Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Science

Presented by

Diplom-Ingenieur (FH) Peter Bewerunge

born in Nastätten

Oral examination: 24.07.2009

Integrative Data Mining and Meta Analysis of Disease-Specific Large-Scale Genomic, Transcriptomic and Proteomic Data

Referees: Prof. Dr. Roland Eils

Prof. Dr. Peter Lichter

Abstract

During the past decades, large-scale microarray technologies have been applied to the field of genomics, transcriptomics and proteomics. DNA microarrays and mass spectrometry have been used as tools for identifying changes in gene- and protein expression and genomic alterations that can be linked to various stages of tumor development. Although these technologies have generated a deluge of data, bioinformatic algorithms still need to be improved to advance the understanding of many biological fundamental questions. In particular, most bioinformatic strategies are optimized for one of these technologies and only allow for an one dimensional view on the biological question. Within this thesis a bioinformatic tool was developed that combines the multidimensional information that can be obtained when analysing genomic, transcriptomic and proteomic data in an integrative manner.

Neuroblastoma is a malignant pediatric tumor of the nervous system. The tumor is characterized by aberration patterns that correlate with patient outcome. aCGH (array comparative genomic hybridization) and DNA-microrarray gene expression analysis were choosen as appropriate methods to analyse the impact of DNA copy number variations on gene expression in 81 neuroblastoma samples. Within this thesis a novel bioinformatic strategy was used which identifies chromosomal aberrations that influence the expression of genes located at the same (*cis*-effects) and also at different (*trans*-effects) chromosomal positions in neuroblastoma. Sample specific cis-effects were identified for the paired data by a probe-matching procedure, gene expression discretization and a correlation score in combination with one-dimensional hierarchical clustering. The graphical representation revealed that tumors with an amplification of the oncogene MYCN had a gain of chromosome 17 whereas genes in cis-position were downregulated. Simultaneously, a loss of chromosome 1 and a downregulation of the corresponding genes hint towards a cross-relationship between chromosome 17 and 1. A Bayesian network (BN) as representation of joint probability distributions was adopted to detect neuroblastoma specific *cis*- and *trans*-effects. The strength of association between aCGH and gene expression data was represented by markov blankets, which where build up by mutual information. This gave rise to a graphical network that linked DNA copy number changes with genes and also gene-gene interactions. This method found chromosomal aberrations on 11q and 17q to have a major impact on neuroblastoma. A prominent trans-effect was identified by a gain of 17q.23.2 and an upregulation of CPT1B which is located at 22.q13.33.

Further, to identify the effects of gene expression changes on the protein expression the bioinformatic tool was expanded to enable an integration of mass spectrometry and DNA-microrarray data of a set of 53 patients after lung transplantation. The tool was applied for early diagnosis of the Bronchiolitis Obliterans Syndrome (BOS) which occurs often in the second year after lung transplantation and leads to a repulsion of the lung transplant. Gene expression profiles were translated into virtual spectra and linked to their potential mass spectrometry peak. The correlation score between the virtual and real spectra did not exhibit significant patterns in relation to BOS. However, the meta-analysis approach resulted in 15 genes that could not be found in the seperate analysis of the two data types such as INSL4, CCL26 and FXYD3. These genes constitute potential biomarkers for the detection of BOS.

Zusammenfassung

In den letzten Jahrzenten wurden unterschiedliche Mikroarray-Systeme entwickelt und in den Bereichen Genomik, Transkriptomik und Proteomik eingesetzt. Dabei finden sie ihren Einsatz, um Veränderungen der Gen- sowie Proteinexpression und des genomischen Materials insbesondere mit unterschiedlichen Phasen der Tumorentstehung zu verknüpfen. Die große Menge an Daten die dabei anfällt, müssen mittels bioinformatischer Algorithmen ausgewertet werden. Allerdings liegt bei derzeitigen Verfahren die Optimierung und Fokussierung auf eine Mikroarray-System im Vordergrund, was zu einer eindimensionale Betrachtung der biologischen Fragestellung führt. Deshalb war Ziel dieser Arbeit, einen bioinformatischen Algorithmus zu entwickeln, der mehrdimensionale Informationen kombiniert, die sich aus einer integrativen Betrachtungsweise von genomischen, transkriptomischen und proteomischen Daten ergibt.

