1 Introduction

In cancer diagnosis one goal is to characterize the individual tumor as detailed as possible. Recently, not only immunohistological or pathological methods have been applied but also molecular gene expression profiling. For this purpose many studies use whole genome microarrays, which measure thousands of genes simultaneously. From these expression measurements classifiers are trained and a diagnostic signature is derived. However, many of the genes on the large, whole genome microarrays are not needed for cancer classification. For diagnostics, the use of a few dozen genes is sufficient (Li, 2005). These can be represented onto a smaller, custom diagnostic microarray. In this thesis, I propose a computational framework for the design and analysis of small diagnostic microarrays holding only few genes. A diagnostic microarray study is composed of the following steps: acquiring of the patient samples, extracting of the mRNA, measuring of the mRNA quantities using microarrays, normalization of the expression data, deriving of a diagnostic signature, and validation of the results. I start with a chapter highlighting the necessary biological, statistical, and clinical background. In the first section, I introduce basic characteristics of molecular genetics. Then, in section 1.2, I describe techniques for measuring gene expression with special emphasis on microarrays. In section 1.3, I review microarray based gene expression studies. I outline that gene expression has already been successfully used for tumor classification. In section 1.4, I introduce machine learning techniques for classification and clustering. For properly assessing the classification performance, I review cross validation and stress its importance in classifier evaluation (section 1.4.2). In order to select relevant genes for the diagnostic microarray and provide better classification performance I describe gene selection and compare various techniques in section 1.4.3.

1.1 Molecular genetics

DNA – Human cells store the genetic information in deoxyribonucleic acid (DNA). DNA is a long, double stranded molecule. Both strands are composed of a sugar phosphate backbone and nucleic bases. These bases can either be adenine (`A`), thymine (`T`), guanine (`G`), or cytosine (`C`). The two DNA strands are complementary and run in antiparallel directions allowing for an easy way of replication. In eukaryotic cells the DNA is stored in a special compartment of the cell, the nucleus. Here, the DNA molecule is tightly packed into thread-like structures, the chromosomes.

Genes – A gene is an information coding region on the DNA with a certain structure. It is composed of a promotor region that controls the transcription of the gene, and the

actual protein coding region. Proteins are required for the structure, transport, function, and regulation of the body's tissues and organs. They are made up of hundreds to thousands of amino acids, which are attached to one another in long chains. The sequence of amino acids determines each protein's unique three-dimensional structure and its specific function.

The process from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. First, a gene has to be transcribed into ribonucleic acid (RNA), which then has to be processed and translated into a sequence of amino acids by ribosomes.

Transcription – In eucaryotes, the transcription of genes happens in the cell nucleus. Here, the part of the DNA coding for a gene is copied to RNA by a protein complex called polymerase. Biochemically, RNA is similar to DNA but it is single stranded, uses uracil instead of thymine and is less stable. The resulting RNA that encodes the information for making a protein is called messenger RNA. However, RNA does not only store information but it can be also biochemically active itself.

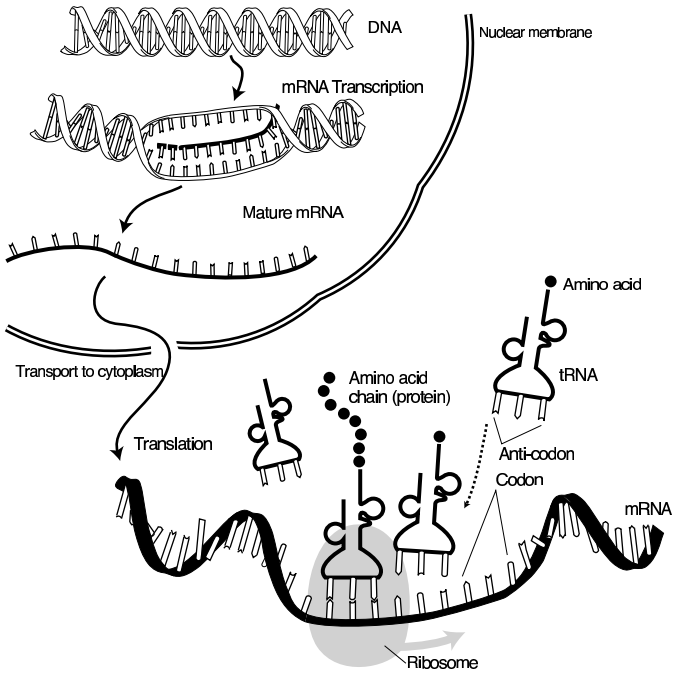

Figure 1.1: Central dogma of molecular biology showing the copying from DNA to RNA (transcription) and the translation of the mRNA into an amino acid chain that is later folded into a protein. *Image from NHGRI*

Translation – The second step from a gene to a protein in eucaryotes takes place in the cytoplasm. The mRNA sequence is translated into a sequence of aminoacids by a specialized complex called a ribosome. Each nucleotide sequence of three bases, called a codon, codes for either one particular amino acid or for the start or the stop of the amino acid sequence. In total there are 20 different types of amino acids. The assignment of an aminoacid to a specific codon is deterministically done using a universal translation table, the genetic code. The whole amino acid sequence is assembled at a ribosome. Here, transfer RNAs (tRNA) that match to the codons of the mRNA are recognized. A tRNA

carries a specific amino acids and at the ribosome these amino acids are connected one by one to form an amino acid chain. This chain later folds into a protein.

Transcription and translation together are the steps from DNA to protein. This flow of information from DNA to RNA to proteins is the fundamental principle of molecular biology and for that called the central dogma (Fig. 1.1).

Gene regulation – The genomic DNA of each cell holds all information on which proteins are synthesized in which cell under which condition. However, only a fraction of genes are expressed at a certain time. The rest of the genes are repressed. The amount of the expressed transcript of a given gene is called its gene expression. Expressed genes include those that are transcribed into mRNA and then translated into protein as well as those that are transcribed into RNA but not translated into protein (e.g. ribosomal RNAs). Many genes are strongly regulated and only transcribed at certain times, in certain environmental conditions, and in certain cell types. The cell controls if and how strongly genes are expressed through cell signals. When the cell receives a signal it can change the mRNA expression level of certain genes and thus react to changes in the environmental conditions.

The process of controlling the mRNA expression level of each gene is known as gene regulation. Gene regulation is especially important in cell development and differentiation. The regulation of genes can occur at any level: transcription, translation, or protein modification and degradation. Signals from the environment or from other cells are detected by cell receptors and activate signal cascades in the cell. In the end this triggers proteins called transcription factors. These proteins bind to regulatory regions of a gene and increase or decrease the level of transcription. By controlling the level of transcription, this process determines the amount of mRNA and therefore the amount of protein that is made by a gene at any given time. In order to produce new protein, first the corresponding gene has to be expressed. In general, the more transcript is expressed, the more protein is produced. Thus, gene expression has a direct influence on protein production.

Certain biological processes go along with a change of the expression of many genes. In cell division the whole genetic material has to be duplicated and systematically distributed into two daughter cells. When the normal regulation of the cell cycle is disrupted cells can divide without order. During this uncontrolled proliferation genetic defects can accumulate. In the end, this can lead to cancer (Bissell and Radisky, 2001). The analysis of gene expression changes of a tumor cell allows to characterize the tumor in detail leading to more accurate tumor diagnosis.

