 cases of IGHV3-21 CLL.

The idea is to use the popular and thriving SAM[84] algorithm as starting point (in particular the modified SAM version of Broberg [17]. In the next subsection a brief introduction to the algorithm will be given before to illustrate the approach to correct possible pitfalls rising from uneven samples size.

3.3.1 Significance Analysis of Microarray Data (SAM statistic)

With small samples size the t test statistic tends to be highly correlated with the standard error term that appears in its denominator. As a result the test has a propensity for picking up significant findings at a higher rate from among those genes with low sample variance than from among those genes with high sample variance. This property of the t statistic is especially troubling because it is difficult to estimate standard errors well when the sample is low and small standard errors can occur purely by chance. Since the sample sizes used in microarray experiments are typically very small, the small sample effect of the t test tends to manifest itself in such experiments as false negative rates for genes whose variability is high.

One solution to the problem was suggested by Tusher et al. [84]. They add a carefully chosen constant, a so-called *fudge factor*, to the denominator of the t statistic.

Lets suppose that we are comparing the expression levels of a gene in two groups of microarrays, which we will refer as Group 1 and Group 2. There are n_1 arrays in Group 1 and n_2 arrays in group 2. Let x_{ij} denotes the intensity measurements of the gene in the $i-th$ microarray in the $j-th$ group, where $i = 1, ..., n_j$ and j=1,2. It is assumed that the data has already been transformed and normalized. In addition let $\overline{x}_{ij}$ and s_j denotes respectively the mean and the standard deviation. Recall that the $t_g$test statistic, for the g^{th} gene, has the form $t_g = \frac{r_g}{s_g}$, where $r_g = |\overline{x}_{1g} - \overline{x}_{2g}|$ and $s_g = s_{pg}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. The adjusted t statistic is:

$$t_g(c) = \frac{r_g}{s_g + c} \tag{3.1}$$

where c is the fudge factor. This test statistic is often called the SAM t statistic, where SAM stands for "significance analysis of microarrays". $t_{g(0)}$ is, of course, the ordinary t statistic. with a very large value of c is equivalent to the t statistic without its denominator, namely to r_g. The plan is to choose an intermediate positive value of c while, given c, the dependence of $t_g(c)$ from s_s is as small as possible. The simplest way to do this, in practice, is to study the relationship of $t_g(c)$ versus s_g for a number of different values of c, with the intention of retaining as the fudge factor, c, the one for which the dependence $t_g(c)$ of on s_g is least.

Tusher et. al. implement this as follows:

Let s^α be the α^{th} percentile of the $\{s_g\}$ values, and let $t_g(s^\alpha) = \left(\frac{r_g}{(s_g + s\alpha)}\right)$. Compute the percentiles, $q_1 < q_2 < ...q_{100}$ of the s_g values. For $\alpha\epsilon\{0, 5, 10, ..100\}$, coimpute the MAD (median absolute deviation from the median), $v_j(\alpha)$, of the $t_g(s^\alpha)$ values within

the interval $[q_j, q_{j+1}]$ for $j = 1, 2, ..., n$. Then compute $cv(\alpha)$, the coefficient of variation of the $v_j(\alpha)$ values. Choose as $\hat{\alpha}$ the value of α that minimises $cv(\alpha)$. Fix as $\hat{c}$ the value $s^{\hat{\alpha}}$.

3.3.1.1 Assessing Significance with the SAM t Statistic

Once the SAM t statistics, $t_g(\hat{c})$, are calculated, the critical value of $t_g(\hat{c})$ that separates significance from non-significance must be set. For the ordinary t statistic this is done by looking up the quantiles of the Student's distribution. However, the null distribution of the SAM t statistic, $t_g(\hat{c})$, is not a t-distribution, so this is no longer correct. In fact the null distribution is intractable. Therefore Tusher et al. assessed the significance of the observed values via a permutation procedure.

Suppose that a suitable $\hat{c}$has been identifies and that the values have been calculated and sorted into increasing order: $t_{(1)}(\hat{c}) \leq t_{(2)}(\hat{c}) \leq ... \leq t_G(\hat{c})$. The permutation procedure proceeds by permuting the columns of the data matrix, X, and assigning the first columns to Group 1 and the remaining columns to Group 2. A total of B such permutation will be done. For the b^{th} permutation, compute the statistics $t_g^{*b}(\hat{c})$, and the corresponding order statistics: $t_{(1)}^{*b}(\hat{c}) \leq t_{(2)}^{*b}(\hat{c}) \leq ... \leq t_G^{*b}(\hat{c})$. From the set of B permutations, the expected order statistics of $t_g(\hat{c})$ can be estimated by $\bar{t}_{(g)}(\hat{c}) = \sum_{b=1}^{B} \frac{t_g^{*b}(\hat{c})}{B}$. Any gene g that is such that its $t_g(\hat{c})$ substantially exceeds its $\bar{t}_g(\hat{c})$value is possibly differentially expressed.

This can be examined further by plotting the $\bar{t}_g(\hat{c})$ values versus $\bar{t}_g(\hat{c})$ values. The central part of this plot lies along the identity line, where $\bar{t}_g(\hat{c}) = t_g(\hat{c})$, indicating genes that are not differentially expressed; the ends tail away from this line. The further a gene is located from the identity line, the more likely it is that the gene is significantly differentially expressed.

The procedure to declare significance is as follows:

for a fixed threshold, Δ, starting at the origin and moving up to the right, find the first i_1 genes such that $t_g(\hat{c}) - \bar{t}_g(\hat{c}) > \Delta$ and call all genes past i_1 "significant positive". Similarly, starting at origin and moving down to the left, find the first i_2 genes such that $t_g(\hat{c}) - \bar{t}_g(\hat{c}) > -\Delta$ and call all genes past i_2 "significant negative". For a given value of Δ, call the smallest value of $t_g(\hat{c})$ among the significant positive genes the "upper cut point", $cut_{up}(\Delta)$, and the largest value of $t_g(\hat{c})$ among the significant negative genes the "lower cut point", $cut_{low}(\Delta)$. This process can be carried out for a series of Δ values.

To determine the number of falsely significant genes generated by SAM, for a chosen value of Δ, count how many genes are reported as differentially expressed in each permutation, that is $t_g^{*b}(\hat{c}) > cut_{up}(\Delta)$ or $t_g^{*b}(\hat{c}) < cut_{low}(\Delta)$ (false positives). Then, calculate the median number of false positives across all the B permutations. Finally calculate the FDR as the number of false positives divided by the number of genes declared significant differentially expressed, having compared $t_g(\hat{c})$with $cut_{low}(\Delta)$ and $cut_{up}(\Delta)$. By evalu-

ating FDR for several values of Δ, a suitable strategy can be devised to decide which genes are significantly differentially expressed (see [84]).

3.3.2 Broberg's modification of SAM method (SAMroc)

Broberg [17] propose to modify SAM methods to optimize with respect to false-positive and false-negative rates. The idea is to jointly minimize the number of genes that are falsely declared positive and the number of genes falsely declared negative (FP, FN) by optimizing over a range of values of the significance level a and the fudge factor c. How well this is achieved can be judged by a receiver operating characteristics (ROC) curve, which displays the number of false positives against the number of false negatives expressed as proportions of the total number of genes.

Assume that we can, for every combination of values of the significance level α and the fudge constant c, calculate (FP, FN). The goodness criterion is then formulated in terms of the distance of the points (FP, FN) to the origin (which point cor- responds to no false positives and no false negatives, see figure 3.3), which in mathematical symbols may be put as

$$C = \sqrt{FP^2 + FN^2} \tag{3.2}$$

The optimal value of (α, c) will be the one that minimizes equation 3.2. It is for practical reasons not possible to do this minimization over every combination, so the suggestion is to estimate the criterion over a lattice of (α, c) values and pick the best combination.

This method (SAMroc) implements, also, the possibility to calculate a uncorrected p-values for each gene. The data matrix X has genes in rows and arrays in columns. Consider the vector of group labels fixed. The permutation method consists of repeatedly permuting the columns (equivalent to rearranging group labels), thus obtaining the matrix $X*$, and calculating the test statistic for each gene and each permutation. Let $d(j)^{*k}$ be the value of the statistic of the jth gene in the kth permutation of columns. Then the p-value for gene i equals

$$P_i = \frac{\#\left\{d(j)^{*k} : \left|d(j)^{*k}\right| \geq d(j)\right\}}{B \times M} \tag{3.3}$$

where M is the number of genes, $d(i)$ the observed statistic for gene i, B the number of permutations and # denotes the cardinality of the set. These are obviously not corrected for multiple comparisons.

Broberg [17]concludes that the proposed method comes out better than or as good as the original SAM statistic in most tests performed. The samroc statistic is robust and flexible in that it can address all sorts of problems that suit a linear model. The methodology adjusts the fudge constant flexibly and achieves an improved performance.

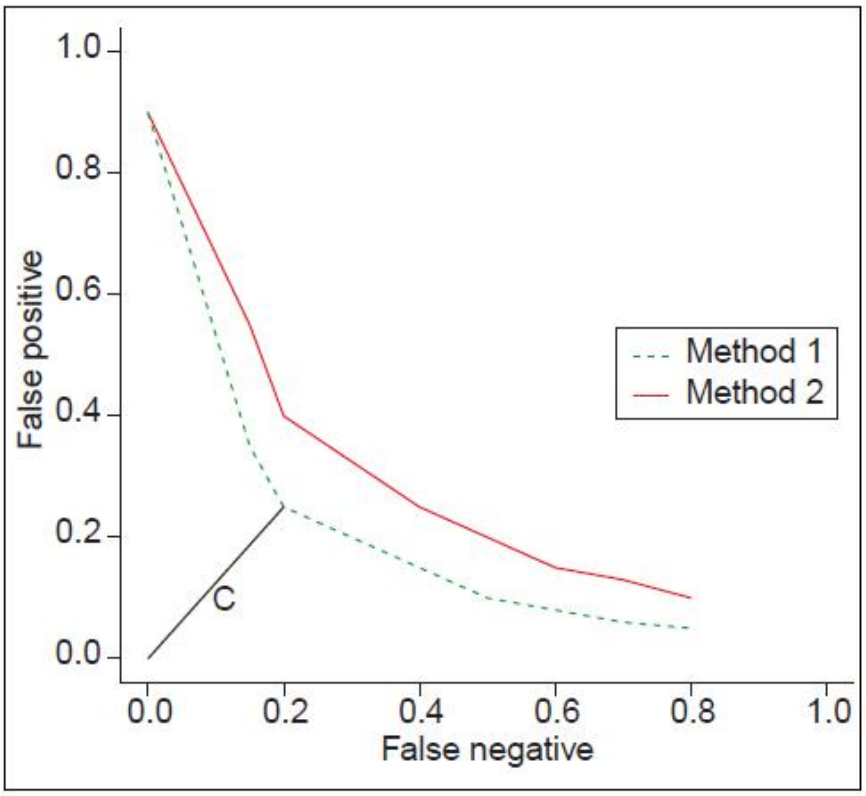

Figure 3.3: ROC curve. This graph displays the number of false negatives (differentially expressed genes (DEGs) not included) versus the number of false positives (non-DEGs included) found on top lists of increasing sizes, expressed as proportions of the total number of genes. The distance C gives an optimal value of equation 3.2. A method whose ROC curve lies below that of another method is preferable, as it will give more DEGs and fewer non- DEGs on any top list of any size, as explained in Additional data. Hence method 1 is preferable to method 2.

The algorithm gives fewer false positives and fewer false negatives in many situations, and was never much worse than the best test statistic in any circumstance. However, a typical run with real-life data will take several hours on a desktop computer. To make this methodology better suited for production it would be a good investment to translate part of the R code, or the whole of it, into C.

Moreover, in the framework to arrange a resampling procedure to have as output a p-value would be an advantage. Also the availability of the R source code (Bioconductor R package called SAGx), permits an easy implementation of this methods in our workflow.

3.3.3 MultiSAM

As stressed in section 3.1 the principal problem of SAM (and in general of modified t statistics) concerned the treatment of uneven sample size data set is that the assumption over the equal variance are likely to be not verified anymore. To correct this pitfall the idea is to reiterate the application of SAMroc[17] analysis made by comparing the less populated class (LPC) of size k with a series of random sampling[24] combination from the more populated class (MPC) of size n, each of the same size of the LPC.

