Analytische Chemie

Chemometric methods for microarray data analysis and their application to leukemia subtype identification

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Chemie und Pharmazie
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von

Eric Andrés Frauendorfer

aus Caracas

- 2004 -

Dekan: Prof. Dr. J. Leker
Erster Gutachter: Prof. Dr. K. Cammann
Zweiter Gutachter: Prof. Dr. J. von Frese
Tag der mündlichen Prüfungen: 12., 14., 16.07.2004
Tag der Promotion: 16.07.2004

Table of Contents

Abbreviations

A	adenine
ACS	American Cancer Society
ALL	acute lymphocytic leukemia
ALM	auto leveling method
AML	acute myeloid leukemia
ASR	analyte-specific reagent
C	cytosin
CEL	data format of Affymetrix
CELL	a feature on an Affymetrix microarray
CLL	chronic lymphocytic leukemia
CML	chronic myeloid leukemia
CVUA	Chemischen und Veterinär Landesuntersuchungsamt
DBMS	database management system
DML	data manipulation language
DNA	desoxyribonucleic acid
EST	expressed sequence tag
FDA	Food and Drug Administration
FISH	fluorescence in situ hybridisation
G	guanine
GAPDH	glyceraldehyde-3'-phosphate dehydrogenase
GPL	general public license
ICB	Institut für Chemo- und Biosensorik
IM	ideal mismatch
IVAT	in vitro analytical test
LOO	leave one out
MAS	Microarray Suite
MDMS	microarray data management system
MIAME	Minimum Information About a Microarray Experiment
MM	mismatch
MRD	minimal residual disease
NIH	National Institutes of Health
NSF	National Science Foundation

mRNA	messenger RNA
PAM	Prediction analysis of microarrays
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PM	perfect match
PMA	premarket approval
QCM	quartz crystal microbalance
R	a statistics program
RMA	robust multi chip analysis
RNA	ribonucleic acid
SAM	significance analysis of microarrays
SD	standard deviation
SNP	single nucleotide polymorphism
SPR	surface plasmon resonance
SQL	structured query language
SVM	support vector machine
T	thymin
U	uracil
U133A and B	Affymetrix microarray types

1 Introduction

Cancer is the second leading cause of death in the western world after heart disease. Classical cancer treatments, including radiation- and chemotherapy, can have many unwanted side effects, often weakening the patient tremendously and reducing the patient's quality of life [1, 2]. The efficiency of drugs used in chemotherapy, and therefore also the amount of the active agent needed, is greatly influenced by the molecular and biochemical properties of the cancer to be eradicated. Different cancer subtypes contain their own subset of abnormalities in the genetic code which change signaling pathways, create proteins in wrong amounts or even proteins lacking any useful structure [3]. It is therefore essential to choose the right medication for a patient with a certain type of cancer to achieve best results, and also to minimize the amount of the drug that has to be administered.

The overall rate of survival has risen constantly due to improvements in diagnosis and treatment made possible by the advances of cell, molecular, computational, developmental, and structural biology as well as biochemistry, genetics, molecular biophysics, bioinformatics and chemometrics. These once separate fields of activity have molten together in the last couple of years, urged by the need for an interdisciplinary approach to understand the complex patterns behind cancer cell biology. One example for this is the molecular characterization of a tumor to determine which drug or combination of therapies is the most effective for a patient. Research is done with state of the art DNA microarrays [4] (chapter 3.4), increasing the knowledge of the genetic abnormalities responsible for certain subtypes of cancer. This information can then be used to design small, affordable diagnostic microarrays for medical applications. The first DNA based diagnostic biosensors will be approved for use in these fields in the very near future (chapter 3.4.5). These new devices have the promise of helping to further increase the effectiveness of anti-cancer therapies and to increase overall survival rates [5-7].

Once a cancer type is recognized, it has to be treated with the right drug. The use of classical chemotherapy drugs including the classical DNA alkylating agents like cis-platin or triethylmelanine, the antimetabolites like pyrimidin- or pyrin-analogues and enzymes like L-asparaginase is often followed by many side effects as these drugs can also affect normal cells. When cis-platin, the most used chemotherapy agent, enters the cell nucleus, the chloro ligands are substituted by two adjacent guanine bases on a DNA strand. This makes the DNA duplex bend and unwind at the site of cisplatin attachment. The high-mobility-group domain (HMG) proteins then become attached to the structural damage, hereby preventing cancer cell replication [8]. As was reviewed in the *Current Medicinal Chemistry – Anti Cancer Agents*

journals (e.g. [9-11]), targeted therapies using novel drugs reduce side effects as scientists try to design these drugs so that they target properties unique to cancer, e.g. they disrupt certain signaling pathways [12], and thus avoid normal cells. One example is Gleevec™, a drug designed to work against a certain, deadly type of leukemia (CML). It was introduced by Novartis in 2001 and revolutionized the treatment of CML. Gleevec works as a signal transduction inhibitor that interferes with cell signaling pathways in tumor development, blocking the abelson-tyrosinkinase without interfering with other tyrokinases abundant in all cells. Other so called *smart drugs* followed, most of them far less successful. Iressa™ for example, which received approval from the U.S. Food and Drug Administration in 2003, is a drug targeted at the epidermal growth factor receptor (EGFR), a protein involved in cancer cell growth. However, chemotherapy together with Iressa™ did not achieve better results than chemotherapy alone during phase III of clinical trials. This example, among others, has shown that a lot of research is still needed to truly understand the complex machinery of cancer creation and proliferation [13, 14].

The role of chemometrics and bioinformatics in this research is to design and select optimal measurement procedures and experiments and to maximize the information which can be extracted from data. The application of a multi step analysis of the very complex multivariate data gathered in microarray experiments has to be done in an optimal way to obtain informative results that can be used to interpret the biological background of the data. The elements which are defining the speed at which progress is made in the bioanalytics sector are now the analysis and interpretation of this complex data and far less often technical reasons [15]. The optimal application of chemometrics is important during the research, the development and the design of DNA biosensors as well as during analysis of actual samples from patients as it should never be forgotten that the data is gathered using tumour samples obtained from individual cancer patients with their own lives and hopes.

2 Aims & Scope

The aim of this work was to create and enhance chemometric methods for the analysis of microarray data, to apply these methods, in order to obtain information on relevant genes useful for the characterization of different cancer subtypes and to use these genes for the creation of classifiers for the discrimination of unknown cancer samples.

The main focus was put on the development of quality control routines for Affymetrix microarrays, which are the best developed DNA biosensor platform and will probably be the first technology applied in real world diagnostics in the very near future. Routines include quality control procedures for the processing of inhomogeneous background signals and procedures for obtaining information on the genetic traits of pediatric acute lymphocytic leukemia.

Several hundred Affymetrix U133 microarrays were analyzed to create novel methods for the automatic detection of signal artifacts and to process these chips to remove inhomogeneous signal background and differences in signal scaling. These methods were applied on replicate measurements to show their efficiency in raising the quality of the obtained signals. Further, the methods were tested on microarrays with known and unknown defects to evaluate their ability to detect them. Different tools have been created for analysis, management and processing of data. A tool was created to facilitate the design of probe sequences for a custom made microarray.

A pediatric leukemia dataset was analyzed with the intention of selecting genes best suited for discriminating different leukemia subtypes. This process was also used to compare different gene selection algorithms as well as different methods for the calculation of gene specific expression signals.

3 Theory

3.1 Nucleic Acid: Structure and Function

Nucleic acids as carriers of the genetic information can be subdivided into two classes:

- deoxyribonucleic acid (DNA) with the sole purpose of storing information;
- ribonucleic acid (RNA) with a role in gene-expression and protein biosynthesis.

The genomic DNA is located nearly solely in the chromosomes of the nucleus of eukaryotic cells, whereas RNA can be found in the nucleus as well as in the cytoplasm [16].
DNA is an unbranched biopolymer composed of nucleotides and can reach considerable length. Each nucleotide consists of sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. DNA can consist of the two purin-bases adenine (A) and guanine (G) and the two pyrimidine-bases cytosine (C) and thymine (T). RNA contains ribose as sugar component, and the structurally similar uracile (U) instead of thymine. The polymer is created by bondage of the sugar groups through the phosphate groups. The genetic information of the organism is stored in the sequence of the four bases, read in the same direction as it was synthesized, that is, from the 5' to the 3' – end. DNA is synthesized by recurrent attachment of an incoming deoxynucloside triphosphate to the free 3' OH-group of the growing DNA sequence.
DNA has a double helix structure in a natural surrounding, in which two complimentary DNA single strands are wound around the same axis.

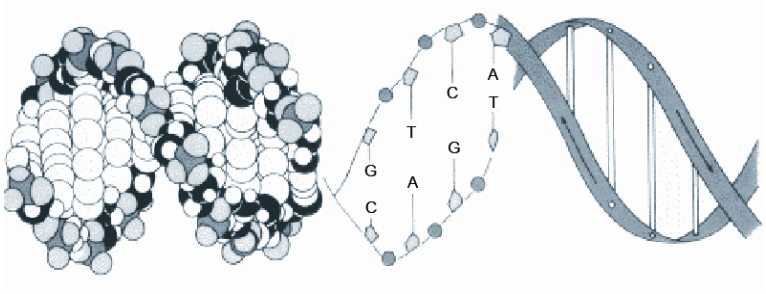


Fig. 3.1 DNA structure [*from chromosome.com*]

The resulting macromolecule has a polar and a negatively charged surface created by the outside lying sugar-phosphate backbone. The interactions between the two DNA strands are

based on hydrogen bonds between the nucleic acid bases adenine and thymine on one hand, and guanine and cytosine on the other.

3.2 Gene Expression

Functional regions of the DNA are called genes, most genes coding proteins. Each amino acid component of a protein is coded by three base pairs in the DNA (codon). The information of a gene is transferred into a messenger RNA (mRNA). The RNA is then transformed multiple times and transferred to the ribosomes. Here it is read and the proteins are synthesized. Proteins are the building blocks of the organic tissues and fluids; they provide most of the molecular machinery. Different genes are expressed in different cell and tissue types and at different developmental stages. Cells at certain locations thus produce the proteins they need at certain times [16].

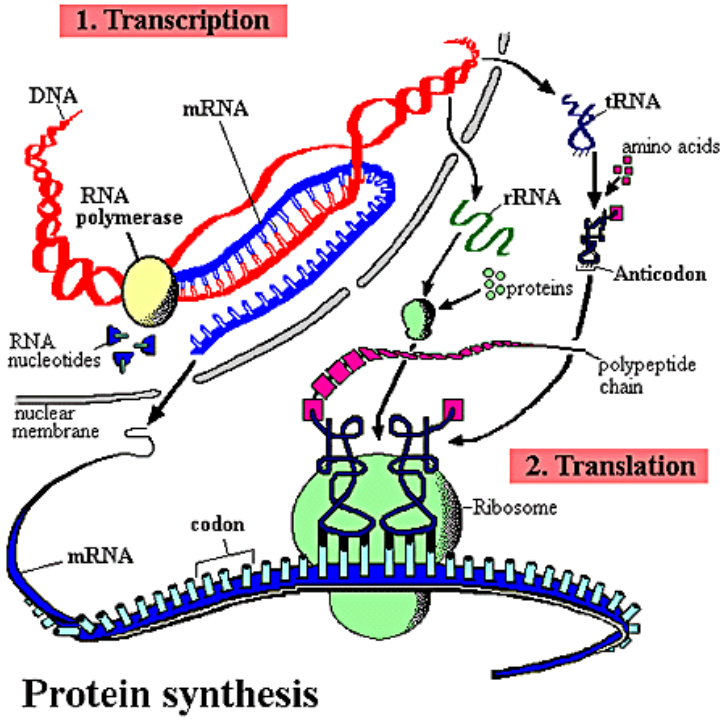


Fig. 3.2 Transcription of information from DNA to mRNA and synthesis of polypeptides in the ribosome [*source: Genentech*].

Analysis of variations in gene expression can lead to an understanding of disease states, targeting of drugs to specific cells, tissues or individuals, development of agricultural products, etc. [17, 18]. Gene expression can be quantified by using microarrays in order to analyze simultaneously the amount of a large multitude of different mRNA in a sample.

This can then be the basis to gain more information on certain tissue types like tumors [19, 20]. More information on nucleic acid analytics is given by Haberhausen *et al*. [21] , Pingoud and Urbanke [22], and Christopoulos [23].

3.3 Cancer

3.3.1 Overview

The word "cancer" is a generic term describing more than 100 forms of the disease that can arise in most tissues [24]. Five general subgroups can be defined [*American Cancer Society, ACS*]:

Carcinoma - a tumor derived from epithelial cells - those cells that line the surface of the skin and the organs, also the surfaces of the digestive tract and the airways. This is the most common cancer type and represents about 80-90% of all cancer cases reported.

Sarcoma - a tumor derived from muscle, bone, cartilage, fat or connective tissues.

Leukemia - a cancer derived from blood cells or their precursors. The cells that form both white and red blood cells are located in the bone marrow.

Lymphoma - a cancer of bone marrow derived from cells that affect the lymphatic system.

Myelomas - a cancer involving the white blood cells responsible for the production of antibodies.

Each form of cancer can have very different properties, although the processes through which these diverse tumors arise are quite similar [25, 26]. The human body consists of 30 trillion cells which work together, they only proliferate when get the signal to do so. It is essential for certain tissues to retain their size and properties in order to work in unison with the rest of the body. Cancer cells, on the other hand, leave this strict ruling, following their own schemes for reproduction. They can leave their site of origin and invade other tissues, often disrupting their function and thus becoming lethal [27]. Every part of the body can develop a primary

tumor, the most prominent regions being the lung (through smoking), the breast in women and the prostate in men (see figure 3.3). The probability of developing a certain cancer type is linked to many different factors like environment, lifestyle, genetic makeup and especially age.

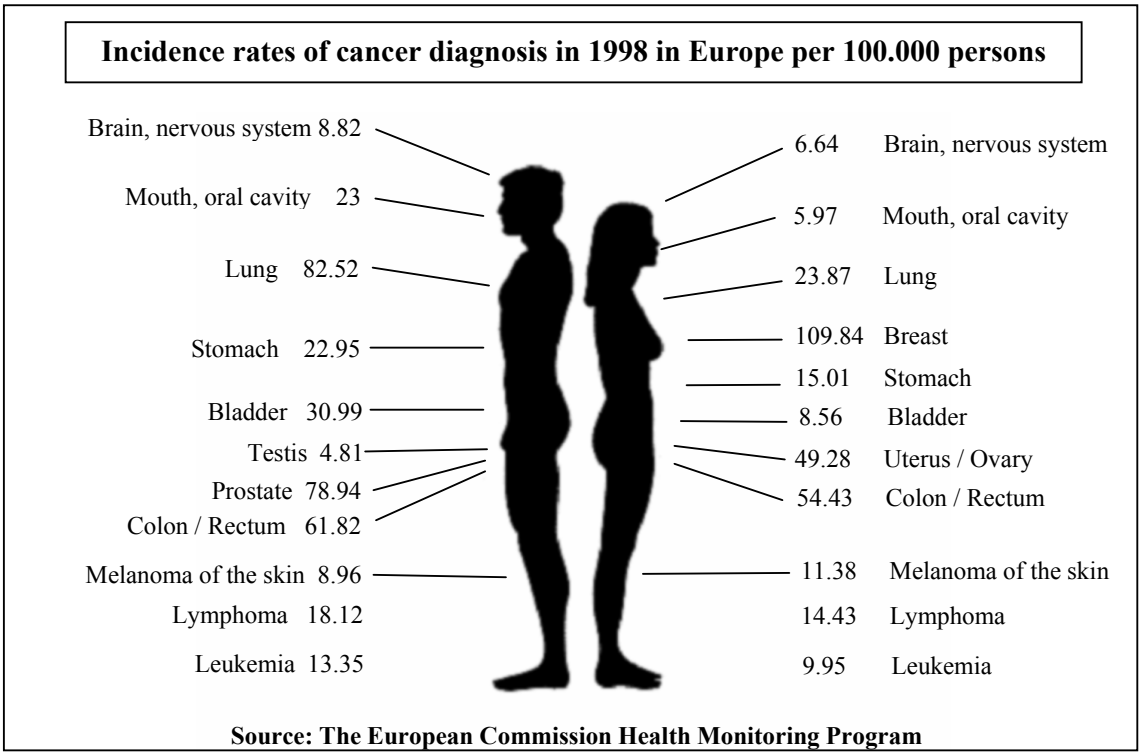


Fig. 3.3 Incidence rates of certain cancers diagnosed in Europe. Rates are calculated by division of the number of new cancer cases observed during the year 1998 by the corresponding number of people in the population at risk. Results expressed as an annual rate per 100.000 persons at risk [28].

These facts have been known for a long time, but the research during the last couple of years has shed light into the molecular sources of these characteristics. It is now common knowledge that the malignant transformation of a cell is the product of an accumulation of mutations of certain genes within it. If a gene is mutated, its function may be disrupted; it may be present at the wrong numbers producing a wrong amount of a protein; it may be located at the wrong area of the genome or it may be missing completely. Proto-oncogenes encourage the proliferation of the cell, tumor suppressor genes inhibit it. Together, these two classes of genes keep the cell in balance, making it a functioning part of the body [28, 29]. Mutated proto-oncogenes can become carcinogenic oncogenes, driving excess multiplication. Tumor suppressor genes can be switched off by mutation, also contributing to a forming malignancy. At least half a dozen growth-regulating genes have to be affected so that a malignant development can start.

Signals from outside are forwarded into the cell by means of a pathway built by many different genes. As members of this chain become deregulated, the net signal the cell receives can be distorted, leading, for example, to excessive multiplication. One example is the *ras* family of genes which are members of one certain signaling chain. Hyperactive *ras* proteins are found in about a quarter of all human tumors [30, 31].

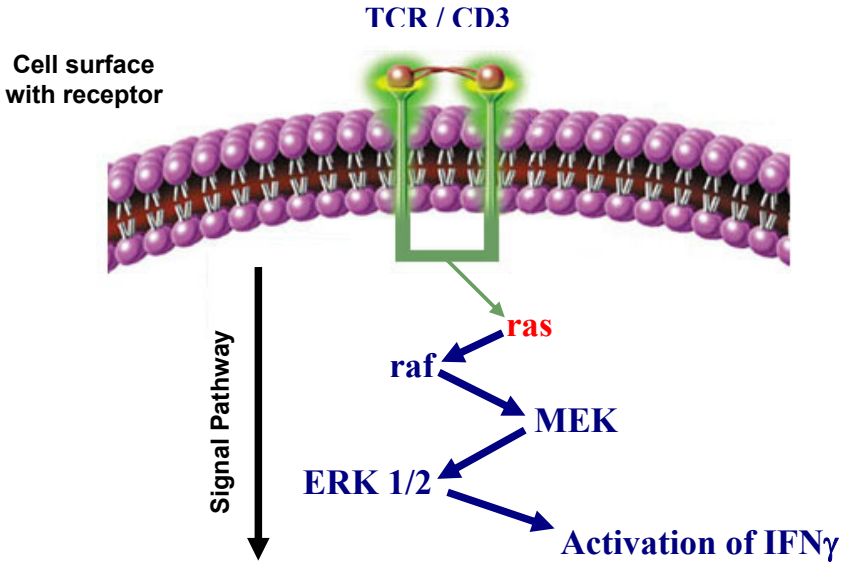


Fig. 3.4 Signal cascade from a cell receptor through a pathway built by many different genes down to the activation of interferon gamma. Hyperactive *ras* proteins are found in a quarter of all human tumors [*Novartis*]

Signals between cells can be forwarded by means of certain proteins acting as growth-factors, docking onto the receptors of cells surrounding the one emitting these molecules. The *myc* proteins are normally only created when growth-factors dock onto the cell. But many cancers, especially those of the tissues producing blood, have a constantly high level of *myc* which then urges the cell to proliferate. Oncogenes can also stimulate a cell into producing too much of these growth-factors, thus affecting all cells surrounding it. Examples are sarcomas and gliomas. Genes creating the receptors of a cell may also be mutated, forming receptors that forward a signal into the cell that has never actually been received.

Besides interpreting wrong growth signals, a cell must also become deaf to growth breaking signals from its surrounding neighbor cells. These inhibitory signals are also transmitted through the cell starting at receptors, just as growth signals do. This signal processing chain has to be compromised, too. One example are colon cancer cells inactivating a gene which create receptors for the *transforming growth factor beta* (TGF-β) which can stop a cell from growing. It could be shown through experiments that the introduction of a missing gene

related to the stopping of growth can restore the function of the cell in this regard, changing it into a healthier one [32]. The identification of a deregulation of a certain gene using the microarray technology can therefore help in designing novel genetic therapies and in the development of signaling molecules which can be applied to the tumor for signal induction [33].

3.3.2 Leukemia

3.3.2.1 Introduction

Patients with leukemia produce abnormal blood cells, called leukemia cells. These cells may function almost normally but change the general composition of the blood and disturb the function of the other blood components like red blood cells, normal white blood cells and platelets. *Chronic leukemia* is a form of leukemia where patients may not feel any immediate symptoms. The symptoms arise slowly together with the increasing number of leukemia cells in the blood. In *acute leukemia,* the leukemia cells are so abnormal that they cannot perform their original function. Their number increases rapidly, clear symptoms arise [3]. The types of leukemia are also grouped by the type of white blood cell that is affected; leukemia can arise in *lymphoid* and *myeloid* cells.

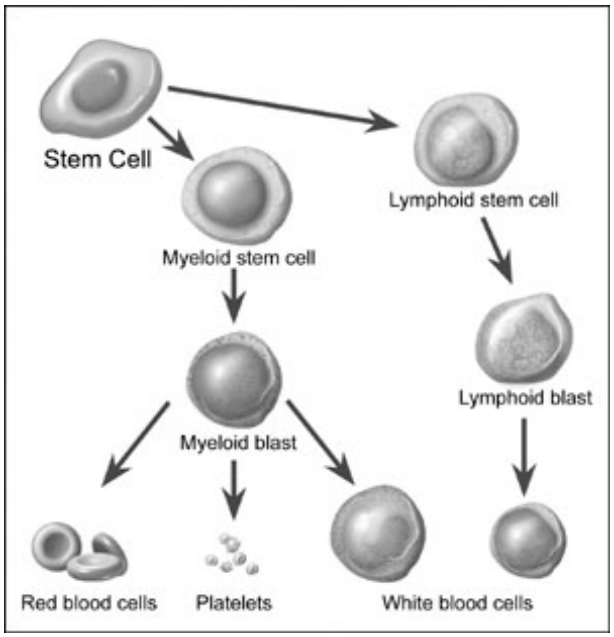


Fig. 3.5 Blood cells maturing from stem cells.

Pluripotent stem cells branch off into myeloid and lymphoid stem cells, becoming blast cells, and then becoming red blood cells, white blood cells, or platelets. [*Source*: *American Cancer Society*].

The four common types of leukemia are:

- **Chronic Lymphocytic Leukemia (CLL)**
  Most people diagnosed with this disease are over age 55.

- **Chronic Myeloid Leukemia (CML)**
  This illness affects mainly adults.

- **Acute Lymphocytic Leukemia (ALL)**
  It is the most common type of leukemia in young children. About 650 children below the age of 15 fall ill with leukemia each year in Germany each year, 83% of these having ALL and 15% AML.

- **Acute Myeloid Leukemia (AML)**
  This illness affects mostly adults; children are only affected in seldom cases [34].

There are several other rare leukemia types like the *hairy cell leukemia*. Myeloid and lymphoid leukemia can be discerned based on morphological, cytochemical and immunological differences.
Lymphocytes originate in the bone marrow and are involved in the immune system. Depending on the follow-up transformation of these cells, they either become T-cells or B-cells. Those cells passing through the thymus become T-cells, responsible for cell-mediated immunity. Other lymphocytes pass through the bursa equivalent organ becoming B-cells which produce antibodies.
Risk factors for falling ill with leukemia include high levels of radiation (as experienced in Japan at the end of WWII and in Chernobyl 1986 [35]), certain chemicals like benzene and certain genetic traits. Patients with pediatric ALL have a mean age of 4 years [36]. The majority of the other leukemia cases arise at a higher age as seen in figure 3.6.

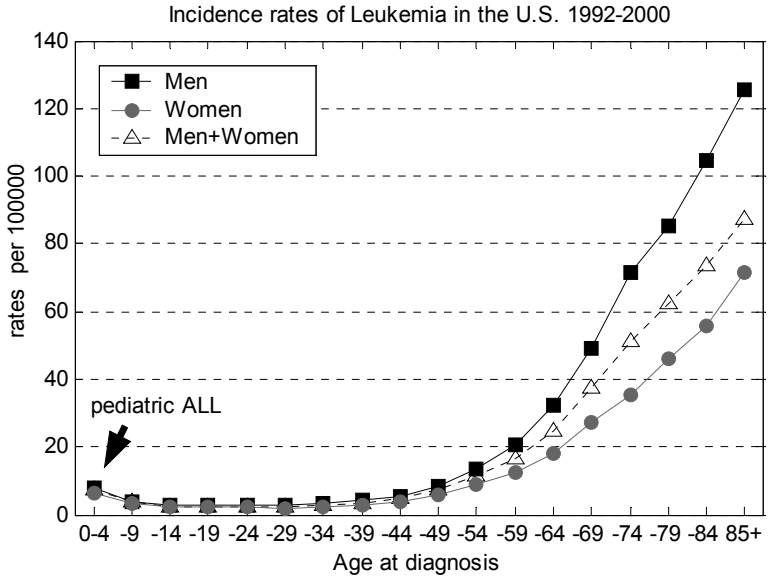


Fig. 3.6 Incidence rates of leukemia in the U.S.A. 1992-2000. Pediatric ALL incidences are clearly visible as are the higher incidence rates for men. [*Source: Surveillance, Epidemiology, and End Results (SEER) Program*]

3.3.2.2 Diagnosis

Patients may only feel weaker than usual if they have chronic leukemia, or they may feel several symptoms equal to having a bad flu (fevers, headaches, pain in the bones and joints); if patients have acute leukemia, they might experience easy bleeding and bruising, swelling or discomfort in the abdomen and swollen lymph nodes, especially in the neck or armpits.
The examination by the physician will include the analysis of several different parameters, for example:

- swelling of the lymph nodes and the liver,
- blood composition,
- antibody status,
- coagulation rate,
- x-ray of the thorax, the hand (bone status).
- results of biomolecular and staining techniques.

The physician might want to take a biopsy, removing some bone marrow from the patient's hipbone. A hematologist will examine the swab under the microscope, studying the

morphology and cytochemistry [37, 38]. Surface antigen expression methods might be used to make leukemia cells visible under the microscope (figure 3.7).

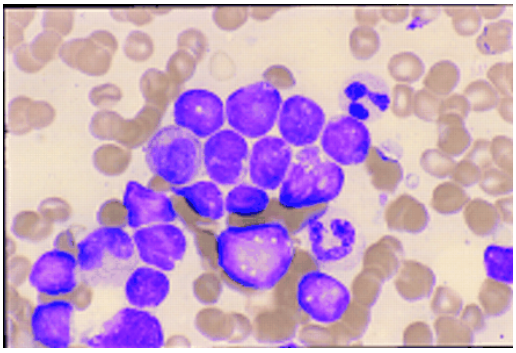


Fig. 3.7 Microscope image of a blood film from a patient with acute lymphoblastic leukemia (ALL). Leukemia cells are immunostained with the CD10 antigens. Application of different antigens gives information on the lymphoid leukemia deriving from the T-cell or B-cell lineage.

The number of chromosomes is also studied, looking for diploidity, hypodiploidity and hyperdiploidity. Different methods might be used to further classify the sample like cytogenetics, interphase-FISH [39] and 24-color-FISH [40, 41].

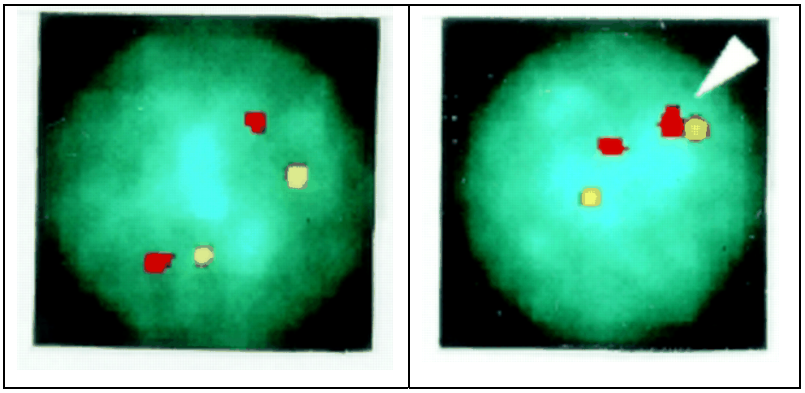


Fig. 3.8 Interphase fluorescence in situ hybridisation (FISH) using probes for BCR and ABL genes. Translocation t(9;22)[BCR-ABL] becomes visible.

Left: Normal cell showing two red dots (two normal copies of the BCR gene) and two yellow dots (two normal copies of the ABL gene). Right: Cell from a child with a certain acute lymphoblastic leukemia with translocation of chromosomes 9 and 22 (gene switching position from one chromosome to the other) [42].

In this way, prognostically unfavorable kinds of leukemia e.g. with translocation t(4;11) and t(9;22) can be detected. The aberration t(9;22) is rarely present in children with leukemia (2.2%) in contrast with hyperploidity (25 %) whit a more favorable prognosis.

These bimolecular methods have gained importance during the last couple of years, especially within the detection of minimal residual disease (MRD). Examinations in the future will include the utilization of microarrays for the molecular diagnostic to gain more information on the subtype of the diagnosed leukemia [43].

3.3.2.3 Treatment and the significance of molecular differences of ALL subtypes

An overall long-term event-free survival can only be achieved by using risk-adapted therapy that involves tailoring the intensity of the treatment to each patient's risk of relapse. Pediatric leukemia for example is a heterogeneous disease consisting of several genetically distinct leukemia subtypes. B-cell leukemias can contain translocations (swapping of chromosome parts between different chromosomes) like t(9;22)[BCR-ABL] (swapping of parts between chromosome 9 and 22), t(1;19)[E2A-PBX1], t(23;21)[TEL-AML1], rearrangements in the MLL gene on chromosome 11 or a hyperdiploid karyotype (i.e. > 50 chromosomes). The different genetic lesions in these leukemia subtypes have a great impact on the efficacy of cytotoxic drugs. Leukemias with t(1;19)[E2A-PBX1] for example respond poorly to conventional antimetabolite-based treatments. They have to be treated more intensively, this way reaching cure rates of about 80% [44].

It is difficult, time consuming and costly to accurate assign of a patient to a certain subgroup. Many hospitals do not have the facilities to perform all the tests, especially those in poorer countries. Even if the facilities are available, highly qualified personnel with much experience has to be found to do the analysis in the best possible way. Enhancing the prognostic criteria with results from microarray analysis will not only reduce the importance of this person-bound quality criteria but give the medical examiners information that previously was not available. Examples would be the detection of novel cancer subgroups that might have to be treated in a different way and the understanding of genetic pathways playing a role in a certain cancer type.

3.4 DNA Biosensors

3.4.1 Definition and Overview

Biosensors rely on the direct spatial coupling of a selective biological component with a physical transducer [45, 46]. The analyte recognition is made possible through the use of the biological component while the transducer converts the resulting chemical or physical changes into an electrical signal, thus strongly influencing the sensitivity of the biosensor. The biological component can be composed of enzymes, antibodies, nucleic acids, receptors, microorganisms or even living cells. The most important classes of transducers are optical, electrochemical, piezoelectric and acoustic devices. The special advantage of biosensors is the reversible interaction with the analyte so that the biosensor can be regenerated and used multiple times. Many biosensors also characterize themselves as being easy to use, having low acquisition and operational costs, being small and of reducing the time needed for a measurement.

DNA-biosensors are based on the specific hybridization between complimentary DNA single strands and make it thus possible to detect this bonding reaction. Single stranded probe-oligonucleotides are immobilized on the surface of the transducer, hence reacting with the complimentary target-sequences in the solution to be analyzed. Optical transducers are favored, many based on the exploitation of the phenomenon of surface plasmon resonance (SPR) and the evanescent field [47]. Also used in nucleic acid analytics are electrochemical transducers, implemented in potentiometric, voltametric and amperometric detection systems. The probe-oligonucleotides can be immobilized on many different kinds of electrode materials. Many different methods for the creation of electrical signals for the detection of hybridization events have been implemented. Markers are either electro-active ligands that intercalate between the created double stranded DNA [48], or enzymes (e.g. peroxidases), that catalyze the reaction of electro-active substrates (e.g. H_2O_2) [49]. Immobilization of inosin-substituted probe-oligonucleotides makes a labeling obsolete due to the detection of the guanine-signal after hybridization [50, 51].

Measurements without prior labeling can be accomplished by a variety of optical DNA-biosensor-systems and piezoelectric methods like the quartz crystal microbalance (QCM) which registers the mass-change after hybridization on the surface of the sensor through a frequency change. This method is used less often as it exhibits a deficit in sensitivity compared to methods using other types of transducers. An overview on the subject of DNA-biosensors is given by Bier and Fürste [52] as well as by Vercoutere and Akeson [53].

3.4.2 DNA Microarray Technology

3.4.2.1 Overview

Huge advances have been made in the sector of the DNA-chip-technology during the last ten years. It has become a leading technology in nucleic acid analysis alongside the polymerase chain reaction (PCR) [54]. DNA-chips, also known as DNA microarrays, are miniaturized slides with known probe-oligonucleotides immobilized in high density on their surface as a grid (array). The probes are hybridized with complimentary, fluorophore labeled target-DNA from a sample-solution. Unbound target molecules are then removed by means of different washing steps. The bound fluorophores can be detected using a fluorescence scanner. The acquired signal pattern contains the information which sequences where present in the sample-solution.

Depending on the target application, DNA microarrays can either be used for gene expression analysis or SNP (single nucleotide polymorphism) detection [55, 56]. DNA microarrays used for the former task are not only able to detect an expressed gene but, they can also quantify the expression.

3.4.2.2 Spotted microarrays

Probe-oligonucleotides can be immobilized on the surface of a slide by being spotted onto it. The resulting spots of oligonucleotides are round with an area without immobilized oligonucleotides between spots. This area can be used to query the background signal.