Textbooks like Griffiths *et al.* (2002) and Alberts *et al.* (2002) give detailed descriptions of the actual mechanisms of molecular genetics and gene regulation.

1.2 Measuring gene expression

Measuring the mRNA expression levels of genes enables us to gain insight into the cell's activity and regulation. However, in order to find diagnostically relevant genes it is necessary to quantify and compare the mRNA levels of samples from different conditions. In the following section, we describe methods for assessing mRNA expression levels of genes often also called the gene expression level. The measured gene expression levels for many genes in one tissue sample is called a gene expression profile.

All methods for gene expression measurement rely on the same principle. They make use of the preferential binding of complementary nucleotide sequences to each other. When screening for a target gene, a specifically labeled probe, which is complementary to the target genes mRNA, is used. The probe preferentially binds to the target genes mRNA and can be detected by its label.

Originally, Northern blotting (Alwine *et al.*, 1977) was used to detect specific mRNA. Here, the RNA is separated by size using an electrophoresis gel. Then the specific RNA is detected using a hybridization probe. But Northern blots are not sensitive enough to small mRNA quantities and time consuming (Dvorák *et al.*, 2003).

1.2.1 PCR (Polymerase chain reaction)

In the 1980s the American biochemist Kary Mullis developed a technique for making an unlimited number of copies of any piece of DNA, the polymerase chain reaction (PCR, Saiki *et al.* (1985)). In a matter of hours, PCR produces millions of copied DNA molecules. Since the amount of copies made depends on the number of initial templates, PCR can be used to quantify gene expression.

The core of the PCR is a polymerase that replicates a piece of DNA flanked by primers. Primers are small pieces of complementary DNA that bind to the DNA strands. The polymerase extends from the primers and copies the region between a primer pair. In each round of replication a cooling step is followed by a heating step. During the cooling step the primers bind and the DNA is doubled by transcription. During the heating steps the assembled strands are separated again. Thus, the specific DNA material enclosed by the flanking primers is doubled every round, resulting in an exponential growth.

In order to measure the original amount of transcript relative to a normalization gene, a dye is used. For both genes the number of replication rounds are counted that are needed before the dye can be detected. PCR can also be used to measure gene expression. First, the RNA has to be reverse transcribed to DNA. Then, the complementary DNA is multiplied with standard PCR. The whole process is called RT-PCR (reverse transcription - polymerase chain reaction). Today, quantitative RT-PCR (qRT-PCR, Wang *et al.* (1989)) is routinely applied to measure mRNA abundance. The assay is cheap, versatile, and requires only small amounts of starting mRNA (Robison *et al.*, 2004).

1.2.2 Microarrays

Particularly for diagnostics it is beneficial to measure more than a single marker gene. Looking at a panel of genes together helps to provide a more accurate and robust molecular diagnosis (Taback *et al.*, 2001; Chen *et al.*, 2005; Schneider *et al.*, 2005). However, PCR experiments are tedious and time consuming for measuring a large number of genes. Here, microarrays offer a solution. With microarrays one can monitor gene expression for tens of thousands of transcripts simultaneously in a single experiment. This high-throughput method was a breakthrough in experimental biology and is used extensively since then. In the year 1996 only 10 articles using microarrays were published. But within 10 years this has increased to more than 4000 articles published per year (Fig. 1.2).

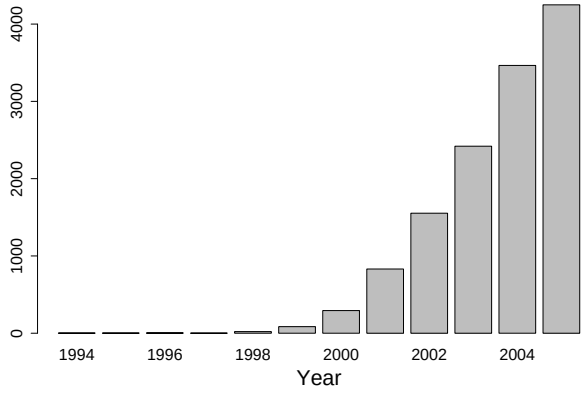

Figure 1.2: Number of articles published per year that contain the keyword "microarray" in the title or abstract. *Source: www.pubmed.org*

Part of the success of microarrays is their versatility. They can be used for the detection of differentially expressed genes of different biological entities (Chee *et al.*, 1996), the detection of changes of correlation in gene groups (Kostka and Spang, 2004), the identification of a subset of genes that provide discrimination between tissue types for diagnosis or prognosis (van 't Veer *et al.*, 2002), the detection of novel subgroups of lymphomas (Alizadeh *et al.*, 2000), time-course analysis in yeast (Spellman *et al.*, 1998), and the analysis of dose response effects on gene expression in colon carcinoma (Hu *et al.*, 2005). Additionally, specially designed microarrays have been used for polymorphism analysis (Wang *et al.*, 1998) and sequencing (Pease *et al.*, 1994).

In the following, target genes are the genes of interest in the sample tissues for which the expression levels are measured. Probes are pieces of DNA that are complementary to the target genes. Each probe is brought onto the microarray at a certain position so that it can be localized later on. When using cDNA clones as probes they are immobilized on the array by spotting the clones mechanically onto their prespecified positions. When using oligonucleotides the probes are immobilized by synthesizing them directly on the microarray at their positions. Typically, between several thousand and several tens of thousands of different probes are immobilized on one array. One probe location does not only hold one single DNA oligonucleotide or cDNA but in the range of millions of copies of the same molecule.

The basic idea behind all gene expression arrays is to extract the mRNA from a tissue, reverse transcribe it to cDNA and amplify it in a way that conserves proportions of molecular abundance. Following this, the cDNA is labeled and used as a target to

bind to complementary DNA. The target is detected using single-stranded cDNA or oligonucleotide probes, attached at fixed spots on a support surface. As DNA binds to complementary DNA sequences, the probes bind to their complementary targets. As the targets are physically bound at known positions the DNA is now also affixed there. Finally, the mRNA abundance can be indirectly measured by measuring the label intensity of the spots lighting up. These are the probes that had complementary sequences to the mRNA under examination.

I now distinguish two phases of the microarray analysis: First, the production of the microarray itself, where the probes have to be produced and spotted onto the microarray. Second, the hybridization of the microarray with the target sample (Fig. 1.3).

Microarray production – There exist several types of microarrays and different techniques to produce them. They can be distinguished by the support system they use (nylon or glass), the way the probes are spotted (mechanical griding, inkjet, or photolithography), and the labeling system (fluorescent or radioactive).

Macroarrays are based on nylon or nitrocellulose filters and measure up to 22cm×22cm. The most commonly used label is radioactivity. Microarrays typically use glass, which allows further miniaturization and the use of fluorescent dyes. Often microscopic slides are used as a support medium for microarrays. Affymetrix uses arrays with a coated quartz surface of 1.5cm×1.5cm.

Macroarrays are spotted using mechanical griding, whereas microarrays can also be produced with inkjet or photolithography techniques. Due to higher standardization in manufacturing and hybridization photo and inkjet technology produce more reproducible results (Li *et al.*, 2002). However, compared to radioactively labeled targets fluorescently labeled microarrays require larger RNA amounts and therefore often the use of mRNA amplification techniques.