Recalling topics of combinatorial mathematics, $k-combination$ (or $k-subset$) is a subset with k elements. The number of k-combinations (each of size k) from a set S with n elements (size n) is the binomial coefficient (also known as the "choose function"):

$$C_k^n = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{3.4}$$

Being exhaustive, applying SAMroc analysis with all the sampled combinations, is prohibitive because of its extremely time consuming implications. Moreover there are no guarantees of improving the gained information. To have an idea, the number of possible combinations, in the case of IGHV3-21 data set (13 LPC samples and 52 MPC samples) C_{13}^{52} is greater the 10^{11}, so to explore all the combinations means to perform more than 10^{11} SAMroc analysis . Empirical solutions in the field of random sampling [24] suggest us the possibility to stop the random sampling at a certain level(fixed at the moment to 10^3) and to accompany the gained information with some quality scores.

In the following more details will be given about these scores.

3.3.3.1 List Score

This quality score aims to filter the inter-combination reproducibility, in the following we explained how it is build.

One of the output of each SAMroc analysis (performed between the LPC and a random sampled subset of MPC) is a list of genes and their related p-values. To obtain the List Score (LS) proceed as follows:

1. For each list of 1000, Correct the p-values with False Discovery Rate (FDR) method (see subsection 2.3.4 on page 49).
2. Order each list accordingly to their adjusted p-values.
3. Fix a quite conservative level of confidence for the adjusted p-values (a value of 10^{-3} seems to be a reasonable choice).
4. With the fixed cut-off we obtain for each of the 1000 SAMroc analyses, a list of differentially regulated genes
5. Check for each genes (that was part of the input data matrix) its presence in the lists o differentially regulated genes.
6. At the end of the check each genes will be labeled with a value ranging from 0 to 1000. based on the times that this probe was present in the 1,000 lists as differentially regulated. The higher the list Score the higher the the probability of this gene to be really differentially expressed between LPC vs. MPC.

To asses the significance of these LSs is important to perform a resample based mock experiment. I.e., considering the original data matrix, assign randomly the class labels and perform all the previous steps (from 1 to 6) for this new mock data set. This allow to obtain list of genes that are differentially expressed by chance accordingly with the previous method. The Maximum Mock List Score (MMLS) will be ten times the greater List Score obtained from the mock experiment and will be a LS cut-off for to select differentially regulated genes.

3.3.3.2 Median Log Fold Change

This filter resemble the strategy adopted Fold Change(see 2.3.2 on page 45) analysis The Median Log Fold Change (MLFC) is calculated as the $median(log_2(FC_g)$. In our experience, as the Operon platform with a reference design have very low level of Signal to Noise ratio, the popular cut-off of fold change greater than 2 (i.e. log fold change greater than 1) is too optimistic. Thus a clustering-guided choice will be made, evaluating the ability of the filtered genes to separate the two classes.

3.3.3.3 Gene Stability Score

Once the genes with a List Score greater than MMLS are identified an important point is to evaluate the stability of the identified lists. For each To this aim a measure similar to relative errore is kept, defined as:

$$GSS = \frac{MAD(r_g)}{\sqrt{median(r_g)}} \tag{3.5}$$

where $MAD(r_g)$ is the median absolute deviation of ranks of the gene, calculated over all the 1000 lists. The GSS of a gene is considered valid if is greater than the 20^{th} percentile of GSS's distribution.

3.4 Benchmarking MultiSAM with simulated data

In order to asses the efficacy of this algorithm, the first point is to use simulated data sets to compare the performance of this new method comparing it with other standard algorithms. For the comparison we choose to compare the MultiSAM performance with:

- one linear model based approach, LIMMA[74]
- one standard modified t statistics, implemented in SAMroc [84].

3.4.1 Simulation of microarray data

It is impossible to model gene expression data precisely since the true nature of such data is not well understood. It is possible however to capture enough of the nature of the data to perform meaningful tests of the algorithms described above. Public data sets elucidate many of the properties of expression data, for example that one-channel data intensities are exponentially distributed, or that gene intensities are not normally distributed as is often assumed (see [44]). The model is intended to have enough flexibility to elucidate how the results depend on many different parameters.

In this dissertation the model described by Grant et al.[44] was used as our purpose is to test high level analysis. In this model, a *run* will mean the generation of K data sets. Each data set will consists of n replicates in group 0 and m replicates in group 1, each replicate being the intensity levels of N "genes".

For each run, M genes are picked to be differentially expressed. For each of the $N-M$ other genes, a mean intensity level $\mu(g)$ is chosen from a fixed distribution. The distribution type and its parameters are specified in a configuration file. In this model they implemented the exponential (as is typical of one channel data), or a beta (to model intensities for ratios in two-channel data), however any distribution could be used.

Once a mean intensity is chosen, the distribution of intensities for each gene is modeled with a beta distribution. Beta distributions were used because unlike Gaussian distributions, they have finite range, and their shape can be varied widely by adjusting the two parameters, $\alpha(g)$ and $\beta(g)$.

The parameters $\alpha(g)$ and $\beta(g)$ are chosen uniformly in a range $[a;b]$, where a and b are set in the configuration file and are fixed for each run. Any desired percentage of the $\alpha(g)$'s and $\beta(g)$'s can be chosen so that the distribution is symmetric. The left endpoint $L(g)$ of the range of the distribution is chosen uniformly in $[0, \mu(g)]$. The right endpoint

is then set to be $2\mu(g) - L(g)$. This allows for a heterogeneous set of gene intensity distributions.

The parameters are fixed throughout the run, so that the intensities for a given gene in a given group are generated by the same distribution for each replicate of each of the K data sets. For the non-differentially expressed genes the same distribution is used for all replicates, regardless of which group it is in. Most of the parameters for the M differentially expressed genes are chosen by hand, and are not randomly generated (however the intensities are randomly generated in each replicate, according to the chosen distributions).

This gives a mechanism for generating as many replicates as desired of a given "experiment," with the differentially expressed genes known a priori. This data can then be used to test the performance of algorithms. Any algorithm that claims to determine differentially expressed genes from array data should be expected to work on this simulated data. If they cannot, then they should not be expected to work well on real data. Of course if they do work on this data, that is no guarantee they will work on real data exactly as well, given that real data will inevitably have more complexities that we will manage to model. The model described above generates gene intensities in a gene-independent fashion. In reality, gene intensities will contain many dependencies between them, and algorithms for predicting differential expression might take these dependencies into account. The step-down method described above, for example, is designed to exploit gene dependencies. As a simple test of such claims, the model has been extended to generate intensities with dependence in the following extreme way: if there are G non-differentially expressed genes, and A is chosen to be less than G then A genes are generated independently, and each of the remaining $G - A$ genes intensities are taken to be equal to some one of the A independent genes. For example gene 100 might be linked to gene 350, so that they always have the same expression level for every replicate. This is a very strong kind of dependence, so that if dependencies are affecting the performance of an algorithm, it should become apparent when varying the parameter A from 0 to G.An example of two genes generated with the software of Grant et al. [44] are reported in figure3.4 ù

With this model simulate 20 two-color microarray data sets and 20 different one-color microarray data sets was simulated, each data sets was forced to be unbalanced: 15 samples of group 1 vs. 65 samples of group 2 (with 10000 features/genes), is the chosen size in order to resemble the dimensionality of the IGHV3-21 data set.

We apply to all these forty data sets the three different algorithms: LIMMA, SAM-roc, MultiSAM in order to investigate the ability to identify the differentially expressed genes. In table 3.1 summary of performance in terms of average sensitivity and average specificity (over the 40 sample) is reported.

As showed in the table, the performance of this three methods, even if the data sets are unbalanced, are quite satisfactory and comparable. Is important to stress however

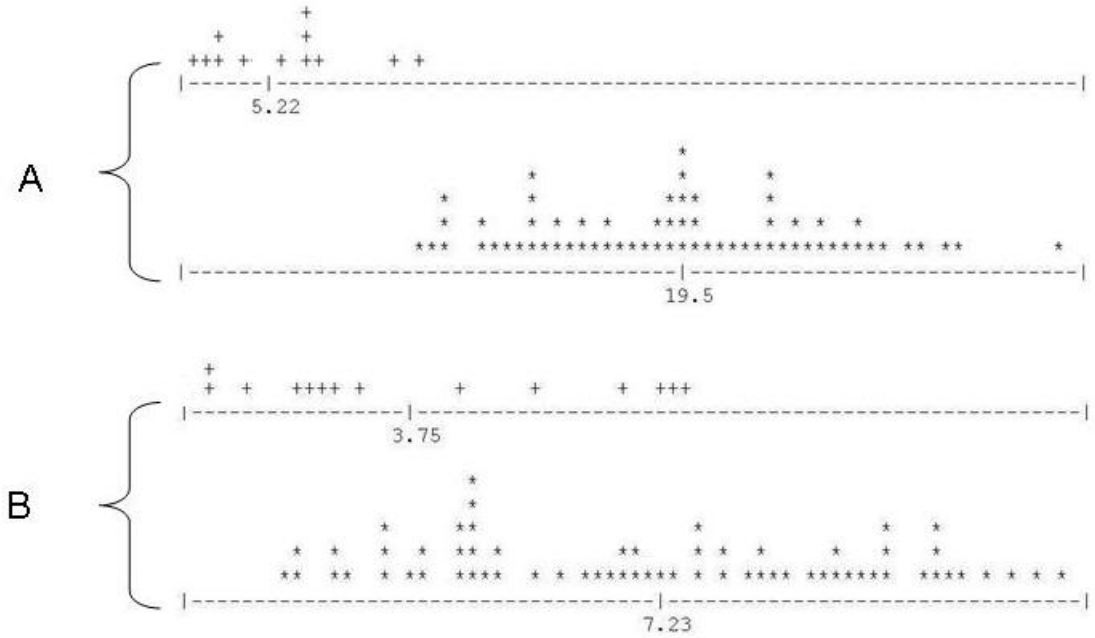

Figure 3.4: Example of gene plot generated with the software of Grant et al.[44]. A) An example of distribution of probably significant differential expression (in an unbalanced simulation). B) An example of a simulated gene with no probably no significant differential expression.

	MultiSAM	LIMMA	SAMroc
Sensitivity	0.7 ± 0.1	0.67 ± 0.2	0.7 ± 0.3
Specificity	1.0 ± 0.1	1.0 ± 0.2	0.97 ± 0.1

Table 3.1: Summary of performance of LIMMA, SAMroc, MultiSAM over 40 simulated data sets in terms of average sensitivity and average specificity.

that this kind of simulation usually cannot reveal the robustness of the method against the noise.

3.5 Results on real IGHV data set

In an analogous fashion LIMMA, SAMroc and MultiSAM was applied to the IGHV3-21 pre-processed data matrix (28513 genes $\times$ 65 samples). For pre-processing details see Appendix.

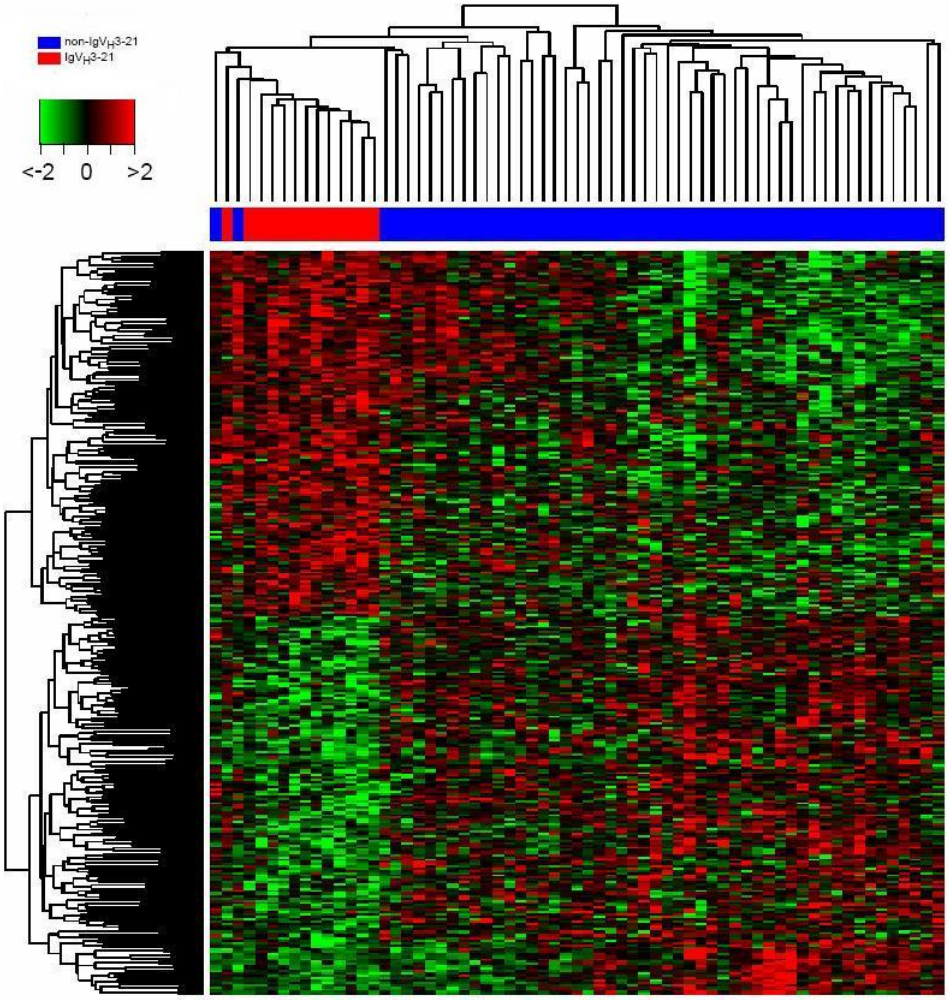

Figure 3.5: Supervised clustering (euclidean distance and average linkage method) of the 319 probes with a List Score >300. The class separation is not satisfactory.