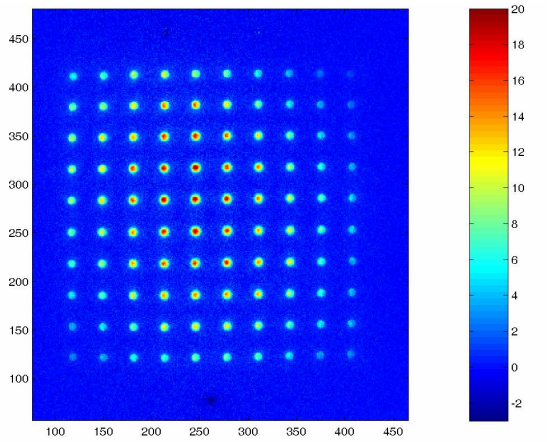


Fig. 3.9 Image of a spotted microarray created at the ICB. Colors denote fluorescence intensities. The area of low intensity between spots can be used to query the background.

A dual gene expression analysis is made possible by the use of two sample-solutions, each labeled with a different fluorophore and applied onto one and the same chip. Signals of each sample can be detected by scanning the array using a wavelength according to the fluorophores used. Ratios of gene expression signals between both samples can be calculated, making it possible to compare them. This approach facilitates the data-analysis, as both samples are measured in one experiment and the effect of e.g. different hybridization efficiencies are taken into account. As the rate of incorporation of the fluorophores into the oligonucleotides also depends on the type of fluorophore used, the two signal channels in this kind of experiment therefore also have to be scaled to be comparable. Further details on data preprocessing are discussed in the chemometrics chapter.

3.4.2.3 Microarrays created by photolithography

Oligonucleotides on microarrays produced by Affymetrix are not immobilized through spotting but created on the surface of the chip itself by photolithography. The oligonucleotide spots created by this process are square and reside tightly packed, so that there might not even be an area between spots that can be used to query the background signal (e.g. U133 microarrays). The used technology was adapted from the semiconductor industry allowing these arrays to hold over half a million cells. This very high density makes it possible to create so called *whole genome chips* which can screen samples for gene expressions of more than 22000 different genes, each one sampled by several cells. The oligonucleotide probe synthesis occurs in parallel, adding the different nucleotides to growing chains simultaneously. A length of about 15-25 nucleotides (15-25 - mer) is used in the common commercial chips.

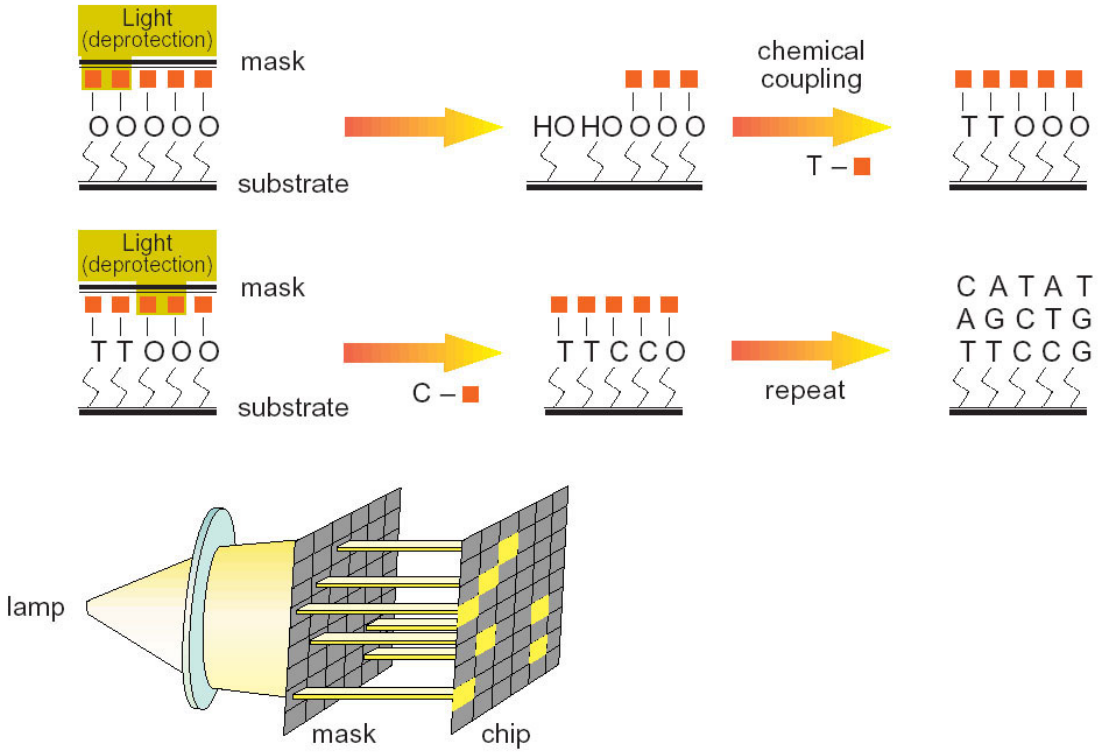


Fig. 3.10 Photolithographic steps in the creation of oligonucleotides directly on the surface of a microarray slide [*source: Affymetrix, Inc.*].

3.4.2.4 Medical Applications of DNA-biosensors

DNA-biosensor-systems have been developed over the last ten years and have only been used for research purposes so far. DNA-biosensors are now at the brink of becoming true medical diagnostic devices [57, 58].

In the last couple of years, high-density chips have increasingly been used to analyze the gene-expressions of several thousand genes in biological samples simultaneously in one experiment. The U133a chip of Affymetrix for example, can screen over 22000 genes. Researchers were thus able to analyze genetic pathways and to identify the role of genes faster and easier. A disadvantage of these chips lies in their cost. Thus, it is not feasible to perform a high-throughput screening of many samples. They are intended to help researchers and were not developed for medical applications, largely because as the interpretation of the data gained can be very tedious and difficult.

Following the high-density chips are the so called low-density chips. These chips can detect the gene-expressions of only a few genes, these having been selected a priori with the

intention of obtaining certain information about the sample. It is therefore necessary to use high-density chips to design the low-density ones. A low-density chip normally costs less and can be evaluated using procedures adapted to this specific chip, thus making it possible to be used by non-experts.

One example is the DNA-biosensor *AmpliChip CYP450* developed by Roche Diagnostics in cooperation with Affymetrix [59] which could become a landmark test case. It is referred to by Roche as the company's first microarray for clinical applications. It detects variations in two genes that affect the rate at which many drugs used, to treat cardiovascular diseases are metabolized. The information gained by using this device could help physicians to select the right drugs for their patients. This kind of tailor-made therapy is seen as the future of medical care by most pharmaceutical companies.

Fig. 3.11 : AmpliChip

Roche started marketing the *AmpliChip CYP450* as an analyte-specific reagent (ASR) and a class 1 medical device which does not necessitate clinical demonstrations of effectiveness and safety. The Food and Drug Administration (FDA; Rockville, MD, USA) believes that the device has the potential of being used improperly, thus endangering patients through misdiagnosis, and should therefore be classified differently. As the FDA believes that genomic data and technologies have increasing importance in pharmaceutical research, a novel regulation of microarray devices for diagnostic purposes is expected to be put into action in the very near future.

3.4.3 Affymetrix Microarrays

3.4.3.1 Physical construction

Affymetrix microarrays are created by using a photolithography process (see figure 3.10). The created spots on these microarrays are square and called *cells* by Affymetrix. The U133a [60] microarrays contain 712 x 712 cells, each cell composed of about 6 x 6 pixels in the raw data provided by the microarray scanner.

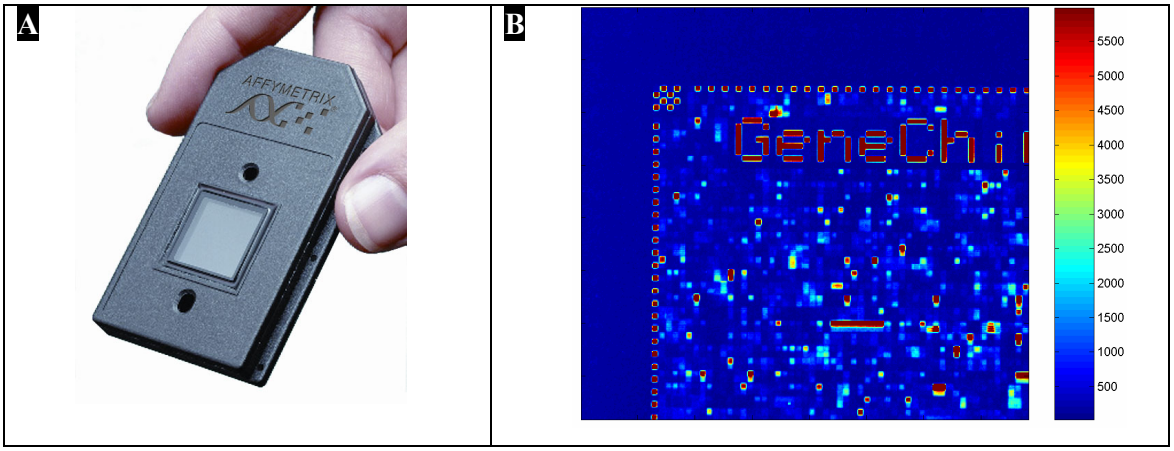


Fig. 3.12 U133 microarray from Affymetrix (A) [*source: Affymetrix Inc.*] and upper left corner of the image created by scanning the chip (B). Colors code the fluorescence intensity. Some cells are used to provide alignment help (checkerboard pattern at the boarder) and give information on the type of chip (GeneChip U133).

The U133a microarray enables a researcher to quantify the gene expressions of 22285 genes. The gene expression of a certain gene is queried by several cells. These cells are called a probeset as they contain the probe oligonucleotides designed to hybridize with one certain gene. Most probesets are composed of 22 cells, 11 perfect match (PM) cells, and 11 mismatch (MM) cells [61].

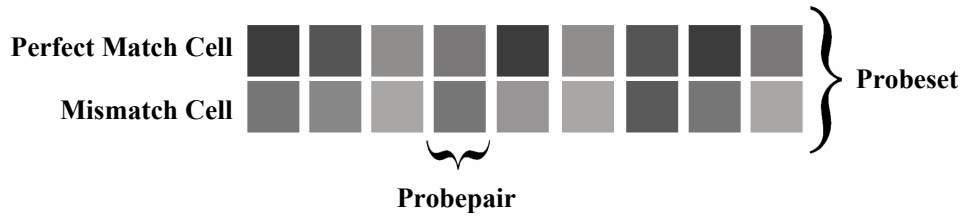


Fig. 3.13 Configuration of a probeset containing several cells. One probeset is designed to query the gene expression of one certain gene.

A perfect match cell is thus always paired with a mismatch cell. The oligonucleotides used in such a MM cell are the same as those in the PM cell, except for one mismatch in the center of the sequence. The mismatch base is the complimentary base of the corresponding Watson-Cricks basepair (e.g. A interchanged with T, G interchanged with C). Sequences in different probepairs differ from each other, designed to match the sequence of the target gene at different positions.

Both cells of one probepair are always located besides each other on the microarray, different probepairs of one probeset are randomly spread throughout the chip to minimize the risk of loss of a complete probeset due to artifacts (see x- and y-pos in table 3.1).

Tab. 3.1 Perfect match probe sequences for the gene 212019_at (Affymetrix identification code).

Name	X-pos	Y-pos	Probesequence
212019_at	138	131	AACACTCAGC**TTTT**CGCAACATAAT
212019_at	124	179	CACTCAGC**TTTT**CGCAACATAATCC
212019_at	4	187	CAGC**TTTT**CGCAACATAATCCCAGC
212019_at	18	319	GC**TTTT**CGCAACATAATCCCAGCAC
212019_at	581	211	C**TTTT**CGCAACATAATCCCAGCACT
212019_at	426	333	GCACTTTGGAACGCTGGGTGGATTG
212019_at	98	669	TTTGGAACGCTGGGTGGATTGCTTG
212019_at	668	589	TGCAAT**GGG**CCATGATCACGATCCT
212019_at	366	167	CAAT**GGG**CCATGATCACGATCCTGC
212019_at	231	155	AAT**GGG**CCATGATCACGATCCTGCA
212019_at	206	131	AACAGAGAGAGACAGTCCCTGGCCC

Notice the sliding character of the sequence selections. X- and Y-pos denote the position of the cell on the microarray. Mismatch sequences are the same as the PM sequences except for one mismatch in the exact center of the sequence. Corresponding MM cells have an Y-position increased by one.

The Affymetrix rational for using PM and MM cells is as follows: the role of a perfect match cell is to provide the signal of the hybridization reaction of the probe oligonucleotides with the matching target oligonucleotides from the gene the sequence was designed for plus any additional signals from cross hybridizations with similar target oligonucleotides and background; the mismatch cell in contrast provides the signal of cross hybridizations and background only, therefore making it possible to calculate the signal of the selective hybridization.

3.4.3.2 Calculation of gene expressions

An experiment using Affymetrix microarrays yields many signals corresponding to all the cells (e.g. 506944 for the U133A chip). These signals have to be translated into gene expression values. The procedure is as follows:

1. background correction of the entire microarray,
2. application of a gene expression summary algorithm to gain a single value for each probeset,
3. normalization of signals from different microarrays of one study.

The MicroArray Suit (MAS) calculates probeset values using PM and MM cells of each probeset. Researchers have discussed the necessity of using MM cells, arguing that these cells may also detect a signal corresponding to specific binding as well as non-specific binding and should therefore be omitted [62-64]. New gene expression summary algorithms have been created, achieving good reproducibility in signal calculation without using the MM cells (e.g. RMA [65, 66]). Some of these methods have also been used in chapter four.
Different background correction and gene expression summary algorithms are discussed in the chemometrics chapter.

3.5 Chemometrics

3.5.1 Introduction

Chemometrics is the application of mathematical, statistical, graphical or symbolic methods to design or select optimal measurement procedures and experiments and to maximize the information which can be extracted from data. Chemometric procedures can prove useful at any stage in an analysis, from the first conception of an experiment, to the extraction of all information [67-69].
Pattern recognition approaches are used to identify similarities and regularities present in data. The most frequently used techniques have traditionally been those in the area of cluster analysis and classification. These methods help extract information out of data. Classification methods perform pattern matching and comparisons, helping to associate an unknown sample to known clusters of data. One of the primary goals of chemometrics is to reduce the number of dimensions needed to accurately portray the characteristics of a data set. A wide variety of methods are available to achieve this, either by selecting an important subset of the original variables (feature selection), or by creating a set of new composite variables, which are more

efficient than the original variables in describing the data (feature reduction). The creation of new variables can be approached in several ways; two of these are projection (e.g. principal component analysis) and mapping.

It is critical for the data to be of good quality and to have been gathered using informative samples. Chemometrics should therefore be applied beforehand to design the experiment or measurement process. One important application is the analysis of how many experiments will be needed to obtain information that is statistically significant, and to optimize experimental parameters to improve the sensitivity and precision of analysis (e.g. simplex algorithm). Application fields include calibration techniques, resolution enhancement techniques (e.g. Fourier transformation, deconvolution), signal processing (e.g. dealing with noise), modelling and parameter estimation. Data verification can be done by learning about the characteristics of the data like means, standard deviations, ranges, modes of distribution and other statistical measures. Process parameters for quality control can be estimated using this knowledge and possible outliers detected.

3.5.2 The role of chemometrics in microarray experiments

The use of microarray experiments has become popular due to its advantages for large-scale studies of gene expression and mutation analysis. This technology makes it possible to study complex systems in less time and to derive comprehensive information. Microarrays in combination with chemometrics and bioinformatics make it possible to tell, for instance, which genes behave differently in cancerous cells compared to healthy cells of the same tissue type. Researchers can hereby learn more about the processes in the cell and about factors that are involved in turning a healthy cell into a cancerous one. Moreover, the technique makes it possible to classify the cancer more precisely and to make a better decision on the treatment that should be applied.

Many different steps are involved in the analysis of microarray data. At the beginning the relative gene expression level has to be obtained. Not only is noise inherent to the measured data, systematic errors can also be added during measurement and analysis alike. Errors must be detected and corrected before any downstream analysis, such as identification of differentially expressed genes and sample classification, can be performed [70]. The identification of noise components as well as systematic errors and artifacts is one major task of chemometrics. The data has to be pre-processed so that a comparison of different experiments becomes possible. The comparison itself also includes chemometric methods, for

example the application of t-tests, cluster analysis, principal component analysis and a whole range of other statistical calculations. Chemometrics provides the building blocks for the entire processing chain, starting from the processing of raw data up to the creation of models for the classification of samples and diagnosis.

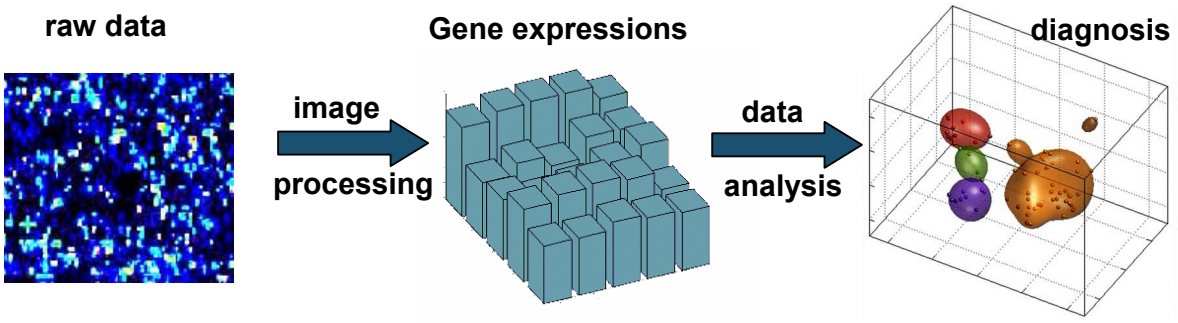


Fig. 3.14 Chemometrics provides means to process data and to create models for diagnosis.

Different methods are explained in the chemometrics chapter, including basic techniques like PCA, clustering and cross-validation. Furthermore, this chapter includes explanations on different gene expression summary techniques that have been used to gain gene expression values for each probeset (see the Affymetrix technology chapter) as well as gene selection techniques that have been applied in order to select genes used for the classification of samples of different cancer subtypes. One major task of bioinformatics is to provide means for storing data reliably so that easy and fast access is possible. A databank system interfacing with Matlab was created for project management which is also explained in this chapter. Details on the application of methods for the analysis of a leukemia dataset can be found in chapter five.

3.5.3 Pre-processing

A common pre-processing step consists of mean centring, i.e. subtracting the mean over all samples for each variable:

$$x_i^c = x_i - m_x \text{ with } m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A centred data cloud lies around the origin of the coordinate system (figure 3.15 II).

If the absolute values of the measured data are important, then centralisation should not be performed [71].

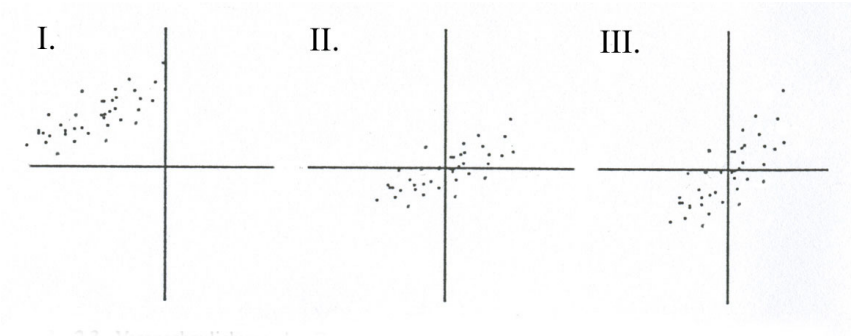


Fig. 3.15 (I) Raw data, (II) Centered data, (III) Standardised data.

The covariance, variance and standard deviation can easily be calculated by using two centred vectors:

Covariance: $\text{cov}(\mathbf{x},\mathbf{y}) = \dfrac{\mathbf{x}^{cT}\mathbf{y}^{c}}{n-1}$

Variance: $\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x},\mathbf{x})$

Standard deviation: $s(\mathbf{x}) = \sqrt{\text{var}(\mathbf{x})} = \dfrac{\left\|\mathbf{x}^{c}\right\|}{\sqrt{n-1}}$

One important aspect of pre-processing data is its standardisation.

Standardisation: $\mathbf{x}^{S} = \dfrac{\mathbf{x}^{c}}{s(\mathbf{x})}$

The influence of arbitrary units (e.g. 1000m or 1 km) is eliminated through standardisation (figure 3.15 III); all axis have the same scale.

3.5.4 Cluster analysis – hierarchical clustering

Cluster analysis can be defined as the unsupervised classification of similar objects into groups of which members and assignments are unknown a priori. The shape of a cluster can be characterised by its cluster-specific means, variance, and covariances that also have a geometrical interpretation [72, 73]. Objects in one subset will have similar properties. The clustering algorithm defines by what kind of similarity measure objects are clustered together. Clustering is done by using an algorithm containing a distance calculation (Euclidean,

cityblock distance and other), and by using an updating part. Both parts can vary, depending on the used algorithm. The result of a cluster analysis can be displayed in a dendrogram.

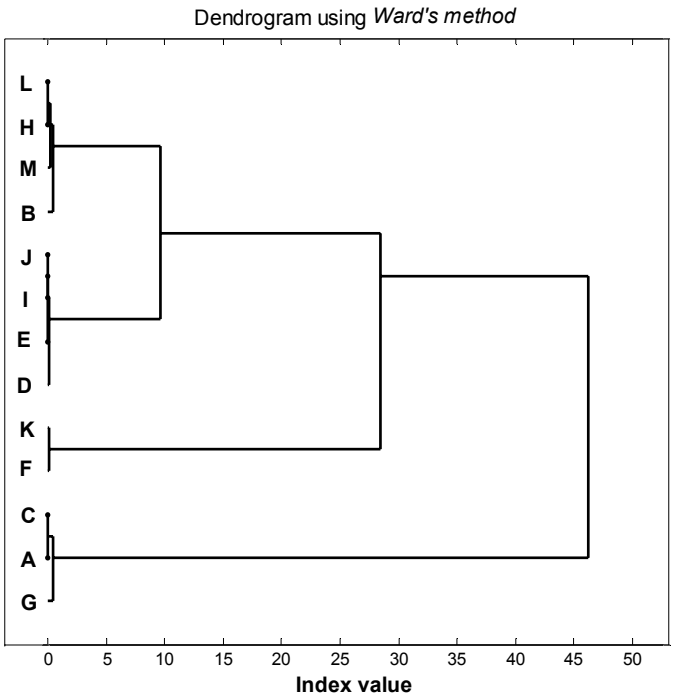


Fig. 3.16 Exemplary dendrogram using Ward's method on 13 objects.

The dendrogram shows the distance between objects (index value where branches meet). Objects that lie near to each other (similarity measure) are connected through a branch at small index value. The used algorithm picks two objects that lie near to each other, connects them and continues working up to higher index values.
It can be formulated as follows:

1. Choose a distance measure.
2. Calculate the distance matrix **D** for the dataset.
3. Identify the smallest value d_{ij} in **D** (other than in the diagonal).
4. Merge objects i and j. The so created new object is regarded as one new single entity.
5. Update the distance between all other objects in **D** and the new entity.
6. Return to number two and repeat until all objects are interconnected.

The used Ward's method calculates the distance between a new entity [AB], created through the merger of objects A and B, and an object C with the formula

$$d([AB],C] = \frac{(n_c + n_a)d(A,C) + (n_c + n_b)d(B,C) - n_c d(A,B)}{n_a + n_b + n_c}$$

with

n_i number of objects in entity i.

It takes the number of samples in each of the united groups into account when updating the distance matrix, thus creating the new cluster center nearer to the group with more objects [74].

3.5.5 Principal component analysis

3.5.5.1 Introduction

Principal component analysis (PCA) seeks to maximize the variance information present in a data set in as few new dimensions as possible [75, 76]. Graphically, principal components analysis rotates the axes of the data to conform to axes which contain a maximum amount of variance information; this can be thought of as looking at the data from a different perspective in hyperspace. Principal component analysis is one of the most important multivariate data analysis techniques as it allows a good visualisation of underlying structures present in the data. It also allows for a dimensionality reduction, which can be used beneficially in subsequent tasks (classification, regression, and cluster analysis). Figure 3.17 shows a plot of 30 objects for which variables x_1 and x_2 were measured.

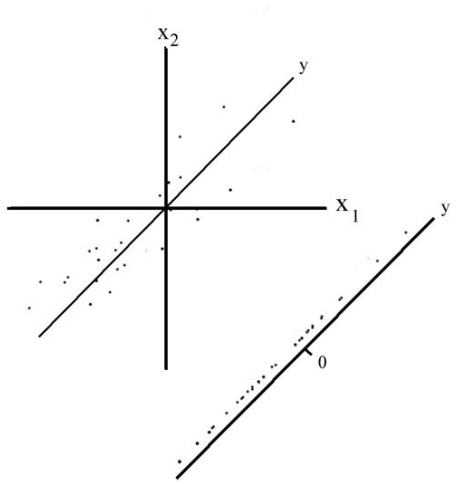


Fig. 3.17 Plot of 30 objects with two variables (x_1, x_2) and first principal component y.

The axis y represents the first principal component (PC1). Of all one-dimensional subspaces (lines), this one covers best the variation of the data. Further principal components can be added, each successive component being orthogonal to the prior ones and covering as much as possible of the remaining variance. In figure 3.17 a second PC could be added thus reaching the maximal dimensionality possible for this data. Through the omission of insignificant low variance principal components a noise reduction is achieved.
Principal components are linear combinations, i.e. a weighted sum of the original p measured variables [77]. If **X** is the data with n objects and $p \leq n$ variables, then a $r \leq p$ -dimensional linear subspace is looked for that covers the variance of the data the most. In order to achieve this, the eigenvectors and the eigenvalues of the covariance matrix of the data have to be determined. Normally, data will be pre-processed, undergoing centring and standardisation.

3.5.5.2 Eigenvalues, eigenvectors

Taken **X** is a (n,n) symmetric matrix, then a **v** and λ exist so that

$$\mathbf{X}\,\mathbf{v} = \lambda\,\mathbf{v} \qquad (1)$$

λ are called eigenvalues, **v** eigenvectors. If (1) is written as $\mathbf{Xv} = \lambda\mathbf{I_n v}$, the equivalent formulation $\mathbf{(X - \lambda I_n)\,v = 0}$ is reached. In short: $\mathbf{Av = 0}$ with $\mathbf{A = X - \lambda I_n}$ which is a linear equation. If **A** is not regular, then the only solution would be **v** = 0:

$$\mathbf{Ax = b} \Rightarrow \mathbf{A^{-1}Ax = A^{-1}b} \Rightarrow \mathbf{I_n x = A^{-1}b} \Rightarrow \mathbf{x = A^{-1}b} \quad ; \quad \mathbf{0 = A^{-1}\,0}$$

The null-vector is not seen as an eigenvector and is discarded. There is not only one eigenvector corresponding to an eigenvalue; they span a subspace (normally a line) because if a vector **v** solves equation (1), then also multiples $\alpha\mathbf{v}$ that do so, too:

$$\mathbf{X}(\alpha\mathbf{v}) = \alpha\mathbf{Xv} = \alpha\lambda\mathbf{v} = \lambda\,(\alpha\mathbf{v}).$$

Not every matrix has eigenvectors / eigenvalues, though symmetric (n,n)-matrices like covariance-, correlation- and distance-matrices that are often used in data-analysis always have n, not necessarily different, eigenvalues / eigenvectors forming n „Eigen-pairs“ (λ_i , $\mathbf{v}_i$) that solve the equation:

$$\mathbf{Xv}_i = \lambda\mathbf{v}_i$$

It can also be written as a matrix-product, with the eigenvectors as columns of an eigenvector-matrix and the eigenvalues constituting the diagonal of an eigenvalue-matrix:

$$\mathbf{X}\left(\mathbf{v}_1|...|\mathbf{v}_n\right) = \left(\mathbf{v}_1|...|\mathbf{v}_n\right)\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

or, in short:

$$\mathbf{XV} = \mathbf{V\Lambda}$$

The λ_i are normally sorted to be in decreasing order $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$.
Different eigenvectors of symmetric matrices are orthogonal to each other.

3.5.5.3 Calculations in Principal Component Analysis

After pre-processing the (n, p)-data-matrix **X,** the $(n-1)^{-1}\mathbf{X}^T\mathbf{X}$-Matrix is calculated. If **X** were centred, then this newly calculated matrix would be the covariance matrix. If **X** were also standardised, it would be the correlation matrix. The main step in principal component analysis is the calculation of the p eigenvalues and eigenvectors of this matrix. The components of the eigenvectors contain the weights of the original variables needed for each principal component while the eigenvalues contain the variance it covers. Different algorithms can be used for the calculation of the eigenvalues / vectors [76].

3.5.6 Cross Validation

The number of training observations usually is much larger than the number of features in a standard discriminant analysis. In microarray studies however, the number of tissue samples measured is typically far smaller than the number of gene-expression detected. It becomes necessary to use cross validation to verify the credibility of a model, as the number of available samples is constricted. In cross validation samples are left out to create a training dataset. A model is created by using this subset and applied to the data points that were left

out. Several different methods for cross validation are available. In the leave-one-out (LOO) method each point of the dataset is left out successively to create the model. Several data points can be left out, either in a Venetian-blind-, a block- or random-order-pattern. Cross validation prevents overfitting a created model to the dataset used, as it is validated using samples unknown to the model because they have been left out. It has to be taken into account that using all samples to select those features (genes) best suited for the construction of the model will introduce a prediction error. This selection bias can be evaded if the gene selection is performed in the cross validation process [78].

3.5.7 Support Vector Machines (SVM)

In a binary linear classification problem, a linear hyperplane which divides space into two regions corresponding to the two classes is looked for [79-81]. The class of an object i is defined by y_i being +1 or -1. The optimal hyperplane is orthogonal to the shortest line connecting the convex hulls of the two classes (see figure 3.18) and intersects it half-way between the two classes.

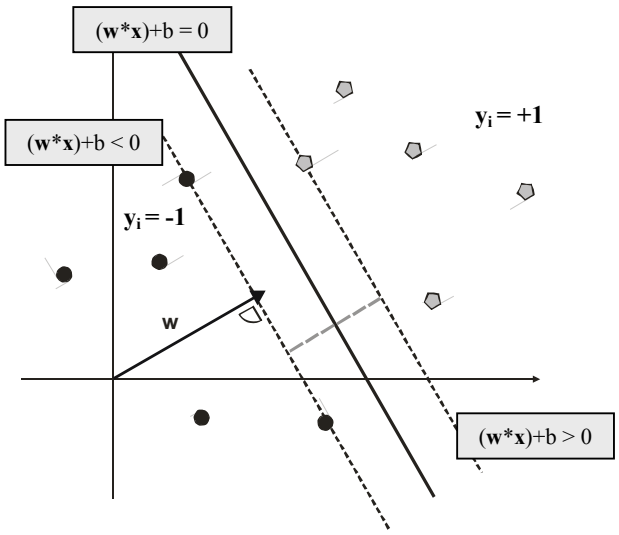


Fig. 3.18 Binary linear classification using SVM.

Diamonds are separated from circles by a hyperplane (solid line). Objects are classified using the weight vector w and the threshold b satisfying (**w*x**)+b ≥ 0 for all diamonds and (**w*x**)+b<0 for all circles.

The hyperplane is defined by the weight vector w and the threshold b. These are defined so that

$(w*x)+b \geq 0$ for all $y_i = +1$,

$(w*x)+b < 0$ for all $y_i = -1$.

If w and b are scaled so that the distance between points closest to the hyperplane measures $1/||w||$ then all points with $y_i [(w*x_i)+b] = 1$ are called support vectors, all other objects having $y_i [(w*x_i)+b] > 1$. The broadness of the separating region is then $2/||w||$. The aim is to maximize the minimal distance of the training data to the separating hyperplane [82, 83].
As data might not be perfectly separable, such a separating hyperplane may not always exist. It is thus necessary to use methods that deal with misclassification. A slack variable ξ_i is introduced so that

$(w*x)+b \geq +1-\xi_i$ for all $y_i = +1$,

$(w*x)+b \leq -1+\xi_i$ for all $y_i = -1$.

$\xi_i \geq 0$.

There are three possibilities:

$\xi_i = 0$: x_i was classified correctly and lies at the border or outside of the separating region.

$0<\xi_i<1$: x_i was classified correctly but lies inside of the separating region.

$\xi_i \geq 1$: x_i was classified incorrectly.

Note that all samples with $\xi_i \neq 0$ are considered as margin errors.
The aim is to reduce the rate of misclassification and to maximize the separating region. The new optimization of $y_i [(w*x_i)+b] \geq 1-\xi_i$ is

$$\min_{w}\left(\frac{1}{2}\|\mathbf{w}\|^2 + \mathbf{C}\sum_{i=1}^{1}\xi_i^k\right)$$

The trade-off parameter C penalizes margin errors for the training data. It can be optimized e.g. through cross-validation techniques.

3.5.8 Gene selection Methods

3.5.8.1 Gene shaving

Gene shaving extracts coherent, typically small clusters of genes that vary as much as possible across samples [84]. The shaving procedure is as follows (see also figure 3.19):

1. Use the entire expression matrix **X,** row centred to have zero mean.
2. Calculate the first principal component.
3. Discard a certain percentage of genes with the smallest absolute inner-product with the principal component.
4. Repeat steps two and three until only one gene remains. A nested sequence of clusters $S_N \supset S_k \supset S_{k1} \supset S_{k2} \supset S_1$ where S_k denotes a cluster of k genes is created. The optimal cluster size $\hat{k}$ is calculated using the gap statistic described in [84].
5. Orthogonalize each row of X with respect to $\bar{x}_{S_{\hat{k}}}$, the average gene in $S_{\hat{k}}$ to promote the discovery of different (uncorrelated) clusters in further iterations of the procedure.
6. Steps 1-5 are repeated with the orthogonalized data to find the second optimal cluster. This process is continued until a maximum of M clusters is found. The value of M is chosen a priori.