Affymetrix uses photolithography techniques with masking. Wherever ultraviolet light can pass through tiny openings (few micrometers) in the mask it removes a protective cap inducing oligonucleotide synthesis at this spot. Depending on the GeneChip generation Affymetrix uses between 11 and 20 pairs of probes (HU6800: 20 probes, HG-U95: 16 probes, HG-U133: 11 probes) representing one gene (Lipshutz *et al.*, 1995). Each probe is 25 nucleotides long. The probes screening for one DNA sequence are included twice at two different positions: once as a perfect matching oligonucleotide sequence (PM) and once with a mismatch at the central 13th nucleotide (MM). The intention of the mismatch is to measure unspecific binding. Other arrays usually only use one longer oligonucleotide or a complete cDNA.

Hybridization – From the tissue of interest mRNA is extracted and copied into cDNA. The cDNA is radioactively or fluorescently labeled. Then, the labeled cDNAs are hybridized to the microarray where they bind to the complimentary probes on the array during an annealing time. The rest of the probe mixture that did not bind to the targets is washed away. Only the cDNA complimentary to the immobilized genes remains on the array. The

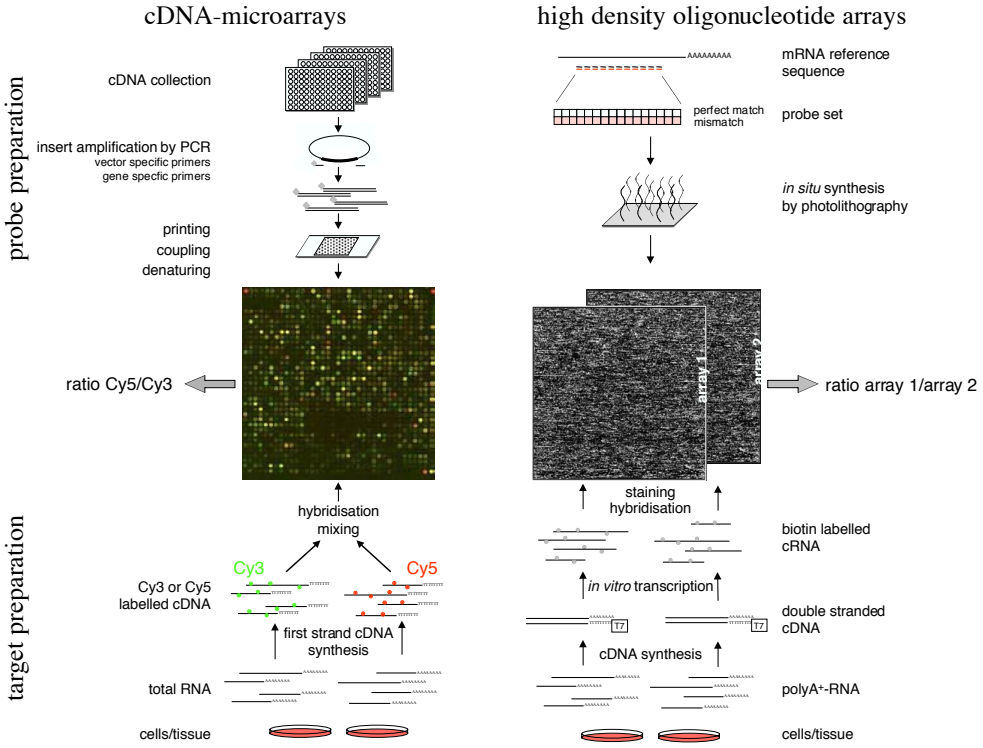

Figure 1.3: Comparison of the production and hybridization steps in cDNA and high density oligonucleotide arrays *Reprinted by permission from Macmillan Publishers Ltd: Nature Cell Biology (Schulze and Downward, 2001), copyright 2001*

more mRNA is present in the original tissue, the more complimentary cDNA is produced by the PCR, the more cDNA binds and the brighter is the spot. The microarray is then put in a scanner or on a screen and a digital image of the microarray is taken. Image analysis software imports the picture, identifies the spot locations and boundaries, and outputs the intensities and colors for these spots. The intensity values still need to be corrected for the surrounding background or spilled over spots (Yang *et al.*, 2000). Finally, in order to encounter systematic variation and to allow a comparison between different microarrays, data normalization has to be performed (Yang *et al.*, 2002).

cDNA microarrays usually use two channels, where simultaneously a tissue of interest is labeled with one dye (e.g. Cy3 or Cy5) and a control or universal reference (Novoradovskaya *et al.*, 2004) is labeled with another dye of different color. The measured intensity of each dye after hybridization reflects the mRNA abundance in the corresponding tissue. Since mechanically spotted probes can have different shapes and sizes it is hard to compare intensities as absolute values directly. Bigger spots show stronger intensities than smaller spots. When two dyes are used a ratio of the intensities can be calculated. This ratio is independent of the spot geometry as both channels are affected equally. Thus, the two dye color approach allows an easy way for a spot size independent normalization.

Affymetrix uses only single channel arrays, where one tissue is labeled with biotin and then hybridized. More detailed information about microarrays can be found in Brown

and Botstein (1999).

1.2.3 Data processing

After the image processing of the microarray we have an expression intensity measurement for each probe or gene. The data is stored in a gene expression matrix $X = (x_{ij}) \in \mathbb{R}^{p \times n}$, where p is the total number of probes or genes and n the total number of arrays.

Normalization – Microarray intensities do not directly reflect the absolute gene expression but are also influenced by experimental artifacts. These factors can stem from sample preparation and variability during the production or the processing of the arrays (Hartemink *et al.*, 2001). Therefore, in order to be able to compare results between different microarrays the data has to be normalized.

One way of normalization is the use of housekeeping genes. Here, a predetermined set of genes is selected that is assumed not to be biologically regulated and ubiquitously expressed. A change in the expression of these genes reflects experimental variation and should be compensated through normalization.

A simple normalization is to subtract the mean intensity of the housekeeping genes and divide by the standard deviation of their intensities. A more robust normalization version is to genewise subtract the median intensity of all housekeeping genes and divide by the inter-quartile range of their intensities (Pan, 2002). In whole genome microarrays most genes are not changed and the number of up- and down-regulated genes are similar. In this case, it is possible to use the data of all genes in the normalization procedure instead of restricting normalization to housekeeping.

Another method is quantile normalization (Bolstad *et al.*, 2003). Here, the expression matrix is sorted numerically for each column (microarray). Then, for each row (gene) the mean is calculated and assigned to every element in this row. Finally, the columns are arranged back to their original order. This makes the distribution of the gene expression intensities for each array in a set of arrays the same.

VSN - Variance stabilizing normalization – Rocke and Durbin (2001) proposed a model for gene expression incorporating additive and multiplicative noise. The gene intensity is modeled by $X = \alpha + \beta e^{\eta} + \nu$, with α being an offset, β the mRNA expression level and η and ν independent, $N(0, 1)$ distributed error terms. Based on this model the variance of a probe intensity depends on its mean via a quadratic function, $\mathrm{Var}(X) = \mathrm{Var}(e^{\eta})\beta^2 + \mathrm{Var}(\nu)$ (Rocke and Durbin, 2001). In order to make the variance of the gene expression intensity approximately independent of the mean Huber *et al.* (2002) have proposed to use a variance stabilizing transformation. Using this together with an affine-linear normalization they obtained $\hat{x}_{ij} = \mathrm{arsinh}(a_j + b_j \tilde{x}_{ij})$. The parameters a_j and b_j are derived using a robust variant of a maximum likelihood estimation calculated on a subset of genes. This subset are those genes that show the least variation in the normalized intensities and are determined by a least trimmed sum of squares regression. For large

intensities the variance stabilizing transformations are equivalent to the logarithm but they do not have the inherent problems with negative or zero valued intensities.