LIMMA Applying LIMMA with a cut-off of 10^{-3} for FDR-adjusted p-values, no gene

result significantly changed. This is a well known problem of linear models with high-noisy data sets [85].

SAMroc Applying SAMroc with a cut-off of 10^{-3} for FDR-adjusted p-values, only 23 genes (over 28513 features) resulted as differentially expressed. Trying to visualize the 23 genes by means of supervised hierarchical clustering[36] methods no one of the different methods explored could separate the two classes (figures not shown).

MultiSAM Applying MultiSAM as described in previous section. From the mock experiment, a value for the Maximum Mock List Score, MMLS, of 30 is obtained so the cut-off for the LS is 300. Multi-SAM find 319 probes with List Score > 300. Application of supervised hierarchical clustering algorithms to those 319 probes results in a not satisfactory separation of the two classes, as shown in figure. So the next step is to filter over use the median log fold change. Setting a cut off for the absolute value of the median log fold change equal to 0.7, the analysis end up with 122 probes which are able to separate the to classes with different algorithm, in figure the heatmap generated with Ward's hierarchical cluster (see 4 on page 54) algorithm with euclidean distance is represented. All the 122 probes selected with this method showed a Gene Stability Score (GSS) less than 1.2, i.e. the 20^{th} percentile of the GSS' s distribution.

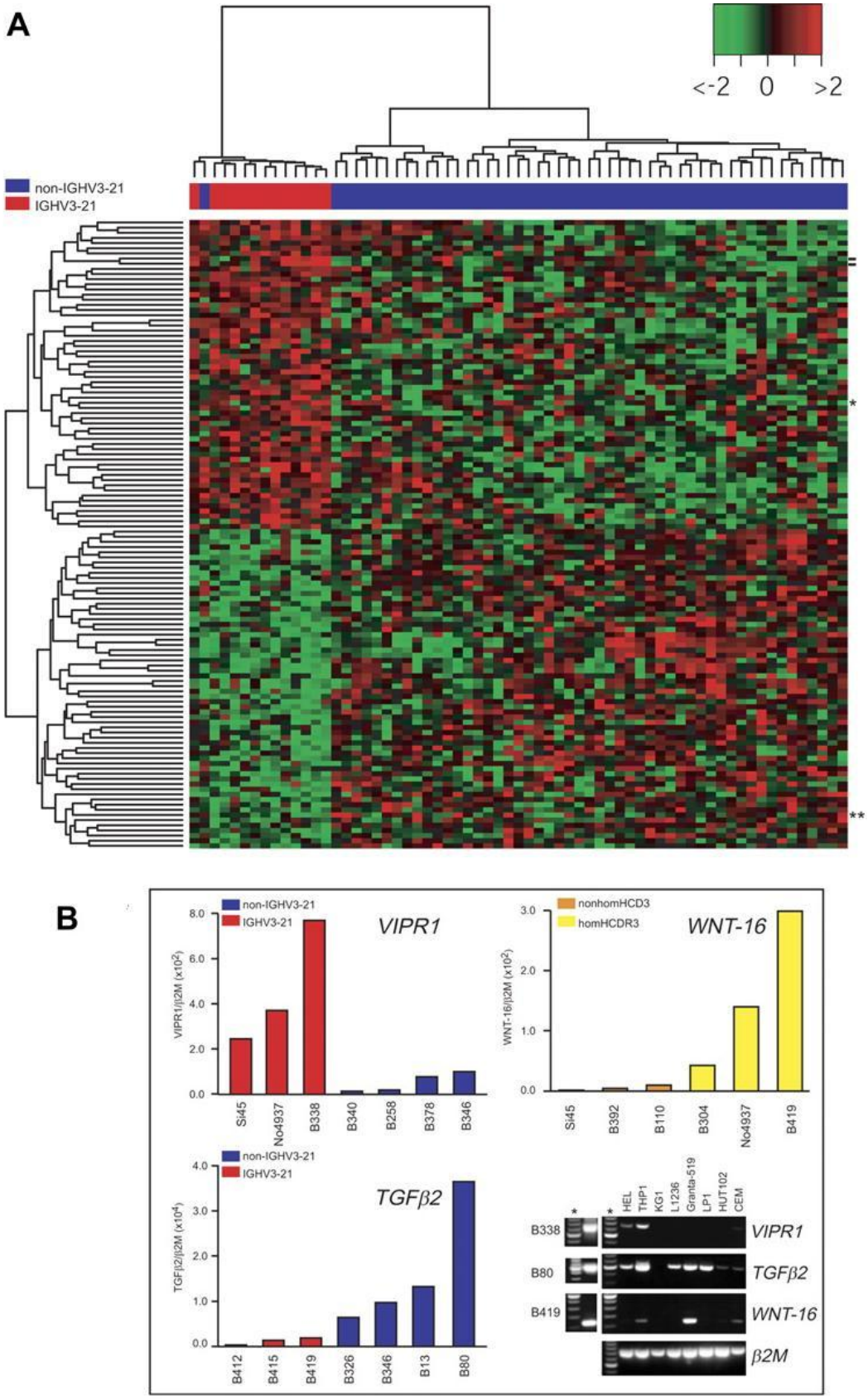

Figure 3.6: A)Ward's Hierarchical clustering of the 122probes selected by MultiSAM (with LS>300 and MLFC<0.7).

3.6 Discussion and Validations

MultiSAM (with LS>300 an MLFC>0.7) select 122 probes (115 genes) correlated with IGHV3-21 usage by CLL: 60 probes were up-regulated and 62 down-regulated in IGHV3-21 CLL. A hierarchic clustering generated by using these 122 best correlated probes clearly split the 13 IGHV3-21 CLLs from all but 1 of the other non–IGHV3-21 cases (figure 3.6). Not surprisingly, IGLV3-21(V2- 14) was the gene up-regulated in IGHV3-21 CLLs with the highest score. In this regard, by checking the portion of heat maps corresponding to both copies of IGLV3-21(V2-14) genes (figure 3.6, bottom), an almost complete consistency between molecular and hybridization data could be detected.

To elucidate the biologic functions of genes representing the differential expression signature of IGHV3-21 CLL, the gene identifiers for the 122 best-correlated probes were linked to 2 web-based bioinformatic tools for global analysis of gene function: "Onto-Express" and "Gene-Ontology Tree Machine." [89, 51, 3] Biologic process and molecular functions showing a significant enrichment for the genes found to be differentially expressed between IGHV3-21 and non–IGHV3-21 CLLs are summarized in 3.7.

As an example, a group of 4 genes (RERG, ABI1, PMP22, TGFB2), all identified as negative regulators of cell proliferation, was selected among genes down-regulated in IGHV3-21 CLL cells along with a regulator of transcription (RUNX1). Conversely, genes upregulated in IGHV3-21 CLLs were involved in positive regulation of cell proliferation (VIPR1), negative regulation of cell adhesion (RND1), as well as regulation of transcriptional (BRCA1) and oxidoreductase activities (HSD17B3, CYP3A5).

The differential expression of 2 of these genes, 1 up-regulated (VIPR1) and 1 down-regulated (TGFB2) in IGHV3-21 CLLs, was investigated by real-time quantitative PCR (QRT-PCR) in representative CLL samples from both CLL subsets. As shown in Figure 5C, these experiments confirmed a higher expression of VIPR1, along with lower levels of TGFB2 transcripts, in IGHV3-21 CLL. Of note, both VIPR1- and TGFB2-specific mRNAs were detected in other human cell lines of different hematopoietic origin, although expression of VIPR1 seemed to be more restricted than that of TGFB2 (Figure 3.6).

Taken together, present data and data reported by Falt et al. [39] collectively demonstrate that the TGF growth inhibition pathway is significantly hampered in IGHV3-21 CLL.

By reviewing the list of the 122 probes differentially expressed by IGHV3-21 CLL,1 we found SEPT10 among the 62 probes (59 genes) significantly downregulated [7]. Down-regulation of SEPT10 in IGHV3-21 CLL was an unexpected finding, because SEPT10 expression is considered a poor prognostic marker and IGHV3-21 CLL typically has poor clinical outcome. To investigate this issue in detail, a quantitative real-time PCR (QRT-PCR) procedure was specifically devoted to investigate the expression levels of SEPT10

GO Category*	ID	Gene Name†	Gene Symbol	UP/DOWN‡	OE§	GOTM‖
Negative regulation of cell proliferation	NM_032918	RAS-like, estrogen-regulated, growth inhibitor	RERG	Down	x	x
	NM_005470	Abl-interactor 1	ABI1	Down	x	x
	NM_000304	Peripheral myelin protein 22	PMP22	Down	x	x
	NM_003238	Transforming growth factor, beta 2	TGFB2	Down	-	x
Transcriptional activator activity	NM_007295	Breast cancer 1, early onset¶	BRCA1	Up	x	-
	S76346	Runt-related transcription factor 1 (aml1 oncogene)	RUNX1	Down	x	-
Secretin-like receptor activity	NM_004624	Vasoactive intestinal peptide receptor 1#	VIPR1	Up	x	x
	NM_000823	Growth hormone releasing hormone receptor	GHRHR	Down	x	x
Oxidoreductase activity	NM_000197	Hydroxysteroid (17-beta) dehydrogenase 3	HSD17B3	Up	x	-
	AK055879	Cytochrome P450, family 3, subfamily A, polypeptide 5	CYP3A5	Up	x	-
Negative regulation of cell adhesion	NM_014470	Rho family GTPase 1	RND1	Up	x	-

* gene ontology (GO) categories with significantly enriched gene numbers have been determined by assaying the gene set of interest, i.e. the 122 genes found to be differentially expressed in IGHV3-21 vs non-IGHV3-21 CLLs, with two web-based bioinformatics tools for global analysis of gene function, i.e. "Onto-Express" (OE) and "Gene-Ontology Tree Machine" (GOTM);

† genes names are those reported in the Operon 2.0 GAL (gene array layout) file; if possible, gene names were updated according to the "Onto-Translate" and "Onto-Miner" tools available in the context of the "Onto-Tools" open access software suite for analysis of microarray dataset (http://vortex.cs.wayne.edu/Projects.html);

‡ UP/DOWN, upregulated/down-regulated in IGHV3-21 CLLs, as reported in Fig. 5A and Table S1;

§ the "Onto-Express" (OE) tool is available in the context of the "Onto-Tools" open access software suite for analysis of microarray dataset;

‖ the "Gene-Ontology Tree Machine" (GOTM) bioinformatics tool is available online (http://genereg.ornl.gov/gotm/);

¶ also identified in the GO category "Regulation of programmed cell death";

also identified in the GO category "Positive regulation of cell proliferation".