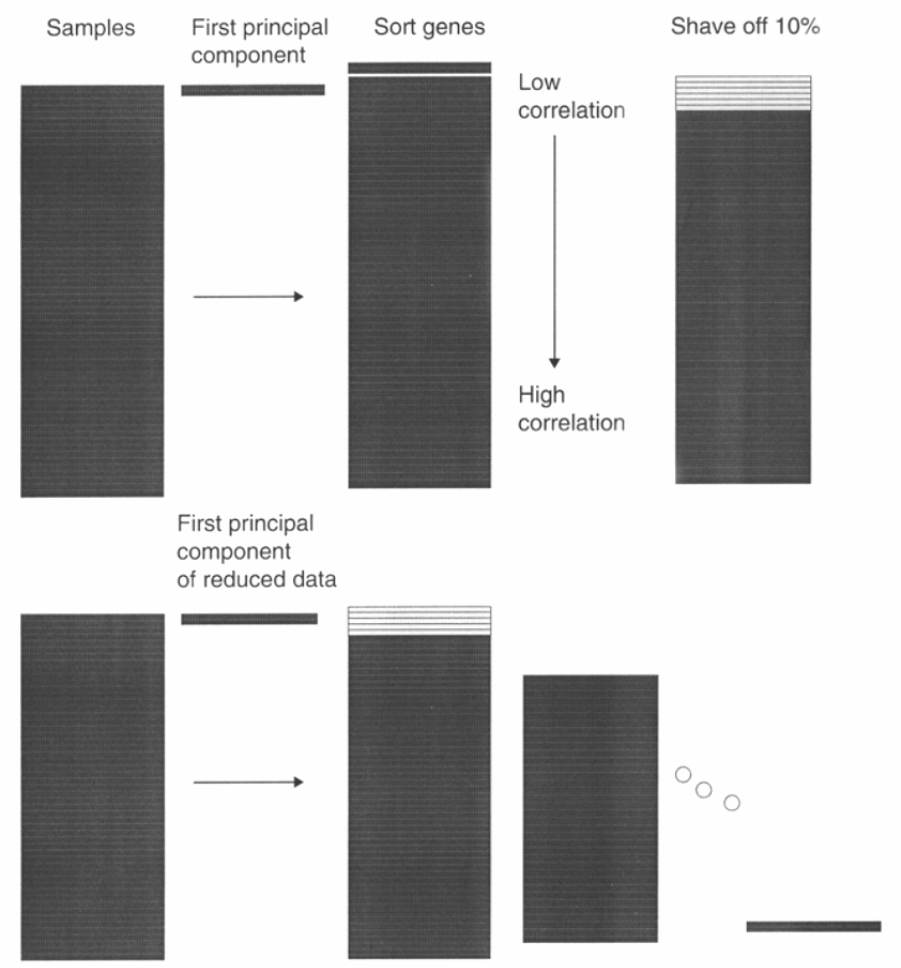


Fig. 3.19 Schematic of the gene shaving process (*source:* [84]).

3.5.8.2 Significance analysis of microarrays (SAM)

Coherent patterns of gene expression can be found by using cluster analysis. This method does not provide much information on the statistical significance of the selection of genes. Conventional t-tests are often used but have the drawback that, applied to microarray data with thousands of genes, several genes will be identified by chance even when using p=0.01. This led to the development of SAM by Virginia Goss Tusher, Robert Tibshirani and Gilbert Chu [85]. Genes with statistically significant changes in expression are selected with SAM by assimilation of a set of gene-specific t-tests. A score is assigned to each gene based on the change in gene expression relative to the standard deviation of repeated measurements for that gene. A gene is rated as potentially significant if its score is higher than a threshold. The false discovery rate is identified by analyzing permutations of the measurements. Although the signal to noise ratio decreases with decreasing gene expression, the fluctuations for a specific

level of expressions are gene specific. These fluctuations are assessed with the relative difference d(i):

$$d(i) = \frac{\bar{x}_A(i) - \bar{x}_B(i)}{s(i) + s_0}$$

with $\bar{x}_A(i)$ and $\bar{x}_B(i)$ defined as the average levels of expression for gene (i) in groups A and B respectively, and s(i) defined as the standard deviation of repeated expression measurements. To ensure that the variance of d(i) is independent of gene expression, a small positive constant s_0 is added to the denominator (see [85] page 5117).

Significant changes in gene expression are found by ranking genes by the magnitude of their d(i) values, so that d(1)>d(2)>…>d(i). Relative differences $d_p(i)$ are also calculated for permutations of measurements balanced for different subgroups, and genes again ranked in such way that $d_p(1)>d_p(2)>\ldots>d_p(i)$. The expected relative difference $d_E(i)$ is then defined as the average over all Np balanced permutations: $d_E(i) = \sum_p d_p(i) / Np$. Changes in expression which might be significant can be identified using a scatter plot of the observed relative difference d(i) vs. the expected relative difference $d_E(i)$ as seen in figure 3.20.

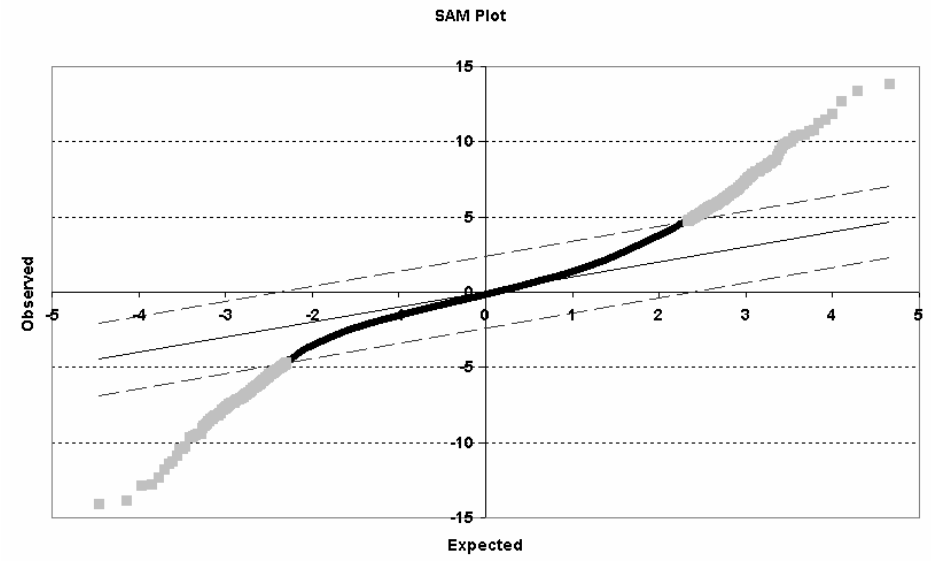


Fig. 3.20 Scatter plot of the observed relative difference d(i) vs. the expected relative difference $d_E(i)$. Genes with significant changes in expression are highlighted as gray squares. The threshold (dashed lines) was set to a fold change of three. TheE2A-PBX1 and T-ALL sample subgroups from the leukemia dataset [86] were compared.

3.5.8.3 Predication Analysis of Microarrays (PAM)

Class prediction is based on an enhancement of the simple nearest prototype (centroid) classifier which works as follows: for each class a training-set nearest-centroid (average expression of each gene in the subgroup) is computed and the overall gene expression centroid subtracted. The resulting values are differences from the overall centroid. The squared distance from the gene expression profile of every test sample to each of the class centroids is computed. The predicted class is the one whose centroid is closest to the expression profile of the test sample. This method uses all genes in the dataset. A modification of this nearest-centroid method, called *nearest shrunken centroid* was developed by Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan and Gilbert Chu, which only uses a subset of the genes [87, 88].

Let x_{ij} be the expression for genes i=1,2, … *p* and samples j= 1, 2, …. *n*. Classes are 1, 2, … *K* and C_k are indices of the n_k samples in class *k*. The *i*th component of the centroid for class *k* is $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} / n_k$, the mean expression value in class *k* for gene *i*; the *i*th component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$ [87]. The class centroids are shrunken towards the overall centroids after having been standardized by the within-class standard deviation for each gene. Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k * (s_i + s_o)}$$

with s_i the pooled within-class standard deviation for gene *i* , $m_k = \sqrt{1/n_k + 1/n}$ and s_0 be a positive constant. The method shrinks each d_{ik} toward zero, giving d'_{ik} and yielding shrunken centroids $\bar{x}'_{ik} = \bar{x}_i + m_k (s_i + s_0) d'_{ik}$. Each d_{ik} is reduced by an amount Δ in absolute value and is set to zero if its absolute value is less than zero. It is defined by $d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$ with + meaning positive part ($t_+ = t$ if $t>0$ and zero otherwise). The threshold parameter Δ is selected through cross validation.

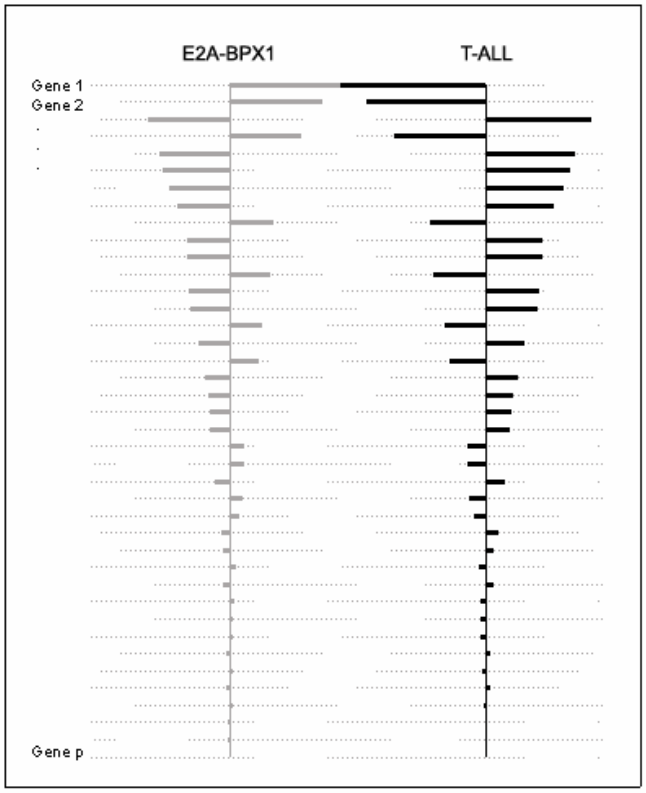


Fig. 3.21 Shrunken differences d'$_{ik}$ for 40 genes having at least one nonzero difference. These genes are the PAM selection for the discrimination of E2A-BPX1 and T-ALL samples of the leukemia dataset [86].

3.5.8.4 Fisher's Ratio

Let $\mathbf{X_1} = \{x_1^1,...,x_{l_1}^1\}$ and $\mathbf{X_2} = \{x_1^2,...,x_{l_2}^2\}$ be samples from two different classes. Fisher's linear discriminant is given by the weight vector **w** which maximizes

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

where $S_B = (m_1 - m_2)(m_1 - m_2)^T$ and $S_W = \sum_{i=1,2} \sum_{x \in \mathbf{X}_i} (x - m_i)(x - m_i)^T$ are between and within class scatter matrices respectively and m_i is defined by $m_i = \frac{1}{l_i} \sum_{j=1}^{l_i} x_j^i$.

The rational behind maximizing J(**w**) is to find a direction which maximizes the projected class means (the nominator) while minimizing the class variance in this direction (the denominator). The weight vector **w** can thus be used to select those variables contributing the most in separating the two classes, making it possible to choose a subset of variables best suited for class discrimination.

3.5.9 Gene expression summary algorithms

3.5.9.1 Affymetrix MicroArraySuit (MAS) 5.0 algorithm

A signal is calculated by subtraction of an ideal mismatch IM value from the PM value. This IM value is a corrected MM value to assure that the resulting difference is larger than zero. The calculated signals after subtraction for each probepair are used to calculate the gene expression value for that probeset. This is done by means of a biweight estimator to provide a robust mean: signal = Tukey Biweight. The exact algorithm is described in the Statistical Algorithm Description Document [89], page 3. Further on, a detection call is calculated, giving the information of whether a transcript of a particular gene was deemed present or absent in an experiment. First, a discrimination score R_i, a measurement of the difference between the PM and MM intensities, is calculated for the *i*th probe pair:

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

If the median(R_i) is greater than a threshold τ (default $\tau = 0.015$), the hypothesis that PM and MM are equally hybridizing to the sample can be rejected. Wilcoxon's rank test is used to calculate a significance or *p*-value. This *p*-value is then compared with preset significance levels to make a call.

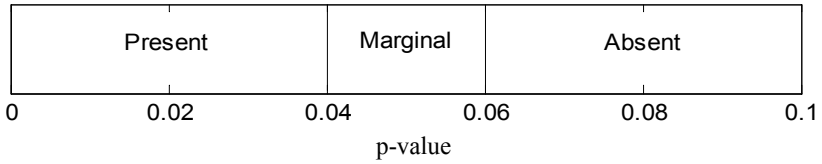


Fig. 3.22 Comparison of the calculated p-value with preset margins to make a call.

3.5.9.2 MAS, perfect match only

The same steps are used as in the MAS approach, except for the subtraction of an IM value from the PM one. MM cells are thus not used at all.

3.5.9.3 Li – Wong Model

For each probeset n, Li and Wong's measure [90, 91] is defined as the maximum likelihood estimate of the Θ_i, i = 1,...,I obtained from fitting

$$PM_{ij} - MM_{ij} = \Theta_i \Phi_j + \varepsilon_{ij}$$

with j number of probepair,
i number of sample,
Φ_j representing probe-specific affinities and
ε_{ij} assumed to be independent normally distributed errors.

The estimation procedure includes rules for outlier removal.

3.5.9.4 RMA Model

The robust mutli-array average (RMA) is a summary measure of background-adjusted, normalized and log-transformed PM values.

For probeset n, fit a linear additive model

$$\log_2(PM_{ij} - BG) = a_i + b_j + \varepsilon_{ij}$$

with i: label of chip,
j: label of probe,
BG: background

ε_{ij}: errors.

a_i: gives the expression for probe n on array i.

RMA estimates a_i for chip i using a robust method, such as median polish (fit iteratively, successively romoving row and column medians, and accumulating the terms, until the process stabilizes).

4 Method development

4.1 Databases

4.1.1 Introduction

Storing data in file systems is accompanied by several disadvantages:

- Possible data redundancy and inconsistency. Data is often stored in many different file formats and data structures, especially if several different persons are working on the same project.
- Concurrent access of the data is made difficult as new routines to access the information might be needed for each new task. Users might not be able to access or store files as these are used by others at the same time.
- Supplementing certain information is difficult as many different files (with different formats) might have to be changed.
- Integrity problems. The meaning of the data and vocabulary used in one file (created by a certain scientist) might not be used the same way by someone else creating a different file [92].

In databases, data is stored as tables. The database management system (DBMS) interacts with these tables on the behalf of the researcher, making it possible to sort, search and fetch subsets of the data in the tables. The strength of DBMSs:

- Provide fast access to selected parts of large databases.
- Powerful ways to summarize and cross-link information in databases.
- Concurrent access from multiple users running on multiple hosts while enforcing security on access to specific data.
- Ability to act as a server to a wide range of clients, independent of their geographical location.

The user applies a Data Manipulation Language (DML) to interact with the database. The most widely used nonprocedural language is the Structures Query Language (SQL).
For a nonprocedural language the user specifies what data is required without specifying how to get that data.

4.1.2 Microarray Data Management System

Microarray data obtained in large studies can reach a size of several gigabytes. The data from the microarray measurement itself is usually stored on file servers so that they can be easily accessed by different users. In addition, the following corresponding clinical information for each sample has to be stored:

- patient age,
- patient prognosis,
- classification of illness (e.g. ALL-lymphoma),
- illness information (e.g. tumor stage),
- sample processing information,
- name of technician making the measurement,
- data gained by other methods (e.g. pathological information gained by immunostaining, cytogenetics),
- filenames of the raw data, preprocessed data etc.,
- data gained in the microarray-data processing.

The Microarray Data Management System (MDMS) was designed to provide easy access to the data without requiring the user to know the exact name of the data files or their location

on the fileserver. Interaction in Matlab® [78] with the data was achieved using MySQL [72]. It is available under the *GNU General Public License (GPL)* and was in use more than four million times world wide by the end of the year 2003. The used MySQL interface in Matlab® was written by Robert Almgren from the University of Toronto [93]. The MDMS structure is depicted in figure 4.1.

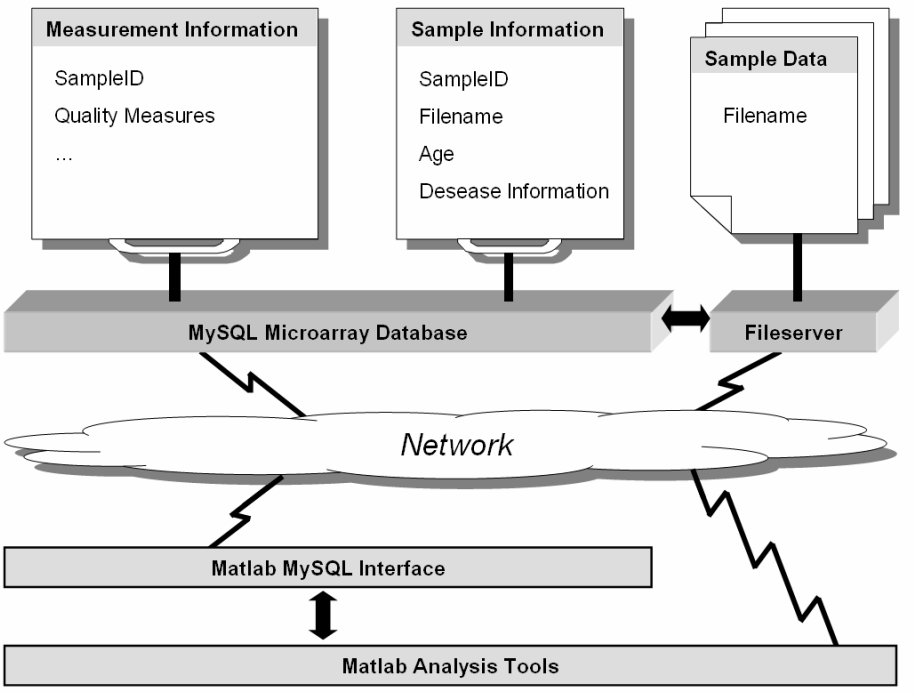


Fig. 4.1 Structure of the Microarray Data Management System.

Routines were written in Matlab® to easily process the data, starting from the raw data files. Different processing steps could be applied to certain files by executing routines provided with specific database selection criteria.

The Minimum Information About a Microarray Experiment (MIAME) project [94], approved by the Microarray Gene Expression Database group (MGED) in a Stanford University meeting in 2001, deals with the concept of a clear and structured storage of information. A similar mode of operation has been adopted by the macromolecular structure community (see, for example http://msd.ebi.ac.uk) where most journals require submission of a well-defined minimum of raw data associated with publications. The corner stones of MIAME are:

1. The recorded information about each experiment should be sufficient to interpret the experiment and should be detailed enough to enable comparisons to similar experiments.
2. The information should be structured in a way that enables useful querying as well as automated data analysis and mining.

The developed microarray data management system has been developed in concordance with these corner stones.

4.2 Analysis of the raw data provided by the scanner

A part of a raw microarray image provided by the scanner is seen if figure 4.2. The images often are skewed, i.e. the borders of the actual array in the center of the image form a parallelogram. One example of the left, vertical array boundary is shown in figure 4.2. The shift in position of the boundary to the left measures 5 pixels in this example.

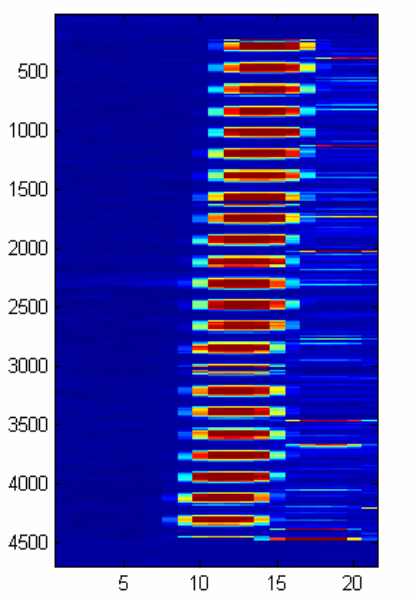


Fig. 4.2 Part of the image of a U133A chip provided by the scanner

The left border of the microarray is shown. The microarray itself is positioned in the center of the chip (and thus in the center of the image). Many cells merge together in this view, creating the illusion of there only being a few cells along the height of the image. The image is skewed, the microarray borders do not form a rectangle.

The exact position of a cell in the microarray does not coincide with discrete pixel coordinates. The exact coordinates of all the features have to be linearly interpolated using the determined pixel-level coordinates of the corner features. If the coordinates of the corner features are determined incorrectly, erroneous cell values will be extracted from the image [90, 95]. A tool for the exact localization of the corner feature coordinates was developed in this work. The localization of the microarray boarders is possible by analysis of the first deviation of horizontal and vertical intensity profiles. An accurate detection of the position of the corner features is done within ±1 pixel. The interpolation leads to a certain amount of signal overlay in the border values of neighboring features. Further, it should be noted that the lithographic process used for the synthesis of the oligonucleotides also infers uncertainties concerning the exact sequence of the oligonucleotides at the border of each feature. As diffuse light shines onto the border of a neighboring feature that lies besides a cell that is illuminated, sequences might be created in this region which have not been explicitly designed.

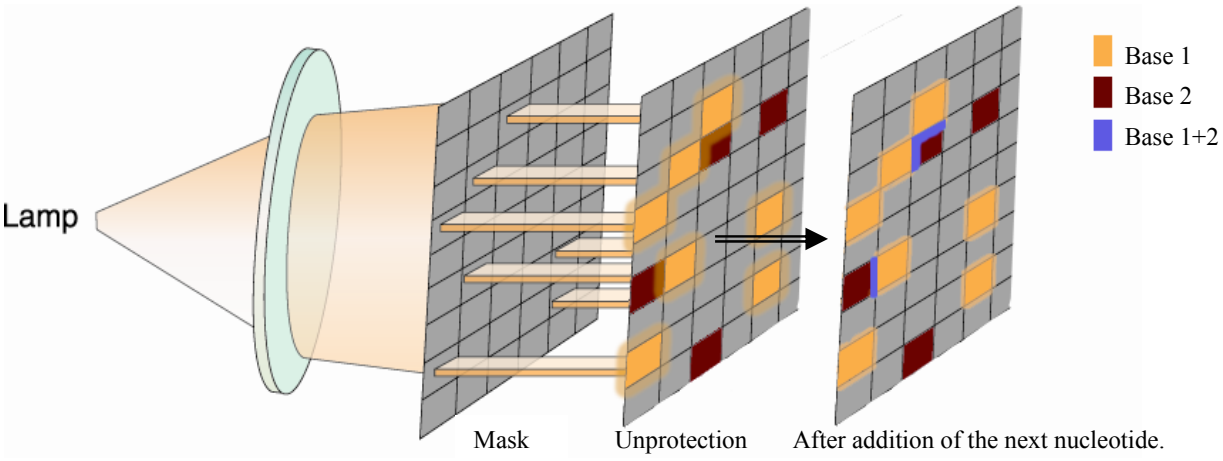


Fig. 4.3 Photolithographic process in which oligonucleotides are created on the surface of the microarray.

Oligonucleotides are created on the surface nucleotide after nucleotide: light eliminates protecting groups at the end of the oligonucleotides, making it possible to extend those molecules with another nucleotide. As light also partially shines onto neighboring cells, oligonucleotides will be created in the border regions which have not been explicitly designed.

This means that the border values of each feature must not be used to calculate the overall signal of a cell. It is therefore necessary to discard the outer perimeter of each feature; an average of 16-24 pixels from the center are used (from a total of ~36 pixel). The values of the remaining pixels also differ from one another. As can be seen in figure 4.4, pixel intensity value are not homogenous.

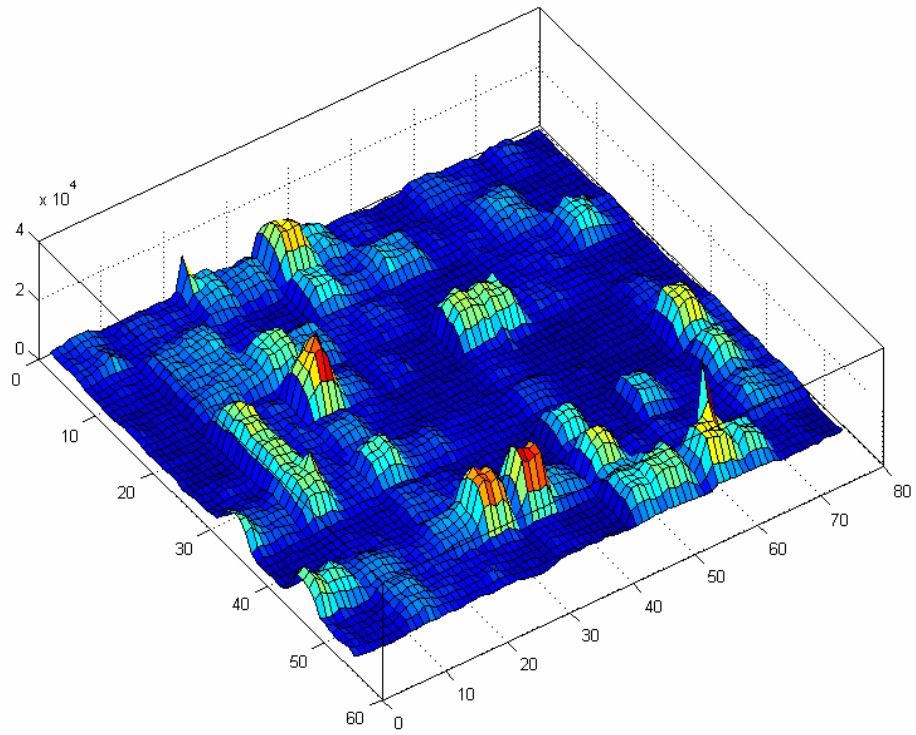


Fig. 4.4 Outtake of a raw U133A microarray image provided by the scanner

Each cell has a dimension of about 6x6 pixels; pixel intensities in a cell deviate from one another. Z-axis values denote fluorescence intensities.

The signals of these central pixels can deviate up to 30% from the mean fluorescence intensity of that feature (figure 4.5).

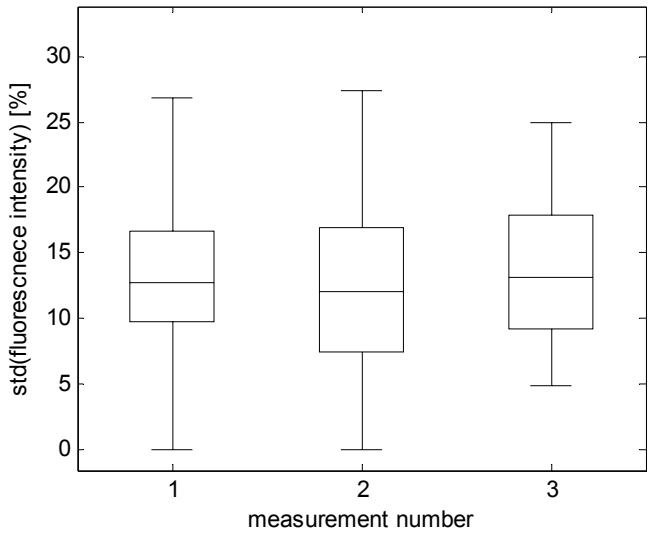


Fig. 4.5 Boxplot of standard deviations of signal intensities in features from three different measurements.

Each feature consists of 16-24 pixels. The signal intensities of these pixels can deviate up to 30% in one feature. Measurements of three different studies were analyzed: (1) CD4 lymphocytes from peripheral blood dataset [96] , (2) Ross pediatric leukemia dataset [86], (3) proprietary dataset [97].

The primary source of these in-cell signal deviations is thought to be the different hybridization efficiencies at different locations [98-100] . Heterogeneous hybridization can be improved by creating conditions enhancing the surface-near diffusion. The surface density of the probes also influences the ability of the immobilized probes to capture solution-phase targets. Bondage kinetics is lower in regions with high probe density, probably due to variations in target capture capabilities [98-100]. The surface density may also depend on the spatial positioning of the oligonucleotide (e.g. upright or parallel to the surface) and also on its geometrical conformation (e.g. linear or creating loops).

4.3 Artifact detection

Features on a U133 microarray from Affymetrix are of the size of 18 μm. With over half a million features packed side by side in one microarray, the detection of artifacts becomes essential. Artifacts can appear in very different shapes and sizes. The microarray analysis software of Affymetrix provides a tool for the user to mask cells per hand. As a detection of artifacts in raw signal images is often impossible and a selection by hand also likely to be imperfect due to the exhaustiveness of checking an area of 712 x 712 cells, an automatic

scheme for artifact detection was developed. This algorithm uses multiple microarrays to create a *median-chip-image* and a thresholding scheme to detect artifacts which become detectable as deviations from the median signal.

4.3.1 Medianchip images for artifact detection.

It is possible to detect artifacts reliably using several microarray chips. First a *medianchip* is created, by calculating the median of the signal of same cells over all preprocessed chips. The preprocessing takes account of background and scaling differences.

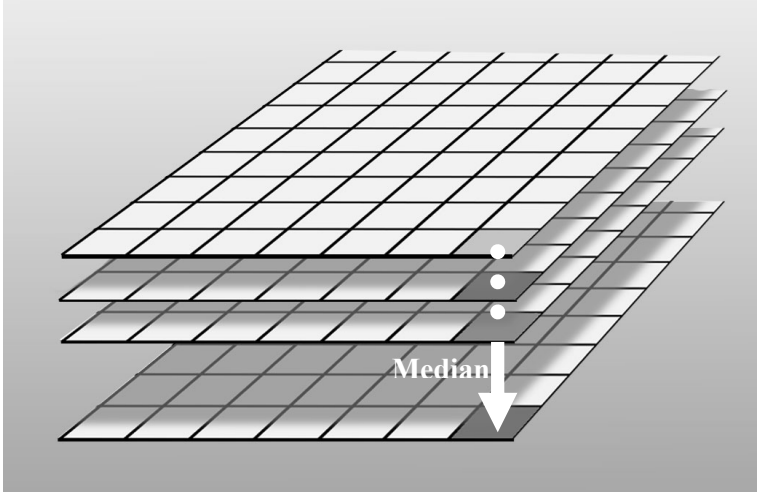


Fig 4.6 Creation of a medianchip from several microarrays. The median of the signal of same cells over all preprocessed chips is calculated.

This *medianchip* is used to calculate a *ratiochip* by dividing the signal of each cell from a specific microarray by the value of the same cell on the medianchip and taking the logarithm. As cells with the same probe oligonucleotides are compared, only those cells which are differentially expressed should show a great deviation from the median. The rest of the cell values on the ratiochip should be more or less the same. The thresholding algorithm was tuned so that a small group of neighboring, differentially expressed cells is not masked but only larger areas of anomalies are, as seen in figure 4.7.

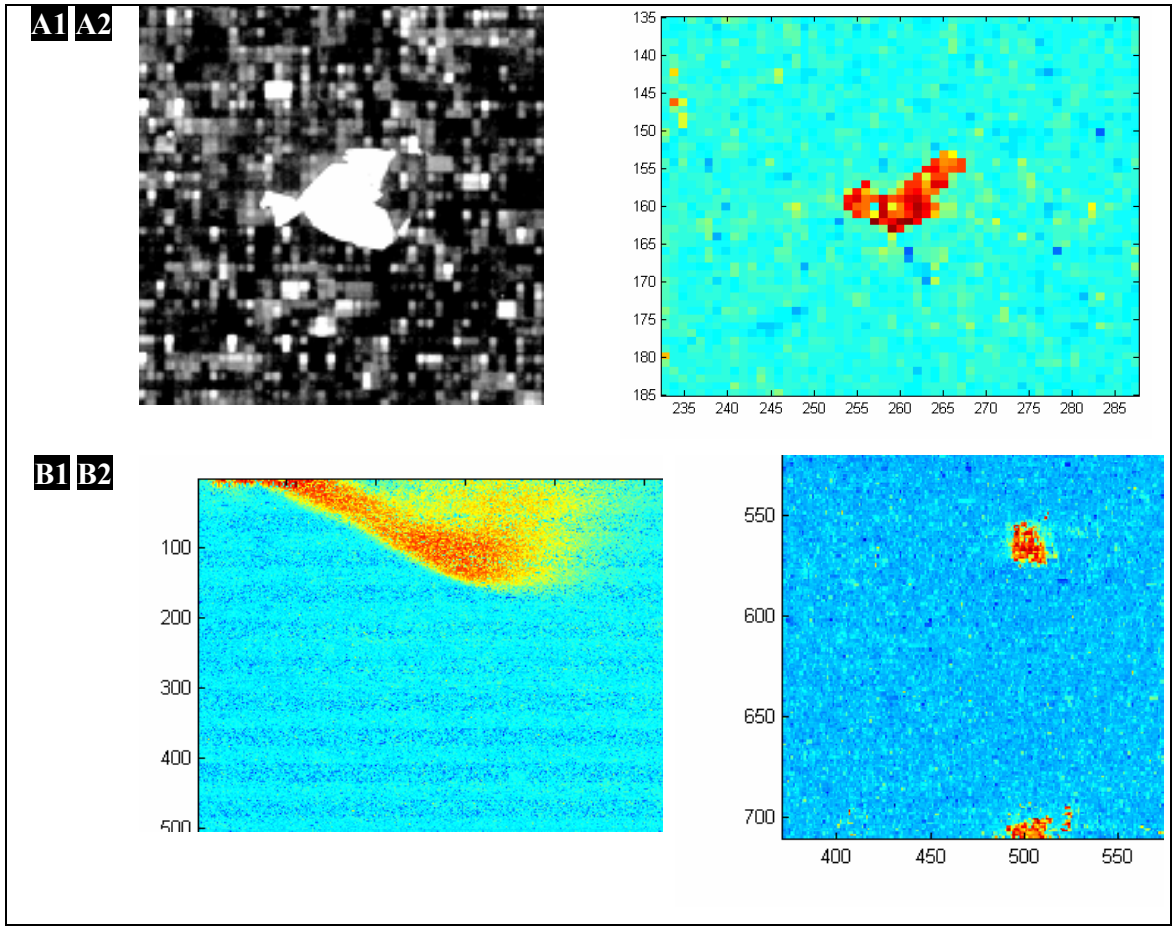


Fig. 4.7 Artifacts on U133a microarrays seen in a raw image (A1) and ratio images.

A1: raw image as taken by the CCD camera; A2 – B2: good recognition of artifacts in ratio images. Ratio images show log(cell signal/median signal). The horizontal pattern in B1 is a product of downscaling the image for presentation and not present in the data. Different colors denote different fluorescence signal values.