Das Neuroblastom ist ein maligner frühkindlicher Tumor des Nervensystems. Charakteristisch sind die Muster der chromosomalen Veränderungen, die mit der Entstehung und/oder Progression des Tumors korrelieren. aCGH (array Comparative Genomic Hybridization) und DNA-Mikroarray Genexpressionsanalysen wurden ausgewählt, um den Einfluss chromosomaler Veränderungen auf die Genexpression von 81 Neuroblastom-Patienten zu untersuchen. Im Rahmen dieser Arbeit wurde eine neue bioinformatische Strategie entwickelt, die chromosomale Veränderungen identifiziert, die die Expression von Genen sowohl an der gleichen (*cis*-Effekt) aber auch an anderen chromosomalen Positionen beeinflusst. Tumorspezifische *cis*-Effekte wurden unter anderem durch eine Korrelationsanalyse in Kombination mit einem eindimensionalen, hierarchischen Verfahren zur Gruppenfindung ermittelt. Die graphische Darstellung zeigte, dass Tumore mit einer Amplifikation des Onkogens MYCN durch einen chromosomalen Zugewinn auf Chromosom 17 charakterisiert sind, während Gene in *cis*-Position eine geringe Expression aufwiesen. Gleichzeitig ging der Verlust des Chromosom 1 mit einer niedrigen Expression der *cis*-lokalisierten Gene einher. Um Neuroblastom-spezifische *cis*- und *trans*-Effekte über das gesamte Datenset zu identifizieren, wurden Bayessche Netzwerke eingesetzt. Das Maß des Zusammenhangs zwischen der DNA-Kopienanzahl und der Genexpression wurde mit Hilfe von "Markov Blankets" und "Mutual Information" berechnet. Das graphische Netzwerk zeigte die Verbindungen zwischen chromosomalen Veränderungen und der Genexpression wie auch mit Gen-Gen-Interaktionen. Hieraus resultierte, dass Veränderungen auf Chromosom 11q und 17q als ursächliche Faktoren für das Neuroblastom verstanden werden können. Auffällig war der *trans*-Effekt zwischen dem Zugewinn auf Chromosom 17q23.2 und der hohen Genexpression von CPT1B (22q13.33).

Weiterhin wurde der bioinformatische Algorithmus um die Eigenschaft erweitert, eine integrative Analyse von Genexpressions- und massenspektrometrischen Daten durchzuführen. Dies wurde auf einen Datensatz angewendet, der die Entstehung des Bronchiolitis Obliterans Syndroms (BOS) untersuchte. BOS wird häufig im zweiten Jahr nach einer Lungentransplantation diagnostiziert und führt in den meisten Fällen zu einer Abstoßungsreaktion. Die zugrundeliegenden Genexpressionsdaten wurden in virtuelle Spektren überführt und den entsprechenden massenspektrometrischen Kurvenverläufen zugeordnet. Eine Korrelationsanalyse zwischen den virtuellen und realen Massenspektren konnte keine Korrelation erfasssen. Hingegen konnte ein integrativer Meta-Analyseansatz 15 Gene

identifizieren, die bei einer separaten Betrachtung der Daten nicht gefunden wurden. Auf diese Weise stellen die Gene potentielle Biomarker für die Früherkennung des BOS dar.

Acknowledgements

Special thanks go to Prof. Roland Eils, who gave me the chance to do my PhD in his renowned laboratory at the DKFZ in Heidelberg. Furthermore, he was always open for scientific discussion and, not to be scoffed at, took care of my financial support.