Figure 3.7: Summary of GeneOntolgy analysis as in supplemental material of [13]

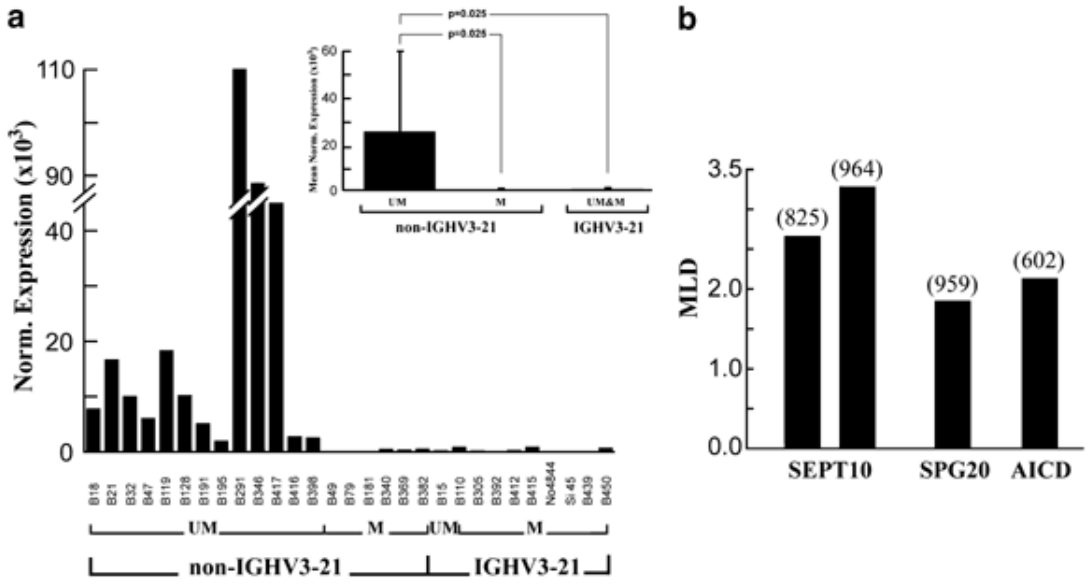

Figure 3.8: SEPT10 transcript levels in CLL expressing or not expressing the IGHV3-21 gene segment. For more details see [7]

in highly purified CLL cells from a series of IGHV3-21 (10 cases) and non-IGHV3-21 (19 cases) CLL, with either UM (14 cases) or M (15 cases) IGHV gene configuration (Figure 1a).

Therefore, overall, QRT-PCR experiments were consistent with GEP results, indicating the lack of SEPT10-specific transcripts in the poor prognosis CLL subset expressing the IGHV3-21 gene.

The above extra-assay validation procedures demonstrate that MultiSAM approach seems to be the only one able to reveal subtle differences in GEP on real unbalanced data set, that have been overlooked by standard algorithms like SAM and LIMMA.

Chapter 4

Assessing Gene Expression Similarities

4.1 The Open Problem: Similarities and Three Class Comparison

In two-class gene expression profile experiments the attention is always focused at exploit the differential expression between the two conditions. Facing problems with more than two classes request a different kind of approach. As seen in subsection 2.3.5 problems with more than two classes can be faced with the ANOVA or model-based approach, but in this case we can just obtain the bulk of genes that are affected at least by one of the factors (or conditions) involved in the experiments, but we cannot discover how they are related to each others.

During the discussion with clinicians at Clinical and Experimental Hematology Research Unit (CRO) it appears quite clear that today in oncological research the attention is focusing on hybrid of poor-characterized classes with severe prognosis, and many efforts has to be invested in characterize this kind of pathologies, that sometimes stand in the middle between two well clinical characterized class.

The question posed by clinicians sounds like: suppose that there are three classes A, B, and C. Class A is really far from class B from a clinical or a molecular point of view. On the contrary patients of class C result as "something in the middle" between A and B patients. Typically class C presents several aspects that make it difficult to characterize it:

- From a clinical point of view, often clinicians and biologists could encounter objective difficulties in doing the classification. Tumor grade is an example of this kind of problems.
- From a molecular point of view sometimes this classes are not yet well known or characterized; markers gene are yet known and mechanisms of development are not

known too. Pathologies with a particular type of mutational status could be an example of this kind of classes.

So the clinician's question is: The gene expression profiling of class C is more similar to GEP of class A or to GEP of class B? With which level of confidence?

At the moment, even if the scientific comunity demands to solve this issue with a statistical answer there are few publications that face this problem with a systematic approach that can give to the clinician also a statistical level of confidence about the answer.

The approaches applied since today are essentially clustering oriented. However, as seen in section 2.4, clustering algorithms are too sensitive with respect to many different variables (for example normalization, gene selection methods ecc.) and even if bootstrapping validation can give a statistical significance, this methods remain too empirical.

Starting from this considerations, supervised machine learning approach could have been a better solution for the robustness; but most of them, having a binary classification output, are not able to associate a statistical level of confidence in case of a test set with samples taken from classes not explicitly related to the classes used for the training. Another essential problem remains the dimensionality of the training samples, with a small number of training samples .

The only reliable approach was presented in Klein et al. [53, 19, 18]. They tried to overcome the intrinsic limitation of supervised classification methods, developing a new classification algorithm that is partially described in [19, 18]. This algorithm claims to able to associate a p-value (to express the level of confidence) to each sample of the test set. In figure you can see the figure reported in [53] that show the significance level for each sample in the test set.

The algorithm is implemented in a software called Gene@Work. This software is available as a stand alone package at the web address indicated in [57]. However, the feature that allow to calculate the p-value for the samples in the test set is not implemented in last version of Gene@Work (neither in older ones). Contacting the authors they told that is impossible to obtain this feature at the moment, more over the source code of Gene@Work is not available. For us is then impossible to use the software for our purposes. Re-implement the software from the beginning is difficult because the papers are not sufficiently detailed. In this case considerations about the availability of source code (see 2.6.3) appears quite obvious.

In the next sections a clustering and machine learning approach will be used to show their potentiality to solve the open question respectively with a small data sets and with a bigger one.

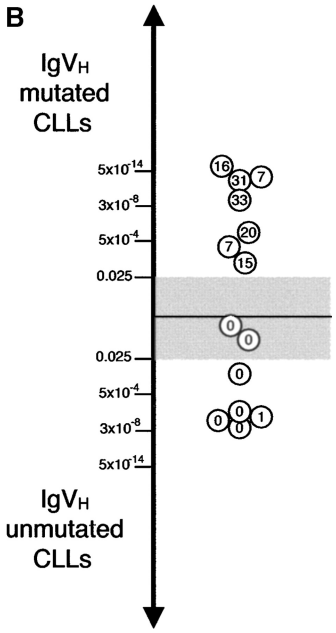

Figure 4.1: Gene@Work output as presented in the Klein et al. gene expression profile of B-CLL

4.2 Supervised Clustering Approach in a Three-Class Problem

At *Clinical and Experimental Hematology Research Unit* in Aviano they started to collect samples of Chronic Lymphocitic Leukemia of B cells (CLL) characterized by a peculiar deletion of chromosome 17 (17p minus). The interest for this particular kind of CLL derive from its extremely severe prognosis and the lack of knowledge about its molecular portrait, inrespectively of the mutational status (M CLL and UM CLL) as described in section 3.2. The study will aim at investigate if the gene expression profiling GEP) of 17p minus CLLs (that usually are characterized by an UM mutational status) is different from the GEP of non-17p-minus UM CLLs, and in this case compare the identified GEP with the one of non-17p-minus M CLLs. At the moment the data sets consists of :

- 9 17p-minus UM CLLs
- 9 non-17p-minus UM CLLs
- 7 non-17p-minus M CLLs

The collection of samples is still in progress from different centers in Italy, however at the moment it seems too small (details about the data sets in Appendix) to be handled with machine learning approach [86].

For an exploratory study we apply a supervised clustering technique

First, to identify the differentially regulated genes between 17p-minus UM vs non-17p-minus UM we apply SAMroc analysis[17] as described in section 3.3.2. A cut off for the FDR adjusted p-values was set at 10^{-3}, with this cut off 300 probes resulted as differentially expressed.

As in the Agilent platform only some genes have more then one probes (see for discussion sub section 1.3.2.4), to avoid any functional or literature bias, we decide to maintain only one probe for each gene selecting the probe (for genes with multiple replicates) with the best adjusted p-values. This selection end up with 286 genes that are used to perform a bootstrapped supervised hierarchical clustering over the entire data set to evaluate the behavior of 7 non-17p-minus patients. In figure 4.2observe one of the obtained clustering (Ward's method with Pearson's Correlation).

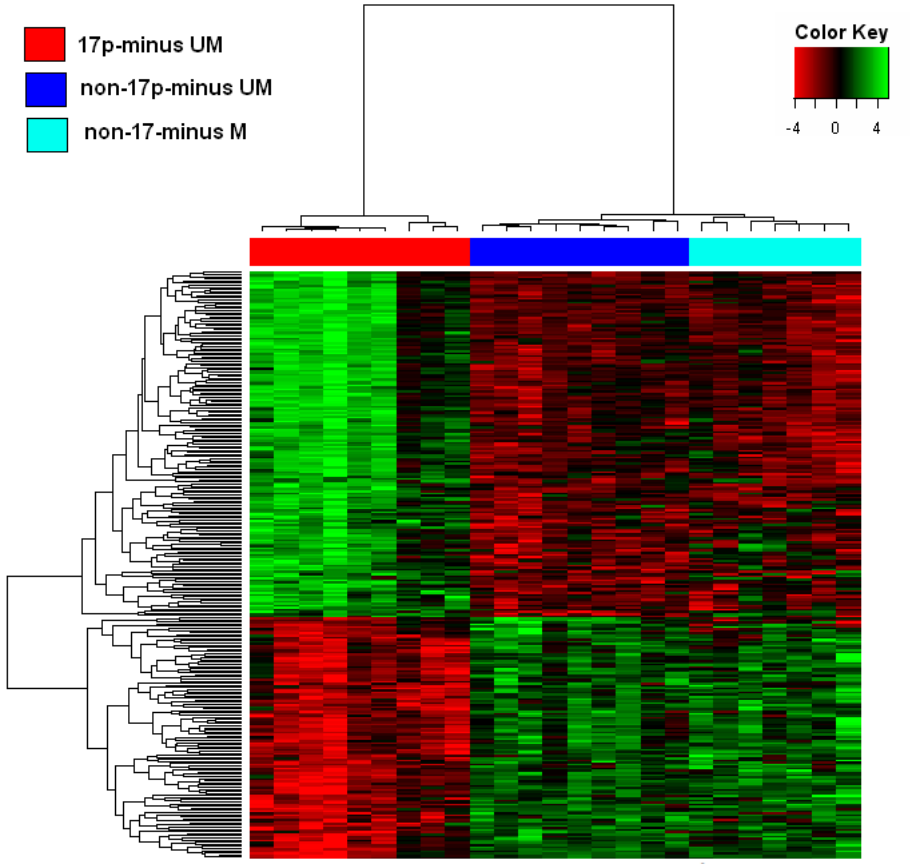

Figure 4.2: Differential gene expression profiling of 9 17p-minus UM CLLs with respect 9 non-17p-minus UM CLLS, clustered to evaluate the relative behavior of non-17p-minus M CLLs.

4.2.1 Discussion

The clustering suggest that 17p-minus CLLs seems to have, despite their mutational status (UM), a gene expression profile that stands apart from other type of CLLs (UM and M). Is important to stress that a clustering technique (for the mentioned intrinsic characteristics), when try to evaluate the grade of similarities give only some hypothesis of work that biologists and clinicians must validate with intensive wet-laboratory work.

Moreover using Onto-Express [51] tools, chromosomal position of the 286 genes are identified. In figure is shown the position of the genes identified: as an internal control some of the genes are located on chromosome 17.

Further validations and collection of more samples are actually in progress at Clinical and Experimental Hematology Research Unit in Aviano.

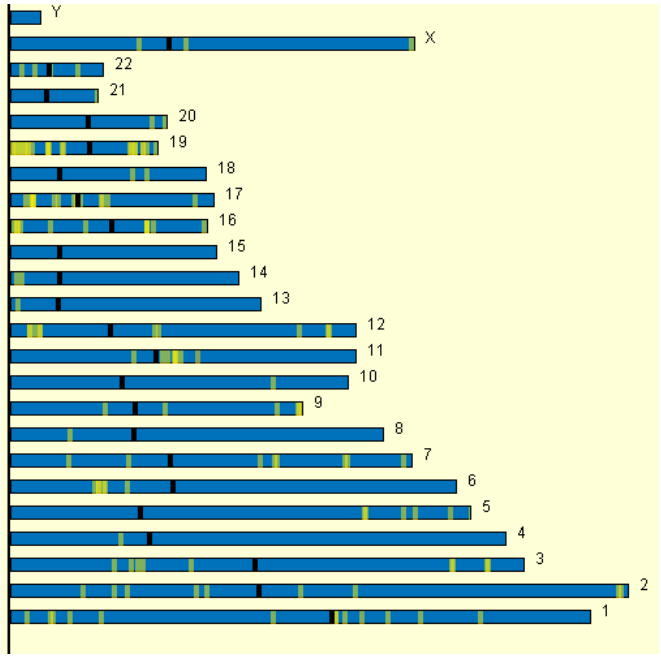

Figure 4.3: Chromosomal mapping of the 286 genes of gene expression profiling of 17p-minus CLLs

4.3 Looking at similarities through Relevance Vector Machine

Relevance vector machines (RVM) have recently attracted much interest in the research community because they provide a number of advantages. They are based on a Bayesian formulation of a linear model with an appropriate prior that results in a sparse rep-

resentation. As a consequence, they can generalize well and provide inferences at low computational cost. In this section a brief introduction of the Relevance Vector Machine (RVM) [82] will show its potentiality to solve some of the limitation sof standard supervised classification methods in order to investigate the similarities in a three class problem.