4.3.2 Artifact detection algorithm

A lowpass filter is applied on the ratioimage, eliminating all small features that might represent differentially expressed cells [101-103]. The lowpass filter was implemented using a moving window 9 by 9 pixel median filter. Its size was optimized to smooth out differentially expressed cells but not known artifacts. This lowpass image (see figure 4.8 B) is then converted into a mask image by thresholding (figure 4.8 C). The level of the threshold is calculated using a larger area of the image, thus gaining information of the overall signal in this image. This is done with another, larger moving window used to calculate the median

signal in that area. The chosen threshold is then chosen relative to this signal level, thus taking account of signal slopes etc.

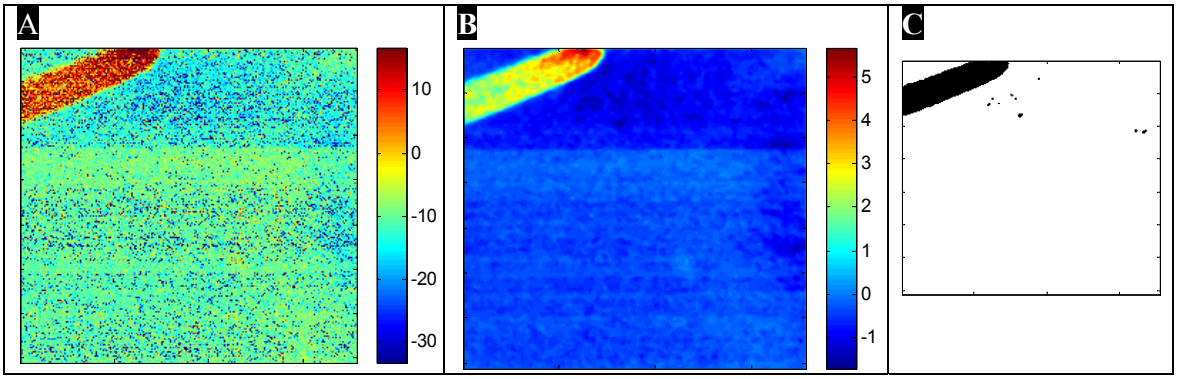


Fig. 4.8 Process of masking. Left: ratioimage; Center: thresholded image using a lowpass filter; Right: masked areas. Shown measurement is JD-ALD416, *E2A-PBX1* subgroup of the Ross pediatric leukemia dataset [86]. It is further analysed in the leukemia data analysis chapter.

The MAS software of Affymetrix also masks some cells using pixel-statistics, i.e. pixel information for a cell in the raw images provided by the scanner [95].

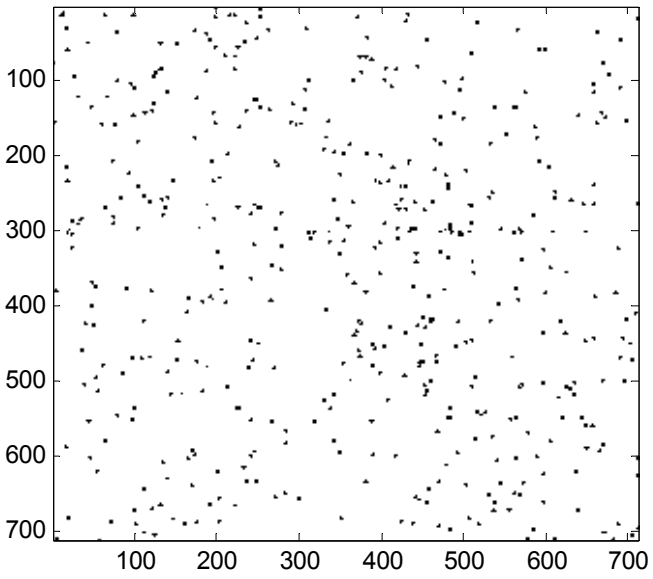


Fig. 4.9 Position of cells masked by the MAS software in the measurement shown in figure 4.8A . As the masking is done using the pixel statistic of each cell, larger artifacts are neither recognized nor masked.

The larger artifacts are not recognized by this MAS masking algorithm. In this sample, 23500 cells where masked by the newly developed thresholding algorithm. 45% of all probesets had at least one probepair in the masked areas, 80% of these having one probepair, 20% having two probepairs in this region. The distributions of cells masked by the MAS software seemed to be random in all analyzed studies (e.g. [86, 96, 97, 104, 105]). Figure 4.10 displays the position of all masked cells in the entire Ross pediatric leukemia dataset [86].

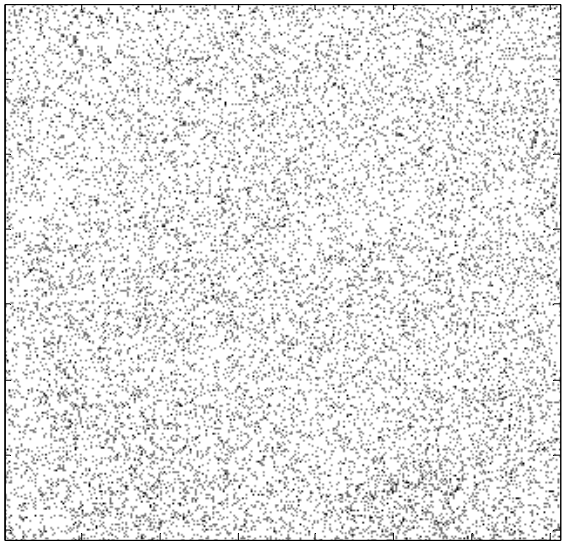


Fig. 4.10 Homogeneous distribution of cells flagged as outliers by the MAS software in all the U133A microarrays in the Ross pediatric leukemia dataset.

The masking algorithm is capable of recognizing and masking all kinds of artifacts commonly seen in U133 microarrays. Some artifacts in the pediatric leukemia dataset [86] and their masking can be seen in figure 4.12.

Artifacts can also often be recognized in the scatter plots of a measurement against the medianchip. Figure 4.11A shows a plot of cell intensities of a sample without artifacts against the cell intensities of the medianchip. Figure 4.11B on the other hand, shows the scatter plot of the sample shown in figure 4.8A.

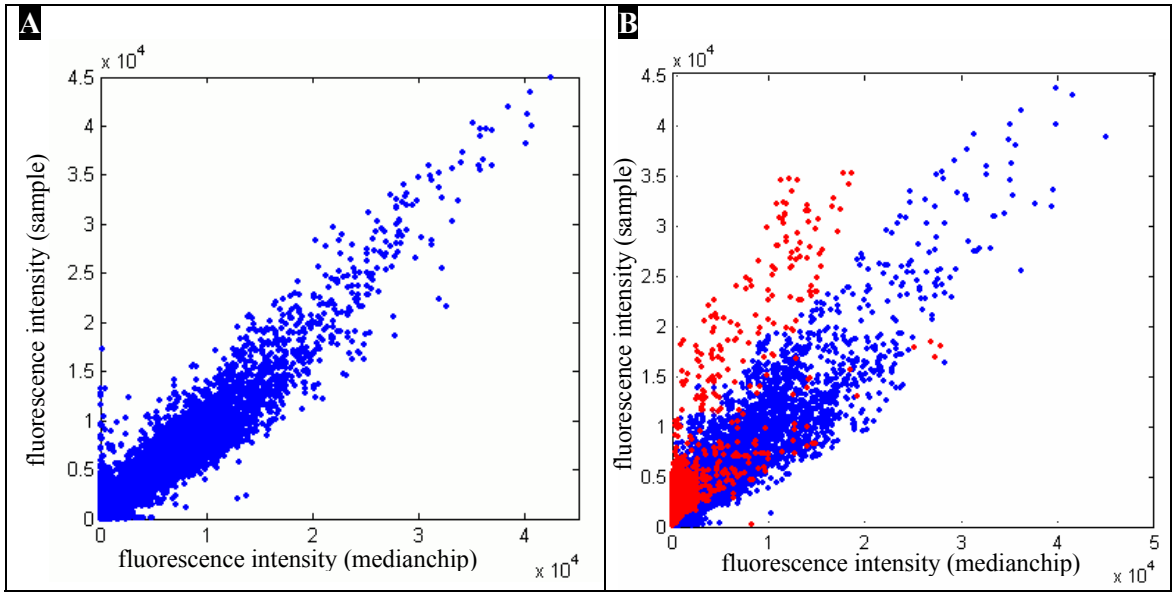


Fig. 4.11 Scatter plots of a sample without artifacts (A) and with artifacts (B , sample as shown in figure 4.8A) against the medianchip. The presence of an artifact can be seen in the abnormal form of the plot (red datapoints).

The number of artifacts in one microarray varied from no artifacts at all to such an amount of artifacts that a measurements had to be discarded. All the analyzed studies [86, 96, 97, 104, 105] had a similar global amount of artifacts; no study was clearly worse then the others.

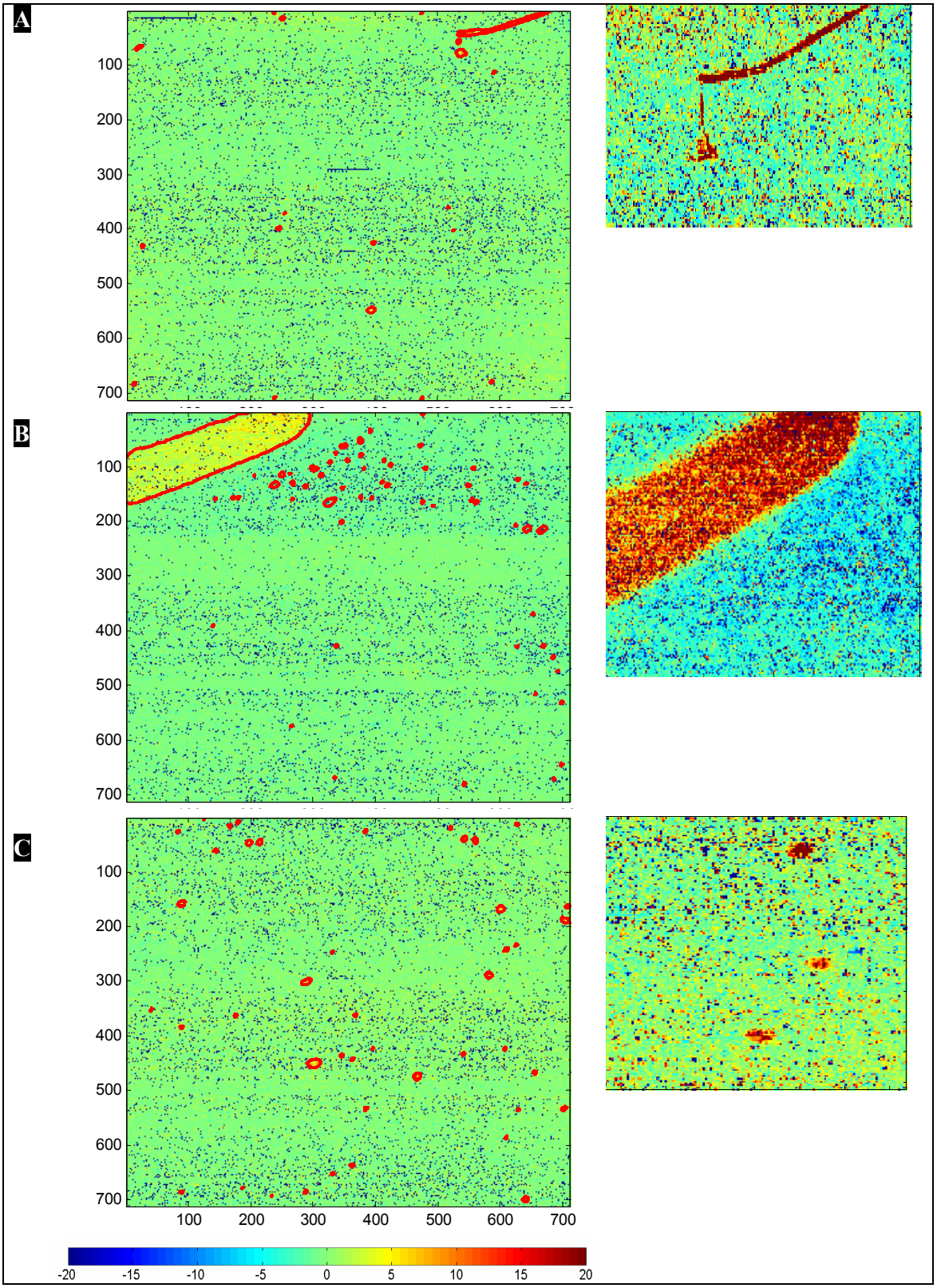


Fig. 4.12 Ratioimages with masked artifacts (defined by red contours). A close-up of the artifacts is visible at the right. All images were taken from the pediatric leukemia dataset [86]. Colors denote signal values, color range of right side images was chosen for best artifact display.

4.4 Background correction

4.4.1 Background in Affymetrix Microarrays

The background in a microarray measurement consists of signal intensity caused by non specific binding and autofluorescence of the array surface. Since probes are so densely packed on Affymetrix microarrays, there are no regions on the chip designed to query the background signal. Further, signals on U133 microarrays are not distributed in a homogeneous way: horizontal stripes exist on the chip containing high-intensity cells, brighter than their surroundings.

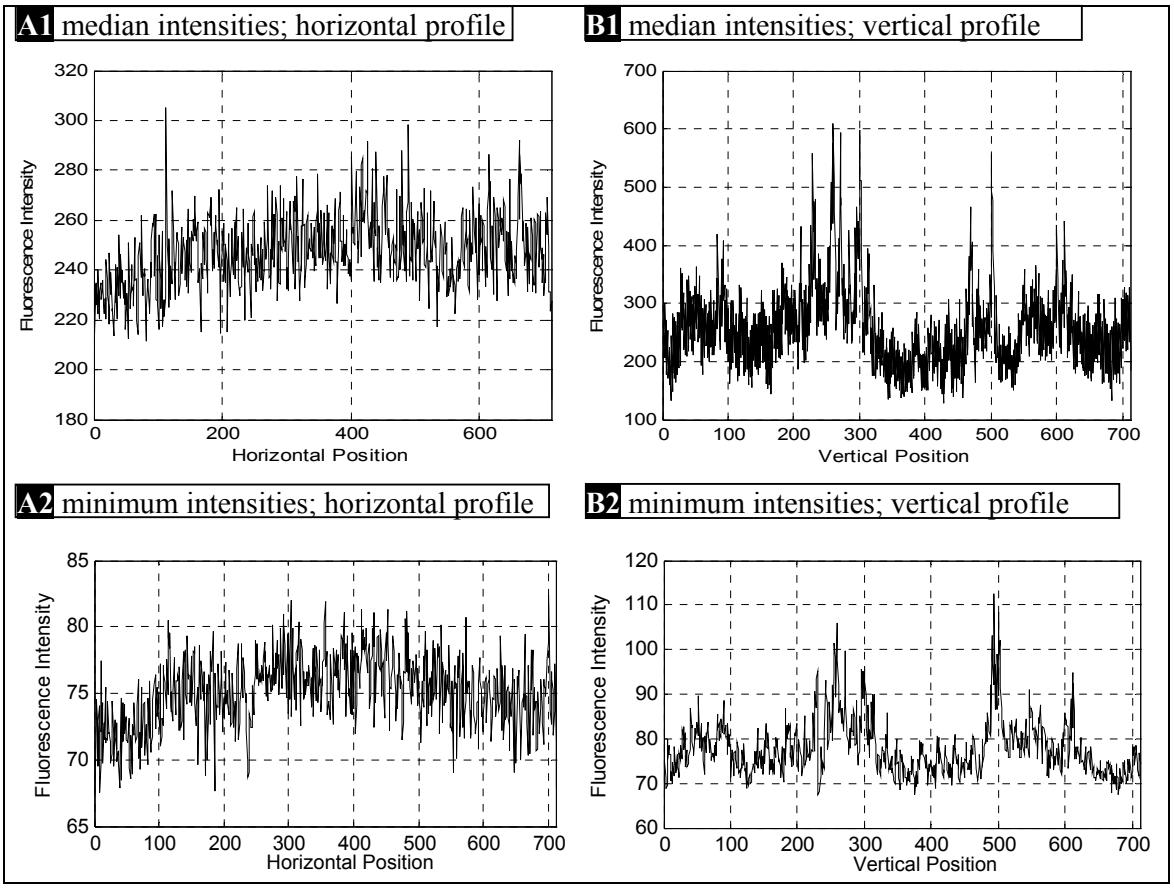


Fig. 4.13 Median and minimum intensities distributed along the horizontal (A1) and vertical (B2) profile of an exemplary U133 microarray.

Regions with different intensities can be seen along the vertical profiles, also when regarding the minimum intensities along the vertical profiles (B2).

Without regions on the chip designed to measure the background signal, Affymetrix uses the lowest 2% of the signals in zones of the chip to define it. It is assumed that there are always cells not showing any signal except for the background, as they cannot hybridize with complimentary target oligonucleotides which are lacking in the sample solution.

Affymetrix calculates the background as follows:

- The array is split up into K rectangular zones Z_K (k=1,…,K, default K=16);
- Control cells and masked cells are not used in the calculation.
- The cells are ranked and the lowest 2% are chosen as the background for that zone.

Distances are then computed for each cell to the center of each zone. The background of this cell is computed using the zone-backgrounds and the distances to the centers of these zones as a weighting function. Weights are calculated based on the reciprocal of a constant plus the square of the distance to all the zone centers.
As can be seen in figure 4.13, it is vital to use zones that are large enough not to be influenced by the stripe pattern of the chips. Using large zones on the other hand limits how accurate the background can be sampled.

4.4.2 Background estimation using the checkerboard pattern

A different method for the calculation of the background is the use of the checkerboard pattern at the border of the chip (see chapter 3.4.3.1). The cells of the checkerboard with low signals are a good representation of the background as there is a linear correlation between the median signal of these cells and the mean signal of the lowest 2% in the microarray as can be seen in figure 4.14.

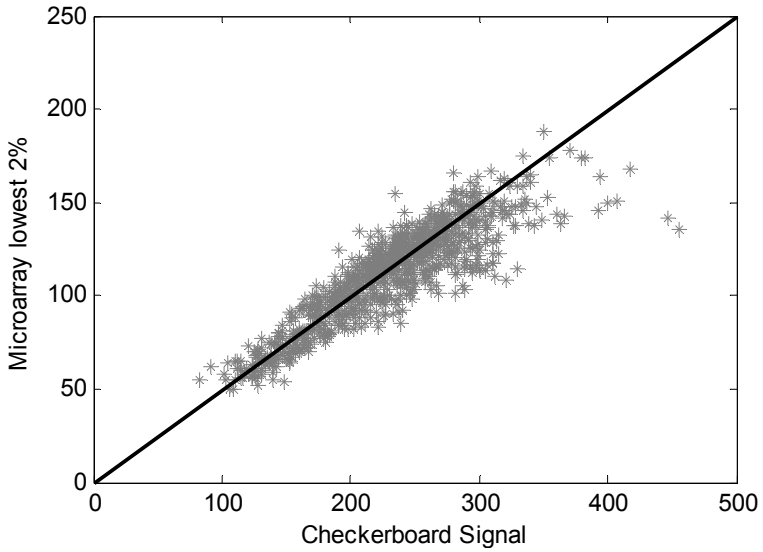


Fig. 4.14 Linear dependency between the median signal of checkerboard cells and the mean signal of the lowest 2% in the microarray using several hundred chips of the type U133a of one study [97].

The ratio between the lowest 2% and the checkerboard signal can be easily obtained by measuring the lowest 2% at a few, large areas in the microarray and comparing it with the checkerboard signals. Using large areas eliminates any local effects. The checkerboard data points are a good estimate for the background after adaptation to the ratio (1:2 in this example) as seen in figure 4.15. A horizontal view was chosen, as the irregular distribution seen in figure 4.13 B2 is not perceived when dealing with columns (figure 4.13 A2).

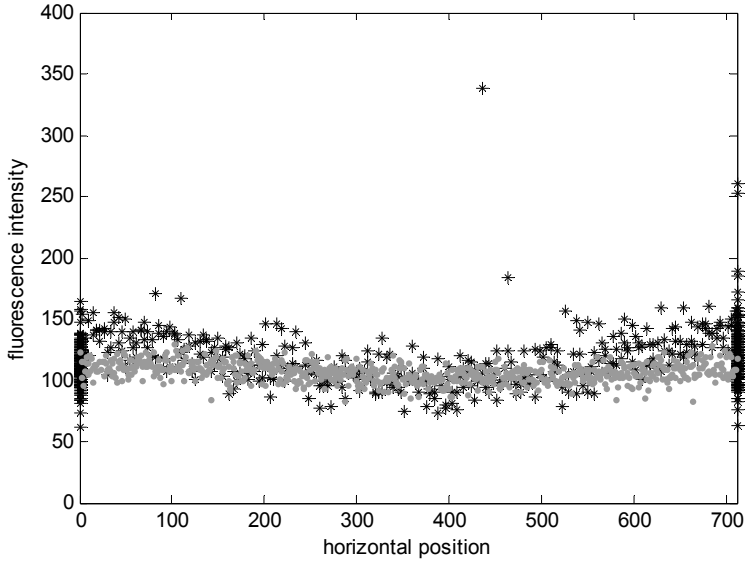


Fig. 4.15 Lowest signals in columns (gray dots) and estimated background using the checkerboard method (black stars).

These data points can be used for interpolation of the background in the microarray, the advantage of this method being that the background is sampled at many more points than with the zone approach described earlier.

As can be seen in figure 4.16, the ratios between the mean lowest 2% signals and the median signals (measured at different locations of a chip and the entire chip) yield similar values in all analyzed U133A microarrays (raw, unscaled data).

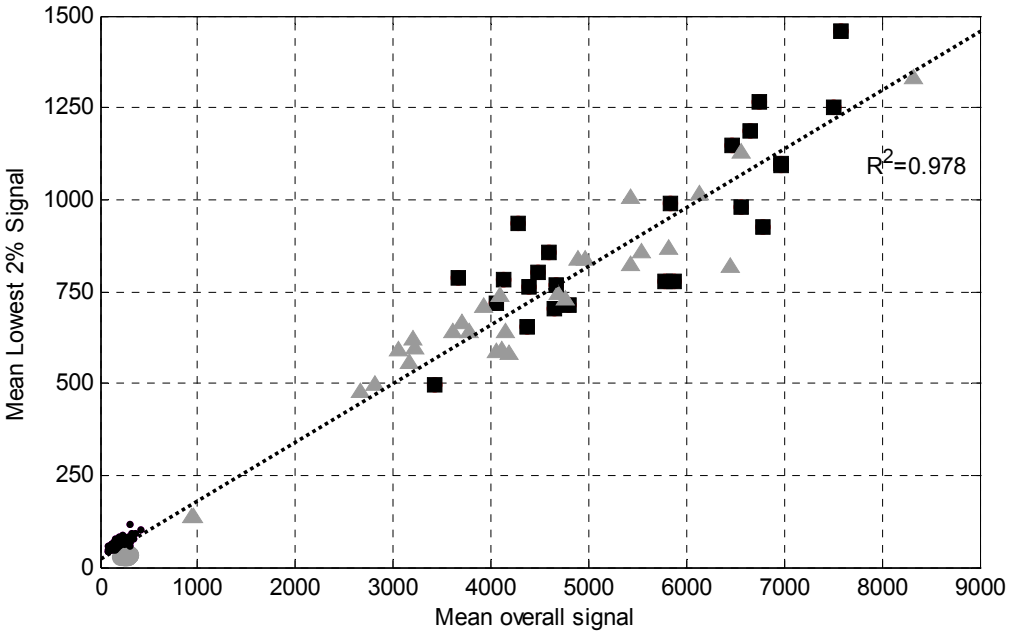


Fig. 4.16 Mean lowest 2% signal against the mean overall signal of several microarrays from four different studies.

The different studies are: Affymetrix spike-in dataset (gray dots) [105], Ross pediatric leukemia dataset (black dots) [86], PGA Human CD4+ Lymphocytes (gray pyramids) [96] and a PGA Human Muscle Obese dataset [104]. Shown are the signals gathered using all cells (excluding Affymetrix control cells).

Raw signals of microarrays from one study can have very different scaling, independent of scanner-settings. Cells that should not show any specific hybridization, e.g. cells containing targets for probes not included in the sample-solution) and are therefore considered as the signal background seem also to be subjected to this scaling factor.

One possible source of different scaling might be the amount of probe-molecules in the sample-solution used for analysis. Capture of the complementary target is modeled by Chan, Graves and McKenzie [98] using two different mechanisms by which targets can hybridize with the complementary probes:

1. direct hybridization from the solution,
2. nonspecific adsorption followed by surface diffusion to the probe.

If the surface diffusion is taken into account, this model might explain a correlation between low-level background cell signals and overall signal intensity.

4.4.3 Interpolation using the Auto-leveling Method (ALM)

An iterative fitting process discards points above a threshold and fits the remaining points with a plane. First, a least squares plane is fitted to the anchor points. If there are fewer points above this plane than below it, they are classified as being outliers and are discarded. This process is repeated until the amount of anchor points above the plane is larger than or equal to those below the plane. The resulting plane which automatically reaches the level of the majority of the data points in this iterative procedure represents the fitted background [106].

4.4.4 Thin-plate interpolation

4.4.4.1 Theory

The interpolation of the background defined by the checkerboard cells is done by using a thin-plate spline. Used correctly, this function will fit the data like a thin plate of metal, able to twist a little bit without overfitting the data. The spline function has to approximate many different data points in space that also contain noise [107, 108].
The surface can be defined by

$$s(\mathbf{x}) = p(\mathbf{x}) + \sum_{n=1}^{N} \lambda_n \phi(\mathbf{x} - \mathbf{x}_n)$$

with

$\phi(\mathbf{x})$ a fixed radially symmetric, basic function,

λ_n a set of N weights corresponding to the N centers,

$p(\mathbf{x})$ a polynomial of degree *k*.

A datapoint f_n can be decomposed into a signal component y_n and a noise component ε_n , i.e.

$$f_n = y_n + \varepsilon_n ,$$

with the ε_n being independent and normally distributed. Interpolation involves minimizing a penalty function.
The penalty function in thin plate splines is

$$J(s) = \int_{\Re^2} \left[\left(\frac{\partial^2 s}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 s}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 s}{\partial y^2} \right)^2 \right] \partial x \partial y$$

As it is not wanted to interpolate noisy data points exactly, a tradeoff is sought, minimizing the penalty function against the mean square error in fitting the data. This leads to the regularized least-squares problem:

$$\min_s \sum_{n=1}^{N} \left[f_n - s(\mathbf{x}_n) \right]^2 + vJ(s)$$

The parameter v attunes the tradeoff between goodness of fit and smoothness. The problem now is to find the best value for v. Exact solutions to the minimization problem can be found in [109]. The selection procedure for v is described further down.

4.4.4.2 Background subtraction

Background data can be fitted best with a nonlinear method. Using a linear background produces fits as seen in figure 4.17.

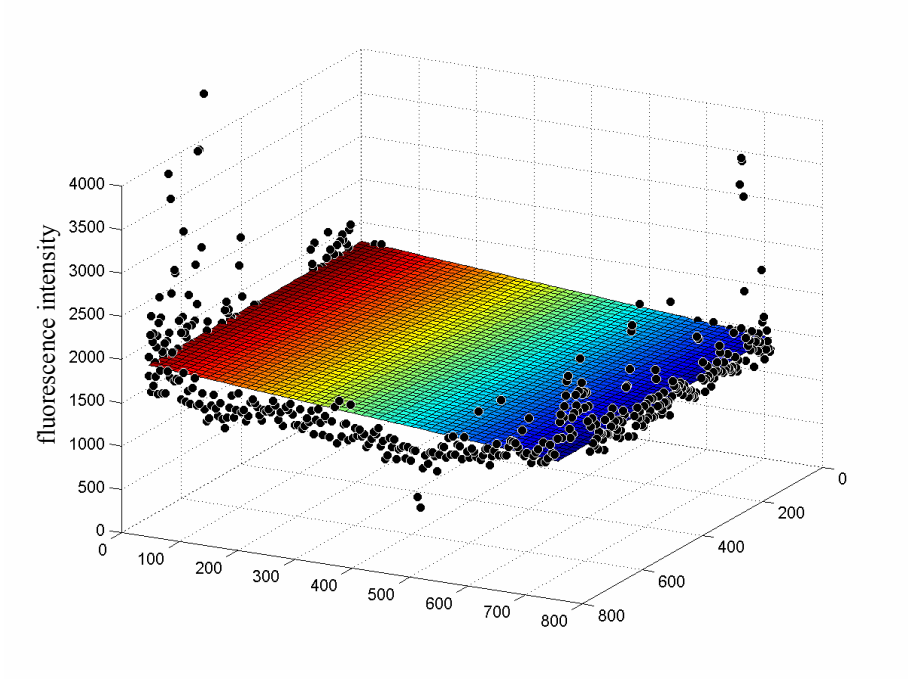


Fig. 4.17 Linear fit to noisy data points.

A good fit or overfitting is achieved depending on the parameter *v*.

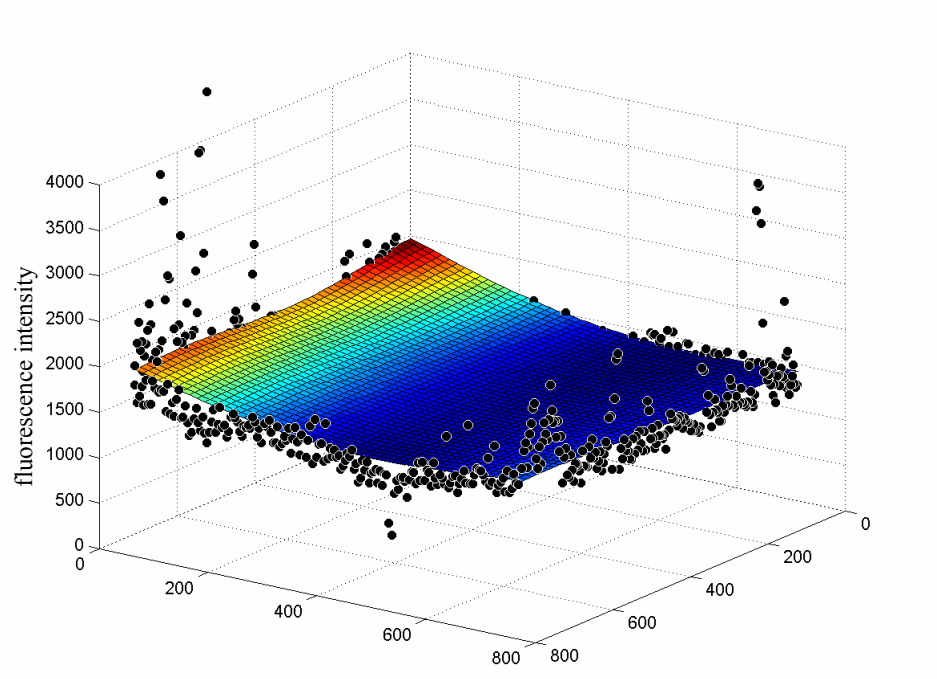


Fig. 4.18 Nonlinear fit to noisy data using a thin-plate spline.

As the checkerboard data is noisy, setting the thin-plate spline too soft can have immense overfitting effects as seen in figure 4.19.

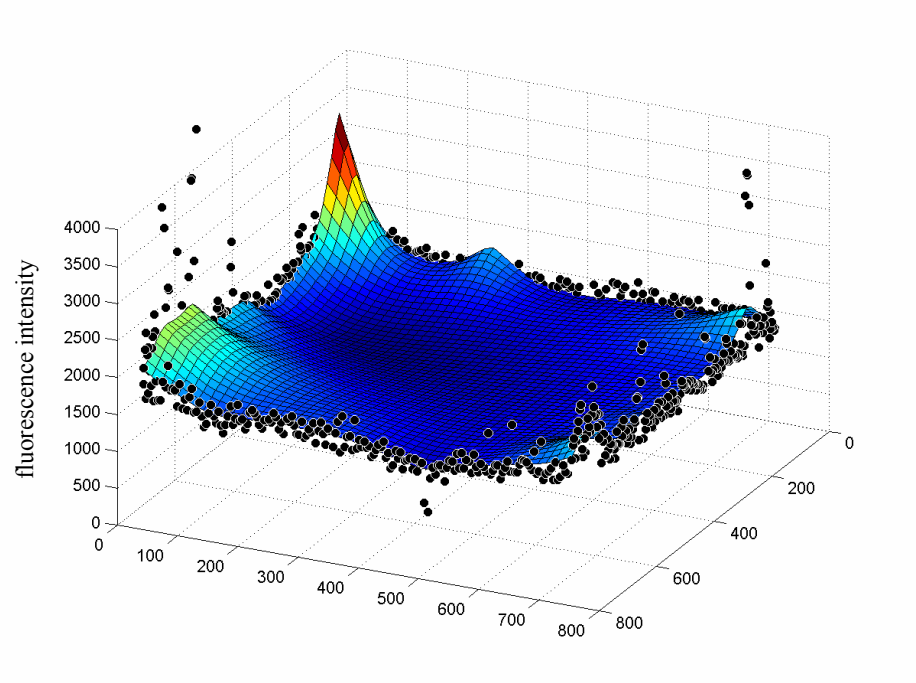


Fig. 4.19 Nonlinear fit to noisy data using a very soft thin-plate spline resulting in overfitting.

A clear overfitting of the soft thin-plate spline is visible. The peak of the thin-plate spline in the back-left corner is created by the influence of a large amount of datapoints that are elevated in relation to the surrounding points. Sparse outliers as seen in the back-right corner only lead to overfitting effects once the thin-plate spline is set much softer. It is therefore not possible to counter the overfitting effect by eliminating some of the visible outliers at this stage.

Although the smoothness parameter can be chosen using the technique of cross-validation to measure the predictive error of a surface [108], an approach using replicate microarray data was chosen. The Latin Square dataset from Affymetrix [105] was used to evaluate the goodness of fit of the calculated background spline. There are 42 very homogeneous replicate measurements of a complex human sample in the dataset, spiked with certain target oligonucleotides. These spike-in cells were discarded and the measurements were overlaid with an arcustangens function, creating a step in the microarray data image that should be fitted.

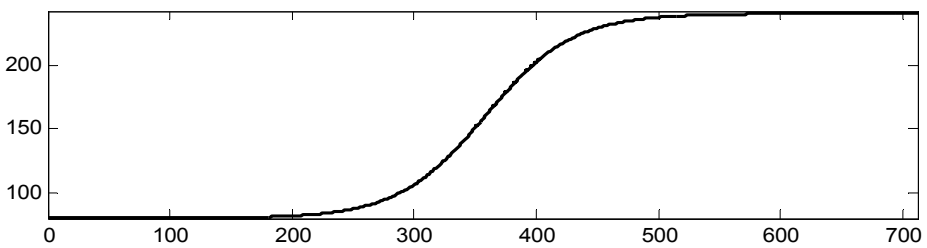


Fig. 4.20 Profile of the simulated uneven background.