I am grateful to Prof. Peter Lichter for his willingness to be the second supervisor of my PhD thesis, to PD. Dr. Karsten Rippe for being a referee of this thesis and also to PD. Dr. Stefan Wiemann for being one of the examiners for my defence.

Thanks to Dr. Benedikt Brors, for supervising my PhD and introducing me into the exciting field of bioinformatics. In addition, I am grateful for his excellent proof reading of this thesis.

Dr. Marc Zapatka is gratefully acknowledged for his selfless commitment and perfect team play in regard to our shared work in the field of mass spectrometry.

Dr. Nils von Neuhoff, Tonio Oumeraci and Prof. Dr. Brigitte Schlegelberger, Department of Pathology at the Hannover Medical School (MHH) deserve my very gratitude for being excellent cooperation partners, their openness for discussions at all times and assisting me in all issues regarding the pathology of the Bronchiolitis Obliterans Syndrom. I very much enjoyed this successful collaboration.

Many thanks go to Mirjam Maierm, a very eager diploma student who felt herself responsible to discover the underlying mechanism of BOS.

Lars Kaderali and Mike Hallet are acknowledged for introducing me to Bayesian Networks. I thank Yvonne Koch for all the fruitfull discussions, new ideas related to my work and her warm accommodation she granted me during my PhD.

I wish to thank the members of the Division of Theoretical Bioinformatics and especially the "Computational Oncology Group" for the friendly working atmosphere and the unforgettable moments we shared together during my stay at the DKFZ in Heidelberg.

My special gratitude goes to my family, Doris and Dr. Dieter Bewerunge, Luis Yancy, Dagmar Yancy and Danny Yancey and my girl friend Melanie Hudler for their continuous support, encouragement, and gentle love. They always believe in me, stand by me in any difficult situation in my life, and remind me to stay down-to-earth. Exceptionally thanks to Melanie for her love and trust in me.

Contents

1 Introduction **13**

1.1 Neuroblastoma . . . 13

1.2 Bronchiolitis Obliterans Syndrom . . . 14

1.3 High-dimensional omics-approaches . . . 15

1.3.1 Omics-Bioinformatics . . . 15

1.3.2 Array-based comparative genomic hybridisation . . . 16

1.3.3 Array-based monitoring of gene expression . . . 19

1.3.4 Mass spectrometry profiling . . . 20

1.4 Integrative bioinformatic analysis of omics-approaches . . . 26

1.4.1 Correlation of chromosomal aberrations and gene expression . . . 26

1.4.2 Bayesian networks and computational biology . . . 28

1.4.3 Separated and integrative analysis of BOS specific gene and protein expression . . . 31

1.5 Aims . . . 32

2 Material and Methods **33**

2.1 Integration of Neuroblastoma specific copy number changes and gene expression by BN . . . 33

2.1.1 Data . . . 33

2.1.2 BNtegrative. A comprehensive toolbox to screen genomic *cis*- and *trans*- effects . . . 34

2.2 Integration of BOS specific gene and protein expression . . . 43

2.2.1 Data . . . 43

2.2.2 Transcriptome analysis . . . 46

2.2.3 Proteome analysis . . . 48

2.2.4 Integrative analysis of gene and protein expression . . . 54

3 Results **57**

3.1 Impact of DNA copy number changes on gene expression in neuroblastoma 57

3.1.1 Distribution of gene expression data after discretization . . . 57

3.1.2 Chromosome aberrations in neuroblastoma . . . 58

3.1.3 Patient related cis-effects . . . 59

3.1.4 Identification of genomewide cis- and trans-effects via Bayesian Modeling . . . 62

3.2 Meta-analysis of genes and proteins identifies potential biomarker for BOS 64

3.2.1 Differently expressed genes and functional domains . . . 64

3.2.2 Detection of significant proteomic patterns . . . 66

3.2.3 Integrative analysis results in novel peaks . . . 74

4 Discussion **80**
4.1 Integrative analysis of genomic and transcriptomic data related to neuroblastoma 80
4.2 Integrative analysis of transcriptomic and proteomic data related to BOS 83