RMA - robust multi-array analysis – Here, preprocessing consists of three steps: background correction, normalization and probeset summarization. In Affymetrix arrays several probes are measuring the same gene and have to be combined into a single gene intensity value during the probeset summary step. For normalization, first, a robust average of the differences was used: $PM_{ij} - MM_{ij} = \theta_i + \epsilon_{ij}$, but this model is only accurate for similar variances of the error term for all probes. Since probes with larger mean intensities tend to have larger variances, a log transformation was used to reduce the dependency between mean and variance. Furthermore, Irizarry *et al.* (2003b); Naef *et al.* (2002) found that subtracting the mismatch probes MM to correct for unspecific binding produces less reliable results compared to not using MM at all. Therefore, Irizarry *et al.* (2003a) suggested a new model, called robust multi-array analysis (RMA). RMA uses a log scale linear additive model, $T(PM_{ij}) = e_j + a_i + \epsilon_{ij}$, where e_j is the log scale expression value, a_i is the log scale probe affinity effect, and T takes the log of the background corrected and quantile normalized intensities. In spike in and dilution experiments RMA showed better precision and provided more consistent estimates of fold change compared to other normalization methods like dChip (Li and Wong, 2001) and MAS5.0 (Irizarry *et al.*, 2003a).

A more detailed review and comparison of Affymetrix preprocessing methods can be found in Irizarry *et al.* (2006).

1.3 Gene expression analysis in clinical studies

Transcriptional profiling has emerged as a powerful tool to identify new cancer classes (class discovery (Jones *et al.*, 2005)), assigning tumors to known classes (class prediction (Golub *et al.*, 1999; Alizadeh *et al.*, 2000)), and predicting clinical outcome solely based on gene expression (van de Vijver *et al.*, 2002). The authors state that it has the potential to affect diagnosis, tumor staging, prognosis, and treatment.

Until now, cancer classification is based on the morphology of tumor samples, the presence of metastases, or the degree of differentiation (Ciro *et al.*, 2003). However, morphologically similar tumors can have different clinical outcome or response to treatment (Ciro *et al.*, 2003). Even with immunophenotyping, cytogenetics, or mutation analysis it is not possible to predict tumor outcome. For example, up to 30% of patients with high-risk cancer show long-term disease-free survival even without chemotherapy and regional therapy only but it is not yet possible to predict this a priori (Olson, 2004).

Gene expression profiling can complement traditional diagnostics. It helps to further refine diagnosis by improved prediction of treatment benefits (Chang *et al.*, 2003; Cheok *et al.*, 2003; Holleman *et al.*, 2004). We now review that gene expression profiling has

been successfully applied for a wide variety of diseases. Mostly tumors have been studied, with by far the most studied being breast cancer (Fig. 1.4).

Breast cancer –Breast cancer is the most prevalent non-skin cancer in the world and the second leading cause of all cancer deaths in western women (Hoyert *et al.*, 2005).

Despite an overall similarity in tumor morphology breast cancer patients vary in the responsiveness to treatment and have different clinical outcomes (Loi *et al.*, 2005). Treatment can have severe side effects including cardiotoxicity, neurotoxicity, and secondary cancers (Loi *et al.*, 2005). Therefore, a better understanding of breast cancer biology is needed. For this, gene-expression profiling offers a promising tool (Robison *et al.*, 2004).

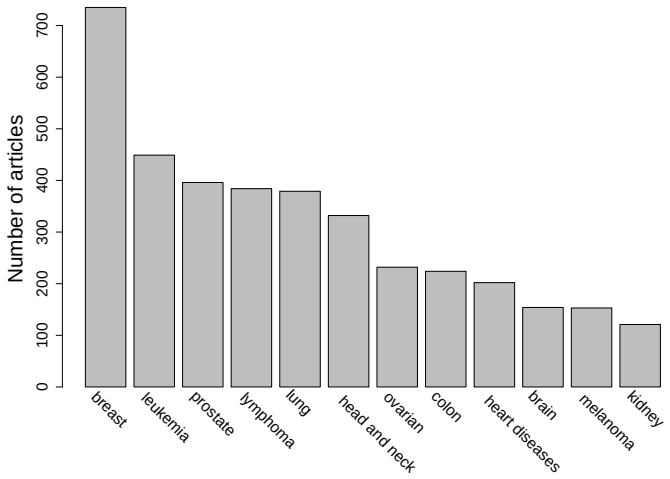

Figure 1.4: Number of published articles using gene expression for disease profiling. Generated on 2006-05-04 querying PubMED with the MeSH (Clarke *et al.*, 1997) term "Microarray analysis" and the corresponding disease term.

Perou *et al.* (2000) showed that gene expression profiling is capable to distinguish breast tumor subtypes. They clustered gene expression data of 65 tumor samples from 42 different patients and derived a 496 gene signature capable of distinguishing ER+ (estrogen receptor positive) from ER- (estrogen receptor negative) breast cancers. When extending the study to more samples they found that the ER+ group further split into five clinically relevant subgroups with different survival (Sørlie *et al.*, 2001). West *et al.* (2001) provided a breast cancer classification by ER status and lymph node status.

Currently, the detection of lymph node metastases at the time of surgery is used to determine whether the cancer has spread and if the patient should receive adjuvant treatment following the removal of the primary tumor (Ciro *et al.*, 2003). This adjuvant treatment can reduce the risk of distant metastases by one third, but for 70-80% of the patients chemo- or hormonal therapy would not have been necessary (Wadlow and Ramaswamy, 2005). van 't Veer *et al.* (2002) found a gene expression profile that predicts the formation of metastases more accurately than standard methods. Examining 117 primary breast tumors from lymph-node negative patients who had not received adjuvant therapy they derived a diagnostic signature containing 231 differentially expressed genes. These gene signatures were significantly stronger disease outcome predictors than other currently used diagnostic methods. Especially, non-relapse patients were identified more accurately by a classifier based on the 231 genes than by using standard methods. Non-relapse patients are those where the cancer did not reoccur within five years. By further refining the gene signature aiming for higher specificity van 't Veer *et al.* (2002) found a subset of 70 genes for prognosis of distant metastases that outperformed standard clinical pre-

dictors like histological grade or lymph node status. The prognostic relevance of the 231 gene signature was verified by Sotiriou *et al.* (2003). They proposed a 93 genes signature that could separate the population into two subgroups with significantly different survival. Glinsky *et al.* (2004) showed that they achieved even better accuracy with a smaller gene set when using gene expression in conjunction with prognostic factors ER status and lymph node status. Huang *et al.* (2003) screened 89 breast cancer samples. They predicted nodal metastatic states as well as relapse for breast cancer with about 90% accuracy. Their classifier may make surgical axillary gland staging unnecessary for the predicted candidates (Sandvik *et al.*, 2005). van de Vijver *et al.* (2002) extended the study by van 't Veer *et al.* (2002) and examined a cohort of 295 lymph-node-negative as well as lymph-node-positive breast cancer patients. Using the original 70-gene signature to predict survival they achieved better accuracy than using standard criteria based on histological or clinical characteristics. On the other hand, Edén *et al.* (2004) compared the 70 gene signature to the traditional NPI (Nottingham Prognostic Index) and found similar performance. They suggested to use both in a combined predictor. A more detailed review of gene expression studies related to breast cancer can be found in Loi *et al.* (2005) and Robison *et al.* (2004).