This function was added to the measurements in different orientations and in different amplitudes. The form of the step function as well as its amplitude range was selected to be as similar as possible to those steps seen in real-life measurements. The calculated fit was then subtracted from the data. The deviation between same cells of different measurements was compared to the unprocessed data and thus used to evaluate the goodness of fit. It was calculated as a measure of the standard deviation between same cells on all microarrays.

$$DeviationIndex = \sum_{x=1}^{X}\sum_{y=1}^{Y} std(cell_{xy}) * 100 / \sum_{x=1}^{X}\sum_{y=1}^{Y} (sum(cell_{xy})/n)$$

with x,y the horizontal and vertical dimension of the array,
$cell_{xy}$ a vector containing the values of all cells at position x,y,
n the number of microarrays in the study (and the length of the vector $cell_{xy}$).

Using only checkerboard cells for a thin-plate nonlinear interpolation shows bad performance, a clear indication of a bad fit.

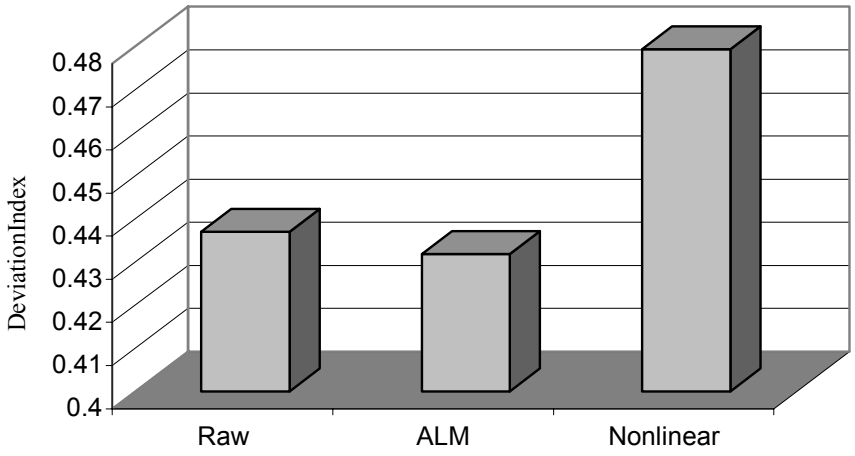


Fig. 4.21 Comparison of goodness of fit of background interpolation methods.

Measurement of deviation between replicate measurements when no background was subtracted (Raw), interpolation using the auto-leveling method (ALM) or a nonlinear thin-plate spline (nonlinear). Only checkerboard data points were used for interpolation.

Here, the strength of the thin-plate spline is also its weakness: it is only defined by border cells, its center is free to be stretched as governed by the smoothing factor.

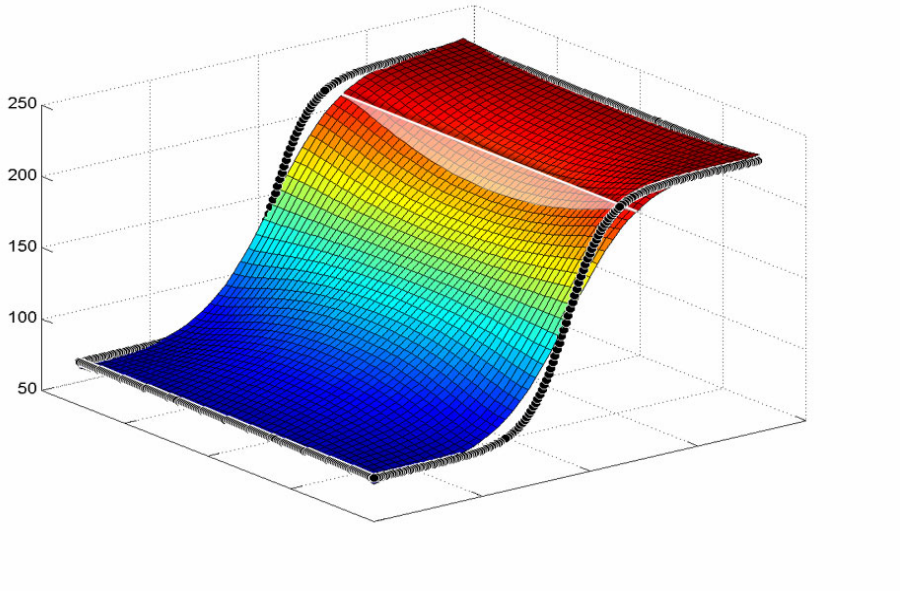


Fig. 4.22 Stretching of a thin-plate spline defined by border anchor points.

Large steps in the data to be fitted result in a stretching of the thin-plate spline, resulting in a bad fit.

This clearly shows that, at least when dealing with larger steps, anchor points within the microarray are needed to compensate for this overstretching. These anchor points were selected with a zone-method as used by Affymetrix. The chip is divided into zones and the mean of the lowest 2% is calculated for this zone. This value is seen as located in the center of the zone. The addition of these anchor points (25 used by default, corresponding to 5 x 5 zones) greatly improves the performance of the nonlinear method.

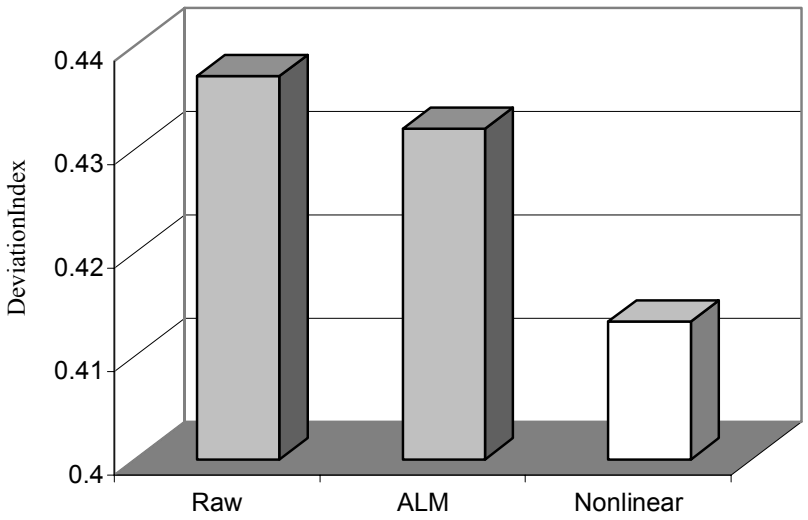


Fig. 4.23 Good performance of the nonlinear, thin-plate spline interpolation.

The thin-plate spline interpolation (Nonlinear), using anchor points at the border of the microarray (checkerboard) as well as in the spotted area, fits the background the best compared with unprocessed data (Raw) and the auto-leveling method (ALM).

The optimization of the parameter v, achieved by this method using the dataset changed by the addition of the tanh-background, lead to a setting of $v = 10^{-5}$.

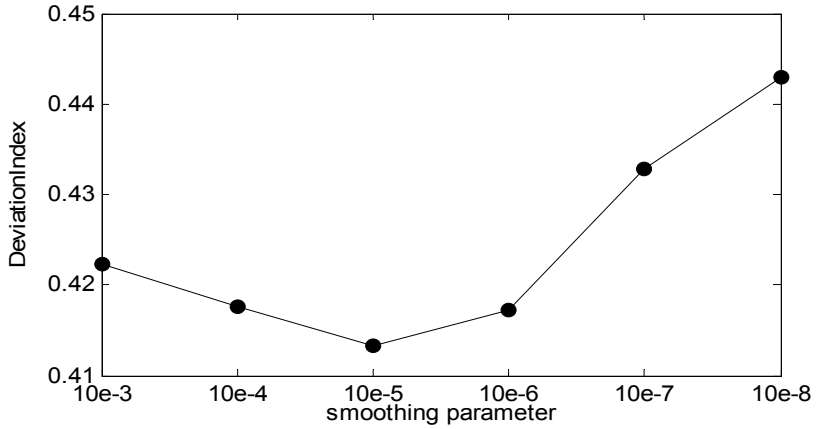


Fig. 4.24 Best results when fitting the surface using the thin-plate spline are achieved when the smoothing parameter is set to 10^{-5}. Higher values make the spline too inflexible, lower ones too flexible, leading to overfitting.

This is the optimal value for background interpolation of the used dataset. The value selected using the tanh-changed measurements is a good choice for most measurements, as the dataset contained a non-uniform background with an intensity difference of a magnitude often seen in

U133 measurements. Application of this method on measurements from different studies [86, 96, 97, 104] showed its performance to be as expected.

4.4.5 Application of a scaling factor

Different measurements are scaled before comparison. These factors compensate for different signal ranges due to different amounts of sample RNA, different gain settings and hybridization efficiencies. The necessity of scaling can be seen in the effect on the histogram of the measurements [110]. As most genes are *not* differentially expressed in different tissues, the overall signal distribution in different experiments should be the same. This is achieved through scaling as seen in figure 4.25.

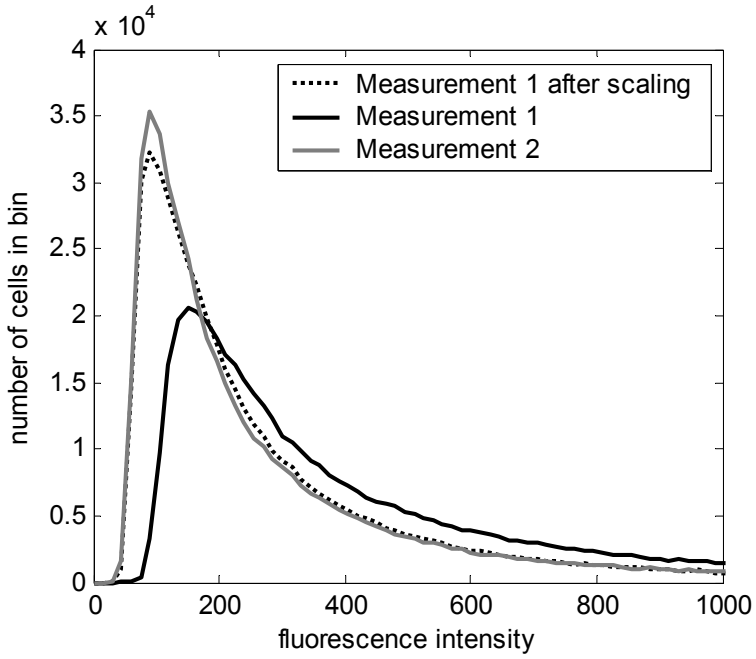


Fig. 4.25 Histograms of different measurements before and after scaling.

Histogram of an experiment before (black line) and after scaling (dotted line) to match the signal distribution of another experiment (gray line).

A homogenous microarray experiment exhibits the same overall signal distribution in different zones, as long as the zones that are chosen are large enough so that they are not subjected to the influence of the stripes with higher intensities as seen in figure 4.13.

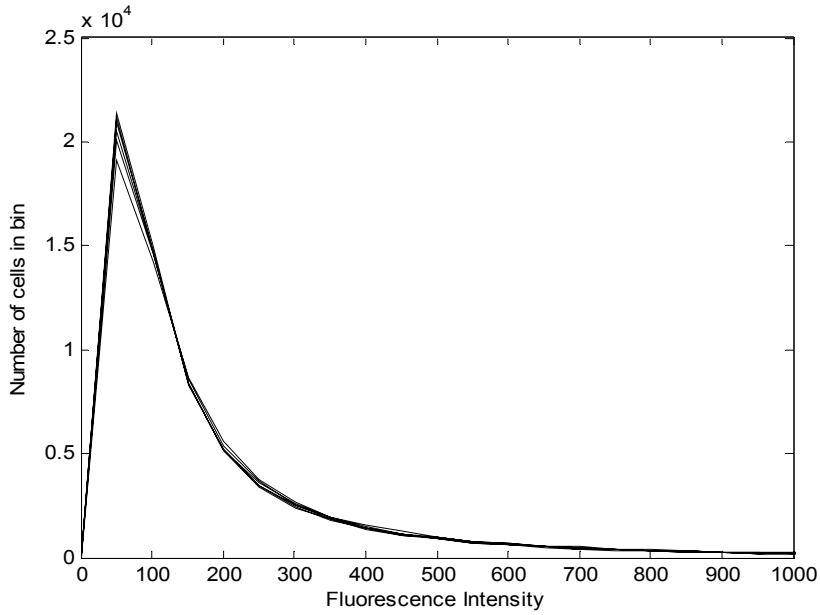


Fig. 4.26 Histograms from different areas in one microarray experiment.

Histograms at different areas (vertical fields of the width of 100 cells) taken from one experiment in the Latin Square dataset. Data not scaled.

Many microarrays are not homogeneous, having different signal distributions in different areas. This can be due to a thermal effect, incomplete washing, diverse hybridization performance etc. One example is seen in figure 4.27. A clear intensity drift from left to right can be observed. The majority of the U133 microarrays of the leukemia dataset analyzed in chapter 5 had an uneven intensity profile, although not as pronounced as seen in figure 4.27. Another example is shown in the leukemia dataset analysis chapter.

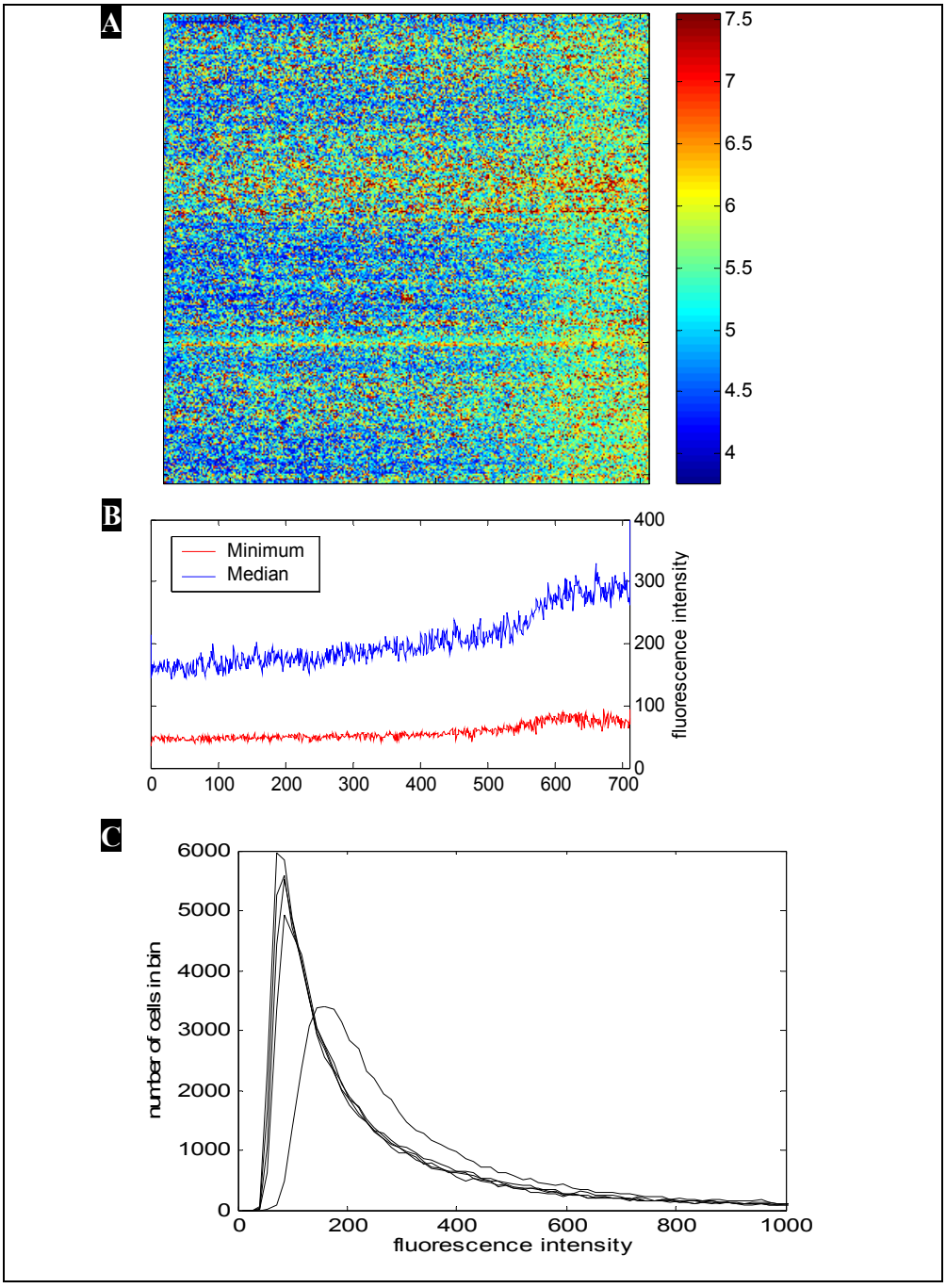


Fig. 4.27 Microarray with horizontal intensity-drift.

A) Microarray CEL-image, the colorbar denotes fluorescence intensities in logarithmic scale. An intensity drift from left to right can be observed, as well as the horizontal stripes common to U133 chips. B) Minimum intensity (red) and concurrent median intensity (blue) horizontal profiles. C) Histograms taken at different areas of the microarray (see figure 4.26).

The histograms in figure 4.27 C and analysis of microarrays from the leukemia dataset used in chapter 4 make it clear that an in-chip scaling is necessary. The intensity of the background in the microarray is subject to the same local scaling as the overall signal, as can be seen in figure 4.16, where overall mean signals are plotted against overall signal minima.

It is thus possible to use the same algorithm to calculate a background surface, and a surface used to deduce the local scaling factors. The result of this local scaling of the microarray experiment shown in figure 4.27 is seen in figure 4.28 B.

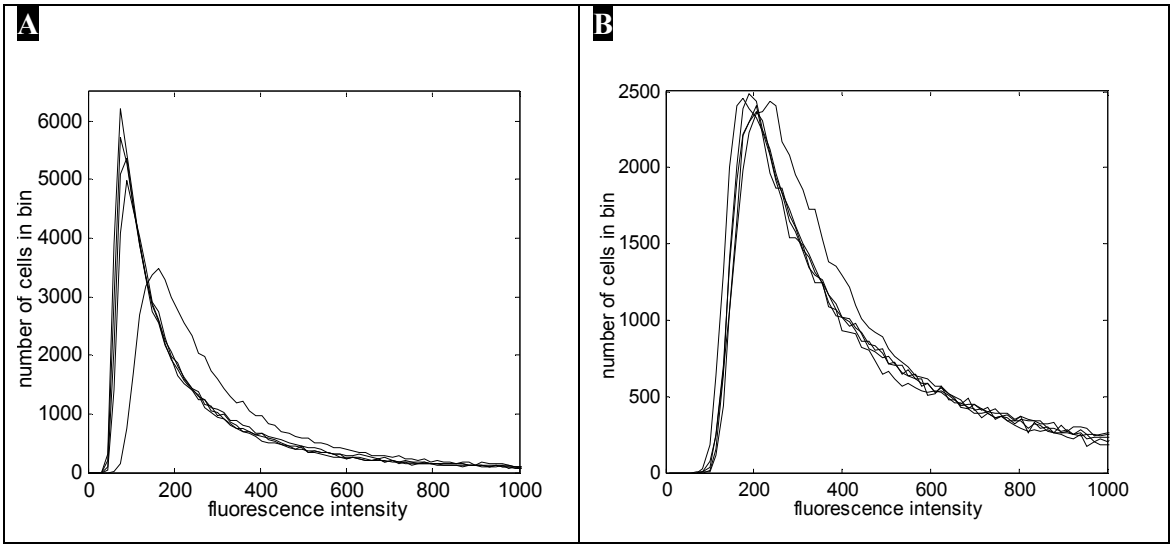


Fig. 4.28 Histograms of different zones on one microarray before (A) and after local scaling (B), see also figure 5.2.

Microarray data in (A) was baseline corrected but not scaled. A clear deviation between histograms is visible but disappears through local scaling (B).

4.5 Probe Sequence Development

4.5.1 Introduction

To create a DNA microarray, oligonucleotides are immobilized onto the surface of a slide. These probes then hybridize with complimentary, fluorophore labeled target-DNA from a sample-solution. It is therefore essential to design a certain immobilized oligonucleotide so that only a target-DNA molecule with a matching sequence will hybridize with it in a highly selective way. It is necessary to find a perfect match sequence for the hybridization with the target-DNA molecule that exhibits as many mismatches as possible towards other target sequences.

A tool was created to facilitate the design of probe sequences for an animal species differentiation microarray under development at the Institute for Chemical and Biochemical Sensor Research (ICB). This microarray was developed by V. Podsadlowski [111] for a fast genetic analysis of food samples using a real-time microarray hybridization and scanning device also developed at the ICB.

4.5.2 Process of probe sequence development

The first step in the development of optimal probe sequences is the alignment of the target sequences which are obtained after a PCR and meant for analysis. This is done using Clustal W, a general purpose multiple sequence alignment program for DNA [112]. This program aligns the sequences, introducing gaps if needed. In a second step, a region in the aligned sequences is searched for which is the most different between the one sequence the probe molecule is meant to hybridize with and all the other sequences. If four samples are to be differentiated, the aligned sequence of sample 1 is taken and compared with the sequences of the other samples, then sample 2 is chosen and so forth. Tables 4.1 and 4.2 show subsequences (20 bases long) derived from the entire aligned sequence of two samples. A mismatch score is calculated for the 20 bases long excerpt by the algorithm in a moving-window approach .

Tab. 4.1: Sequences of two samples aligned using Clustal W. Mismatch scores are calculated in a 20-bases long moving window scheme.

```
Sample 1      CCATGAGGACAAATATCATTCTGAGGAGCAACAGTCATTACCAACCTTCTCTCAGCAATT
Sample 2      CCATGAGGACAAATATCATTCTGAGGAGCAACAGTTATTACCAATCTTCTCTCAGCAATC

Sample 1      CCATATATTGGGACAAACCTAGTCGAATGGATCTGAGGGGGCTTTTCAGTAGACAAAGCA
Sample 2      CCATACATTGGTACAAACCTAGTTGAATGAATCTGAGGAGGCTTTTCAGTAGACAAAGCA

Sample 1      ACCCTAACCCGATTTTTCGCTTTCCACTTTATTCTCCCATTTATCATCGCAGCACTCGCT
Sample 2      ACCTTAACTCGATTCTTCGCTTTCCACTTTATTCTACCATTCATCATTGCGGCACTTGCT

Sample 1      ATAGTACACTTACTCTTCCTTCACGAGACAGGATCTAATAACCCAACAGGAATTCCATCA
Sample 2      ATAGTACATTTACTCTTTCTTCACGAGACAGGATCCAATAACCCAACAGGAATCCCATCA

Sample 1      GACGCAGACAAAATCCCCTTTCATCCTTATTATACCATTAAAGATATCTTAGGCATCTTA
Sample 2      GATGTAGATAAAATTCCCTTTCATCCCTACTACACCATTAAAGATATTTTAGGCATCCTA

Sample 1      CTTCTAGTACTCTTCTTAATATTACTAGTATTATTCGCACCAGACCTACTTGGAGACCCA
Sample 2      TTCCTATTTCTCTTCTTAATAACACTAGTACTATTTGCACCAGACTTGCTTGGAGACCCA

Sample 1      GATAACTACACCCCA
Sample 2      GACAAATACACTCCA
```

Tab. 4.2: Mismatch scores for five different 20 bases long excerpts from the sequence in table 1 at different positions.

Score	Sequence 1	Sequence 2	Mismatch position
1	TCCTTCACGAGACAGGATCT	TTCTTCACGAGACAGGATCC	-*------------------*
8	ACAGGATCTAATAACCCAAC	ACAGGATCCAATAACCCAAC	--------*----------
13	CCTAGTCGAATGGATCTGAG	CCTAGTTGAATGAATCTGAG	------*-----*-------
17	CTCCCATTTATCATCGCAGC	CTACCATTCATCATTGCGGC	--*-----*-----*--*--
29	ATCTTA CTTCTAGTACTCT	ATCCTA TTCCTATTTCTCT	---*---*-*---*-*----

The weighting of the mismatch score for each mismatch is primarily based on an analysis of the position of the mismatch in the 20 bases long subsequence. It is known that mismatches at the border of the sequence affect a hybridization less than a mismatch in the center of the sequence. As can be seen in table 4.2, an aligned subsequence with a mismatch in its center scores higher than a subsequence with two mismatches near its borders. The location dependant weighting function has to take into account the above mentioned behavior. The function used in table 4.2, for example, rises linearly with growing proximity to the subsequence center; the score $b_{sc,p}$ for each basepair at the position p of the subsequence of length L is

$$b_{sc,p} = (p-1) \quad \text{if } p \le L/2,$$

$$b_{sc,p} = (L-p) \quad \text{if } p > L/2$$

for a mismatch and $b_{sc,p} = 0$ for a matching basepair. The score is the sum of all basepair scores

$$score = \sum_{p=1}^{L} b_{sc,p} \ .$$

One further consideration is the so called GC-content of an aligned subsequence. The GC base pair ineraction is stronger than the AT one, as one more hydrogen bridge is created (moving the equilibrium of the hybridization reaction farther in direction of the hybridized species). The chosen sequence should therefore also have a high GC content.

Figure 4.29 shows the result of the mismatch score analysis of the comparison of a coney CytB-gene sequence with the CytB sequence derived from five different animals (pork, fallow deer, horse, rabbit, human) [111].

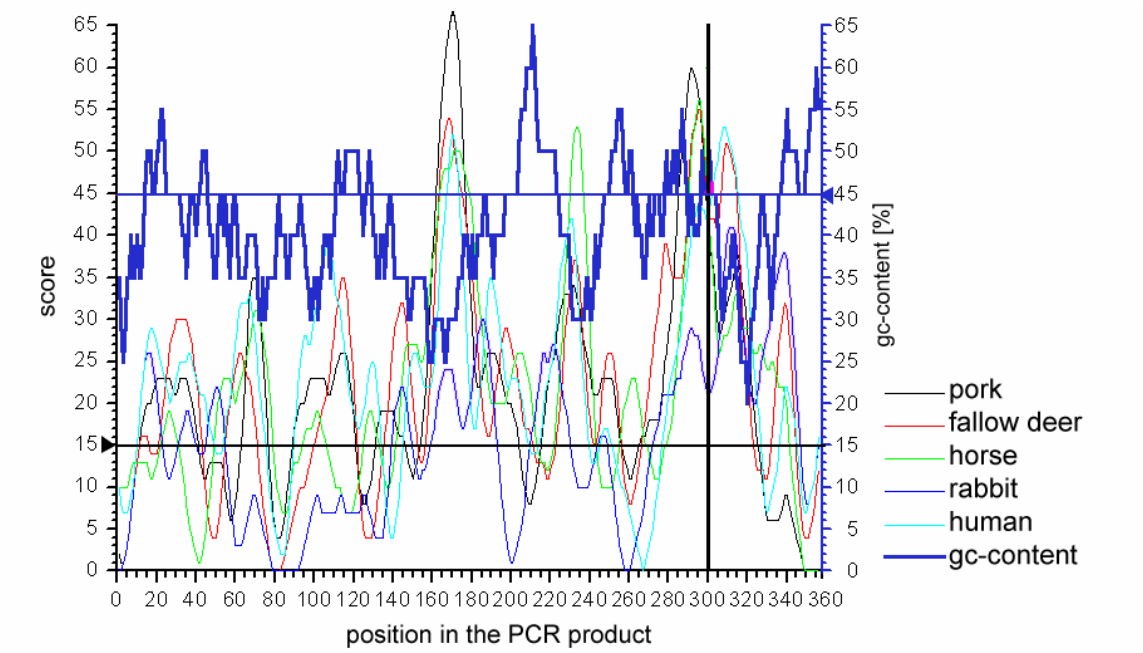


Fig. 4.29 Mismatch scores for the comparison of a CytB coney sequence with the same gene from five other species.

Compared were subsequences of a 360 bases long CytB gene sequence. The optimal subsequence region is marked with a line at base number ~300.

Application of the tool showed an useful selection to have a mismatch score of at least 15 in all coney / other species comparisons and a GC content of at least 45 %. It was shown during the design phase of the animal species differentiation microarray that the algorithm worked as predicted, facilitating the microarray development. Table 4.3 shows the results of food samples analyzed by the CVUA (*Chemischen und Veterinär Landesuntersuchungsamt*, Münster) and by means of the microarray developed at the ICB [111]. All animal species were detected correctly.

Tab. 4.3 Results of food samples analzed by the CVUA and by means of the microarray developed at the ICB [111]

Sample source identification by CVUA	Identification using the microarray developed at the ICB
haunch of venison	venison
wild boar roast	wild boar
haunch of hare	hare
deer medallions	deer
haunch of deer	deer

Some of the developed sequences are for the identification of origin for samples from the species horse, goat, buffalo, fallow deer, deer, sika deer, red deer, hare, coney, human and dog.

4.6 Discussion of signal processing methods

Almost all of the several hundred U133 microarrays analyzed contained artifacts. Their size and form varies substantially, from small round speckles or fine scratches up to large areas containing ten thousands of cells. Detection of these cells is important as their utilization in the calculation of the expression value for a certain probeset can have drastic consequences depending on the gene expression summary algorithm used. As the MicroArray Suite of Affymetrix does not provide for an automatic artifact detection scheme, the method developed in this work proves to be a good way to tackle this task in an easy and reliable way. The algorithms have been fine-tuned using hundreds of microarrays so that a minimal number of differentially expressed cells are flagged as being outliers. Outlier detection is done by calculating the deviation of a cell from the median signal of this cell over several microarrays. This implies that several microarray experiments have been made to gain the necessary amount of data. Median chips created by using measurements of different studies can differ from one another, the best and most reliable method therefore being the utilization of data from one and the same study for analysis and the creation of the medianchip. To provide for a

medianchip when only very few measurements are made, a library of medianchips can be created. A medianchip that has the closest fit to the new measurements can then be selected. The storage of medianchips can be easily done by using the developed microarray data management system. This management system was developed in accordance with the rules of the MIAME project approved by the Microarray Gene Expression Database group.

Detailed information on the nature of the artifacts further provide quality measures giving information on the entire analysis process. One example is the absence of small round artifacts, previously found very often on microarrays, the reason being that scientists have switched from using powdered to non-powdered laboratory gloves when doing microarray experiments. These measures therefore provide information that can be helpful for the entire research process, the ultimate goal being the enhancement of the quality of the entire process and the beforehand reduction of the possibility of artifact creation.

Unfortunately, U133 microarrays do not show a homogeneous signal distribution. There are horizontal stripes with higher signal intensities. These stripes make a more direct identification of abnormalities impossible. This includes the calculation of the background signals, as the U133 microarrays do not provide any direct means to measure it. A disadvantage of the Affymetrix method for background correction is the small number of anchor points available for the interpolation of the background signal, as the zones defined have to be large enough to evade the effect of the intensity stripes previously discussed (4x4 zones are normally used). The developed method, on the other hand, uses the information of these anchor points as well as several hundred additional points provided by the checkerboard area of the microarray. The number of available data points makes a thin-plate interpolation possible. This method has shown to be well suited in interpolating noisy data in many sectors like meteorology and geology. The importance of a good area dependant background correction scheme becomes visible when regarding histograms from different areas on the microarray before and after background correction (see figure 5.2). The second step is the scaling. The MicroArray Suite scales the entire microarray by one value, ignoring area dependant differences in overall signal values. This area dependant discrimination is essential to match the different signal distributions present at different areas on a microarray. The developed method for the interpolation of background signals can also be used for the area-dependant scaling of the data. When applied to several different microarray measurements from different series a good matching of histograms from different areas was observed. Application of the methods on replicate measurements provided by Affymetrix showed a clear improvement in signal quality.

5 Analysis of Leukemia Data

5.1 Introduction

A public pediatric leukemia dataset [86] was selected and analyzed to find a list of genes best suited to differentiate several subgroups of cancer and to analyze the effect of using different gene selection methods. Classification was done using a SVM classifier. The selection algorithms used were Fisher ratio, gene shaving, significance analysis of microarrays (SAM) and prediction analysis of microarrays (PAM). The different gene expression summary algorithms used include the Li-Wong model, RMA and the MAS perfect match only model.

5.2 Data Source and Composition

The data used was first presented in the paper „*Classification of pediatric acute lymphoblastic leukemia by gene expression profiling*“ published in BLOOD, 15 October 2003, Vol. 102, Number 8 [86].

As described in the leukemia chapter, the exact classification of the disease into several subgroups is essential for a tailored therapy and good survival prognostics. The pediatric acute lymphoblastic leukemia (ALL) dataset that has been analyzed can be subdivided into several sample subgroups. These groups differ in their cellular and molecular characteristics and also in their response to therapy. The leukemia subgroups contained in the dataset are shown in table 5.1.

Tab. 5.1 Subgroup distribution of samples in the dataset

Sample Type	Numbers
T-ALL	14
B-Cell lineage:	
BCR-ABL	15
E2A-PBX1	18
Hyperdiploid with more than 50 chromosomes	17
MLL	20
TEL-AML1	20
Others:	
Novel	13
Hypodiploid	4
Normal_diploid	3
Pseudodiploid	8
Total Sample Number	**132**

The *Novel*-subgroup was found in an ALL study performed with U95 microarrays. This study is the predecessor of the above mentioned paper and was published in 2002 ("*Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene eypression profiling*", Yeoh EJ, Ross ME, Shurtleff SA, *et al*., Cancer Cell, 2002; 1:133-143) [113]. The new study analyzed here uses a subset of the data analyzed by Yeoh *et al.* The *Novel* subtype can also be seen in the data obtained using U133 microarrays.

5.3 Preprocessing

The gene expression value for each probeset was calculated using different methods. The data preprocessing consisted of the following steps:

1. The same background correction and scaling methods were used as described by the authors of the different gene expression summary techniques.
2. Data of one sample was scaled independently for each U133A and U133B chip to a trimmed mean of 250 and then combined.
3. Genes having the following characteristics were discarded:
    - expression below 100 in all samples,
    - flagged as not present in more than half the samples (using the MAS flags),
    - being in saturation.
4. Genes were auto-scaled.

The Li-Wong, RMA and MAS pm only models were applied using the statistical software R [114].

5.4 Quality Aspects

5.4.1 Time of measurement

The date at which the sample was scanned is included in each Affymetrix CEL-file. It is possible to analyze if there is a relation between the information extracted from the data or about the data (e.g. sample subgroup) and the measurement time. The ideal measurement procedure would be to measure all samples randomly and as close as possible in time.