4.3.1 Introduction

Linear models are commonly used in a variety of regression problems, where the value $t_* = y(x_*)$ of a function $y(x)$ needs to be predicted at some arbitrary point x_*, given a set of (typically noisy) measurements of the function $t = \{t_1, ..., t_N\}$ at some training points $X = \{x_1, ..., x_N\}$:

$$t_i = y_i(x) + \varepsilon_i \tag{4.1}$$

where ε_i is the noise component of the measurement.

Under a linear model assumption, the unknown function $y(x)$ is a linear combination of some known basis functions $\phi_i(x)$, i.e.,

$$y(x) = \sum_{i=1}^{M} w_i \phi_i(x), \tag{4.2}$$

where $w = (w_1, .., w_M)$ is a vector consisting of the linear combination weights. Equation 4.1 can then be written in vector form as :

$$t = \Phi w + \varepsilon, \tag{4.3}$$

where ϕ is an $N \times M$design matrix,whose $i-th$ column is formed with the values of basis function $\phi_i(x)$ at all the training points, $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)$ is the noise vector.

Assuming independent, zero-mean, Gaussian distribution for the noise term, i.e. $\varepsilon_i \sim N(0, \sigma^2)$, the maximum likelihood estimate for $w = (w_1, ..., w_M)$ is given by:

$$w_{OLS} = arg_w min\left(\|t - \Phi w\|^2\right) = \left(\Phi^T \Phi\right)^{-1} \Phi^T t, \tag{4.4}$$

which is also known as the ordinary least square (OLS) estimate. In many applications, the matrix $\Phi^T\Phi$ is often ill-conditioned, and the OLS estimate suffers from over-fitting, which is typical with maximum likelihood estimates. In order to overcome this problem, constraints are commonly introduced on the parameters $w = (w_1, ..., w_M)$, which are used to imply specific desired properties of the estimated function. The Bayesian methodology provides an elegant approach to define such constraints by treating the parameters as random variables, to which suitable prior distributions are introduced. For example, preference for smaller weight values, which can lead to desirable smooth

function estimates, can be specified by assigning a zero-mean, Gaussian distribution to the weights:

$$p(w) = N(w \mid 0, \lambda I). \tag{4.5}$$

Here, the variance parameter λ is adjusted according to the learning problem in order to achieve good results.

Another desirable property of the unknown function, developed more recently, is sparseness, in which the least number of basis functions are desired in the function representation, while all the other basis functions are pruned by setting their corresponding weight parameters to zero. Sparseness property is useful for several reasons. First, sparse models can generalize well and are fast to compute. Second, they also provide a feature selection mechanism which can be useful in some applications.

There exist different methodologies for sparse linear regression, including least absolute shrinkage and selection operator (LASSO) [1],[2] and support vector machines (SVM) [3]. In a Bayesian approach such as RVM, sparseness is achieved by assuming a sparse distribution on the weights in a regression model. Specifically, RVM is based on a hierarchical prior, where an independent Gaussian prior is defined on the weight parameters in the first level, and an independent Gamma hyperprior is used for the variance parameters in the second level. This results in an overall student t prior on the weight parameters, which leads to model sparseness. A similar Bayesian methodology to achieve sparseness is to use a Laplacian prior [5], which can also be considered as a two-level hierarchical prior, consisting of an independent Gaussian prior on the weights and an independent exponential hyperprior on their variances.

4.3.2 RVM Theory

4.3.2.1 Multi-kernel Relevance Vector Machine

Relevance vector machine (RVM) is a special case of a sparse linear model, where the basis functions are formed by a kernel function ϕ centered at the different training points:

$$y(x) = \sum_{i=1}^{N} w_i \phi(x - x_i) \tag{4.6}$$

While this model is similar in form to the support vector machines (SVM), the kernel function here does not need to satisfy the Mercer's condition, which requires ϕ to be a continuous symmetric kernel of a positive integral operator.

4.3.2.2 Sparse Bayesian Prior

A sparse weight prior distribution can be obtained by modifying the commonly used Gaussian prior in 4.5, such that a different variance parameter is assigned for each weight:

$$p(w \mid \alpha) = \prod_{i=1}^{M} N\left(w_i \mid 0, \alpha_i^{-1}\right) \tag{4.7}$$

where $\alpha = (\alpha_1, ..., \alpha_M)$ is a vector consisting of M hyperparameters, which are treated as independent random variables. A Gamma prior distribution is assigned on these hyperparameters:

$$p(\alpha_i) = Gamma(a, b) \tag{4.8}$$

where a and b are constants and are usually set to zero, which results in a flat Gamma distribution. By integrating over the hyperparameters, we can obtain the "true" weight prior weight prior $p(w) = \int p(w \mid a)\, p(a)\, da$. The above integral gives a student-t prior, which is known to enforce sparse representations, owing to the fact that its mass is mostly concentrated near the origin and the axes of definition.

4.3.2.3 Bayesian inference

Assuming independent, zero-mean, Gaussian noise with variance β^{-1}, i.e.,

$$\varepsilon \sim N\left(0, \beta^{-1} I\right) \tag{4.9}$$

we have the likelihood of the observed data as

$$p(t \mid w, \alpha, \beta) = N\left(t \mid \Phi w, \beta^{-1} I\right), \tag{4.10}$$

where Φ is either a $N \times N$ or an $N \times (N * M)$ "design" matrix for the single and multikernel cases respectively. This matrix is formed by all the basis functions evaluated at all the training points, i.e., $\Phi = [\phi(x_1), ..., \phi(x_n)]^T$.

In order to make predictions using the Bayesian model, the parameter posterior distribution $p(w, \alpha \mid t)$ needs to be computed. Unfortunately, it cannot be computed analytically owing to its complexity, and approximations have to be made. Following the procedure described in [82], we decompose the parameter posterior as:

$$p(w, \alpha, \beta \mid t) = p(w \mid t, \alpha, \beta)\, p(\alpha, \beta \mid t) \tag{4.11}$$

Then, the posterior distribution of the weights can be computed as

$$p(w, \alpha, \beta \mid t) = \frac{p(t \mid w, \beta)\, p(w \mid \alpha)}{p(t \mid \alpha, \beta)} \sim N(w \mid \mu, \Sigma), \tag{4.12}$$

where $\Sigma = \left(\beta \Phi^T \Phi + A\right)^{-1}$, $\mu = \beta \Sigma \Phi^T t$ and $A = diag(\alpha_1, ..., \alpha_M)$.

The posterior of the hyperparameters $p(\alpha, \beta \mid t)$ cannot be computed analytically and approximates by the delta function and its mode:

$$p(\alpha, \beta \mid t) \approx \delta(\alpha_{MP}, \beta_{MP}). \tag{4.13}$$

We can find α_{MP} and β_{MP} by maximizing $p(\alpha, \beta \mid t) \propto p(t \mid \alpha, \beta)\, p(\alpha)\, p(\beta)$ as:

$$\alpha_{MP} = \arg\max_{\alpha} \left(p(t \mid \alpha, \beta)\, p(\alpha)\right) \tag{4.14}$$

and

$$\beta_{MP} = \arg\max_{\beta} \left(p(t \mid \alpha, \beta)\, p(\beta)\right) \tag{4.15}$$

The term $p(t \mid \alpha, \beta)$ is known as the marginal likelihood or type-II likelihood[32] and is computed by marginalizing the weights:

$$p(t \mid \alpha, \beta) = \int\int\int p(t \mid w)\, p(w \mid \alpha)\, dw \tag{4.16}$$

which yields

$$p(t \mid \alpha, \beta) = N\left(0, \beta I + \Phi A^{-1} \Phi^T\right). \tag{4.17}$$

An alternative approach is to follow the variational Bayesian methodology to obtain an approximation to the posterior parameter distribution $p(w, \alpha \mid t)$. This is demonstrated in [32], but it is concluded that the method achieves only slightly improved results at significant additional computations.

4.3.2.4 Marginal Likelihood Optimization

The optimization problem in 4.14 cannot be solved analytically and an iterative method has to be used. Instead of maximizing the hyperparameter posterior, it is equivalent, and more convenient, to minimize its negative log likelihood [82] which for the multikernel case is:

$$L(\alpha) = -\frac{1}{2}\left[\log|C| + t^T C^{-1} t\right] + \sum_{m=1}^{M}\sum_{i=1}^{N}\left(a \log \alpha_{mi} - b\alpha_{mi}\right) + c \log \beta - d\beta, \tag{4.18}$$

where $C = \beta I + \Phi A^{-1} \Phi^T$. This equation when $M = 1$ gives the single kernel case.
Setting the derivative of L(α) to zero gives the following iterative formula:

$$\alpha_{mi}^{new} = \frac{1 + 2a}{\mu_{mi}^2 + \Sigma_{(mi)(mi)} + 2b}, \tag{4.19}$$

where mi μ_{mi} is the $mi - th$ element of the posterior mean weight and $\Sigma_{(mi(mi))}$is the $mi - th$ diagonal element of the posterior weight covariance. At each iteration, both are

evaluated from $\Sigma = \left(\beta\Phi^T\Phi + A\right)^{-1}$ and $\mu = \beta\Sigma\Phi^T t$ using the current estimate of α_{MP} . Similarly, the following formula can be obtained for the variance parameter:

$$\beta = \frac{N - \sum_{m=1}^{M}\sum_{i=1}^{N}\left(1 - \alpha_{mi}\Sigma_{(mi)(mi)}\right) + 2c}{\|t - \Phi\mu\|^2 + 2d} \tag{4.20}$$

Computation of Σ requires $O\left((NM)^3\right)$ computations, which can be very demanding for models with many basis functions. During the training process, basis functions whose corresponding weights are estimated to be zero may be pruned. This will make matrix Σ smaller after a few iterations, and its inversion will be easier. However, there are M basis functions initially at each point, and computation of Σ is time consuming. It is interesting to note that the iterative updates for the hyperparameters in equation 4.19and 4.20 can also be derived using an expectation-maximization (EM) algorithm by treating the weights w as hidden variables and the observations t and the hyperparameters α and β as observed variables.

4.3.3 RVM for Classification

Similar to regression, RVM has also been used for classification. Consider a two-class problem with training point $X = \{x_1, ..., x_N\}$ and corresponding class labels $t = \{t_1, ..., t_N\}$ with $t_i \epsilon \{0, 1\}$. Based on the Bernoulli distribution, the likelihood (the target conditional distribution) is expressed as:

$$p\left(t \mid w\right) = \prod_{i=1}^{N} \sigma\left\{\left(y\left(x_i\right)\right)\right\}^{t_i}\left[1 - \sigma\left\{\left(y\left(x_i\right)\right)\right\}\right]^{1-t_i}, \tag{4.21}$$

where $\sigma\left(y\right)$ is the logistic sigmoid function:

$$\sigma\left(y\left(x\right)\right) = \frac{1}{1 + \exp\left(-y\left(x\right)\right)}. \tag{4.22}$$

Unlike the regression case, however, the marginal likelihood $p\left(t \mid \alpha\right)$ can no longer be obtained analytically by integrating the weights from equation 4.21, and an iterative procedure has to be used.

Let α_i^* denotes the maximum a posteriori (MAP) estimate of the hyperparameter α_i . The MAP estimate for the weights, denoted by w_{MAP} , can be obtained by maximizing the posterior distribution of the class labels given the input vectors. This is equivalent to maximizing the following objective function:

$$J\left(w_1, ..., w_N\right) = \sum_{i=1}^{N} \log p\left(t_i \mid w_i\right) + \sum_{i=1}^{N} \log p\left(w_i \mid \alpha_i\right), \tag{4.23}$$

where the first summation term corresponds to the likelihood of the class labels, and the second term corresponds to the prior on the parameters w_i . In the resulting solution, only those samples associated with nonzero coefficients w_i (called relevance vectors) will contribute to the decision function.