Response to treatment or external stimulus – For a long time it is postulated that tumor stroma generation and wound healing show a similar histological behavior (Dvorak, 1986). In several common epithelial tumors such as breast, lung, and gastric cancers the expression of a wound-response signature predicted poor overall survival and increased risk of metastasis (Chang *et al.*, 2004). Chang *et al.* (2005) used an independent data set of 295 early breast cancer patients to confirm that a wound response gene expression signature (response of normal fibroblasts to serum) is a powerful predictor of clinical outcome in patients with early stage breast cancers. Chang *et al.* (2003) identified transcriptional patterns associated with sensitivity and resistance to the drug docetaxel.

Lung cancer – Bhattacharjee *et al.* (2001) conducted a gene expression study of human lung carcinomas. They screened 186 lung tumor samples (including 139 adenocarcinomas) and found that gene expression profiling allowed to discriminate primary lung adenocarcinomas from metastases of extra-pulmonary origin. They also found distinct subclasses of adenocarcinomas with different patients' survival. Therefore they suggested that integration of expression profiles with clinical parameters has a strong potential in diagnosis of lung cancer patients. Beer *et al.* (2002) predicted patient survival in early-stage lung adenocarcinomas using a 50 gene signature. Additionally, in pulmonary adenocarcinomas a set of eight genes was found that provided independent prognosis and can become a clinical tool soon (Endoh *et al.*, 2004).

Skin cancer – Bittner *et al.* (2000) used gene expression profiling on 31 melanoma samples dividing them into two groups with different metastatic potential. The genes in their signature were mainly involved in cell motility and invasion. Haqq *et al.* (2005) found gene expression signatures differentiating between normal skin, nevi, and primary melanomas. They also detected two novel types of metastatic melanoma.

Prognostic and predictive gene expression signatures were also derived for many **other solid tumors**, like gliomas (Nutt *et al.*, 2003; Freije *et al.*, 2004; Fuller *et al.*, 2002), colon (Alon *et al.*, 1999; Hu *et al.*, 2005), prostate (Lapointe *et al.*, 2004; Singh *et al.*, 2002; Dhanasekaran *et al.*, 2001), head and neck carcinomas (Roepman *et al.*, 2005), and non solid tumors like hematological tumors, which we describe in the next paragraph.

Non-Hodgkin lymphomas – are a heterogeneous group of lymphoproliferative diseases with different prognosis and treatment responses (Sandvik *et al.*, 2005). In diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin lymphomas, long lasting remission after cytostatic drug treatment can only be achieved in 40-50% of the patients (Sandvik *et al.*, 2005). Alizadeh *et al.* (2000) showed that it is possible to subclassify DLBCL in three distinct types with different therapy response and prognosis based on gene expression profiling of 96 patients. Two novel subclasses are GCB (germinal center B-cell like DLBCL) and ABC (activated peripheral blood B-cell like DLBCL). The authors showed that GCB has a better survival than ABC. This was verified by Rosenwald *et al.* (2002). However, Shipp *et al.* (2002) could not verify the GCB signature on their own data. They suggested another gene expression signature for DLBCL outcome prediction that is independent of previously published signatures. Their signature consists of a 13 gene predictor that proved to be superior to traditional risk assessment using the International Prognostic Index. Lossos *et al.* (2004) combined the previous two datasets and derived a six-gene set for predicting survival. Poulsen *et al.* (2005) used a cross platform validation and could also subgroup DLBCL into GCB and ABC. For another group of lymphomas, the follicular lymphoma, Dave *et al.* (2004) established an expression signature that divided 191 specimen of untreated follicular lymphoma into four groups with different survival. These groups could not be detected by standard clinical prognostic parameters.

Leukemia – Golub *et al.* (1999) showed that their class predictor using gene expression distinguished between AML (acute myeloid leukemia) and ALL (acute lymphoblastic leukemia) with an accuracy of more than 85%. Armstrong *et al.* (2002) were further able to separate MLL (mixed-lineage leukemia), which have a particularly poor prognosis, from ALL and AML. Ross *et al.* (2003) identified all prognostically relevant subtypes of ALL with 97% accuracy using gene expression patterns. Ross *et al.* (2004) showed that they could identify and predict all major prognostic subtypes of AML with 93% accuracy. Yeoh *et al.* (2002) conducted a large study where they discovered new subtypes, classified known types, and predicted outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. With unsupervised methods they identified several known biological subgroups of pediatric ALL, namely: E2A-PBX1, MLL, T-ALL, hyperdiploid, BCR-ABL, TEL-AML1, and one novel group. They classified the groups with almost 100% accuracy. Recently, Haferlach *et al.* (2005) published a leukemia study, where they screened 937 individuals including 45 non-leukemic controls and achieved an overall prediction accuracy for the subgroups of 95.1% in cross validation.

Survival prognosis and response to treatment – Bullinger *et al.* (2004) screened 116 adults with AML and found two prognostically relevant subgroups in AML that could

be distinguished using a 133-gene signature for clinical-outcome prediction of survival. Chiaretti *et al.* (2004) and Willenbrock *et al.* (2004) identified distinct subgroups in ALL with different immunophenotype, response to therapy, and survival. Cheok *et al.* (2003) found a gene expression pattern of drug response in human leukemia cells. Holleman *et al.* (2004) showed that drug resistance and treatment outcome could be predicted using a signature with 124 genes. Cario *et al.* (2005) found a 54 gene-signature that distinguishes treatment resistant from treatment sensitive ALL samples with an accuracy of 84%.

Besides cancer research, gene expression profiling has been applied to studies of nonalcoholic steatohepatitis, serious autoimmune diseases, transplant rejection, and heart diseases (Sandvik *et al.*, 2005). Hwang *et al.* (2002), Grzeskowiak *et al.* (2003) and Jiang *et al.* (2002) have reported differential expression of cardiomyopathy and normal tissue as well as differences in dilated and hypertrophic cardiomyopathy.

More detailed reviews of clinical studies using gene expression can be found in (Sandvik *et al.*, 2005; Burczynski *et al.*, 2005; Ciro *et al.*, 2003; Raetz and Moos, 2004; Olson, 2004; Wadlow and Ramaswamy, 2005).

Validation – Simon (2003) found that most prognostic markers are either not medically relevant or not validated on independent test sets. However, only when a prognostic model can be validated on an independent group it can work satisfactory for future patients (Altman and Royston, 2000). Reid *et al.* (2005) was for example not able to validate a predictive model for antiestrogen response after tamoxifen treatment based on the expression ratio of two genes on an independent cohort of 58 patients. However, besides validation on independent samples sound statistical evaluation is indispensable. Ntzani and Ioannidis (2003) analyzed 84 clinical studies of which 30 addressed major clinical outcomes. Only nine of the studies used cross-validation and of these only two used complete cross validation. They conclude that larger studies applying proper clinical design, adjustment for known predictors, and sound validation are essential.

1.4 Machine learning

In this thesis I aim for a diagnosis using gene expression data. This can be seen as a classification problem where gene expression data are features or covariates and disease types are labels. Therefore I review machine learning methods for supervised classification in section 1.4.1. For estimating the prediction error achieved with the classifier I discuss cross validation strategies in section 1.4.2. For further optimizing the classification performance gene selection methods are used and reviewed in section 1.4.3. In chapter 3 I propose a novel gene selection method for classification and show that it improves classification accuracy. I first structure the genes using unsupervised clustering methods and therefore review them in section 1.4.4.