Bibliography **85**

List of Figures **101**

List of Tables **104**

List of Algorithms **105**

Appendix **106**

Chapter 1

Introduction

1.1 Neuroblastoma

Neuroblastoma is a malignant tumor of the sympathetic nervous system in young children . It arises in immature nerve cells and affects mostly infants and children. Often neuroblastoma begins in the nerve tissue of the adrenal glands. The adrenal glands produce hormones that help control heart rate, blood pressure, blood sugar, and the way the body reacts to stress. Neuroblastoma may also begin in the chest, in nerve tissue near the spine in the neck, or in the spinal cord. It sometimes forms before birth but is usually found later, when the tumor begins to grow and cause symptoms. When neuroblastoma is diagnosed, the cancer has usually metastasized, most often to the lymph nodes, bones, bone marrow, liver, and skin.

Neuroblastoma is characterized by diverse clinical courses. This ranges from complete regression of the disease to rapid tumour progression and death [30]. Important factors in determining outcome are the patient age and stage of the disease. The majority of children over 1.5 years of age have metastatic disease at the time of diagnosis, which comes along with a poor prognosis despite intensive therapy. The mechanisms leading to this diverse clinical behavior of neuroblastomas remain largely unclear.

Although the overall survival of current high-risk patients has improved in the last decades [20], there is a need to detect novel markers to identify those high-risk patients with a more favorable biology. For these purposes, additional prognostic indicators have been proposed in recent years. analyzes of DNA copy number alterations resulted in the delineation of three major genetic subgroups with predictive tumour behaviour (subtype 1, 2A and 2B). Subtype 2A Neuroblastoma represents an aggressive subgroup characterised by loss of loss of 1p, 3p [151] and 11q [10], gain of 17q, which independently predicts poor prognosis [26], and *MYCN* amplification [30]. In contrast to MYCN gene amplification, the degree of expression of the MYCN gene in the tumor does not predict prognosis. Additionally extensive microsatellite heterozygosity mapping studies point at various critical regions of loss, located at 11q23.3 [71] and within the chromosomal region 11q14-11q23 [110]. Spontaneous regression of neuroblastoma is a phenomenon that has been well described in infants, especially in those with the 4S pattern of metastatic spread [119]. Regression generally occurs only in tumors with a near triploid number of chromosomes, no MYCN amplification, and no loss of chromosome 1p.

Apart from copy number alterations, expression levels of an growing number of single

candidate genes, e.g. *NTRK1* [117], *FYN* [21], *PRAME* [121] and *PHOX2B* [116, 112] were reported to be indicative of neuroblastoma tumor behavior.

1.2 Bronchiolitis Obliterans Syndrom

The Bronchiolitis Obliterans Syndrom is the most frequent clinical manifestation of chronic repulsion reaction and destruction of lung transplants which occurs frequently in the second year after lung transplantation. The diagnosis of bronchiolitis obliterans is important, as appropriate immunosuppressive treatment may be helpful in the preservation of lung function [36]. The term "obliterans" refers to inflammation of the bronchioles, which partially destroys (obliterates) the small airways.

The auto-immune reaction behaves in such a way that the small respiratory system - the bronchioles - thickens due to a chronic inflammatory process. This leads to fibrosis and cellular deposition in airways, complicating long-term survival [172]. A malfunction of the lung, which can be mild or severe depending on the degree of BOS, often follows [157].

One severe consequence of this repulsion reaction is a remarkably short survival time which is often shorter than after other transplantations [23]. After the first postoperative year, BOS is the main cause of death with a prevalence of 39%. Previous clinical experiences identified that 5 years after lung transplantation half of the patients, and after 8 years even 2/3 of them are affected [57].

Until now, no effective therapy is available for BOS, however, certain immunosuppressive regimens may slow down the progression of the disease [7]. Besides that, no diagnostic markers exists for the detection of this chronical disease.