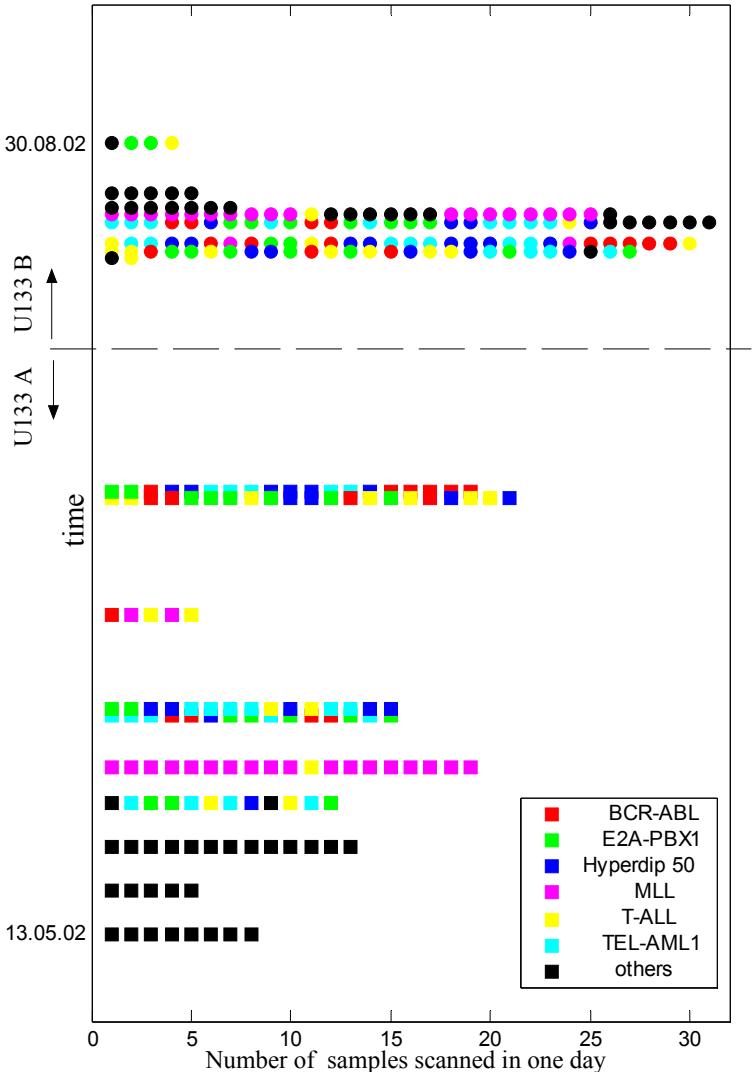


Fig. 5.1 Chart of measurement dates of the samples (color coded) and chip types (squares = U133A; circles = U133B).

Measurement took place throughout three and a half months. A and B chips were scanned separately and some sample groups were scanned nearly completely as one block (*Others* and *MLL* samples U133A microarrays). The ideal would have been a complete random scanning of all samples to evade a possible influence of measurement time, although no apparent influence can be seen in this study. Samples are represented as colored circles or squares, some of these being drawn close to one another (some are not visible in their entire breadth).

Both chips, the U133A and U133B, should be measured together. As can be seen in figure 5.1, some groups of samples were measured nearly completely in one block (e.g. *MLL* or *Others*). U133A and U133B chips were scanned at two different time intervals.

5.4.2 Homogeneity of Chips

In-chip background and scaling can vary greatly. Background correction and scaling are essential, as can be seen in figure 5.2. Signal distributions at different location of chip JD-ALD051 (*MLL* subgroup, B-chip) differ greatly before preprocessing. A background correction as is done by default in the MAS software helps considerably to make the overall signal distribution more homogeneous. An area-dependant scaling is also necessary to match the histograms but is not done by the MAS software and thus not provided for in the original data.

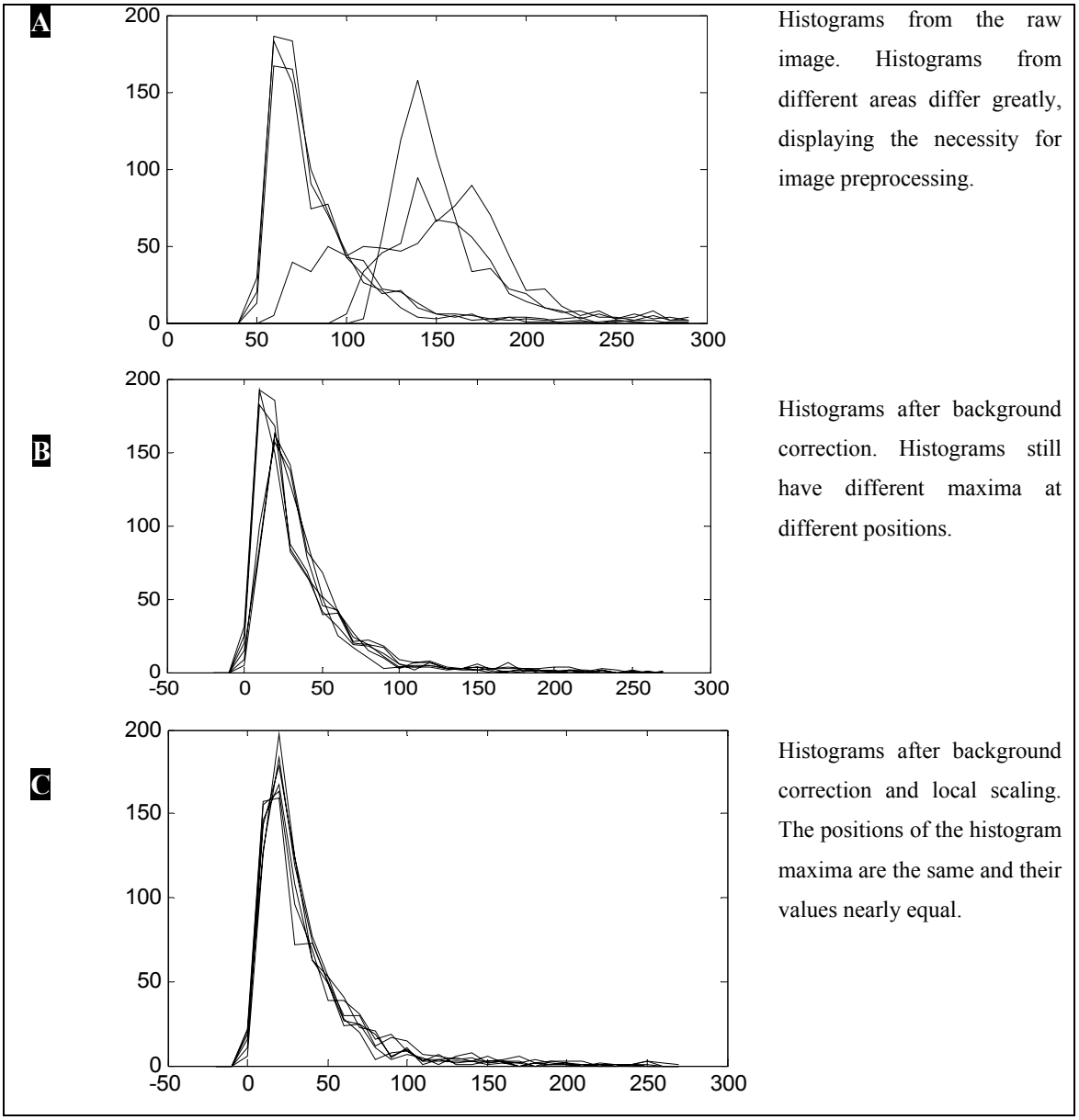


Fig. 5.2 Histograms from different sectors in measurement JD-ALD051 (*MLL* subgroup). Background correction and in-chip scaling are essential for making the data usable.

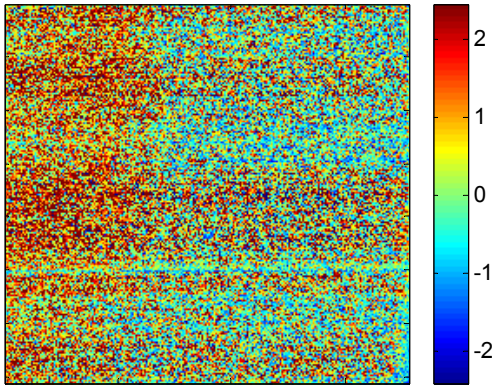


Fig. 5.3 Ratioimage of the microarray (logarithmic scale) discussed in figure 5.2. Clear differences in signal distribution are visible.

5.4.3 GAPDH 3’ / 5’ Ratio

RNA can degrade during cDNA synthesis. The result is first-strand cDNA containing a mix of truncated and full-length transcripts. Measurement of 5’-end and 3'-end fragments of a long RNA, here glyceraldehyde-3’-phosphate dehydrogenase (GAPDH), can give information on the quality of the RNA. The ratio of the 3' to 5' amplified fragments provides a direct indication of RNA integrity since in most cases RNA degradation starts in the 5'-end region of a RNA molecule.

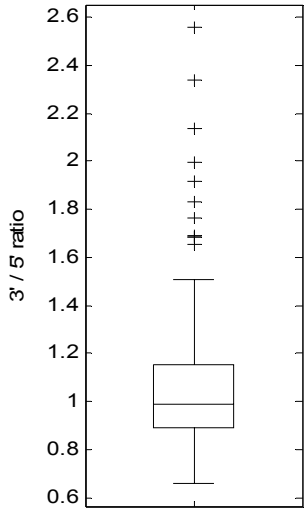


Fig. 5.4 Boxplot of the GAPDH 3’/5’ ratios of all 132 samples. All samples have a ratio below 3.

Quality control during creation of the dataset enforced the 3’ / 5’ ratio to be well below 3. More important, no trend for a specific ratio for a certain sample subgroup can be found.

5.4.4 Present Calls

Analysis of the present calls of the samples showed that all U133A chips have higher present calls than U133B chips. This is usually the case as U133B chips contain more expressed sequence tags (ESTs), while U133A chips contain a higher amount of known gene sequences. Median present calls are 36% for U133A and 21 % for U133B chips in this study. Also, different sample subgroups neither had overall different present calls nor were the values time related (see figure 5.5).

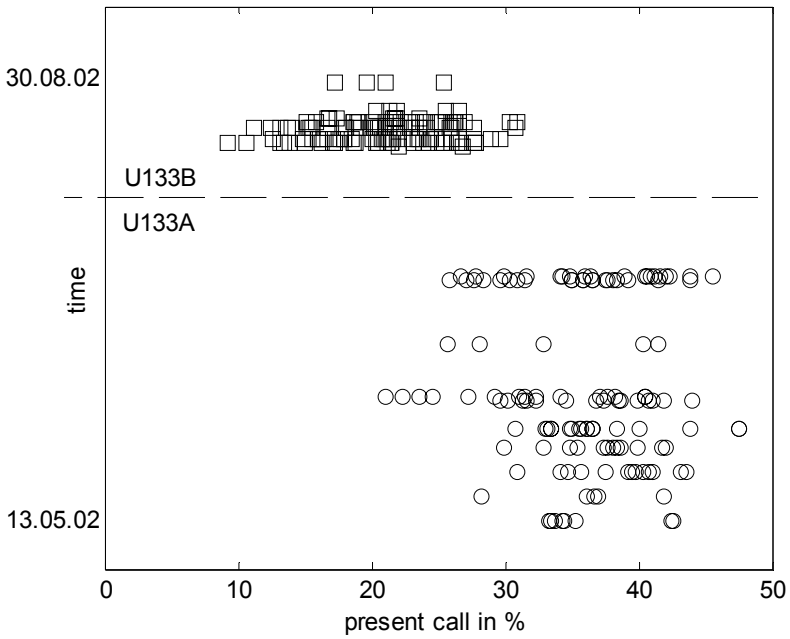


Fig. 5.5 Percent of signals called *present* in U133A chips (circles) and U133B chips (squares). The fraction of present calls is higher in U133A chips as expected, with a mean around 35 % as observed in many other studies. U133B chips have a lower overall percentage as they contain many ESTs.

5.4.5 Number of Affymetrix Outliers and Masked Cells

The number of outliers as flagged by the MAS software is generally higher in U133B chips with a wider distribution. Two U133A samples have a higher number of outliers (JD-ALD109, TEL-AML1 subgroup and JD-ALD-052, MLL subgroup) but a normal amount of masked cells, baseline levels and signal distributions.

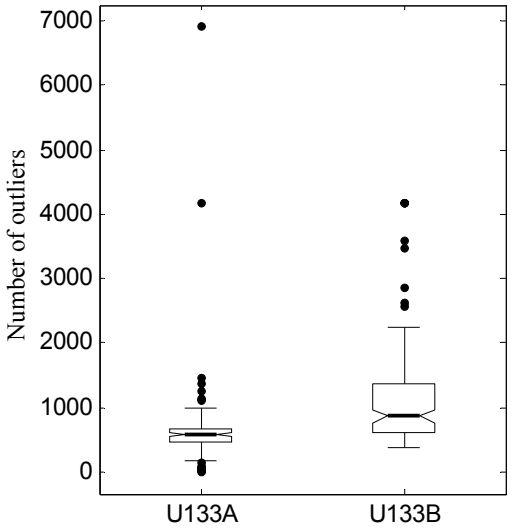


Fig. 5.6 Boxplot showing the number of Affymetrix outlier cells in different microarrays.

The number of cells masked by the new algorithm (see chapter 4.2.2) is also higher in U133B chips. One U133A chip distinguishes itself by having nearly 35000 masked cells (marked as A in figure 5.7). This is a microarray with a very large artifact in the upper left corner (JD-ALD416, *E2A-PBX1* subgroup) which was already analyzed closer in chapter 4.3.2. It is therefore possible to select these kinds of microarrays automatically by monitoring the deviation concerning masked cell numbers from that of the rest of the population.

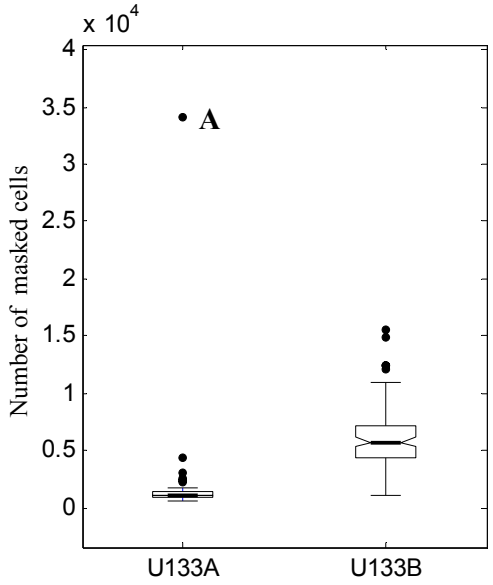


Fig. 5.7 Boxplot showing the number of masked cells in different microarrays. One microarray, marked as (A), contains a very large artifact (JD-ALD416, *E2A-PBX1* subgroup). It is analyzed closer in chapter 4.3.2.

Generally, there is no connection between the number of masked cells and the number of Affymetrix outlier cells.

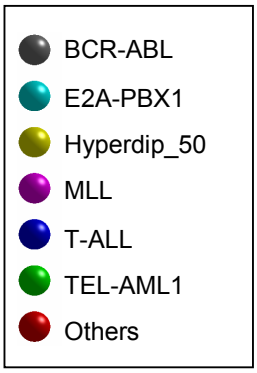
The authors of the original paper wrote “Microarray scan images were visually inspected for apparent defects […].” ([86] page 2952). It is interesting to note that the microarray denoted as (A) in figure 5.7 did not have any flagged cells in the raw data provided by the authors. Cells in the area of the artifact were thus probably also used for the calculation of the probeset specific gene expressions using the MAS software, although it is possible that the authors of the paper had used a different set of files with flagged cells.

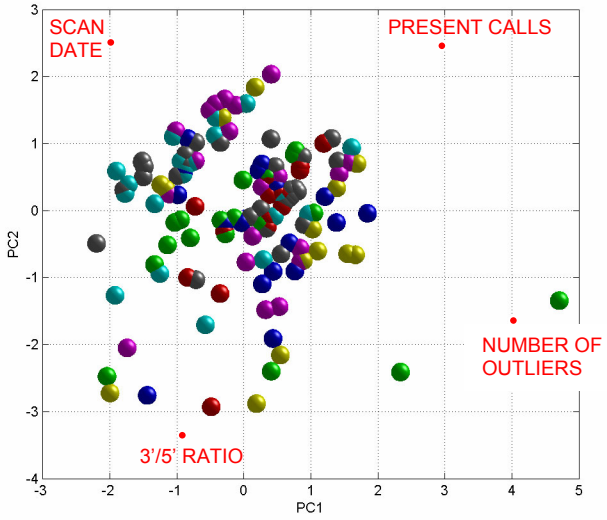
5.4.6 Relation between sample class and sample quality

The relationship between samples as well as the relationship between variables can be visualized by using a principal component biplot in which the scores-plot as well as the loadings-plot of the principal component analysis are diplayed in the same diagram. Figure 5.8 contains the bi-plots of all U133A as well as all U133B samples using their corresponding values of the quality criteria (scan-date, percentage present calls, 3’/5’ ratio, number of outliers). It can be seen that the number of present-calls is anti-correlated with the height of the 3’/5’ ratio, i.e. samples with a low ratio also have a high present call value thus being of overall good quality. Samples with below average quality have both high 3’/5’ ratios and low present call values. It is visible that there is no clear separation of any sample-subgroup exists. No subgroup can be discriminated from the rest of the samples by using quality measure alone or a combination of these.

A

PCA Biplot showing U133A samples and quality criteria





B

PCA Biplot showing U133B samples and quality criteria (legend as seen under A)

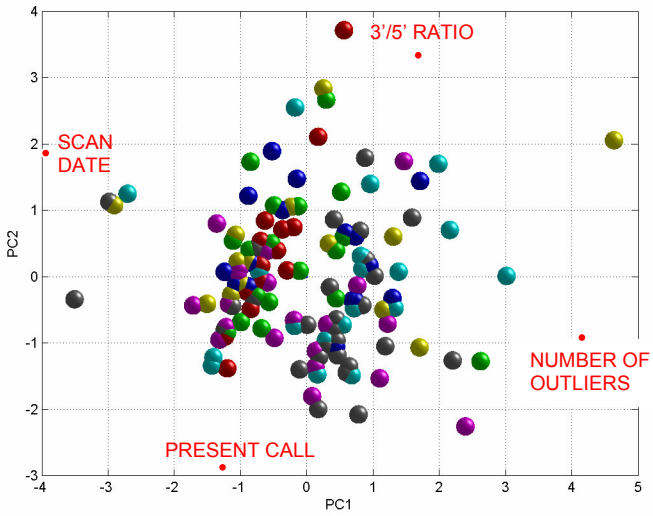


Fig. 5.8 Biplots created using quality measures. No subgroup can be discriminated from the rest of the samples by use of any quality measure alone or a combination of these.

5.5 Gene Selection

5.5.1 Introduction

Not all sample subgroups are separated well using all genes for discrimination as can be seen in figure 5.9. The only subgroup falling entirely into one cluster is *T-ALL*, thus being clearly separated from all other samples, that is, members of the B-cell lineage. Most genes measured are not differentially expressed. The including of these non-informative genes in a classifier would only increase the amount of noise in the system. These genes therefore have to be left out. Even when regarding only those genes that are differentially expressed, it is important not to use too many genes. Although the apparent error rate (the proportion of training tissues misclassified) of a classifier would decrease when using more informative genes, its error rate in classifying tissues outside of the training set would eventually increase. It is therefore necessary to select a few, highly selective genes for better discrimination [115]. This goes hand in hand with the desire to shrink the number of genes used for discrimination to make it possible to design a low-density diagnostic chip. Therefore, the target is to select a minimum number of genes suited to achieve the best separation of sample subgroups [116]. Four methods were used for gene selection, one being the calculation of the Fisher ratios of each sample subgroup against the rest of the samples. This provides genes that have the most different gene expression patterns between these two sample groups. The second procedure is the so-called gene shaving which clusters genes according to similar gene expression patterns across samples. The third and fourth are the significance analysis of microarrays (SAM) [85] and the prediction analysis of microarrays (PAM) [87] methods. The data was preprocessed as described in 5.3 if not otherwise described.

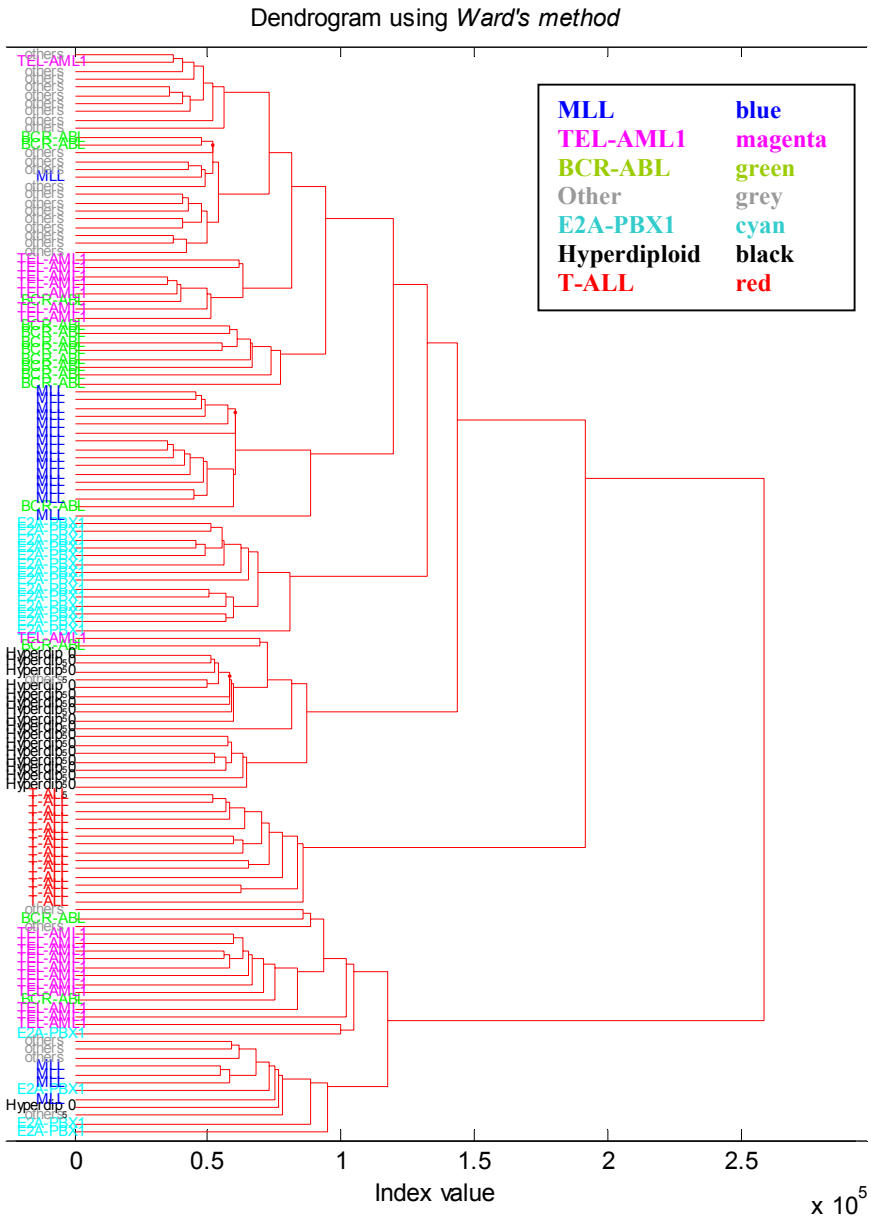


Fig. 5.9 Dendrogram of all samples using all genes. Samples of the same subgroups do not necessarily fall into same clusters. Only T-ALL is clearly distinguished from all B-cell lineage samples. A gene selection procedure is necessary to increase sample subgroup separation.

A SVM-classification (see classification chapter) using all genes provided a classification accuracy of 92.1%. A comparison with accuracies reached using fewer genes (see further below) also illustrates the necessity of gene-selection.

5.5.2 Gene selection using Fisher ratio calculations

Fisher ratios were calculated comparing one sample subgroup with the rest of the samples. As can be seen in figure 5.10, where an exemplary top 100 genes from each comparison are displayed, the overall quantitative pattern of expression of discriminating genes varied significantly between leukemia subtypes. Blocks of genes with clear up-regulation for each sample subgroup compared to the rest of the samples can be distinguished. The *Other*-subgroup was divided into its own subgroups.

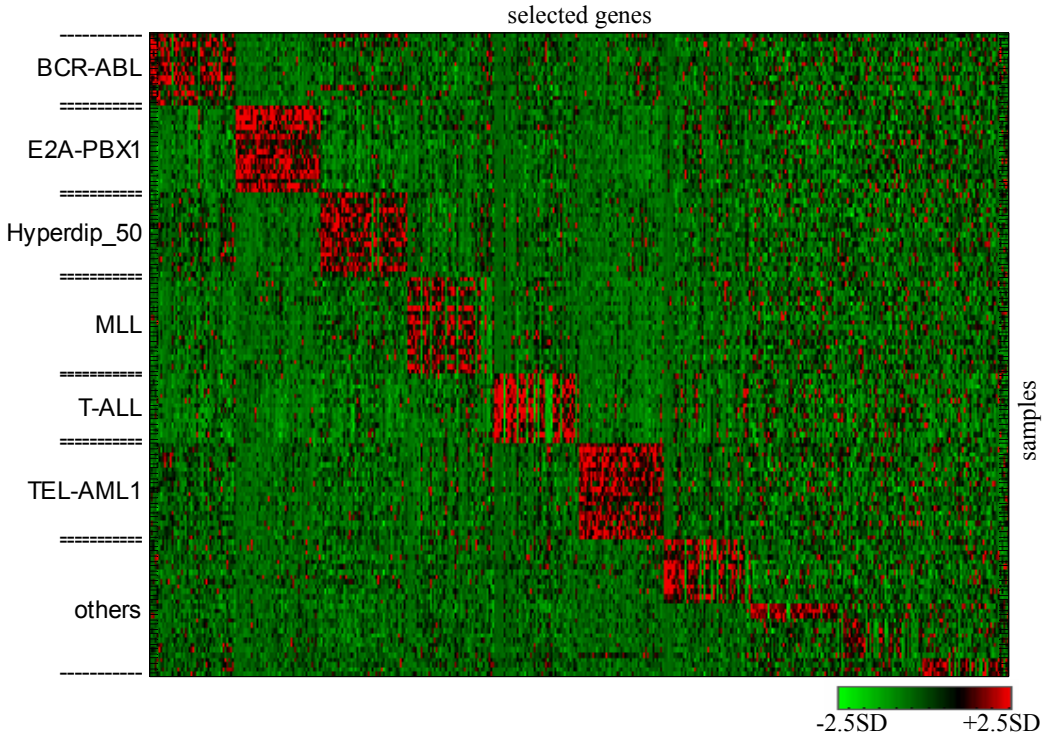


Fig. 5.10 Heatmap showing the gene expressions of the top 100 genes from every cluster selected by Fisher ratio. *Others*-subgroup consists of groups *Novel, Normal_diploid, Pseudodiploid and Hypodiploid* (in that vertical order). Colors denote standard deviation units.

The top five genes from each comparison were selected, yielding a total of 35 genes. The mean signal difference and the Fisher's separability criterion can be used to assess the homogeneity of the sample subgroups, gaining information that shows which subgroup separations could be problematic.

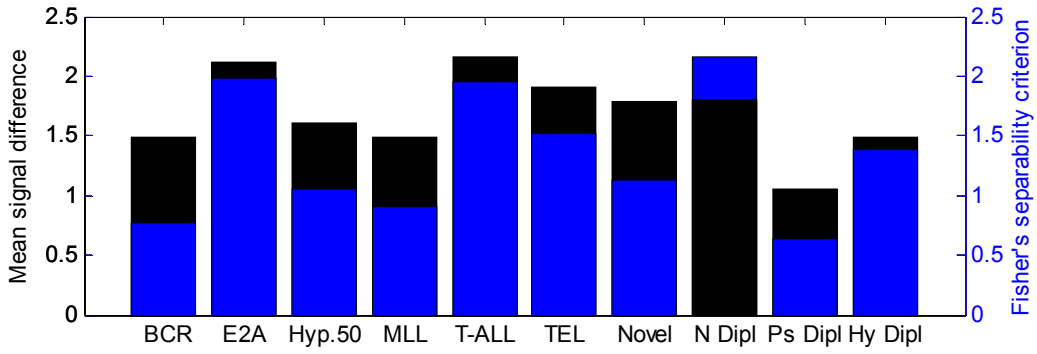


Fig. 5.11 Difference between the median signal of top 100 discriminating genes in one sample subgroup in relation to the rest of the samples and Fisher's separability criterion for the top 100 genes.

The other-sample-group is splitted into its *Novel*, *Normal_diploid*, *Pseudodiploid* and *Hypodiploid* subgroups. High values for the *Normal_diploid* and *Hypodiploid* subgroups are to be regarded with caution as these groups only consist of three and four samples respectively.

Values for the *Other*-subgroups have to be regarded with caution, as some only consist of three or four samples.

5.5.3 Gene selection using Gene shaving

Gene shaving was performed on the data selecting 200 clusters with highly covariant genes. Several groups of genes were found with an expression pattern highly selective for one certain leukemia subgroup (figure 5.12) as well as gene clusters with significantly varying expression patterns for *BCR-ABL, E2A-PBX1, Hyperdiploid>50, MLL, T-ALL* and *TEL-AML1* and the *Others-novel* sample subgroups. The novel group was easily identified. Signal patterns of genes differentiating members of the rest of the *Other* samples found by gene shaving are less distinct.

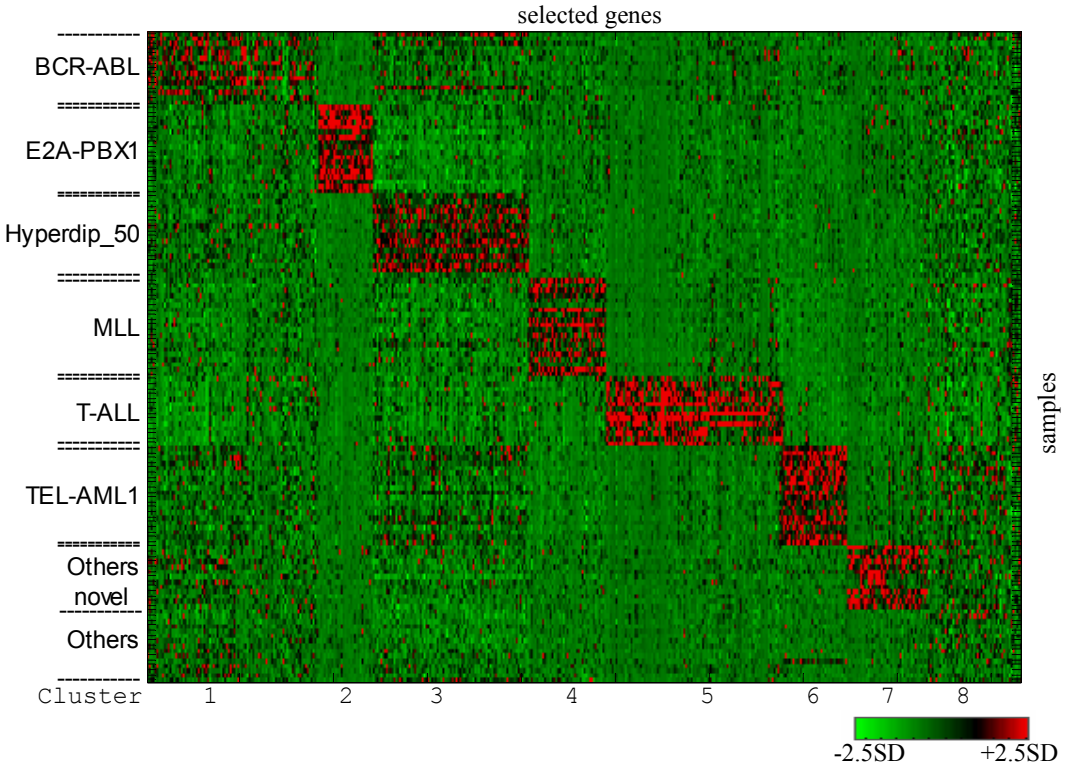


Fig. 5.12 Heatmap of all samples (rows) and 8 groups of genes selected by gene shaving. Colors denote standard deviation units.

Top 15 genes were selected from each cluster having the highest correlation with the cluster expression pattern.

5.5.4 Gene selection using Significance Analysis of Microarrays (SAM)

The utilization of SAM yielded a list of genes whose expression patterns are highly individual for each corresponding sample subgroup (see figure 5.13). Each subgroup was compared with the rest of the samples. The top genes of each comparison were selected using scores provided by SAM, resulting in 98 selected genes in total.

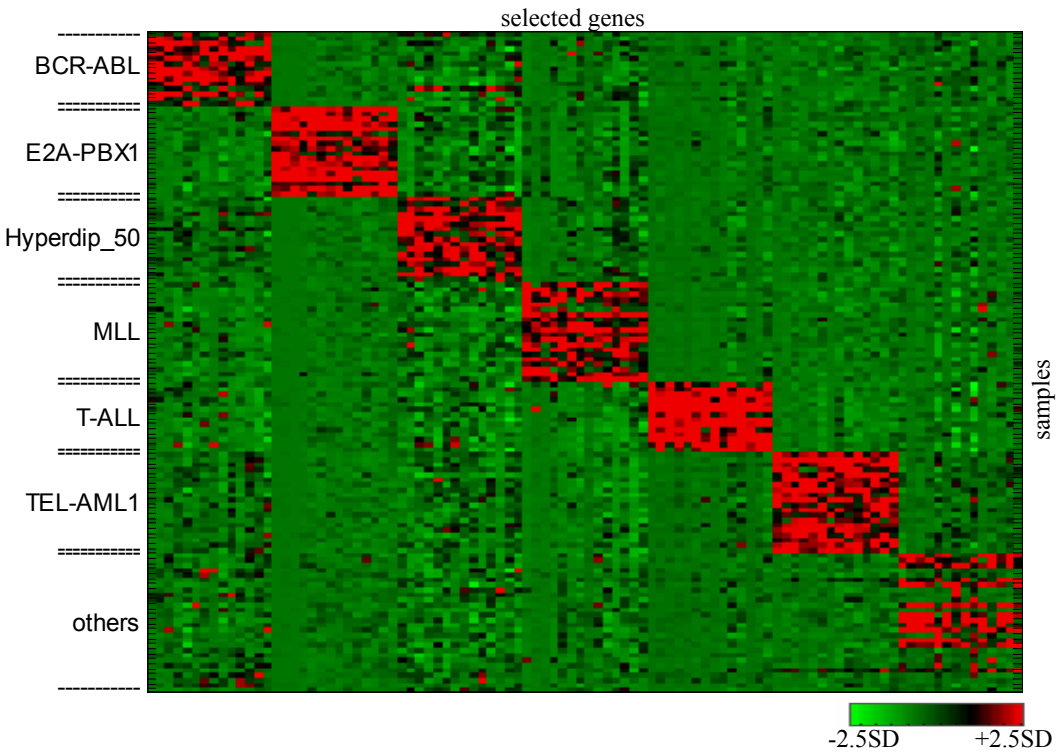


Fig. 5.13 Heatmap of all samples (rows) and 7 groups of genes selected by significance analysis of microarrays (SAM). Colors denote standard deviation units.