1.4.1 Supervised methods

In supervised methods the class label information of each object is given. The task is to build a classifier that can predict the class label of so far unseen objects with high accuracy. Supervised methods are often also called classification, discriminant methods, or class prediction. More formally, n objects are given with feature measurements $\mathbf{x}_i \in \mathbb{R}^p, i \in \{1, ..n\}$ and labels $y_i \in \{1, .., k\}$, where p is the number of features and k is the number of class labels. The data $(\mathbf{x}, y)$ can be seen as realizations of the random variables X and Y. Let the learning set $S = \{(\mathbf{x}_i, y_i)\}, \forall i \in \{1, .., n\}$ be n instantiations of the pair of random variables (X, Y). Given the learning set S the task is to find a classification function $c(\mathbf{x}; \alpha, S)$ that predicts the label from the features of a so far unseen object $\mathbf{x} \in \mathbb{R}^p$ using the learning set S and the parameters of the classifier α. α is a regularization parameter constraining the classifiers model complexity, e.g. number of genes used or kernel parameters used for SVMs. In our setting $\mathbf{x}$ is microarray data measuring p genes, y_i are disease type labels, and n is the number of microarrays.

There have been many classifiers for microarray analysis described in the literature. Nearest neighbor classification assigns the label of the new sample to the label of the closest sample with known label. This can be extended to kNN (k-nearest neighbor) where the labels from the k closest samples are averaged.

Tan *et al.* (2005) reviews several classification methods and concludes that simple decision rules in gene expression analysis are as effective as more complicated classifiers but easier to interpret, and therefore favorable. In the next paragraph we describe support vector machines (SVMs), a robust and fast classification method, in more detail as they are used throughout the following chapters of this thesis. SVMs have been shown by Brown *et al.* (2000) to provide on average the best prediction accuracy when comparing various learning methods. This was verified in a bigger study by Lee *et al.* (2005) who compared classifiers in combination with gene selection methods.

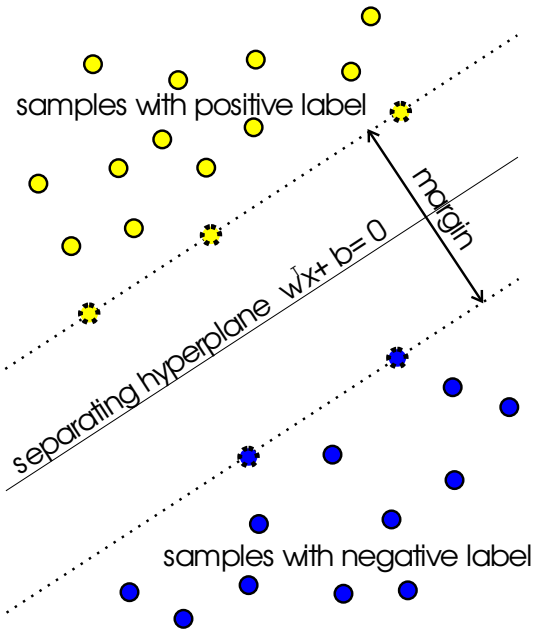

Figure 1.5: SVMs find a separating hyperplane with maximal margin. The objects on the margin (marked by dotted circles) are called support vectors.

Support Vector Machines (Boser *et al.*, 1992) are binary classifiers, that construct an optimal separating hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ with maximal margin. Here, the distance from the boundary to the closest data points, the support vectors, is maximized (Fig. 1.5).

$||\mathbf{w}||$ is minimized under the constraints: $y_i(\mathbf{w}^T\mathbf{x}_i+b) \geq 1 \quad \forall i$, $y_i \in \{-1, 1\}$ being the class label. The size of the margin separating the two classes is then given by $\frac{2}{||\mathbf{w}||}$. As long as the data is linearly separable the construction of the hyperplane is possible. An extension of SVMs, that can also deal with the non-separable case (where no solution for the above optimization problem exists), are soft-margin SVMs (Cortes and Vapnik, 1995). In soft margin SVMs slack variables $\xi_i \geq 0$ are introduced that penalize misclassified examples. The optimization problem then changes to a minimization of $||\mathbf{w}||$ with the constraints: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$, $\xi_i \geq 0$ and $\sum \xi_i \leq c$. If $\xi_i > 1$ misclassification occurs. Thus, the last constraint bounds the total number of training misclassifications to c.

Usually the optimization problem is formulated in its dual form using Lagrange multipliers λ_i and can be solved using standard numerical optimization packages or specially suited quadratic optimization packages. As typically only a small number of Lagrange multipliers are non-zero the hyper-plane is supported only by a few support-vectors. The final decision function $c(\mathbf{x})$ that classifies a new sample $\mathbf{x}$ is the sign of the distance vector from the hyperplane, $c(\mathbf{x}) = \text{sign}[\mathbf{w}^T\mathbf{x} + b]$. SVMs can be extended to non-linear classification strategies when replacing the inner products with a kernel function $K(u, v)$ that fulfills Mercer theorem (Vapnik, 1998). The kernel maps the data into a high dimensional feature space where the hyper-plane classification takes place. Typical kernels include, linear: $K(u, v) = u^T v$, polynomial: $K(u, v) = (1 + u^t v)^\theta$, and RBF (radial basis function): $K(u, v) = \exp\left(-\frac{||u-v||^2}{\theta}\right)$. Supervised learning methods are reviewed in more detail in Hastie *et al.* (2001).

In the next section we review general strategies for assessing the prediction error of a classifier (performance assessment) and how to choose the regularization parameters α for such a classifier (model selection).

1.4.2 Cross Validation

Let P be the underlying, but unknown, population distribution of (X, Y). Let $\hat{P}$ be the empirical distribution in the learning set S. The loss function $L(y, c)$ measures deviations between the predicted label and the true class label. The expected loss, or risk, is defined as:

$$R(c, P) = E_P[L(Y, c)] = \int L(y, c(\mathbf{x}))dP(\mathbf{x}, y)$$

However, the expected loss is a theoretical measure as the population distribution P is needed for the classifier construction and the assessment. Usually P is unknown and the rule is based upon the observations in S. This defines the conditional risk (also known as generalization error):

$$R(c(\cdot; \alpha, S), P) = \int L(y, c(\mathbf{x}; \alpha, S))dP(\mathbf{x}, y)$$

Ideally, an independent dataset can be used to assess the conditional risk. However, typically, one must use the learning set S for model building, selection, and performance

assessment. This leads to the resubstitution or apparent error, which is a simple method for estimating the conditional risk:

$$R(c(\cdot;\alpha,S),\hat{P}) = \int L(y,c(\mathbf{x};\alpha,S))d\hat{P}(\mathbf{x},y)$$

The evaluation of the conditional risk is important for two purposes. First, for finding the appropriate model complexity of the classifier, and second, for estimating the generalization error (the performance on future samples).

The goal in model selection is to find a model that minimizes the conditional risk over a collection of potential models. In general, more complex models fit the learned data better. However, S is only a sample of the underlying population and characteristics learned from one set may not generalize well to the whole population. This leads to less accurate predictions on future samples and is known as overfitting or overtraining (Geman *et al.*, 1992). When the whole learning set S is used for constructing, selecting, and evaluating the prediction error of c the generalization error is underestimated (Efron, 1983).