Despite continuous improvements in surgical methods and other therapy options, the causes of BOS are still complex and so far unsolved. It is experimentally proven that within a few hours after transplantation nearly one third of the lung tissue cells dies via apoptosis [61]. Early detection of BOS is essential because prompt initiation of treatment may halt the progression of the disease and the development of chronic transplant failure [5].

1.3 High-dimensional omics-approaches

During the last 60 years expert knowledge about molecular elements of life has grown like never before in human history. In 1953, James Watson and Francis Crick published their model of the three-dimensional structure and the chemical components of deoxyribose nucleic acid (DNA) [173]. They were pioneers in describing the DNA as a double helix with base pairs as their backbone. Crick postulated in 1970 the "Central Dogma of Molecular Biology" which reads [53].

> *DNA makes RNA, RNA makes protein, and proteins do almost all of the work of biology [66].*

With the structure of DNA and this dogma in hand, researchers started to answer the question of the impact and mechanisms of genes. It became clear that genes do not work in isolation but rather interact with each other. With this demand on a more concise picture of genes and the cell in general the Human Genome Project (HGP) was founded [46, 169]. Initiated in 1990, it took 13 years till in April 2003 the gene-containing part of the human sequence was completely deciphered. So far, about 750 genomes from different organisms have been sequenced, and the sequencing of about 2750 genomes is in progress [80, 81]. Among other facts we learned from the HGP that a great part of the genome does not correspond to any expressed gene.

> *We are probably at the end of the beginning rather than at the beginning of the end because genomics will probably change biology to a greater extent than previously forecasted [79].*

Since the sequencing of many genomes is finished, an increasing number of high-throughput methods have been developed. In this subsection three of this well established molecular biological methods which provide the basis for the data in this thesis will be explained. In addition for each biological approach, a method related bioinformatic background will be given.

1.3.1 Omics-Bioinformatics

There exist several definitions of bioinformatics [19, 127, 69]. One obvious way to look at it, is to take it as a merge of two sciences, namely biology and informatics, into one discipline [11, 64, 68, 96, 97, 113]. The increasing demands on bioinformatics started in parallel with the HGP in the early 1980s, when methods for DNA sequencing became widely available. Data were concentrated in large databases such as GenBank, EMBL or SWISS-Prot and opened up the way for new methods adopted in data retrieval and analysis, structural and functional prediction [18, 22, 131, 154].

> *"Every institution that expects to be competitive in this new era will need to have strengths in high-throughput genomic analyzes and computational approaches to biology," (Francis Collins, director of the National Human Genome Research Institute, Bethesda, Maryland U.S. [32])*

The availability of different data types of high-throughput experimental data in the late 1990s, like DNA-microarrays, aCGH and matrix-based mass spectrometry, has expanded the role of bioinformatics. Having solved the challenges of data storing, establishing of preprocessing steps, like signal detection or normalization, bioinformatics was then expanded in depth. New tools including statistical tests, principal component analysis or cluster analysis to reduce data to a lower dimensionality emerged from this research.

An area called "extragenomics" throws new light on pathways, networks and interactions which affect genes and proteins. The Gene Ontology Consortium has become an important part in understanding of those cellular processes by defining a common vocabulary for protein function. Also pathway databases, for example KEGG, try to define cellular processes and inspire bioinformatics to build up a complete representation of the cell and the organism.

Integrative Bioinformatics today of a single high-throughput method can not fully unravel the complexities of fundamental biology. It takes more than the traditional one dimensional, vertical consideration on the biological dogma, that DNA makes RNA and RNA makes proteins. We need to integrate the knowledge on genomics, transcriptomics, proteomics and even extragenomics at the same time to get a deeper insight into complex human diseases, like cancer.