The gradient of the objective function J with respect to w is:

$$\nabla J = -A^* w - \Phi^T (f - t) \tag{4.24}$$

where $f = [\sigma(y(x_1)) ... \sigma(y(x_N))]^T$, matrix Φ has elements $\phi_{i,j} = K(x_i, x_j)$. The Hessina of J is

$$H = \nabla^2 (J) = -\left(\Phi^T B \Phi + A^*\right) \tag{4.25}$$

where $B = dig(\beta_1, ..., \beta_n)$ is a diagonal matrix with $\beta_i = \sigma(y(x_i))[1 - \sigma(y(x_i))]$.

The posterior is approximated around $w_{MAP}$by a Gaussian approximation with covariance

$$\Sigma = -(H \mid_{w_{MAP}})^{-1} \tag{4.26}$$

and mean

$$\mu = \Sigma \Phi^T B t. \tag{4.27}$$

These results are identical to the regression case (14) and the hyperparameters α_i are updated iteratively in the same manner as for the regression case.

4.3.3.1 Comparison to SVM Learning

SVM is another methodology for regression and classification that has attracted considerable interest [32]. It is a constructive learning procedure rooted in statistical learning theory [86], which is based on the principle of structural risk minimization. It aims to minimizing the bound on the generalization error (i.e., the error made by the learning machine on data unseen during training) rather than minimizing the empirical error such as the mean square error over the data set [3]. This results in good generalization capability and an SVM tends to perform well when applied to data outside the training set.

In the context of classification, an SVM classifier in concept first maps an input data vector x into a higher dimensional space $\mathcal{H}$ through an underlying nonlinear mapping $\Phi(x)$, then applies linear classification in this mapped space. Introducing a kernel function $K(x, y) \equiv \Phi(x)^T \Phi(y)$, we can write an SVM classifier $f_{SVM}(x)$ as follows:

$$f_{SVM}(x) = \sum_{i=1}^{N_s} \alpha_i K(x, s_i) + b \tag{4.28}$$

where s_i ,$i = 1, ...N_s$ are a subset of the training samples x_i, $i = 1, ..., N$ (called support vectors). The SVM classifier in 4.28 resembles in form the RVM classifier in equation 4.6, yet the two classifiers are derived from different principles. For SVM the support vectors are typically formed by "borderline", difficult-to-classify samples in the training set [32], which are located near the decision boundary of the classifier; in contrast, for RVM the relevance vectors are formed by samples appearing to be more representative of the two classes, which are located away from the decision boundary of the classifier.

Compared to SVM, RVM is found to be advantageous on several aspects including:

1. The RVM decision function can be much sparser than the SVM classifier, i.e., the number of relevance vectors can be much smaller than that of support vectors;

2. RVM does not need the tuning of a regularization parameter (C) as in SVM during the training phase.

3. RVM classification incorporate a posterior probability estimation (more details in subsection 4.3.3.2) Essentially similarities classification in gene expression, as addressed in this dissertation, take advantage of this feature.

As a drawback, however, the training phase of RVM typically involves a highly nonlinear optimization process.

4.3.3.2 Posterior Probabilities in Classification with RVM

When performing 'classification' of some test example x , we would prefer the model to give an estimate of $p\left(t_* \epsilon \mathcal{C} \mid \mathbf{x}_*\right)$, the posterior probability of the example's membership of the class $\mathcal{C}$ given the features $\mathbf{x}_*$. This quantity expresses the uncertainty in the prediction in a principled manner while facilitating the separation of "inference" and "decision" [32]. In practical terms, these posterior probability estimates are necessary to correctly adapt to asymmetric misclassification costs (which nearly always apply in real applications) and varying class proportions, as well as allowing the rejection of 'uncertain' test examples if desired.

Importantly, the presented Bayesian classification formulation incorporates the Bernoulli output distribution (equivalent in the log-domain to the 'cross-entropy' error function), which in conjunction with the logistic sigmoid squashing function, enables $\sigma\left\{y\left(\mathbf{x}\right)\right\}$ to be interpreted as a consistent estimate of the posterior probability of class membership. Provided $y\left(\mathbf{x}\right)$ is sufficiently fleexible, in the in infinite data limit this estimate becomes exact[11].

By contrast, for a test point $\mathbf{x}_*$ the SVM outputs a real number which is thresholded to give a 'hard' binary decision as to the class of the target t_* . The absence of posterior probabilities from the SVM is an acknowledged de deficiency, and a recently

proposed technique for tackling this involves the a posteriori fitting of a sigmoid function to the fixed SVM output $y(x)$ [66] to give an approximate probability of the form $\sigma\{A \cdot y(\mathbf{x}) + B\}$. While this approach does at least produce predictive outputs in the range [0; 1], it is important to realize that this does not imply that the output of the sigmoid is necessarily a good approximation of the posterior probability. The success of this post-processing strategy is predicated on the original output $y(x)$ of the SVM, with appropriate shifting and rescaling, being a good approximation to the "inverse-sigmoid" of the desired posterior probability. This is the quantity $\log\{p(t\epsilon\mathcal{C}_{+1} \mid \mathbf{x})\}$, referred to as the "log-odds". As a result of the nature of the SVM objective function, y(x) cannot be expected to be a reliable model of this. On the basis of the last points in this dissertation RVM was used to estimate posterior probability of the third class in a three.class problem

4.3.4 RVM applied to Tumor Grade Breast Cancer Data Set

The three-class problem was posed to address the specific question of three-class problem posed in the context 17p-minus data sets, however the data collection is not yet finished, and there are not enough samples to engage a training phase with only nine samples per class. So the decision was to investigate the reliability of the RVM approach for class similarities through one bigger data set recovered in microarray repositories. To decide the problem to treat we focus on a popular problem addressed in breast cancer microarray research: investigation of histological tumor grade[77, 61, 76] to improve the prognosis.

The histological grade of breast carcinomas has long provided clinically important prognostic information [37]. However, despite recommendations by the College of American Pathologists that tumor grade be used as a prognostic factor in breast cancer [41], the latest Breast Task Force of the American Joint Committee on Cancer did not include histological tumor grade in its staging criteria, because of insurmountable inconsistencies in histological grading between institutions [72]. Concordance between two pathologists has been investigated and found to range from 50% to 85% (6–9). Although about half of all breast cancers are assigned histological grade 1 or 3 status (with a low or high risk of recurrence, respectively), a substantial percentage of tumors (30%–60%) are classified as histological grade 2, which is not informative for clinical decision making because of intermediate risk of recurrence. This high percentage of histological grade 2 tumors is still observed when grading is performed by a single pathologist. Thus, to increase the prognostic value of tumor grading, refinement of histological grade 2 status, perhaps into low- and high-risk categories, and improvement of the reproducibility of the technique are necessary.

The query "breast tumor" in GEO [6](`GEO;http://ncbi.nlm.nih.gov/geo`) results in 330 different items of which 68 are data sets. Exploring MIAME of each single data set, only 4 of them contains enough patients who are classified for "tumor grade" clinical

variable. At the end we select the data set with the GEO Accession GSE3494 (for more details see the Appendix). The tumor grade label divides the patients in 67 samples of grade I (G1), 54 samples of grade 3 (G3) and 100 samples of grade 2 (G2).

The goal is use RVM to separate G1 from G3, then evaluate the third class G2 as test-set to obtain the probability for samples of G2 to be member of class G1 or class G3.We preprocessed the data using rma algorithm [47]. After we filtered the data through SAM regularized t-test with a significance level for fdr adjusted p-value of 0.01. Then we standardize the data with a Z-score normalization, before applying RVM with a distance kernel.

After the trainig phase with the training G1-G3 training set (67 G1 samples and the 54 G3 samples), using 100 G2 samples as test

For this dataset we obtain a double output for each sample of class G2: the binary classification (classmembership) and the probability for the class-membership. To visualize the output, in figure is reported a *probability classification plot*: the color scheme reveals the strength of class membership whereas the y axis report the level of significance (in term of posterior probability evaluated in RVM through sigmoid function). In x axis there are reported the patients. The yellow lines identify a posterior probability of 50% to be member of the classes , i.e. to say if a sample lies in the area between the two yellow lines the conclusion about class mambership in not trustworthy.

The interesting results from a clinical and biological point of view is that the 90% of G2 samples are classified as G1 with a probability greater then 50%. The 66% of the 100 G2 samples have a probability greater then 90% to be classified as G1. The 10% that are classified as G3 have a probability lower then 40%.

The guessed hypothesis is that breast cancer samples of grade II have a molecular profiling more similar to breast cancer samples of grade I, rather then to breast cancer samples of grade III. Moreover looking at other clinical variables, such as survival outcomes, the patients of grade II classified.as more similar to grade III, are those that show the worst survivals outcomes.

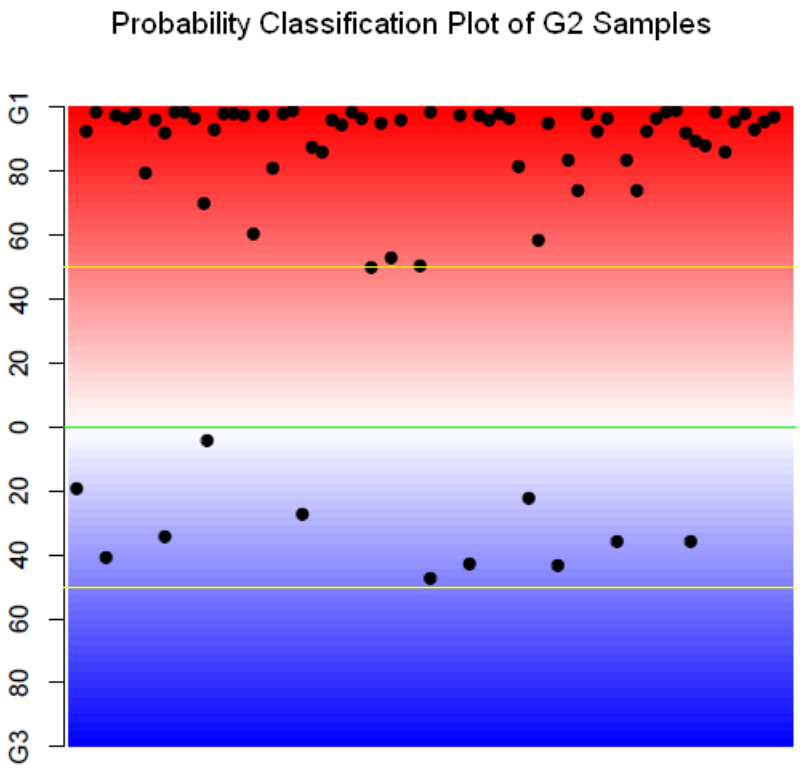

Figure 4.4: Probability Classification plot for samples G2, classified on the basis of G1-G3 training samples. Color scheme reveals the strength of class membership whereas the y axis report the level of significance (in term of posterior probability evalueated in RVM through sigmoid function). If a sample lies in the area between the two yellow lines the conclusion about class membership in not trustworthy

4.3.5 Discussion

The RVM seems to address the tumor grade three-class problem in an original way that can disclose some feature of tumor grade II captured in others gene expression profiling approaches but without scoring the information with a posterior probability estimation. Ma et al. [61]applied a supervised hierarchical clustering approach (similar to the one described in section 4.2), using discriminant analysis, instead of SAM statistics) over a similar structured independent data set. They selects 200 genes as differentially expressed correlating with tumor grade I and tumor grade III. As result they assert:

"This analysis indicates that these 200 genes are all significant at the level of $P < 0.01$. Gene expression values were expressed as ratios and two-dimensional clustering analysis was performed revealing three major gene clusters (figure 4.5). One cluster of genes demonstrates decreased expression in all samples with subtle quantitative differences

between grade I and grade III (green bar). A second cluster of genes (denoted as the grade III signature) shows markedly increased expression in grade III samples (red bar), whereas a third cluster (grade I signature) demonstrates increased expression primarily in grade I samples (blue bar). Most striking is the existence of reciprocal gradients in the intensities of the grade I and grade III signatures (figure . Notably, most grade II lesions exhibit a hybrid of grade I and grade III signatures (e.g., cases 130, 169, and 198). Some grade II lesions, however, show an expression pattern that is most similar to either grade I or grade III lesions (cases 41 and 43, respectively), and a few grade III samples demonstrate coexpression of some genes that are characteristic of the grade I signature (cases 65, 88, and 112)"

But it is reasonable to give another key of interpretation to figure 4.5, above concerning clusters of samples.