5.5.5 Gene selection using Prediction Analysis of Microarrays (PAM)

PAM was used to select the genes best suited for a discrimination of sample subgroups. The threshold was selected by analysis of the training- and misclassification-errors (see chapter 3.5.8.3). The aim was to maximize the threshold and to minimize the errors and number of genes (see figure 5.14).

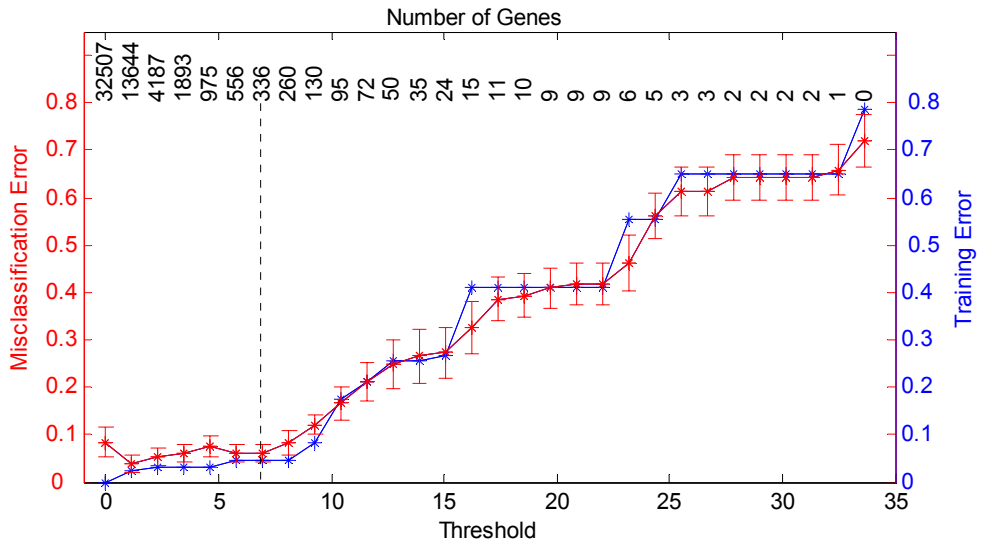


Fig. 5.14 Training- and misclassification errors for different thresholds. A threshold of 7 was selected, yielding a total of 336 genes.

A threshold of 7 was selected, yielding a total number of 336 genes which are suited well to differentiate different sample subgroups. The number of genes was then further downsized to achieve a selection with an amount of genes comparable to the one achieved with the other methods. This was done using the gene-scores for each centroid yielding a total of 89 selected genes.

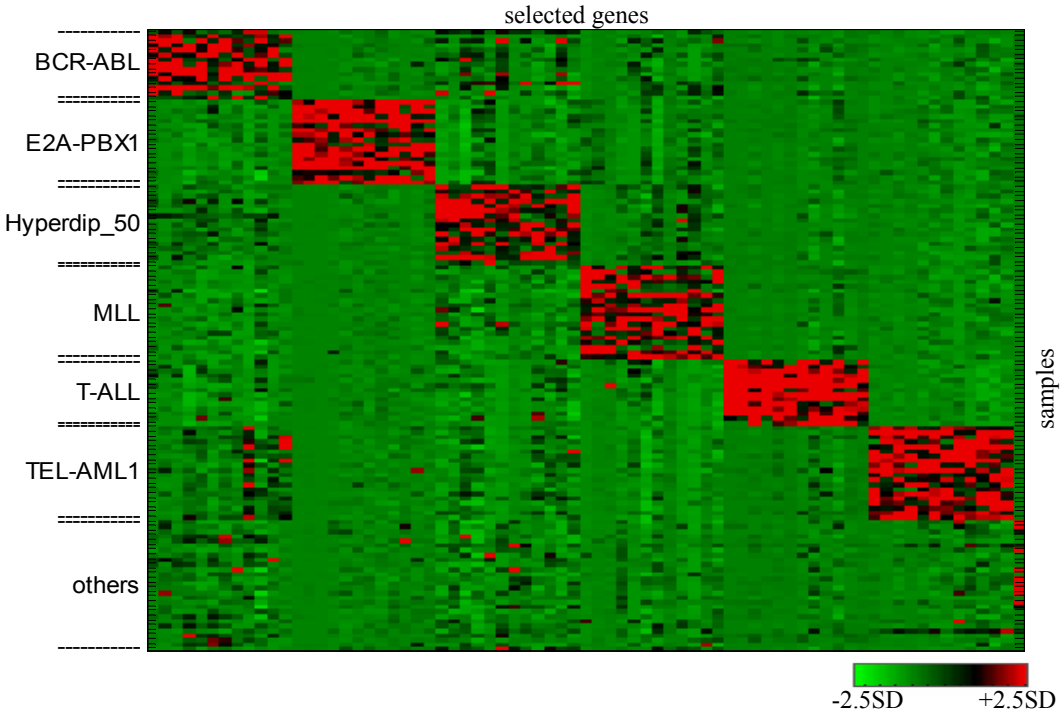


Fig. 5.15 Heatmap of all samples (rows) and 7 groups of genes selected by prediction analysis of microarrays (PAM). Colors denote standard deviation units.

5.5.6 Separation of sample subgroups using selected genes

Highly selective genes for a better discrimination of sample subgroups were selected from the Fisher ratio and gene shaving clusters. The selected top 5 genes of each Fisher ratio cluster where supplemented with certain genes from gene-shaving clusters as these were the most different compared to the Fisher ratio selection as can be seen in table 5.2. Those genes from gene shaving clusters were used that were unique and had references in online databases denoting them as being cancer related. This selection of these 66 genes made a better subgroup separation possible.

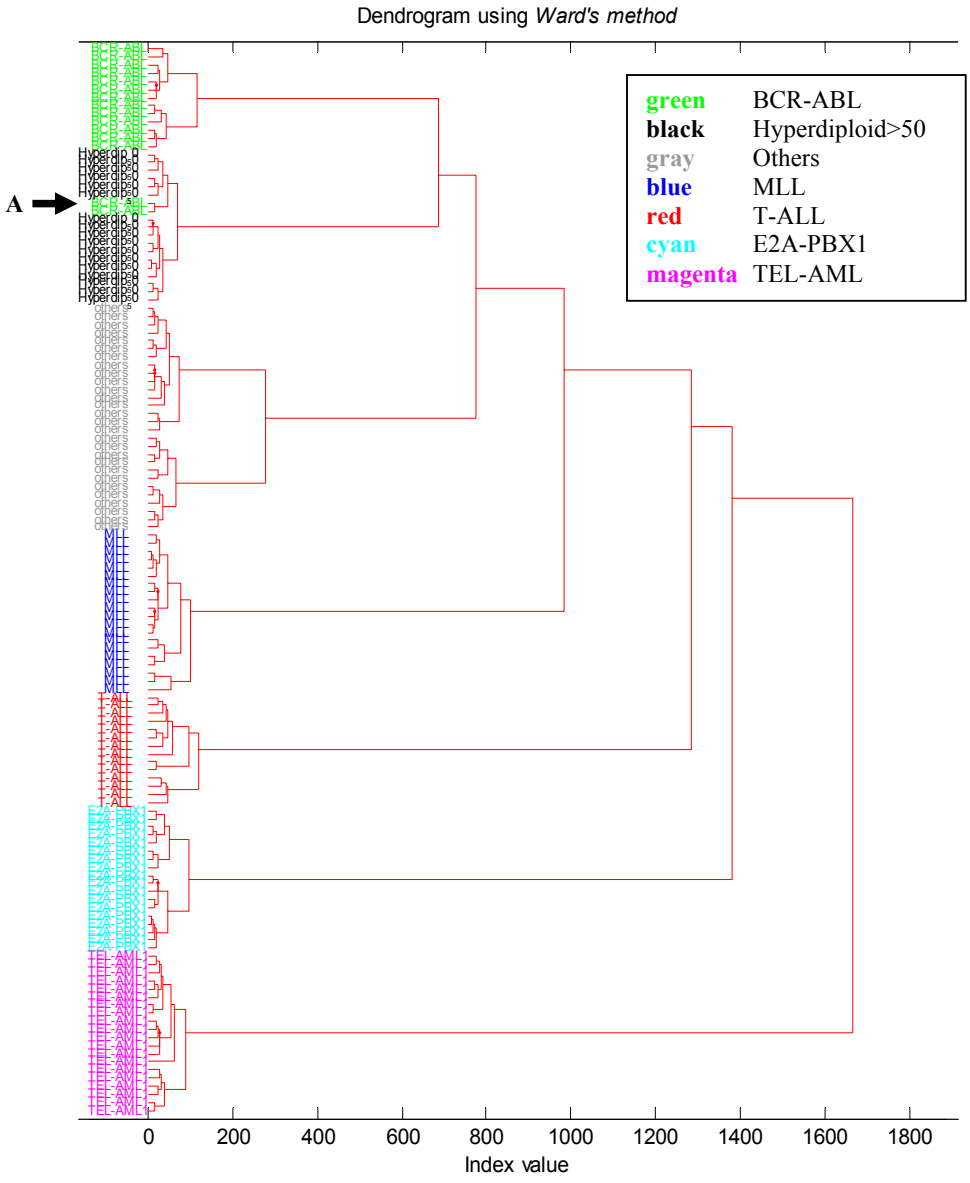


Fig. 5.16 Dendrogram of all samples using 66 genes. Most sample subgroups fall into their own clusters. The two *BCR-ABL* samples marked with the arrow also have a hyperdiploid karyotype and are therefore also clustered correctly.

The heterogeneous character of the *BCR-ABL* subgroup was already mentioned in the original data publication paper [86]. Two cases of *BCR-ABL* (marked with an arrow in figure 5.16) have a karyotype showing the presence of both the Philadelphia chromosome and a hyperdiploid karyotype consisting of more than 50 chromosomes, including trisomy of chromosomes X and 21. They are therefore clustered correctly according to their karyotype. This can also be seen in figure 5.17.

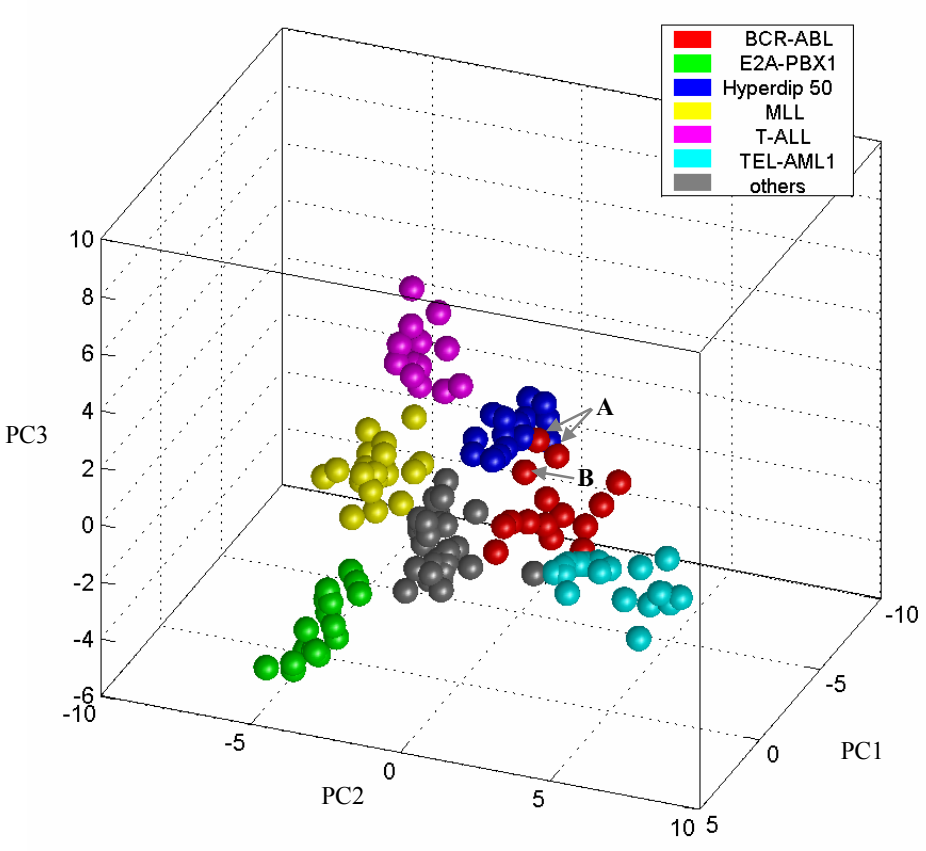


Fig. 5.17 Plot of samples using the first three principal components. The disperse distribution of *BCR-ABL* samples is visible (with the two *BCR-ABL* samples having hyperdiploid karyotype marked with arrows [A], and a third sample often misclassified as member of the *hyperdiploid>50* subgroup, marked with arrow [B]).

5.5.7 Selection of genes for differentiation of the Other-subgroups

Samples in the group denoted as *Others* have normal, pseudodiploid and hypodiploid karyotypes. Some of these samples have a distinct gene expression profile and have been defined as members of the *Novel* subgroup in the original paper [86]. This novel subgroup could also be identified using gene-shaving. A clear grouping of the samples from different *Other*-subgroups is visible using the top 5 genes selected by a Fisher ratio calculation comparing one *Other*-subgroup with the rest of the *Othe*r-samples (see figure 5.18).

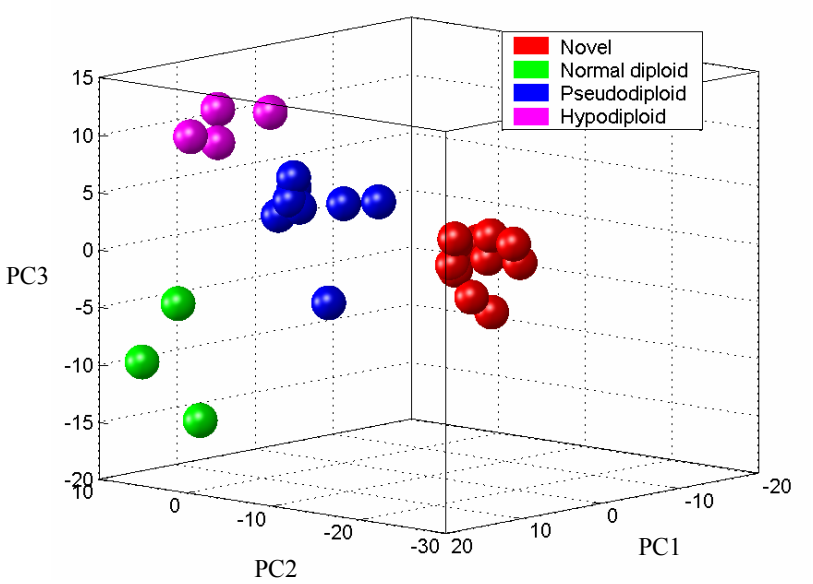


Fig. 5.18 Separation of samples from different *Other*-subgroups using 20 genes in total, selected by one against the rest Fisher ratio calculation.

It is also possible to differentiate between samples from the *Novel* group and all other samples using the genes from the gene shaving cluster 7 (figure 5.12).

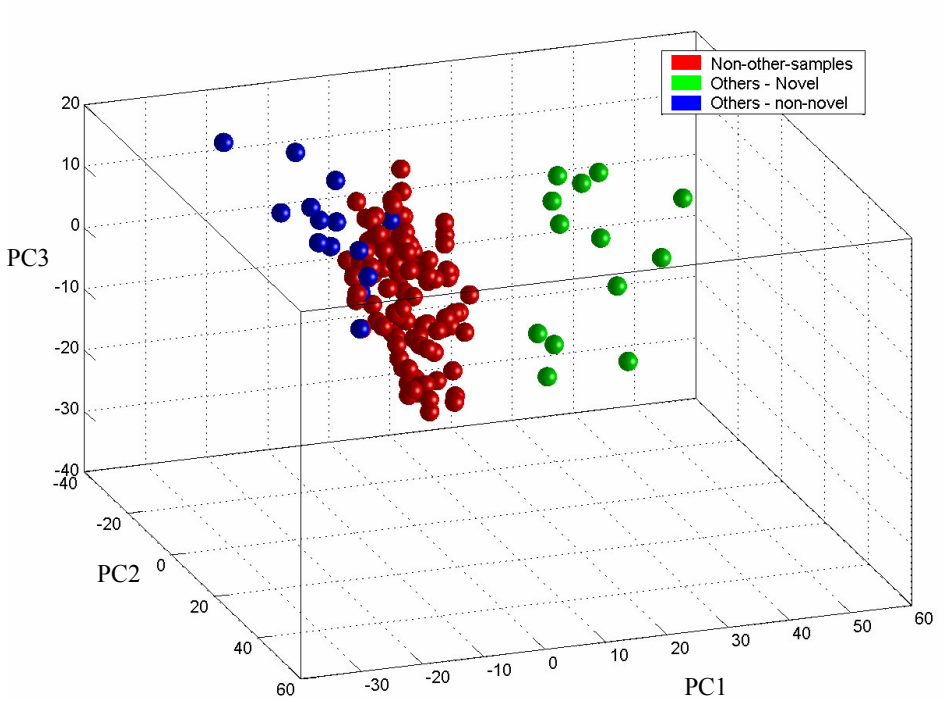


Fig. 5.19 Separation of samples from the *Novel*-subgroup from all other samples using 86 genes selected by gene-shaving (genes from cluster 7 in figure 5.12).

As a differentiation and cross-validation of some subgroups of the *Other* sample-group is not feasible because they only contain three or four samples, only one classifier for the differentiation of the *Novel* subgroup was created. Samples in this subgroup can be separated from the rest of the samples in the *Other*-group using one gene.

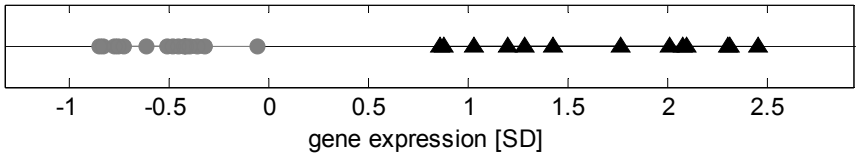


Fig. 5.20 Separation of samples from the *Novel* group from the rest of samples of the *Other* group using only one gene. Pyramids: Samples in the *Novel* group; Circles: rest of the samples in the *Other*-group. Values denote the gene expression in standard deviations.

A good separation of the rest of the *Other*-subgroups (*Hypodiploid, Normal_diploid* and *Pseudodiploid*) seems possible. More samples of this category should be analyzed to determine the feasibility of this discrimination. As can be seen in figure 5.21, *Hypodiploid, Normal_diploid* and *Pseudodiploid* samples are separated using genes selected by Fisher ratio calculations.

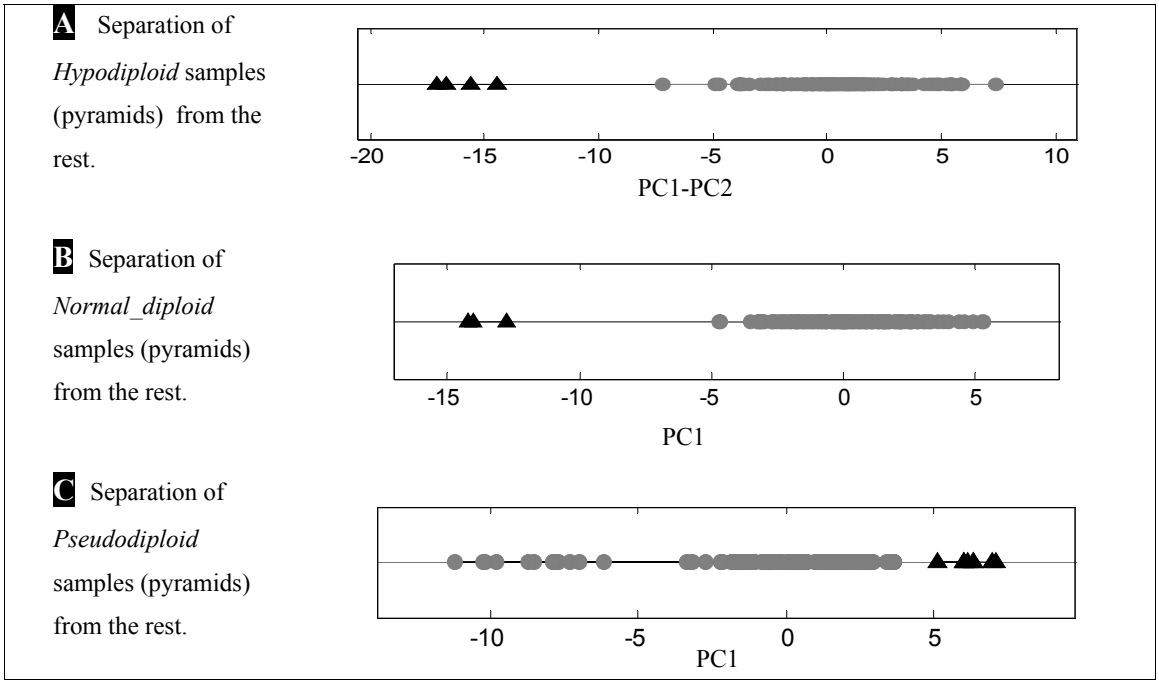


Fig. 5.21 (A) Separation of the four *Hypodiploid* samples from the rest of the samples (128 in total); the diagonal between the first and second principal component is shown. (B) Separation of the three *Normal_diploid* samples from the rest of the samples (129 in total). (C) Separation of the eight *Pseudodiploid* samples from the rest of the samples (124 in total), each time using genes selected by Fisher ratio calculations

5.5.8 Comparison of different selection methods

Each method for gene selection yielded a different gene-list. The top genes from each selection were compared with one another to gain information on the resemblance of the selections.

Tab. 5.2A Resemblance of gene lists selected using different methods (top 100 genes)

A		Percentage of genes contained in			
		Fisher Ratio	SAM	PAM	Gene-Shaving
Genes taken from	Fisher Ratio	100	63	53	37
	SAM	45	100	56	36
	PAM	45	65	100	35
	Gene-Shaving	25	33	30	100

Tab. 5.2B Classification accuracies using the top genes from each selection

B		Fisher Ratio	SAM	PAM	Gene-Shaving
7 groups	Accuracy	98.5%	98.2%	96.8%	93.4%
	Number of genes	70	98	89	105
6 groups	Accuracy	99.4%	99.1%	98.8%	96.0%
	Number of genes	60	84	88	90

The selection process was done using all samples. 30% of the samples were then left out in cross validation for classification. Classification was done with a linear-SVM classifier. The number of genes was selected using selection scores provided by the different algorithms. Classification was done using all groups and only six groups (leaving out the *Others* sample subgroup). This was done as the centroid for this seventh class only yielded one gene.

It is interesting to note that the highest similarity in selection for the SAM list lies in the PAM list and vice versa. The highest accuracy when classifying the samples using a linear SVM classifier (see also chapter 5.6) was achieved by using the Fisher ratio selection, followed by the SAM and the PAM selections. The classification of the *Other*-samples leads to a high misclassification rate using the PAM list, as only one gene was selected by the algorithm in the shrunken centroid of this sample subgroup. To confirm the order of the classification accuracies, classifications were performed using only six sample subgroups and discarding the samples of type *Other*. The same order of accuracies was observed, as can be seen in table 5.2B. It is to be noted that the differences in accuracies are only slight ones (except for the gene shaving selection) and the different methods perform very similarly using this dataset. The classification by PAM itself yielded a classification accuracy of 96.3%.

5.6 Sample Classification

5.6.1 Introduction

Cross validation was used to evaluate the accuracy of a classification model. 30% of each sample subgroup were left out in the training of the classifier and classified afterwards (38 samples in total). This was done 100.000 times. A linear support vector machine classifier was used for classification. The routines used are from the OSU SVM toolbox (v3.0) for Matlab® by Junshui Ma, Yi Zhao, and Stanley Ahalt, Department of Electrical and Computer Engineering, Ohio State University.

5.6.2 Main Classifier

The mean misclassification rate using the 66 genes selected in 5.5.6 was 1.2 %. The main erroneous classifications being

1. classifying a *BCR-ABL* sample as *Hyperdiploid>50*,
2. classifying an *Other* sample as *TEL-AML1*.

96% of the misclassified *BCR-ABL* samples were the two samples that also had the hyperdiploid>50 karyotype. Only one other *BCR-ABL* sample was misclassified (sample bcr-abl#3). To minimize this error, two further classifiers were appended to the classification process.

5.6.3 BCR-ABL classifier

Fisher ratio calculations were performed comparing *BCR-ABL* samples with *hyperdiploid>50* samples and the top10 genes were selected.

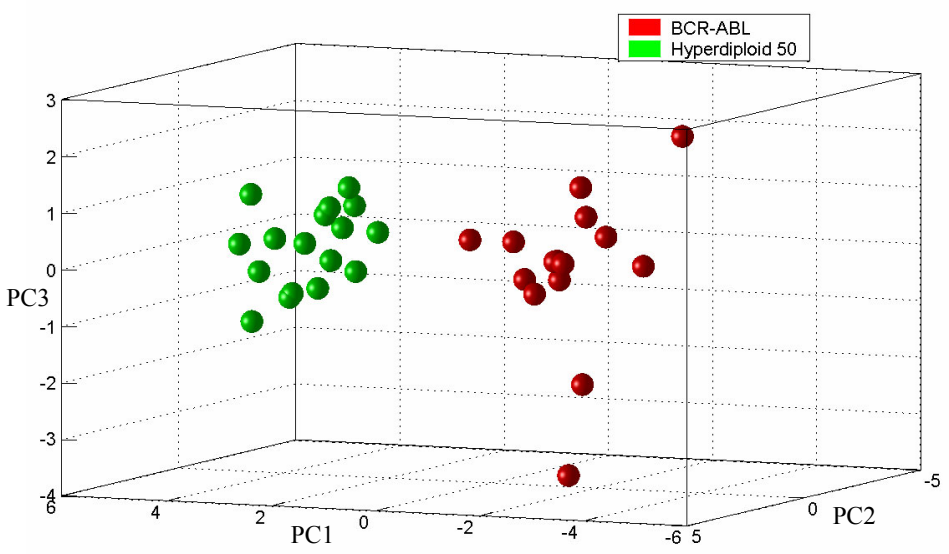


Fig. 5.22 Plot of the first three principal components using ten genes selected by Fisher ratio calculations using *BCR-ABL* and *hyperdiploid>50* sample subgroups only.

All samples previously classified as *BCR-ABL* could now be classified as *hyperdiploid>50*. It is to be noted that a classification of the two *BCR-ABL* samples marked in figure 5.16 and

5.17 as members of the hyperdiploid>50 sample subgroup is also correct but as the BCR-ABL translocation has a worse survival prognostics, a clear identification of a sample as being a member of this group is essential to administer the best treatment possible. The third *BCR-ABL* sample also classified this way should be analyzed further to determine if it does not actually also have these sorts of genetic lesions. The differentiation of *BCR-ABL* from *Hyperdiploid>50* samples in this step had an accuracy of 100%. This makes it possible to define those *BCR-ABL* samples having a hyperdiploid>50 karyotype with an accuracy of 96%.

5.6.4 TEL-AML classifier

Fisher ratio calculations were performed comparing *TEL-AML* samples with the *Other*-subgroup samples. The top genes were selected and supplemented with genes from the gene shaving clusters assigned to these two sample groups.

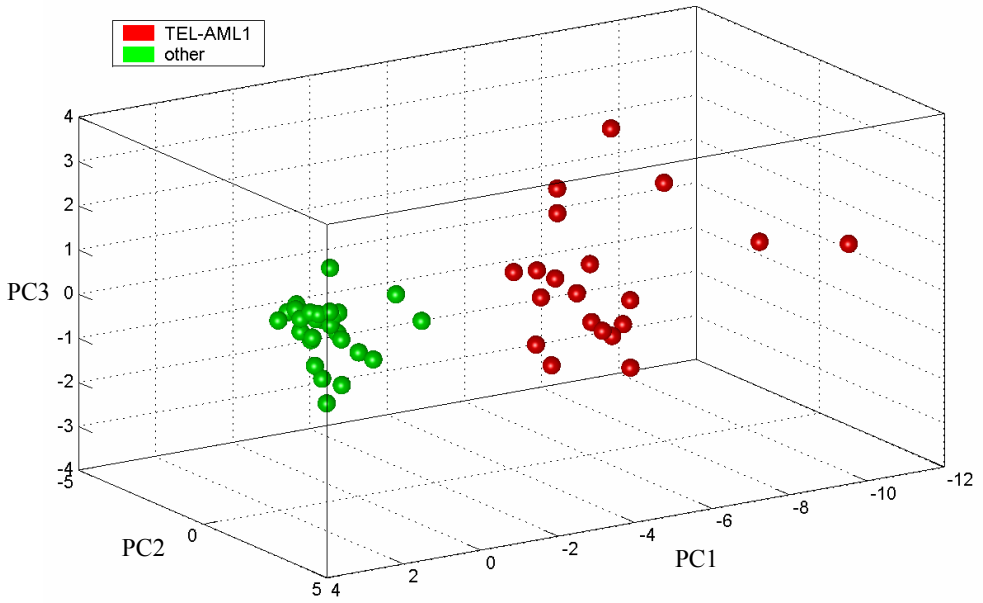


Fig. 5.23 Plot of the first three principal components using 11 genes selected by Fisher ratio calculations and gene shaving clusters using *TEL-AML1* and *Other* sample subgroups only.

Utilization of 11 genes produced an overall classification accuracy of 100%.

5.6.5 Novel-group Classifier

The classification of *Novel*-samples with an accuracy of 100% was added to the classification chain, providing a means to classify a sample into 8 different groups (*BCR-ABL, E2A-PBX1 , Hyperdiploid>50, MLL, T-ALL, TEL-AML1, Novel* and *Others*).

5.6.6 Final sample subgroup classification

The main classifier was used to make a preliminary classification. All samples classified as *hyperdiploid>50* or *TEL-AML1* were then subjected to their respective follow-up classifier. This resulted in an average accuracy of class assignment in cross-validation using a total of 88 genes of 99.97%, with a range from 99.9% to 100%.

Tab. 5.3 Subgroup prediction accuracies using 88 genes selected with Fisher ratio calculations and gene shaving.

Subgroup	Apparent Accuracy	Sensitivity	Specificity
BCR-ABL	100.0	100.0	100.0
E2A-PBX1	100.0	100.0	100.0
Hyperdip50	100.0	100.0	100.0
MLL	99.8	99.8	100.0
T-ALL	100.0	100.0	100.0
TEL-AML1	100.0	100.0	100.0
Others	99.8	99.9	100.0
Novel*	99.8	99.9	100.0

The differentiation into the *Novel* group was done using only the samples previously classified as being members of the *Others*-subgroup. This differentiation of the *Others*-subgroup had an accuracy of 100%.

5.7 Feature selection bias and true accuracies

The performance of a classifier using a subset of genes is assessed by using cross validation as described above. If the subset of genes was selected using all samples, there is a selection

bias, leading to too optimistic classification accuracies. A better estimate of true accuracies can be gained by applying the feature selection itself inside a cross-validation scheme, i.e. using a group of samples for selection of genes suited the best for class separation and applying a classifier using these genes on the rest of the samples that were completely hidden in the selection process [78]. A leave-one-out approach makes it possible to implement this procedure when working with a small number of samples that is common when dealing with microarray data.

70% of the samples of the leukemia dataset were used to select genes through Fisher ratio calculation in each of the 30 performed iterations. These genes were then used to perform the above mentioned classifications of the 30% of the samples completely withheld previously.

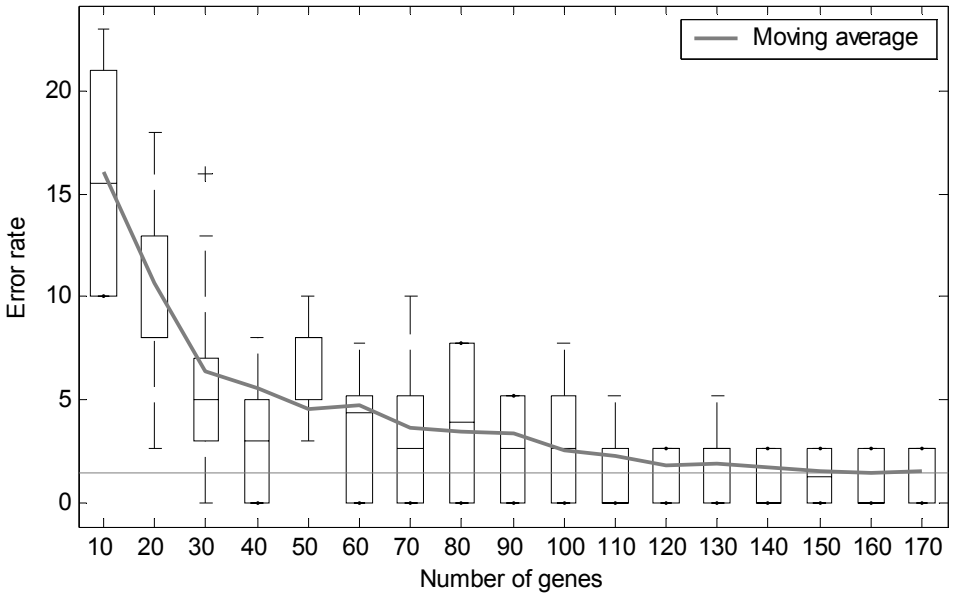


Fig. 5.24 Misclassification rates for the classification of previously withheld samples.