One way to obtain less biased estimates is the hold-out procedure. Here, the dataset S is split into a training and a test set. The training set is used for model fitting and the test set is used for estimating the generalization error.

Lachenbruch and Mickey (1968) and Stone (1974) proposed a method called cross-validation when one is forced to use the same data for model building and its assessment. Cross-validation is a resampling method for estimating the prediction error of a classifier and can provide more accurate estimates than hold-out procedures without reducing the number of training examples. The basic idea is to divide the whole dataset into chunks and repeatedly use all but one chunk to train the model and the held-out chunk to assess the generalization performance.

k-fold CV – In k-fold cross-validation, the data is split into k subsets (or folds) and the classifier is only trained on $k-1$ of the subsets but tested on the left out. This is repeated k times so that every sample was exactly one time in the test-set. The observed errors are averaged. Blum *et al.* (1999) showed that the k-fold estimate is strictly more accurate than a hold-out estimate on $1/k$ of the data based on its variance and all higher moments. A special case of k-fold cross validation is the leave-one-out cross validation. Here, k is equal to the number of samples n. Exactly one sample is left out, the classifier is trained on all others and tested on the left out sample. As the parameters of the classifier strictly depend on the training data set LOOCV needs to learn the classifier n times, once for each left out sample. Therefore, LOOCV tends to be variable, especially when data are noisy. Bengio and Grandvalet (2004) proofed that there exists no universal unbiased estimator of the variance of k-fold cross-validation. In general, cross validations with a small k have a larger bias but offer the advantage of smaller variance (Hastie *et al.*, 2001). Therefore, Geisser (1975) and Zhang (1993) suggested to use k-fold cross validation instead of hold-out procedures and Hastie *et al.* (2001) recommended to use k around 5-10.

While cross-validation error estimation is less biased than resubstitution, it displays excessive variance, which makes individual estimates unreliable for small samples. Bootstrap methods (Efron, 1979) provide less variable error estimates. However, bootstrapping has high computational cost and often an increased bias (Braga-Neto and Dougherty, 2004). Kohavi (1995) compared cross validation and bootstrap and concludes that for large enough real world datasets 10-fold cross validation should be used for model selection.

In this thesis we use the evaluation method proposed by Ruschhaupt *et al.* (2004). They suggested a nested cross validation strategy with an outer and an inner cross validation. In the outer cross validation the held-out test set is used for evaluating the generalization performance. The training set is used to assess the model parameters for the classifier. Therefore, in an inner cross validation, this set is further split up into a training set on which the classifier's parameters are tuned and a validation set on which the performance with this parameter set is assessed. For every possible classifier parameter combination the inner cross validation has to be run. The predictive performance for the best parameter set achieved in the inner loop is evaluated on the outer loop test set. Thus, the nested cross validation serves two purposes. First, finding the optimal model parameters and second, assessing the generalization performance with this parameter set.

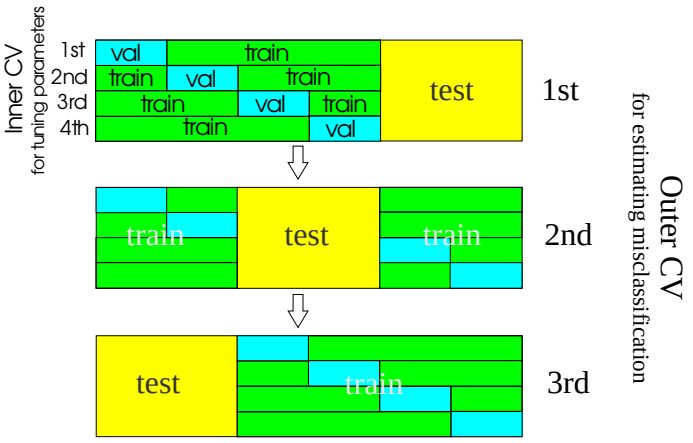

Figure 1.6: Nested cross validation scheme with 3-fold outer cross validation for estimating the misclassification error and 4-fold inner cross validation for tuning the parameters of the classifier. In each outer cross validation round the training set of this fold is further cross validated in an inner round to unbiasedly estimate the best classification parameters.

Overfitting – Applying some form of predictive performance assessment in gene expression analysis is crucial for diagnostics. West *et al.* (2001) stressed the need for cross-validation in order to provide honest assessment of the outcome of future samples. Ambroise and McLachlan (2002) showed that in the past many microarray dataset used gene selection on all available microarrays and thus ran into the problem of overfitting. They underline that feature selection has to be done for every fold separately in the cross validation. Ntzani and Ioannidis (2003) analyzed clinical gene expression studies and found that only 9 applied any form of cross-validation, and only 2 of these avoided the selection bias.

1.4.3 Gene Selection

One strength of microarrays is that they can measure thousands of genes simultaneously. On the other hand many of the genes are noisy, irrelevant, or redundant. Noise in the

measurements stems from experimental variation but can also reflect sampling effects. As we cannot learn population characteristics from noisy genes they should be filtered out. Irrelevant genes are those that are not related to the disease and e.g., constantly expressed or not regulated disease specific. Redundant genes are those that are highly correlated to other genes. This can reflect the fact that the genes are involved in the same pathway or even directly connected. Gene selection strategies aim to identify a relevant subset of the genes. The set of relevant genes can then directly be used for a diagnostic microarray design.

Gene selection not only serves to identify biologically relevant genes, it also improves the classifier's predictive performance. Training the classifier with many more genes than samples makes it hard to learn underlying patterns. The classifier tends to use characteristics specific to the training data and cannot generalize well to future samples (Hughes, 1968). In a setting with more variables than objects a linear separation of any bipartition of the objects is always possible when the vectors associated with the objects are linearly independent. This separation, however, does not necessarily reflect population characteristics but potentially uses features that just by chance allow a good separation. When new examples are classified this can lead to misclassifications. Therefore, sparse models, which restrict the model complexity, are needed to provide better generalization. One way to restrict the model complexity is reducing the number of features through feature selection.

Feature selection is the problem of finding a feature subset with the smallest expected generalization error. Here, the classifier only uses the measurements of these features and disregards all others. The number of features in a classifier can be seen as a classifier parameter controlling model complexity. The classifier then only uses the projection of $\mathbf{x}$ onto the given features and omits all other data.

Feature selection is composed of two steps (Liu and Motoda, 1998): the generation of a suitable feature subset (feature subset generation) and the evaluation of the subset according to some quality criteria (feature subset evaluation).

Feature subset generation – If there are only few features it is possible to completely enumerate and score all potential feature subsets. However, the number of feature subsets grows exponentially with the increase of dimensionality d. Finding a strictly optimal subset is intractable (Kohavi and John, 1997). It has been shown that using brute force approaches are not tractable even for simple classifiers (Blum and Rivest, 1992; Evgeniou and Poggio, 1997). Therefore, many algorithms have been described to find reasonable, suboptimal feature sets. Usually non-deterministic or heuristic strategies are used. Here, the feature generation can either follow a forward strategy (where beneficial features are added to the subset), backward strategy (where noninformative features are removed from the subset), a combination of forward and backward, or a random selection strategy. This can be done one feature at a time or in chunks of several features. Extensions where the size of the chunks is dynamically adapted are called floating forward/backward selection (Pudil *et al.*, 1994).