Looking at the the dendrogram of samples we notice that two major clustering are observable if we stop to high level branching: the first on the left contains more tumor grade III samples, the other one contain more tumor grade I samples. Tumor grade II fall for the 75% of sample in tumor grade I. This resamble the hypothesis obtained via RVM: tumor grade II is more similar to tumor grade I, but unfortunately clustering in spite of RVM is not to able to provide any statistical significance associated to each samples.

As the data set is part of a repository is impossible to perform any extra assay validation as in Chapter 2, we have done for IGHV3-21 data set. However pushed by this promising performance of RVM, we are looking forward to approach originally data set created by our collaborators at CRO in Aviano to analyze new challenging three-class problems data sets.

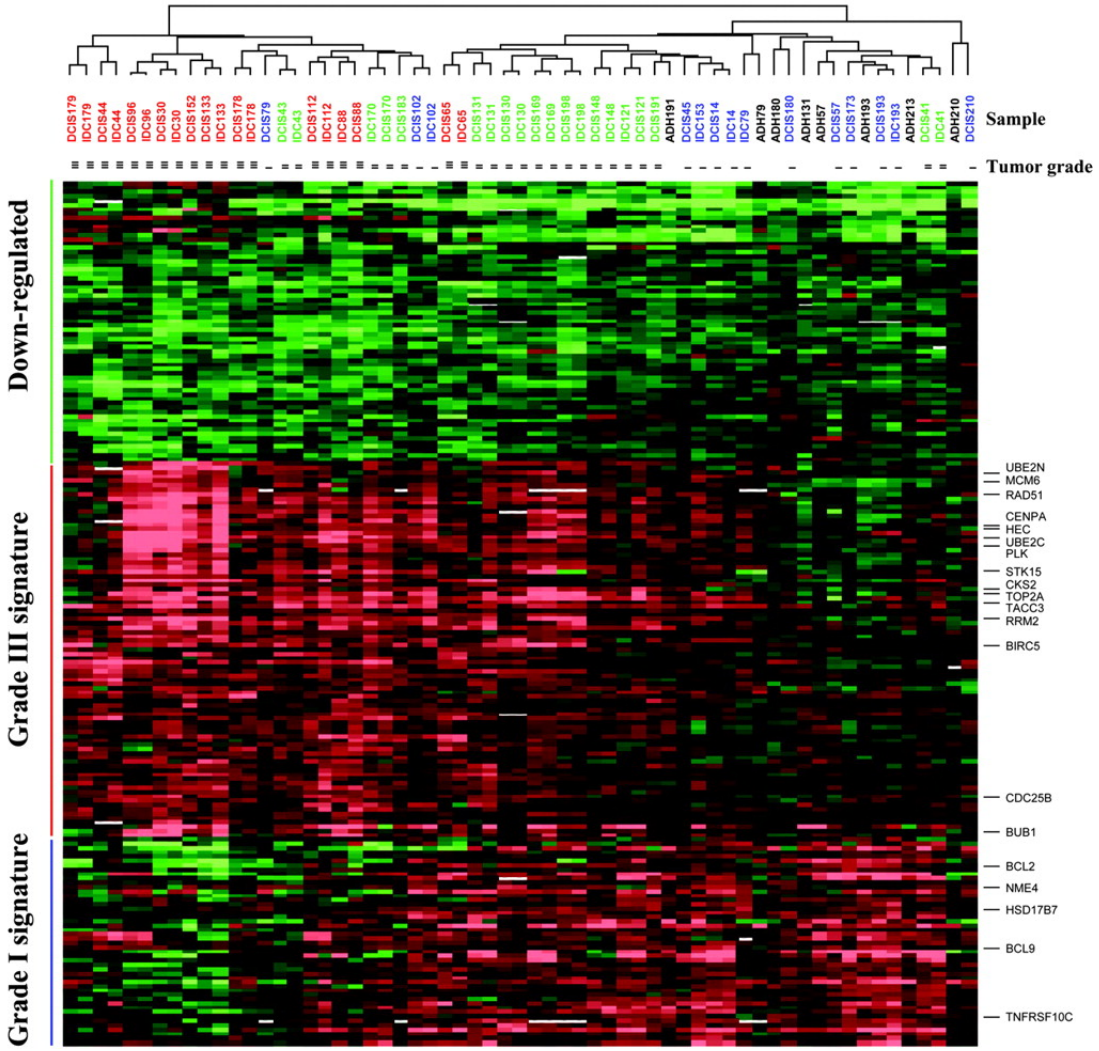

Figure 4.5: Two-dimensional clustering of 61 samples and the top 200 genes correlating with tumor grade. Genes (rows) and samples (columns) were clustered independently by hierarchical clustering. Three main clusters are highlighted by color bars.

OUTLOOK

The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.
[John von Neumann (1903–1957)]

Modeling and data analysis approaches has showed in the last years their potentiality in high-throughput gene expression investigation. One example is cancer diagnosis and prognosis, where a wealth of work has been published. This dissertation has tried to answer some specific questions posed by clinician that do not have yet a compelling methodology able to answer it. The immediate result is a further biological insight into pathologies like chronic lymphocytic leukemia and breast cancer. Most of the results prove to be biological or clinically verifiable, above in case of chronic lymphocitic leukemia.

In the long run the result is to propose and made available to the scientific community these new approaches to face those specific and unresolved issues. To this aim, as member of recently founded (2007) Italian Network for Oncology Bioinformatics (`http://acc.cineca.it/ACC/`), in the next future the source code or a self standing implementation of the methods, will be made available.

Nevertheless challenging questions remain today open and should be addressed in the next future. To give some examples, the integration of a priori functional genomics information into the gene expression modeling and analysis, the establishment of well defined gene regulatory networks related to disease, and finally the integration of clinical and different high-throughput data , are nowadays some of the hottest topics in gene expression that will be addressed in the next future.

Appendix

According to MIAME standarsd, below MIAME details are reported for all the data sets used in this dissertation. Respecting the six point fixed :

- Experimental design
- Array design o array specification for commercial arrays
- Samples: samples used, extract preparation and labeling
- Hybridization procedure and parameters
- Measuremnts :image quantification and specification
- Normalization controls: types, values and specifications.

The IGHV3-21 Data Set

Only the 65 array that passed quality control check performed with ArrayQuality package are included in thios data set. Local Background subtraction was performed. Print-tip Lowess within array normalization and Quantile between array normalization was applied to the median row ntensities. The array contains 21239 gene expression probes in duplicate, prefiltering was applied eliminating 1)all genes whose fluorescence signals were below a given threshold (700 arbitrary units) in at least 80% of the spots at both CY3 and Cy5 channels (bad signal to noise ratio) and 2) all the genes whose interquartile range was below the 20th percentile of distribution (genes whose expression was overall not differentially modulated). Following this filterin 28513 features was left.

All the 65 samples is part of a reference experimental design (labelling and sampling procedure described at point 4). All the 13 IGHV3-21 samples are biological replicates. All the 52 non-IGHV3-21 are bilogical replicates.

The Operon Human Genome Oligo Set Version 2.0 platforms (21,329 70mer, Operon Biotechnologies, Huntsville, AL, USA) were spotted in duplicate onto MicroMax SuperChip I glass slides (Perkin Elmer Life Sciences Inc., Boston, MA) by the MicroCRIBI (University of Padua).

Neoplastic and normal B cells (more than 95% pure), purified by Ficoll-Hypaque gradient (Pharmacia, Uppsala, Sweden) and positive immunoselection with anti-CD19-conjugated immunomagnetic beads (Miltenyi Biotec, Bologna, Italy), were employed as total RNA source. cDNA synthesis, cDNA purification, in vitro transcription of aminoallyl RNA (aRNA), aRNA dye coupling, and purification of dye-labeled aRNA were performed using the Amino Allyl MessageAmp II aRNA Amplification Kit (Ambion, Austin, TX) following manufacturer's guidelines. Cy3/Cy5 was from Amersham Biosciences (Amersham, United Kingdom). Cy3/Cy5 dye incorporation into aRNA yielded incorporation rates of 30 to 60 dye molecules per 1000 nucleotides by spectrophotometric analysis, as requested by the manufacturer. GEP was performed by a dual labeling strategy using Cy3-labeled aRNA from pooled normal PB B cells of healthy donors as common reference and Cy5-labeled aRNA from purified CLL cells as tester. A mixture of 4 μg Cy3-labeled reference aRNA and 4 μg Cy5-labeled aRNA from each CLL was hybridized to Operon Human Genome Oligo Set Version 2.0 platforms (21 329 oligomers of 70 nucleotides; Operon Biotechnologies, Huntsville, AL) and spotted in duplicates onto MicroMax SuperChip I glass slides (Perkin Elmer Life Sciences, Boston, MA). Hybridization was performed for 18 hours at 48C in a buffer containing 5x standard sodium cytrate (SSC), 0.1% sodium dodecyl sulfate (SDS), 25% formamide, and 100 μg/mL salmon sperm DNA using an automated HybArray 12 hybridization system (Perkin Elmer Life Sciences). Following washings, glass slides were analyzed by a Scan Array Lite scanner (Packard BioChip Technologies, Billerica, MA); images and data were analyzed using the GenePix Pro software (Axon Instruments, Foster City, CA).

The 17p-minus Data Set

Only the 25 patients that pass the quality control check made by means of FeatureExtraction QC report are part of the data set. No background subtraction was made on the slides (because the Signal to Noise ratio was extremely low). Lowess Normalization within-array Normalization and Quantile betweeen-array normalization was performed. Agilent's Whole Human Genome Oligo microarray is comprised of approximately 41,000 (60-mer) oligonucleotide probes, some of them has e 10 replicates. All the probes whose interquartile range was below the 20th percentile of distribution (genes whose expression was overall not differentially modulated). After the filtering procedure appriximately 31.000 probes are used in further analysis.

All the 25 samples is part of a reference experimental design (labelling and sampling procedure described at point 4). All the 9 17p-minus UM samples are biological

replicates. All the 9 non-17p-minus UM samples are bilogical replicates. All the 7 non-17p-minus M samples are biological replicates.

Agilent Whole Human Genome Oligo Microarray Kit 4×44K was used in the experiment. For annotation reference refer to `http://www.chem.agilent.com/`. Agilent provides a complete line of RNA isolation, labeling and hybridization reagents as well as Agilent SureHyb-enabled hybridization chambers and accessories that, when used together, enhance the ease-of-use and performance of Agilent's microarrays.

Neoplastic and normal B cells (more than 95% pure), purified by Ficoll-Hypaque gradient (Pharmacia, Uppsala, Sweden) and positive immunoselection with anti-CD19–conjugated immunomagnetic beads (Miltenyi Biotec, Bologna, Italy), were employed as total RNA source. cDNA synthesis, cDNA purification, in vitro transcription of amino-allyl RNA (aRNA), aRNA dye coupling, and purification of dye-labeled aRNA were performed using the Amino Allyl MessageAmp II aRNA Amplification Kit (Ambion, Austin, TX) following manufacturer's guidelines. Cy3/Cy5 was from Amersham Biosciences (Amersham, United Kingdom). Cy3/Cy5 dye incorporation into aRNA yielded incorporation rates of 30 to 60 dye molecules per 1000 nucleotides by spectrophotometric analysis, as requested by the manufacturer. GEP was performed by a dual labeling strategy using Cy3-labeled aRNA from pooled normal PB B cells of healthy donors as common reference and Cy5-labeled aRNA from purified CLL cells as tester. A mixture of μ4g Cy3-labeled reference aRNA and 4 μg Cy5-labeled aRNA from each CLL was hybridized of the Agilent Whole Human Genome Oligo Microarray Kit 4×44K

The Tumor Grade Data Set

The biological tumor samples (ie, breast tumor specimens) consisted of freshly frozen breast tumors from a population-based cohort of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989. Estrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. An experienced pathologist determined the Elston-Ellis grades of the tumors, classifying the tumors into low, medium and high-grade tumors. The clinico-pathological characteristics accompanying each tumor include p53 status, ER status, tumor grade, lymph node status and patient age. All other MIAME details are available at web site `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494&targ=self&form=html&view=quick`

Bibliography

[1] Statistical algorithms description document. *Technical Report, Affymetrix, Santa Clara, CA*, 2001.