To do so, researchers started integrative studies where they included the different levels of cellular information flow. Combined analysis of two popular platforms, DNA microarrays and gel-free proteomics, aims to answer the question to which extent the pattern of gene expression correlates with the corresponding protein levels. The general consensus is currently that the correlation between transcriptomes and proteomes across large datasets is typically modest [40, 51, 72]. Measurement errors and poorly conceived instruments have been considered to contribute, at least in part to this poor correlation between mRNA concentration and protein abundance [51].

Integrative analysis of genomic and transcriptomic data provides additional information on wether changes in the DNA content have functional consequences on the activation or inactivation of genes that play key roles in multiple biological networks. Most studies considered a one dimensional examination of *cis*-effects and try to answer the question of what happens to the gene expression, when the chromosome it is located on, is mutated. More promising are studies where people analyze distant interactions of chromosomal aberrations which impact genes located elsewhere. This is called a *trans*-effect.

1.3.2 Array-based comparative genomic hybridisation

Each gene is localized to a specific site along the length of a specific chromosome. This is often termed a genetic locus. Normal chromosomes of a cell should have two copies of each genomic region, except for the sex chromosomes. The normal configuration of a chromosome is called euploid, whereas aneuploidy describes a change in the number of chromosomes. A missing chromosome from a diploid organism is called monosomy, and an addtional chromosome is called trisomy (e.g. trisomy 21).

DNA copy number aberrations (CNA) frequently occur during tumor progression and are demeed as the driving force of tumorigenesis and of the progression of cancer [104, 103]. Specifc DNA regions of the tumor DNA are lost or gained. For example, when a genomic region of a diploid tumor cell is affected by a loss of DNA we would expect to get 0 or 1 copy, in the simplest case. However, in the case of a gain this will result in 3 or more copies. All genomic aberrations of a sample can be characterised as a genomic profile (Fig. 1.1). Methods like comparative genomic hybridization (CGH), and also the array-based version, aCGH, reveal which regions, and to what extend DNA regions have been gained or lost.

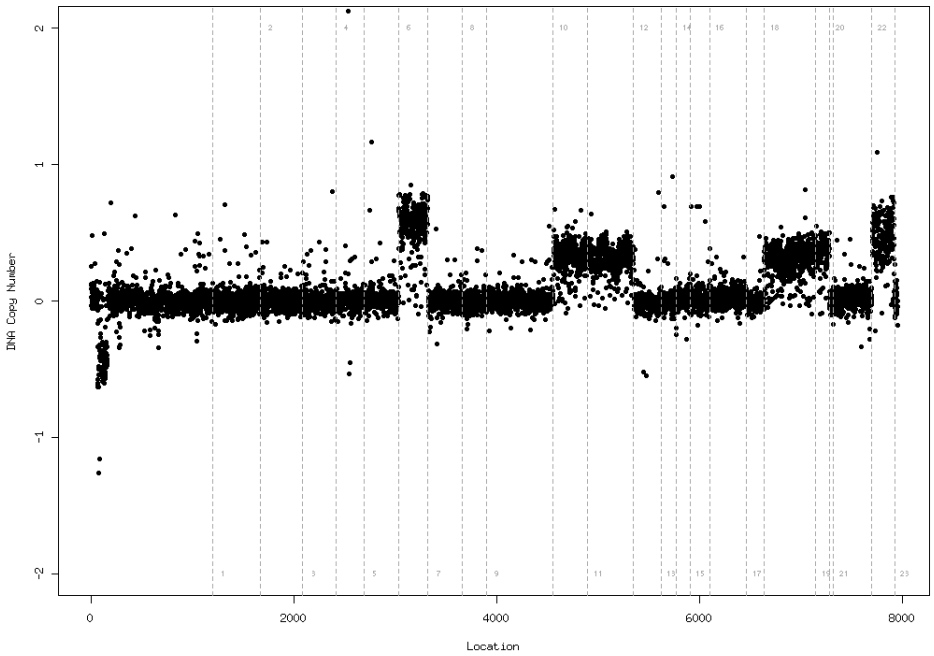


Figure 1.1: Example plot of a genomic profile. The y-axis depicts the copy number ratio of all measured chromosomal regions. The chromosomal positions are displayed at the x-axis. Gray vertical lines define the different chromosomes numbered in the same color.