Boxplots show upper and lower 25th percentile. 30% of the samples were withheld from the gene-selection procedure. A certain number of genes was selected and classification then performed on the 30% of the samples previously left out. This procedure was performed 30 times for each number of genes.

An accuracy of 98.1% using 120 genes could be reached using this method of increased selection bias correction.

Tab. 5.4 Subgroup prediction accuracies using 120 genes selected with Fisher ratio calculations after selection bias correction

Subgroup	True Accuracy	Sensitivity	Specificity
BCR-ABL	88.4	88.4	100.0
E2A-PBX1	100.0	100.0	100.0
Hyperdip50	100.0	100.0	99.1
MLL	100.0	100.0	100.0
T-ALL	97.6	97.6	100.0
TEL-AML1	100.0	100.0	100.0
Others	100.0	100.0	100.0

5.8 Effects of using different gene expression summary algorithms

The different gene expression summary algorithms used were the Li-Wong model, the RMA model, the MAS perfect match only model and the MAS model implemented in the BioConductor packages for R. The same background correction and scaling methods were used as described by the authors of the different gene expression summary techniques.

Several probesets are located on the U133A as well as the U133B chip. The gene expression summary algorithm should calculate very similar values for these replicates when regarding one and the same sample. Replicate pairs were used to assess this similarity of calculated signals. The mean relative difference in signal between replicates was calculated to estimate how good each summary technique performed. Figure 5.25 shows that the RMA method performed best, with a mean deviation of 0.2%. The MAS method performed worst, with a mean deviation of 1.3%.

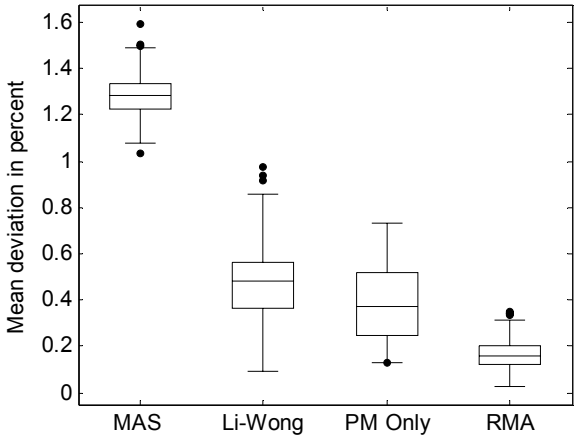


Fig. 5.25 Mean deviation of the signals of duplicate probesets in percent. Boxplots show upper and lower 25th percentile.

These findings concur with the results of Speed *et al.* [117]. The MAS PM only method, which is the same as the MAS method without subtraction of the MM value (or rather the IM, *Ideal Mismatch* value, see [89]), produces data with clearly lower signal deviations between replicates. This concurs with the findings of Irizarry *et al*. that a subtraction of MM signals increases the noise [118]. The Li-Wong method also performs worse than the RMA method. Although this also concurs with the findings of Calogero *et al.* [119] and Irizarry (RMA having better precision), the reason for such strong differences could also be that not enough microarrays for a good convergence of the Li-Wong model had been used. The authors of the method suggest using at least 10 microarrays. Eleven microarrays were used to generate the data, as a higher number was not possible due to memory problems during computation.
Results of a comparison of the genes selected by Fisher ratio can be seen in table 5.5. There are great differences in selection, at least when regarding the top 20 or even the top100 genes for each class.

Tab. 5.5 Percent of genes being the same in the top 20 genes found by Fisher ratio one – against – rest selection using different gene–expression gene expression summary algorithms

		Genes compared with top 20 genes from			
		Li-Wong	MAS PM only	MAS	RMA
Top 20 genes from	Li-Wong	100%	54%	52%	49%
	MAS PM only	54%	100%	62%	51%
	MAS	52%	62%	100%	51%
	RMA	49%	51%	51%	100%

Fisher ratios were calculated using samples from one group and compared with all other samples. The seven resulting groups of top 20 genes were compared individually with the same group calculated, using a different summary algorithm. Results were then combined to create the overall percentage of similarity. This overall percentage is very similar to the percentages in single comparison groups. The percentages stay on the overall same level if the top 100 genes are compared (deviation of ± 2-3 %points).

Samples were classified using a linear SVM classifier using genes selected by Fisher ratio calculations (20 genes for the differentiation of each subclass against the rest) and by PAM (225 genes selected by score). As PAM had difficulties selecting genes for the centroid of the class *Others*, only six sample subgroups were used in this analysis. Genes were selected using the entire dataset.

Tab. 5.6 Misclassification rates using a linear SVM classifier applied to six samples subgroups (*Others* omitted) using genes selected by Fisher ratio criterion and PAM.

	Li-Wong	MAS PM only	RMA	MAS
Fisher Ratio, 60 genes	2.0%	1.2%	1.4%	1.3%
PAM, 225 genes	1.6%	1.5%	2.3%	1.9%

The classifier performed best when using the data generated by the MAS PM only method in both the Fisher ratio and PAM case. The other method's order varies depending on the gene selection procedure used. The MAS perfect match only method yields the best results in classification and the RMA method in signal reproducibility; these methods are thus to be favored.

5.9 Discussion of leukemia data analysis

Data analysis can only yield informative results if the data used has the needed quality [120, 121]. Analysis of quality parameters has shown the leukemia dataset to be of overall good quality. One drawback is the time sequence in which the samples were measured [69]. U133A and U133B chips were measured at completely separate time intervals and some sample subgroups were not shuffled with other samples well enough and can be found in almost completely continuous blocks. Although these facts did not prove to have any detectable effects on the data or class assignment, a correct implementation of the measurement sequence could have easily been done without any further cost.

Signal distributions on the microarrays often showed an area dependant background and signal intensity drift. This drift was more pronounced than in other measurement series like the Latin Square dataset [105], a CD4 Lymphocytes dataset [96] or a proprietary dataset containing several hundred measurements from Haferlach *et al*.[97]. As could be shown with histogram comparisons, a global scaling, as performed by the authors of the original dataset, is not enough to compensate for these scaling differences. These different background and scaling values can be accounted for using the novel methods created in this work.

Further, large artifacts could be found in the data. These cells were not flagged in the original CEL-files and were thus incorporated in the downstream analysis using the MicroArray Suite. The artifacts could be easily found using the novel, automatic detection scheme developed.

As expected, the use of all genes for class separation yielded worse results than when using a subset of genes. This is only logical, as most genes do not show any differential expression. Including them in the class separation procedure means incorporating a higher noise component, barely contributing to the net informative signal [122].

Several different gene selection algorithms were used to select the differentially expressed genes, yielding several groups of genes with an expression pattern highly individual for one certain leukemia subgroup. Gene lists created using these methods differed from one another as the methods are based on different statistical methods. The most different list in comparison with the other lists was created by the gene-shaving procedure. This coincides with the fact that this method was the only one used without the *a priori* class assignment information. The preliminary classification results using a linear SVM classifier showed the gene lists selected by Fisher's ratio calculations, SAM and PAM, to perform similarly. The classification using genes selected by gene-shaving performed worse, showing that the selected genes are less well suited for class separation than those selected by the other methods. This fact can be attributed partially to the lack of *a priori* knowledge [123] but also infers a selection of subsets of genes with coherent expression patterns and large variation across subgroups which were not optimal. In gene-shaving, each shaving sequence starts with the entire expression matrix which is then shaved down to a minimal cluster size. The adequate size of a cluster is derived from the gap estimate [84]. This estimate was not optimal at all the times, leading to clusters that included less predictive genes which might also have contributed to the worse classification results.

The *Novel* sample-subgroup previously defined by Yeoh *et al.* [113] could be clearly seen when using gene-shaving. Using the genes hereby selected made it possible to separate these samples from all other samples in the dataset. A separation of the different *Others* sample-subgroups seems possible but could not be further analyzed due to the lack of enough samples. The promising preliminary class separation results speak in favor of analyzing further samples of these types.

The dataset also included two samples which were classified as *BCR-ABL* samples and also had a hyerdiploidity with more than 50 chromosomes. These samples were primarily classified as members of the hyperdiploid>50 group. Although this classification is molecularly correct, it was defined as a misclassification as the detection of the *BCR-ABL* status is far more important due to less good survival predictions for this leukemia subgroup [3]. Including these two samples into the study therefore made an evaluation of classification rates more difficult but also helped make it more realistic, as these kind of samples can also

appear in real life diagnostics. A detection of these two samples that have a BCR-ABL as well as a hyperdiploid>50 status was only possible whenever the gene selection was done using all samples. This selection procedure introduces a selection bias as those samples which are to be classified are also used in the selection step [78]. This bias can be accounted for by implementing the feature selection step into the cross-validation routine. The overall classification accuracy dropped from 99.9% using 88 genes to 98.1% using 120 genes selected in this unbiased manner. The latter value is a more realistic estimate of the classification accuracy. Although a 100% accuracy is desired, the reality shows that complex cancer subtype classifications like the one at hand frequently fall short of this goal. It will have to be determined what level of diagnostic accuracies is needed to move this biomolecular approach into a clinical setting. One major advantage of these methods over classical diagnostic techniques like cytogenetic analysis is that their accuracies do depend far less on the level of expertise of the practitioner. Although the classification might be improved by the use of more selected genes that are informative, raising their numbers does not only hamper the creation of a low-density chip but it is also difficult to gain the license rights needed to use all of these genes for a certain task or a certain device [124].

Analysis of different gene expression summary algorithms showed the MAS perfect match only approach to yield more reproducible values than the MAS method. This concurs with the opinion of Naef *et al*. [62] and others that the expression measure should only be based on the PM signal.

6 Summary and Outlook

The life sciences are currently undergoing a rapid revolution that is driven by large-scale genome sequencing, advances in structural and chemical biology, bioinformatics, chemometrics, imaging technology and bioanalytics. These disciplines have long existed as a collection of knowledge fields with well-defined territories but are now merging together to form new, powerful interdisciplinary domains. One of these new domains deals with the measurement of gene-expressions by quantitative analysis of the amount of the corresponding mRNA-molecule-content in cells. This bioanalytic technique, combined with state of the art chemometrics and bioinformatics, makes it possible not only to gain information on the role of certain genes in diseases like cancer, but also to classify unknown samples as being, for example, members of a certain cancer subtype. This molecular approach makes it possible to use specially designed therapies to maximize the rate of survival of cancer patients.

Methods were developed in this work dealing with the complete processing chain used for the analysis of Affymetrix™ U133 DNA biosensor data. The aim was to increase the overall quality of the signals derived from microarray experiments, to find indicators applicable in quality management and to select diagnostic markers for the differentiation of pediatric leukemia cancer subtypes as one example application.

A novel method for the calculation and subtraction of background signals and signal scaling was created based on thin plate-spline interpolation. Many microarrays exhibit spatially inhomogeneous signal intensity distributions. This can be due to thermal effect, incomplete washing, diverse hybridization performances etc. It is important to correct these deviations in order to make signals from different parts of the microarray comparable. The developed methods use the information of several hundred data points provided by the border area of the microarray. The number of available data points makes a nonlinear interpolation possible. This method has shown to be well suited in interpolating noisy data without giving rise to overfitting. The application of the interpolation method to calculate the area dependant background signal and the deviations in signal scaling made it possible to improve the overall signal quality of Affymetrix U133 microarrays.

The data used was handled using a newly developed microarray data management system combining a SQL database and a file server with Matlab®. This system made it possible to store all data as well as analysis results in a safe and reliable way and is compliant with the MIAME standard. This data included quality measures calculated using novel methods, one main aspect being the automatic detection of signal artifacts on microarrays. This constitutes a

major enhancement as the MicroArray Suite of Affymetrix does not provide for an automatic artifact detection scheme. Almost all of the several hundred U133 microarrays analyzed contained artifacts. Their size and form varies significantly - from small round speckles or fine scratches up to large areas containing ten thousands of cells. The detection of these cells is important as their usage can have drastic consequences in the calculation of the expression value of the corresponding gene. The developed methods are able to detect these artifacts in a fast and reliable way and were fine-tuned to flag only a minimal number of differentially expressed cells as outliers. Detailed analysis of the artifacts provides information on the quality of the measurement which can be used to raise the quality of the entire analysis process. Artifact detection is achieved by using data from several microarrays, combining these signals to a virtual microarray containing the median values of same cells from all sensors. It is possible to detect anomalies on a microarray by comparing it with this median-microarray. Detection of known and previously unknown artifacts showed this procedure to be very reliable.

Microarray measurements of leukemia samples were used to select a group of genes suited best for differentiation of seven pediatric leukemia subtypes (*BCR-ABL, E2A-PBX1, Hyperdiploid with more than 50 chromosomes, MLL, T-ALL, Others*). A classification accuracy of 98.1% was achieved using 120 genes, performing a rigorous cross-validation by also including the gene selection process in the cross-validation procedure. Otherwise a potentially large selection bias, i.e. an optimistically biased error assessment, might arise. Different selection methods were compared, showing that Fisher ratio calculation, SAM and PAM perform similarly when applied to the pediatric leukemia dataset. Gene-shaving performed worse as it does not contain *a priori* class assignment knowledge. The *Novel* sample subgroup previously defined by Yeoh *et al.* [113] was found using gene-shaving. It was possible to separate all samples of this subgroup from all other samples using the genes selected by this method.

Analysis of different gene expression summary algorithms (*MAS*, *MAS PM-only*, *Li-Wong-model*, *RMA*) showed the *MAS PM-only* approach to yield more reproducible values than the *MAS* method. This concurs with the opinion of Naef *et al*. [62] and others that the expression measure should only be based on the PM signal. Further, the *MAS PM-only* as well as the *RMA* methods are good choices to achieve low deviation between replicate measurements and good classification results.

It will have to be determined what level of diagnostic accuracies is needed to move this biomolecular approach of cancer detection and classification into a clinical setting, especially

as it may also be difficult to gain the license rights needed to use certain genes for a certain task or a certain device. But there is no doubt that DNA biosensors based on the Affymetrix technology will be used in clinical diagnostics in the near future.

7 Literature Index

1. *Long term and late effects of treatment for blood-related cancers.* Fakt Sheet, The Leukemia & Lymphoma Society, 2003.
2. *Side effects of specific cancer drugs.* Cancer Research UK, www.cancerhelp.org.uk, 2003.
3. Haferlach, T., *Manual Leukämien, myelodysplastische und myeloproliferative Syndrome*. 2003, Tumorzentrum München und W. Zuckschwerdt Verlag: München. p. 1-16.
4. Niemeyer, C.M. and D. Blohm, *DNA-Microarrays.* Angew. Chem., 1999. **111**: p. 3939-3043.
5. Bugg, C.E., W.M. Carson, and J.A. Montgomery, *Drugs by Design.* Sci. Am., 1993. **269**: p. 92-98.
6. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**: p. 531-536.
7. Golub, T.R., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* PNAS, 2001. **98**(26): p. 15149-15154.
8. Ohndorf, U.M., et al., *Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins.* Nature, 1999. **399**(6737): p. 708-712.
9. Cozzi, P., N. Mongelli, and A. Suarato, *Recent Anticancer Cytotoxic Agents.* Current Medicinal Chemistry - Anti-Cancer Agents, 2004. **4**(2): p. 93-121.
10. Dabid-Cordonnier, M.-H., et al., *Design of Novel Antitumor DNA Alkylating Agents: The Benzacronycine Series.* Current Medicinal Chemistry - Anti-Cancer Agents, 2004. **4**(2): p. 83-92.
11. Sissi, C. and M. Palumbo, *The Quinolone Family: From Antibacterial to Anticancer Agents.* Current Medicinal Chemistry - Anti-Cancer Agents, 2003. **3**(6): p. 439-450.
12. Marx, J., *Cancer research. Drug candidate bolsters cell's tumor defenses.* Science, 2004. **303**(5654): p. 23-25.
13. Benowitz, S., *As targeted therapies evolve, challenges remain.* J. Natl. Cancer Inst., 2004. **96**(5): p. 351-2.
14. McBride, G., *Are intellectual property rights hampering cancer research?* J. Natl. Cancer Inst., 2004. **96**(2): p. 92-94.
15. Weinstein, J.N., et al., *The bioinformatics of microarray gene expression profiling.* Cytometry, 2002. **47**: p. 46-49.

16. Alberts, B., et al., *Molekularbiologie der Zelle*. 1995, New York: VCH Weinheim.
17. Peter, C., *Evaneszent-Feld-DNA-Biosensor zur schnellen, zeitaufgelösten Detektion multipler Hybrisisierungsereignisse - Einsatz zur Tierartendifferenzierung in Lebensmitteln und für die Identifizierung von Microorganismen.* Dissertation, 2003.
18. Avidor, Y., N.J. Mabjeesh, and H. Matzkin, *Biotechnology and drug discovery: from bench to bedside.* South Med J., 2003. **96**(12): p. 1174-1186.
19. Alizadeh, A.A., et al., *Towards a novel classification of human malignancies based on gene expression patterns.* J. Pathol., 2001. **195**: p. 41-52.
20. Xiong, M., et al., *Gene Selection in Gene Expression Based Tumor Classification.* Mol. Metab, 2001. **73**: p. 239-247.
21. Haberhausen, G., et al., *Bioanalytik*. 1998, Heidelberg: Spektrum Akademischer Verlag.
22. Pingoud, A. and C. Urbanke, *Arbeitsmethoden der Biochemie*. 1997, Berlin: Walter de Gruyter.
23. Chrisopoulos, T.K., *Nucleic Acid Analysis.* Anal. Chem., 1999. **71**: p. 425R-428R.
24. Robert, A., *How Cancer Arises.* Sci. Am., 1996: p. 62.
25. Trichopulus, D., F.P. Li, and D.J. Hunter, *What causes cancer?* Sci. Am., 1996. **9**: p. 50-57.
26. Fey, M.F., *Krebs und Zellzyclus: ein aktuelles Karussell in der Onkologie von klinischer Relevanz.* Schweiz. Med. Wochenschr., 1998. **128**: p. 629-637.
27. Hanahan, D. and R.A. Weinberg, *The Hallmarks of Cancer.* Cell, 2000. **100**: p. 57-70.
28. *Advances in Understanding Genetic Changes in Cancer: Impact on Diagnosis and Treatment Decisions in the 1990s*. 1993: National Academic Press.
29. Cooper, G.M., *The Cell, A molecular approach*. 2000: Sinauer Associates, Inc. Tumor Suppressor Genes Chapter.
30. Ramirez de Molina, A., A. Rodriquez-Gonzalz, and L. J.C., *From Ras signaling to Chok inhibitor: a further advance in anticancer drug design.* Cancer Lett, 2004. **206**(2): p. 137-148.
31. Stryer, L., J.L. Tymoczko, and J.M. Berg, *Biochemistry*. 2002: W. H. Freeman and company.
32. Das, S., J.E. Dixon, and W. Cho, *Membrane-binding and activation mechanism of PTEN.* PNAS, 2003. **100**(23): p. 7491-7496.

33. Li, S., X. Zhang, and X. X., *Regression of Tumor Growth and Induction of Long-Term Antitumor Memory by Interleukin 12 Electro-Gene Therapy.* Journal of the National Cancer Institute, 2002. **94**(10): p. 762-768.
34. Valk, P.J.M., et al., *Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia.* N. Engl. J. Med., 2004. **350**: p. 1617-1628.
35. Ivanov, V.K., et al., *Incidence of post-Chernobyl leukemia and thyroid cancer in children and adolescents in the Briansk region: evaluation of radiation risks.* Vopr. Onkol, 2003. **49**(4): p. 445-449.
36. Bender-Götze, C., et al., *Besonderheiten der akuten Leukämie im Kindesalter.* Manual, Tumorzentrum München und W. Zuckschwerdt Verlag München, 2003: p. 77-96.
37. Pui, C.H. and W.E. Evans, *Acute lymphoblastic leukemia.* N. Engl. J. Med., 1998. **339**: p. 605-615.
38. Bagg, A. and B.V.S. Kallakury, *Molecular pathology of leukemia and lymphoma.* Am. J. Clin. Pathol., 1999. **112**(Suppl 1): p. 76-92.
39. Eils, R., et al., *An optimized, fully automated system for fast and accurate identification of chromosomal rearrangements by multiplex-FISH.* Cytogenetic Cell Genet, 1998. **82**: p. 160-171.
40. Lu, Y., et al., *Evaluation of 24-color multifluor-fluorescence in-situ hybridisation (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47D.* Chromosome Research, 2000. **8**: p. 127-132.
41. Kaeda, J., A. Chase, and J.M. Goldman, *Cytogenetic and molecular monitoring of residual disease in chronic myeloid leukemia.* Acta Haematol, 2002. **107**: p. 64-75.
42. Reeves, B. and H. Kempski, Department of Hematology, Great Ormond Street Hospital for Children NHS Trust, London.
43. Wright, G., et al., *A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma.* PNAS, 2003. **100**(17): p. 9991-9996.
44. Raimondi, S.C., et al., *Cytogenetics of pre-B-cell acute lymphoblastic leukemia with emphasis on prognostic implications of the t(1;19).* J. Clin. Oncol., 1990. **8**: p. 1380-1388.
45. Cammann, K., G.C. Chemnitius, and W. Kleiböhmer, *Instrumentelle Analytische Chemie*, ed. K. Cammann. 2001, Heidelberg: Spektrum Akademischer Verlag.
46. Scheller, F.S., F., *Biosensoren*. 1989, Berlin: Birkhäuser Verlag.
47. Cammann, K., et al., *Optical DNA-sensor chip for real-time detection of hybridization events.* Fresenius J. Anal. Chem., 2001. **371**: p. 120-127.

48. Ju, H.X., et al., *Hybridization biosensor using di(2,2'-bipyridine)osmium (III) as electrochemical indicator for detection of polymerase chain reaction product of hepatitis B virus DNA.* Anal. Biochem., 2003. **313**: p. 255-261.
49. Caruana, D.J. and A. Heller, *Enzyme-amplified amperometric detection of hybridization and of a single base pair mutation in an 18-base oligonucleotide on a 7-μm-diameter microelectrode.* J. AM. Chem. Soc., 1999. **121**: p. 769-774.
50. Wang, J., et al., *Indicator-free electrochemical DNA hybridization biosensor.* Anal. Chim. Acta, 1998. **375**(197-203).
51. Ozkan, D., et al., *Allele-specific genotype detection of factor V Leiden mutation from polymerase chain reaction amplicons based on label-free electrochemical genosensor.* J. AM. Chem. Soc., 2002. **121**: p. 769-774.
52. Bier, F.F. and J.P. Fürste, *Nucleic acid based sensors.* Frontiers in Biosensorics I, Fundamental Aspects, Scheller, F.w., Schubert F, Fedrowitz, J., 1997.
53. Vercoutere, W. and M. Akeson, *Biosensors for DNA sequence detection.* Curr. Opin. Chem. Biol., 2002. **6**: p. 816-822.
54. Mullis, K.B., *The unusual origin of the polymerase chain reaction.* Sci. Am., 1990. **262**: p. 56-65.
55. Kwok, P.Y. and X. Chen, *Detection of single nucleotide polymorphisms.* Curr. Issues Mol. Biol., 2003. **5**: p. 43-60.
56. Dong, S., et al., *Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation.* Genome Res, 2001. **11**: p. 1418-1424.
57. Kling, J., *Roche's microarray tests US FDA's diagnostic policy.* Natur Biology, 2003. **21**(9): p. 959-960.
58. *Food and Drugs Administration, OIVD Requests a Meeting with Roche Diagnostics Regarding the AmpliChip CYP450 Microarray.* FDA open letter to the General Manager of Roche Molecular Diagnositcs, 2003.
59. *Microarray and Roche AmpliChip CYP450 Backgrounder.* Roche Diagnostics Information, 2003.
60. *GeneChip Human Genome Arrays.* Affymetrix Inc. data sheet, 2003.
61. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nature Biotechnology, 1996. **14**: p. 1675-1680.
62. Naef, F., et al., *From features to expression: High density oligonucleotide array analysis revisited.* Tech Report, 2001. **1**: p. 1-9.

63. Naef, F. and M.O. Magnasco, *Solving the riddle of the bright mismatches: hybridization in oligonucleotide arrays.* Physical Review E, 2003. **68**(011906).
64. Affymetrix, *Genechip expression analysis technical manual.* 2003.
65. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res., 2003. **31**(4): p. e15.
66. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics., 2003. **4**(2): p. 249-264.
67. Kowalski, B., *Presentation for the Center for Process Analytical Chemistry (CPAC).* 1997.
68. Sharaf, M.A., D.L. Illman, and B. Kowalski, *Chemometrics*. 1986, New York: John Wiley & Son.
69. Trygg, J. and S. Wold, *Introduction to statistical experimental design - What is It? Why and Where is it Useful?* Homepage Of Chemometrics Editorial, 2002.
70. Tseng, G.C., et al., *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.* Nucleic Acids Res., 2001. **29**(12): p. 2549-57.
71. Höskuldsson, A., *Centring and scaling of data.* Homepage Of Chemometrics Editorial, 2004.
72. Kaufman, L. and P.J. Rousseeuw, *An Introduction to Cluster Analysis.* John Wiley & Sons, New York, 1990.
73. Coleman, D., et al., *Some computational issues in cluster analysis with no a priori metric.* Computational Statistics and Data Analysis, 1999. **31**: p. 1-12.
74. Massart, M.A. and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis.* John Wiley & Sons, New York, 1983.
75. Pearson, K., *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 1901. **2**: p. 559-572.
76. Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* PNAS, 2000. **97**(18): p. 10101-10106.
77. Henrion, R. and G. Henrion, *Multivariate Datenanalyse*. 1995: Springer Verlag.
78. Ambroise, C. and G.J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data.* PNAS, 2002. **99**(10): p. 6562-6566.
79. Duin, R.P.W., *Support vector classifiers: a first look.* Proc. of the Third Annual Conference of the Advanced School for Computing and Imaging, 1997.

80. Vapnik, V., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.
81. Suykens, J., *A (short) introduction to Support Vector Machines and kernelbased learning.* ESANN 2003, Bruges, 2003.
82. Guyon, I. and D. Stork, *Linear Discriminant and Support Vector Classifiers.* Advances in Large Margin Classifiers, 2000: p. 21-36.
83. Belousov, A.I., S.A. Verzakov, and J. von Frese, *A flexible classification approach with optimal generalisation performance: support vector machines.* Chemometrics and Intelligent Laboratory Systems, 2002. **64**(1): p. 15-25.
84. Brown, P.O., et al., *'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.* Genome Biol., 2000. **1**(2): p. 3.1-3.21.
85. Tusher, V.G., R.J. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* PNAS, 2001. **98**(9): p. 5116-5121.
86. Ross, M.E., et al., *Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.* Blood, 2003. **102**(8): p. 2951-2959.
87. Hastie, T.J., et al., *PAM: Prediction Analysis of Microarrays User guide and manual.*
88. Tibshirani, R.J., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proceedings of the National Academy of Science, 2002. **99**: p. 6567-6572.
89. *Statistical Algorithm Description Document.* Affymetrix Inc., 2002.
90. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.* PNAS, 2001. **98**(1): p. 31-36.
91. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.* Genome Biol., 2001. **2**(8): p. 32.1-32.11.
92. *Structured Query Language Introduction Course Material.* OOPSLA Lab., Object-Oriented Programming, Systems, Languages, and Applications Laboratory, Seoul National University, Korea, 2004.
93. Almgren, R., Associate Professor of Mathematics and Computer Science, University of Toronto.
94. Brema, A., et al., *Minimum information about a microarray experiment - MIAME - toward standards for microarray data.* nature genetics, 2001. **29**(4): p. 365-371.
95. Baggerly, K.A., *Detecting and Correcting Misalignment in Affymetrix Data.* Department of Biostatistics, M. D. Anderson Cancer Center, 2002.

96. Diette, G. and N. Hansel, *PGA Human CD4+ Lymphocytes U133 microarray study.* http://microarray.cnmcresearch.org/pgadatatable.asp, 2002.

97. Haferlach, T., et al., *Proprietary dataset; Medizinische Fakultät, Ludwig-Maximilian-Universität München.*

98. Chan, V., D.J. Graves, and S.E. McKenzie, *The biophysics of DNA hybridization with immobilized oligonucleotide probes.* Biophys. J., 1995. **69**(6): p. 2243-55.

99. Hagan, M.F. and A.K. Chakraborty, *Hybridization dynamics of surface immobilized DNA.* Journal of Chemical Physics, 2004. **120**(10): p. 4958-4968.

100. Peterson, A.W., R.J. Heaton, and R.M. Georgiadis, *The effect of surface probe density on DNA hybridization.* Nucleic Acids Res., 2001. **29**(24): p. 5163-5168.

101. Gonzalez, R.C. and R.E. Woods, *Digital Image Processing*. 2002: Prentice Hall, Pearson Education International.

102. Parker, J.R., *Algorithms for image processing and computer vision*. 1997: Wiley Computer Publishing.

103. Seul, M., L. O'Gorman, and M.J. Sammon, *Practical Algorithms for Image Analysis.* Cambridge University Press, 2000.

104. Hulver, M., *PGA Human Muscle Obese Dataset.* Children's National Medical Center, 2002.

105. Affymetrix, *Human Genome U133 Latin-Square Dataset.* http://www.affymetrix.com/support/technical/sample_data/datasets.affx.

106. Thermo Galactic, W., MA, USA.

107. Hegland, M., S. Roberts, and I. Altas, *Finite Element Thin Plate Splines for Surface Fitting.* Computational Techniques and Applications, CTAC97, 1997.

108. Roberts, S., M. Hegland, and I. Altas, *Approximation of a thin plate spline smoother using continuous piecewise polynomial functions.* SIAM J. Numer Anal, 2003. **41**(1): p. 208-234.

109. Billings, S.D., G.N. Newsam, and R.K. Beatson, *Smooth fitting of geophysical data using continuous global surfaces.* Geophysics, 2002. **67**(6): p. 1823-1834.

110. Schuchhardt, J., et al., *Normalization strategies for cDNA microarrays.* Nucleic Acids Res., 2000. **28**(E47).

111. Podsadlowski, P., *Neue Fluorophore für die Bioanalytik – Verwendung in der Tierartendifferenzierung mittels DNA-Biosensor und DNA-Chiptechnologie.* Dissertation, 2004.

112. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res., 1994. **22**(22): p. 4673-4680.

113. Yeoh, E.-J., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002. **1**: p. 133-143.

114. Ihaka, R. and R. Gentleman, *R: A language for data analysis and graphics.* J. Royal. Statist. Soc. B., 1996. **58**: p. 267-288.

115. Kudo, M. and J. Sklansky, *Comparison of algorithms that select features for pattern classifiers.* Pattern Recognition, 2000. **33**(1): p. 25-41.

116. Schadt, E.E., et al., *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* J. Cell. Biochem, 2001. **S 37**: p. 120-125.

117. Speed, T.P., *Summarizing and Comparing Genechip Data.* Talk at the Affymetrix User Group Meeting Redwood City, CA, 2002.

118. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.

119. Saviozzi, S. and R.A. Calogero, *Microarray probe expression measured, data normalization and statistical validation.* Comp Funct Genom, 2003. **4**: p. 442-446.

120. Lambert, D., *Key challenges for statisticians in business and Industry: Another View.* Technometrics, 1998. **40**: p. 201-203.

121. Lambert, D., *What Use is Statistics for massive data?* 2000.

122. Guyon, I. and A. Elisseef, *An Introduction to Variable and Feature Selection.* Journal of Machine Learning Research, 2003. **3**: p. 1157-1182.

123. Di Lecce, C., et al., *Classifier combination: the role of a-priori knowledge.* Proceeding of the Seventh International Workshop on Frontiers in Handwriting Recognition, 2000.

124. Saner, M., *Of Mice and Men: regulating and using patents.* Institute On Governance Backgrounder, 2002.

This work was created in the time span from June 2001 to October 2004 at the Anorganic and Analytical Chemistry Department of the Westfälische Wilhelms-University Münster under supervision of Prof. Dr. Karl Cammann.

I want to thank Prof. Dr. Karl Cammann for the opportunity to work on a task embedded in a very interesting field dealing with state of the art technologies. He and his workgroup provided an ideal environment for concentrated and joyful work.
Special thanks go to Dr. Jürgen von Frese for his great support and assistance with everything related to chemometrics and his supervision of this work. Of course, my thanks also go to the other members of the chemometrics group, Dr. Anton Belousov and Serguey Verzakov, which were not only always helpful but also provided a relaxed and fun atmosphere at work.
The members of the workgroup of Prof. Dr. Karl Cammann, especially Ms. Viola Podsadlowski as well as Ms. Silke Flotho from the workgroup of Prof. Spener, are to be thanked for their collaboration and help in many life-science related topics as well as for the great working atmosphere. I thank Dr. Torsten Börchers from the Institute for Chemical and Biological Sensor Research for helpful insights in molecular biology and microarray technology. My thanks also go to Mrs. Marianne Lüttmann and Mrs. Karin Weißenhorn for continuous support and an organized infrastructure at the university.

Very special thanks go to my parents and my sister Evelyn not only for their idealistic and materialistic support during my entire studies ☺. Thanks for everything!

Lebenslauf

Persönliche Daten

Name:	Eric Frauendorfer
Geburtstag:	09. Januar 1976
Geburtsort:	Caracas, Venezuela
Familienstand:	ledig
Eltern:	Erich Frauendorfer
	Hannelore Frauendorfer, geb. Popp

Schulbildung:

08/82 – 07/89	Collegio Humboldt, Caracas
08/89 – 05/95	Johannes Althusius Gymnasium, Emden.
18.05.1995	Allgemeine Hochschulreife

Studium und Prüfungen

10/95 – 05/01	Studiengang Diplom-Chemie, WWU Münster
24.10.1997	Vordiplom im Studiengang Diplom-Chemie, WWU Münster
10/00 - 05/01	Diplomarbeit am Institut für Anorganische und Analytische Chemie, WWU Münster
28.05.2001	Diplom im Studiengang Diplom-Chemie, WWU Münster
seit 06/01	Promotionsstudium Chemie, WWU Münster; Beginn der Dissertation im Fach Analytische Chemie unter der Leitung von Herrn Prof. Dr. K. Cammann an der WWU Münster.

Tätigkeiten:

04/00 - 07/00	Studentische Hilfskraft am Institut für Wirtschaftsinformatik
10/00 – 04/01	Wissenschaftliche Hilfskraft am Institut für Chemo- und Biosensorik (ICB).
05/01 – 07/04	Wissenschaftliche Hilfskraft am Institut für Anorganische und Analytische Chemie.
SS03 – WS03/04	Leitung des Grundpraktikums Instrumentelle Analytik an der WWU Münster