A heuristic forward strategy is to rank features according to some quality criteria and select the top features. This can be done by looking at each feature separately (univariate) or evaluating the potential of several features together (multivariate). Univariate methods considering each gene separately potentially miss sets of genes that together allow a good separation between the classes. Multivariate methods measure the relative contribution of a gene to the classification by taking other genes into consideration. In general, multivariate methods select fewer genes (Xu and Setiono, 2003). However, the computational cost of multivariate methods is high and the selection may be sensitive to noise or irrelevant features (Lai *et al.*, 2005, 2006). Guyon and Elisseeff (2003) provides a comprehensive review of feature selection methods. A comparison of feature subset selection algorithm can be found in Kudo *et al.* (2000) and Jain and Zongker (1997). Broberg (2003) reviews feature selection in the context of microarrays.

Feature subset evaluation – The evaluation of the feature subset can be on the basis of accuracy, consistency, information, distance, or dependence. If the evaluation of the feature subset uses the predictive accuracy of the classifier itself the method is called a wrapper (Kohavi and John, 1997). If the feature selection is done separately from the classifier based on other criteria, it is called a filter.

In this thesis we propose an improvement to filter based univariate forward feature selection and therefore review now several methods used for feature ranking.

Feature ranking for classification seeks to rank features according to their ability for class separation. A standard method is to use the unequal variance t-score,

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

of the t-test (Student, 1908). Variants of t-like scores such as Fisher $\frac{\mu_1-\mu_2}{\sigma_1^2+\sigma_2^2}$ (Bishop, 1995) and Golub $\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2}$ (Golub *et al.*, 1999) put different weights on the variance and number of samples. When thousands of genes are screened some of them will have low sample variance just by chance. To counterfeit a selection of low variance genes that might show only marginal differential expression, Tusher *et al.* (2001) suggested a modification of the t-score, called SAM, that inflates the variance estimate. The SAM score is defined as

$$z = \frac{\mu_1 - \mu_2}{\sigma_{12} * (1/n_1 + 1/n_2) + s_0},$$

where

$$\sigma_{12} = \sqrt{((n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2)/(n_1 + n_2 - 2)}$$

is the pooled variance estimate and s_0 is a small positive constant that minimizes the coefficient of variation. Also approaches from entropy and information theory, like the information gain (Xing *et al.*, 2001), were used to identify differential genes. However, in t-like scores outliers or scale transformations (e.g. log) strongly influence results. Rank based scores that do not use the numerical values directly but just the rank are more robust. Rank based scores include TNoM (Threshold number of misclassifications, Ben-Dor *et al.* (2000)) and Wilcoxon rank-sum (Wilcoxon, 1945). Both first sort the numerical

gene expression vector. Then, they only look at the class labels of the sorted vector. They assess for every position in the vector how well a split at this position would separate the two class labels. The TNoM score reports the minimal number of misclassification done over all possible split positions. The Wilcoxon score sums up the distance from the misclassified labels to the split boundary and reports the minimal sum of distances. The Wilcoxon score has been successfully used as a scoring function for genes (Park *et al.*, 2001). Pepe *et al.* (2003) proposed a feature selection based on receiver operator curves (ROC) closely related to the Wilcoxon score, which takes sensitivity and specificity into account. A more detailed review of variable selection methods can be found in Dudoit *et al.* (2002), Guyon and Elisseeff (2003) and Liu and Motoda (1998).

In chapter 3 we introduce a novel method for improving gene selection by grouping genes using unsupervised methods an review these methods in the following section.

1.4.4 Unsupervised methods

In unsupervised methods, often also called class discovery methods, the classes are a priori unknown and the task is to automatically group the data. Typically, clustering methods are used to partition the dataset into several groups or clusters, where similar objects are assigned to the same cluster whereas dissimilar objects are assigned to different clusters. One distinguishes hierarchical clustering methods (Eisen *et al.*, 1998) and partitioning clustering methods (Hartigan and Wong, 1979). Hierarchical methods iteratively examine one object at a time and join it to the closest cluster, such producing a tree like structuring of the data. Partitioning clustering methods partition the objects into a prespecified number of clusters, so that the within-cluster distance is minimized and the without-cluster distance is maximized.

Fuzzy clustering (Dunn, 1973; Bezdek, 1973) is a partitioning clustering method. It deals with the problem that there is often no sharp boundary between clusters in real applications. Instead of assigning an object to a specific cluster there is a membership probability for each cluster. The fuzziness of the membership of an object to a cluster can be tuned with a softness parameter. In the extreme case where the softness parameter is close to 1 a hard assignment to the clusters is achieved (i.e., each item belongs exactly to one cluster with probability 1). Viewing

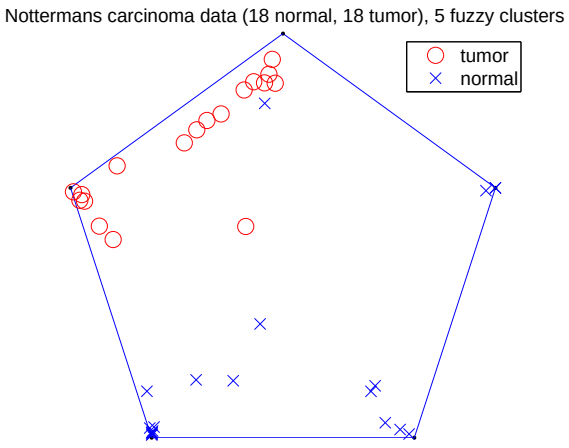

Figure 1.7: Notterman carcinoma data was fuzzy clustered and is visualized on a regular pentagon. Each corner represents a cluster center and the probability of an object's membership to each cluster defines the distance from each corner.

it that way fuzzy clustering is a generalization of k-means clustering. Fuzzy clustering partitions a finite collection of elements $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ into c fuzzy clusters V, with $V = (\mathbf{v}_i), i = \{1, .., c\}$ being the cluster centers, $\mathbf{v}_i, \mathbf{x}_i \in \mathbb{R}^p$. A partition matrix $U = (u_{ij}), i = \{1, .., c\}, j = \{1, .., n\}$ specifies the probability to which the element $\mathbf{x}_j$ belongs to the i-th cluster, $u_{ij} \in [0, 1]$. The fuzzy c-means algorithm minimizes the objective function

$$J(\mathbf{X}, \mathbf{w}; \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m w_j d^2(\mathbf{v}_i, \mathbf{x}_j).$$

d is the distance function between a data vector and a cluster center, w_j is the weight of $\mathbf{x}_j$, and $m \in [1, \infty]$ is the fuzzy softness parameter. After specifying the number of clusters c, the weights $\mathbf{w}$, and the softness parameter m, the following steps are processed during fuzzy clustering: the positions of the cluster centers are randomly set or initialized by a previous hierarchical clustering. Then, the partition matrix U is calculated by minimizing the objective function with the given cluster centers V. With the new U the cluster center positions V are recalculated. These steps are repeated until convergence.

In gene expression analysis, often, the task is to group similar samples together. The Notterman Adenoma (Notterman *et al.*, 2001) dataset contains mRNA expression of approximately 6600 cDNAs and ESTs. These were measured in 4 colon adenomas and 4 paired normal colon samples. When using the FCMeans Clustering MATLAB Toolbox V2-0 for clustering the gene expression vectors of all samples in this dataset, tumor samples can be separated from normal tissue samples (Fig. 1.7).

We now propose a framework for deriving a biomarker panel from clinical microarray studies.