Comparative genomic hybridization (CGH) to metaphase spreads was the first efficient method for the detection of relatively large chromosomal regions ($\sim$ 10 Mb) that are lost or gained in a tumor [95, 101]. DNA preparations from two samples, e.g. a tumor sample and a control sample or different tissue from a single individual, are labeled with different fluorophores, either a red-fluorescent dye (Cy5) or a green-fluorescent dye (Cy3) (Fig. 1.2). Based on changes in signal ratios, gains and losses can be detected. However, CGH has some main limitations, especially with regard to the resolution. Changes in regions smaller than 5-10 Mb are not realiably detectable [62].

Array-based CGH greatly improves the resolution of classical CGH. Solinas-Toldo *et al.* (1997) utilized a microarray-based technology to detect chromomosomal imbalances

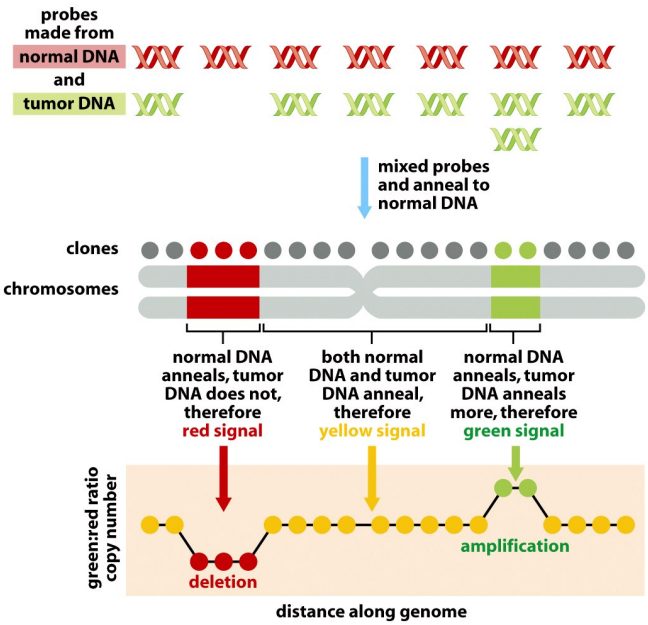


Figure 1.2: Comparative genomic hybridisation. Fragments of normal (red) and tumor (green) DNA are differentially labeled with two different fluorophores. These fragments hybridise to metaphase chromosomes. A red signal indicates that only normal DNA is annealed, but no tumor DNA was present. This is when a loss of DNA in the tumor DNA has occured. In the case of a yellow signal, both normal and tumor DNA are bound in the same amount, i.e. the tumor DNA shows no chromosomal aberration. A gain of tumor DNA is indicated by a a green signal, which denotes that more tumor DNA is annealed than normal DNA. Figure taken from [174].

and improved the detection of altered chromosomal segments to 75-130 kb in size [147]. In a pioneer study, Pollack et al. (1999) presented the first genome-wide array [129]. They used 3195 unique cDNA target clones which where distributed consistently across the genome. The big advantage of cDNA approaches is the potential to analyze changes in DNA copy number and gene expression levels in parallel [130]. Limitations involve the exclusive detection of aberrations in known genes which results in an irregular distribution of measured loci across the genome.

However, the majority of aCGH data today has been generated by the use of Bacterial Artificial Chromosome (BAC) CGH arrays. In 2001, Snijders *et al.* used a microarray with 2400 BACs across the genome. The BACs varied in length from 150 to 200 kb, and the array size varies from 2.400 to $\approx$ 30.000 unique array elements which makes this method outstandingly sensitive and precise [146].