[2] Dhammika Amaratunga and Javier Cabrera. *Exploration and Analysis of DNA Microarray and Protein Array Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1 edition, 10 2003.

[3] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.

[4] F. Azuaje and J. Dopazo. *Data Analysis and Visualization in Genomics and Proteomics*. John Wiley and Sons, 2005.

[5] C.A. Ball and A. Brazma. MGED Standards: Work in Progress. *OMICS A Journal of Integrative Biology*, 10(2), 2006.

[6] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, 35:D760–765, Jan 2007.

[7] D. Benedetti, R. Bomben, M. Dal-Bo, D. Marconi, A. Zucchetto, M. Degan, F. Forconi, G. Del-Poeta, G. Gaidano, and V. Gattei. Are surrogates of IGHV gene mutational status useful in B-cell chronic lymphocytic leukemia? The example of Septin-10. *Leukemia*, 22:224–226, Jan 2008.

[8] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[9] F. Bertucci, P. Finetti, N. Cervera, E. Charafe-Jauffret, E. Mamessier, J. Adelaide, S. Debono, G. Houvenaeghel, D. Maraninchi, P. Viens, et al. Gene Expression Profiling Shows Medullary Breast Cancer Is a Subgroup of Basal Breast Cancers. *Cancer Res*, 66(9):4636–4644, 2006.

[10] F. Bertucci, S. Salas, S. Eysteries, V. Nasser, P. Finetti, C. Ginestier, E. Charafe-Jauffret, B. Loriod, L. Bachelart, J. Montfort, et al. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene*, 23:1377–1391, 2004.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 10 2007.

[12] BM Bolstad, RA Irizarry, M. Astrand, and TP Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, 2003.

[13] R. Bomben, M. Dal Bo, D. Capello, D. Benedetti, D. Marconi, A. Zucchetto, F. Forconi, R. Maffei, E.M. Ghia, L. Laurenti, P. Bulian, M.I. Del Principe, G. Palermo, M. Thorsélius, M. Degan, R. Campanini, A. Guarini, G. Del Poeta, R. Rosenquist, D.G. Efremov, R. Marasca, R. Foà, G. Gaidano, and V. Gattei. Comprehensive characterization of IGHV3-21-expressing B-cell chronic lymphocytic leukemia: an Italian multicenter study. *Blood*, 109:2989–2998, Apr 2007.

[14] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480(1):17–24, 2000.

[15] L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001.

[16] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classiÿcation and Regression Trees. *Wadsworth, Belmont, CA*, pages 75–80, 1984.

[17] P. Broberg. Statistical methods for ranking differentially expressed genes. *Genome Biol*, 4(6):R41, 2003.

[18] A. Califano. SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics*, 16(4):341–357, 2000.

[19] A. Califano, G. Stolovitzky, and Y. Tu. Analysis of gene expression microarrays for phenotype classification. *Proc Int Conf Intell Syst Mol Biol*, 8:75–85, 2000.

[20] HG Chew, DJ Crisp, RE Bogner, and CC Lim. Target detection in radar imagery using support vector machines with training size biasing. *Proc. Int. Conf. on Control, Automation, Robotics, and Vision (ICARCV)*, 2000.

[21] C.H. Chung, P.S. Bernard, and C.M. Perou. Molecular portraits and the family tree of cancer. *Nature Genetics*, 32(supp):533–540, 2002.

[22] W.S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[23] W.S. Cleveland and S.J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[24] William G. Cochran. *Sampling Techniques, 3rd Edition.* Wiley, 3 edition, 1 1977.

[25] David Collett. *Modelling Survival Data in Medical Resea.* Routledge, 3 2003.

[26] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[27] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, Aug 1970.

[28] L. Deng, J. Pei, J. Ma, and D.L. Lee. A rank sum test method for informative gene discovery. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 410–419, 2004.

[29] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences (with Student Suite Online).* Duxbury Press, 7 edition, 1 2007.

[30] R. Díaz-Uriarte. Supervised methods with genomic data: a review and cautionary view. *Data analysis and visualization in genomics and proteomics. New York: Wiley*, pages 193–214, 2005.

[31] S. Draghici. *Data Analysis Tools for DNA Microarrays.* Chapman Hall/CRC, 3 edition, 6 2003.

[32] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition).* Wiley-Interscience, 2 sub edition, 10 2000.

[33] S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–88, 2002.

[34] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1):71–103, 2003.

[35] D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. Expression profiling using cDNA microarrays. *Nat. Genet.*, 21:10–14, Jan 1999.

[36] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

[37] CW Elston and IO Ellis. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[38] E. Espinosa, JA Vara, A. Redondo, JJ Sanchez, D. Hardisson, P. Zamora, F.G. Pastrana, P. Cejas, B. Martinez, A. Suarez, et al. Breast Cancer Prognosis Determined by Gene Expression Profiling: A Quantitative Reverse Transcriptase Polymerase Chain Reaction Study. *Journal of Clinical Oncology*, 23(29):7278, 2005.

[39] S. Falt, M. Merup, G. Tobin, U. Thunberg, G. Gahrton, R. Rosenquist, and A. Wennborg. Distinctive gene expression pattern in VH3-21 utilizing B-cell chronic lymphocytic leukemia. *Blood*, 106(2):681, 2005.

[40] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[41] P.L. Fitzgibbons, D.L. Page, D. Weaver, A.D. Thor, D.C. Allred, G.M. Clark, S.G. Ruby, M.F. O'Malley, J.F. Simpson, J.L Connoly, et al. Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Archives of pathology & laboratory medicine(1976)*, 124(7):966–978, 2000.

[42] P. Ghia, K. Stamatopoulos, C. Belessi, C. Moreno, S. Stella, G. Guida, A. Michel, M. Crespo, N. Laoutaris, E. Montserrat, et al. Geographic patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: the lesson of the IGHV3-21 gene. *Blood*, 105(4):1678, 2005.

[43] TR Golub, DK Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, JP Mesirov, H. Coller, ML Loh, JR Downing, MA Caligiuri, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531, 1999.

[44] G. Grant, S. Sokolowski, and CJ Stoeckert Jr. Performance Analysis of Differential Expression Prediction Algorithms Using Simulated Array Data. Technical report, Technical Report, 2005.

[45] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422, 2002.

[46] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, et al. Gene-Expression Profiles in Hereditary Breast Cancer. *New England Journal of Medicine*, 344(8):539, 2001.

[47] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 31:e15, Feb 2003.

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

[49] M.K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77(02):123–128, 2001.

[50] M.K. Kerr, M. Martin, and G.A. Churchill. Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology*, 7(6):819–837, 2000.

[51] P. Khatri, S. Draghici, G.C. Ostermeier, and S.A. Krawetz. Profiling Gene Expression Using Onto-Express. *Genomics*, 79(2):266–270, 2002.

[52] T.J. Kim, J.J. Choi, W.Y. Kim, C.H. Choi, J.W. Lee, D.S. Bae, D.S. Son, J. Kim, B.K. Park, G. Ahn, et al. Gene expression profiling for the prediction of lymph node metastasis in patients with cervical cancer. *Cancer Science*, 99(1):31–38, 2008.

[53] U. Klein, Y. Tu, G.A. Stolovitzky, M. Mattioli, G. Cattoretti, H. Husson, A. Freedman, G. Inghirami, L. Cro, L. Baldini, et al. Gene Expression Profiling of B Cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells &cjs3566; &cjs3566; The online version of this article contains supplemental material., 2001.

[54] E.S. Lander. The new genomics: global views of biology. *Science*, 274:536–539, Oct 1996.

[55] E.S. Lander. Array of hope. *Nat. Genet.*, 21:3–4, Jan 1999.

[56] M.L.T. Lee, GA Whitmore, and R.Y. Yukhananov. Analysis of Unbalanced Microarray Data. *Journal of Data Science*, 1:103–121, 2003.

[57] J. Lepre, JJ Rice, Y. Tu, and G. Stolovitzky. Genes@ Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data. *Bioinformatics*, 20(7):1033–44, 2004.

[58] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.*, 98:31–36, Jan 2001.

[59] Y. Li, C. Campbell, and M. Tipping. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 18(10):1332–1339, 2002.

[60] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nat. Genet.*, 21:20–24, Jan 1999.

[61] X.J. Ma, R. Salunga, J.T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B.M. Zhang, et al. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences*, 100(10):5974, 2003.

[62] C. Matthews, M. Catherwood, TC Morris, and HD Alexander. Routine Analysis of IgVH Mutational Status in CLL Patients using BIOMED-2 Standardized Primers and Protocols. *Leukemia and Lymphoma*, 45(9):1899–1904, 2004.

[63] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, 35:D747–750, Jan 2007.

[64] Giovanni Parmigiani, Elizabeth S. Garett, Rafael A. Irizarry, and Scott L. Zeger, editors. *The Analysis of Gene Expression Data*. Springer, 1 edition, 4 2003.

[65] E. Pennisi. Human genome. Reaching their goal early, sequencing labs celebrate. *Science*, 300:409, Apr 2003.

[66] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

[67] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–27, 2001.

[68] C. Romualdi, S. Campanaro, D. Campagna, B. Celegato, N. Cannata, S. Toppo, G. Valle, and G. Lanfranchi. Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Human Molecular Genetics*, 12(8):823–836, 2003.

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

[70] T.D. Shanafelt, S.M. Geyer, and N.E. Kay. Prognosis at diagnosis: integrating molecular biologic insights into clinical practice for patients with CLL. *Blood*, 103(4):1202, 2004.

[71] R. Simon, MD Radmacher, K. Dobbin, and LM McShane. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute*, 95(1):14–18, 2003.

[72] S.E. Singletary, C. Allred, P. Ashley, L.W. Bassett, D. Berry, K.I. Bland, P.I. Borgen, G. Clark, S.B. Edge, D.F. Hayes, et al. Revision of the American Joint Committee on Cancer Staging System for Breast Cancer. *Journal of Clinical Oncology*, 20(17):3628, 2002.

[73] G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):3, 2004.

[74] GK Smyth. limma: Linear Models for Microarray Data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 2005.

[75] RL Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *BIOINFORMATICS*, 19(12):1484–1491, 2003.

[76] C. Sotiriou, S.Y. Neo, L.M. McShane, E.L. Korn, P.M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A.L. Harris, and E.T. Liu. Breast Cancer Classification and Prognosis Based on Gene Expression Profiles from a Population-Based Study. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18):10393–10398, 2003.

[77] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *jnci*, 98(4):262–272, 2006.

[78] K. Stamatopoulos, C. Belessi, C. Moreno, M. Boudjograh, G. Guida, T. Smilevska, L. Belhoul, S. Stella, N. Stavroyianni, M. Crespo, et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood*, 109(1):259, 2007.

[79] Dov Stekel. *Microarray Bioinformatics*. Cambridge University Press, 1 edition, 9 2003.

[80] C.J. Stoeckert, H.C. Causton, and C.A. Ball. Microarray databases: standards and ontologies. *Nat. Genet.*, 32 Suppl:469–473, Dec 2002.

[81] M. Suárez-Fariñas, M. Pellegrino, K.M. Wittkowski, and M.O. Magnasco. Harshlight: a" corrective make-up" program for microarray chips. *BMC Bioinformatics*, 6(1):294, 2005.

[82] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(211):244, 2001.

[83] G. Tobin, U. Thunberg, K. Karlsson, F. Murray, A. Laurell, K. Willander, G. Enblad, M. Merup, J. Vilpo, G. Juliusson, et al. Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. *Blood*, 104(9):2879, 2004.

[84] V.G. Tusher, R. Tibshirani, G. Chu, et al. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[85] S.A.F.T. van Hijum, A. de Jong, R.J.S. Baerends, H.A. Karsens, N.E. Kramer, R. Larsen, C.D. den Hengst, C.J. Albers, J. Kok, and O.P. Kuipers. A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. *BMC Genomics*, 6:77, 2005.

[86] V.N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[87] I.A. Wood, P.M. Visscher, and K.L. Mengersen. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23:1363–1370, Jun 2007.

[88] Aidong Zhang. *Advanced Analysis of Gene Expression Microarray Data (Science, Engineering, and Biology Informatics)*. World Scientific Publishing Company, 1 edition, 6 2006.

[89] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16, 2004.

[90] H. Zhang and C.Y. Yu. Tree-based analysis of microarray data for classifying breast cancer. *Frontiers in Bioscience*, 7:c63–67, 2002.