Oligonucleotides are also used in genome-wide screening for genomic imbalances. Different commercial platforms, e.g. from Affymetrix, Agilent Technologies or NimbleGen, contain short oligonucleotides ranging from 25-70mers [15, 28, 35, 189]. These methods claim that the processing is rapid, cost-effective, and easy to handle.

The terms mCGH and aCGH (array-based comparative genomic hybridisation) will be used as synonyms in this thesis.

Several bioinformatic algorithms have been proposed to find aberrations in a genomic profile (Fig. 1.1 on page 17). These methods assign the copy number ratio to all positions in a region of a profile. A common method, implemented in a R[1] package called *GLAD* (Gain and Loss Analysis of DNA), is based on adaptive weight smoothing [86]. It estimates the breakpoints of a piecewise constant function and also assigns losses and gains to each region by a clustering algorithm. Another approach, implemented in the R package *DNAcopy*, recursively split is whole segments of a profile into smaller regions at the breakpoints, and assigns aberrations to each individual segment [123]. The R package *aCGH* is based on a hidden Markov model, where each state in a genomic profile represents a region with similar copy number ratios [92].

1.3.3 Array-based monitoring of gene expression

Genes and DNA-Microarray Driven by the awareness that sequence information alone is not sufficient for a full understanding of gene function, expression and regulation, Schena et al. (1995) presented a spottet cDNA based gene expression array. One year later in 1996 David Lockhart introduced the expession monitoring by hybridization to high-density oligonucleotided arrays [102]. Thus DNA microarrays come into play for the monitoring of large numbers of mRNAs in parallel.

Like aCGH measurements, DNA-microarrays are a powerful tool for the simultaneous analysis of expression of thousands genes on a genome-wide scale. The set of transcripts that are expressed or transcribed from genomic DNA in the cell at the same time, is called the ´expression profile‘ or the transcriptome. It is also called expression signature and can be understood as a "barcode" for a specific phenotype. Differences in the expression profile of a cell are responsible for phenotypic differences as well as indicative for cellular response to an environmental stimulus.

Two methods of microarray-based gene expression monitoring are mainly in use. These are two-color cDNA microarrays and one-color oligonucleotide arrays [139] [102].

cDNA-microarrays are typically custom-printed by spotted, PCR- amplified cDNA clones. These clones are of size of approximately 0.6-2.4 kb and are mostly bound to glass microscope slides, or on porous membranes like nylon [138]. In most experiments, expressed sequence tags (EST) represent the most reliable source of sequences for gene identification [191]. Another characteristic of cDNA microarrays is the use of two different fluorophores. DNA from two samples, e.g. tumor and control, or different tissues from a single individual, are labeled with different fluorophores, either a red-fluorescent dye (Cy5) or a green-fluorescent dye (Cy3) and hybridized together on a single microarray [55]. These two samples on a single microarray allow the direct comparison by determining the relative abundance by a ratio of fluorescence intensities [77, 183]. This minimizes the variability from processing multiple microarrays per assay. A disadvantage

[1] www.r-project.org

lies in the dye-specific biases which can lead to misinterpretation of the results, but can be controlled by performing dye-swap replicates.

Oligonucleotide-microarrays are performed similarly to cDNA- microarrays, except the spotted probes and the used amount of fluorophores. Short probes, 25 nucleotides or longer in length, selected on the basis of their sequence specificity, are synthesized in situ by photolithography or inkjet technology on a solid surface [102]. The signal for each gene-specific mRNA is determined by hybridization to a group of up to 20 pairs of oligonucleotides. Unlike cDNA-arrays, single samples are hybridized to each microarray after they have been labeld with a single fluorophore. Rather than a ratio, an absolute value of fluorescence intensity is determined. This value is compared with other experiments to detect transcriptomic changes. A key issue, and a problem of oligonucleotide-based arrays, is how to select probe sequences with high sensitivity and specificity. On the other hand, these arrays are commercially available, have a high density and are well standardized [144].

Bioinformatic strategies for both cDNA- and oligonucleotide- microarrays require several pre-processing steps including image analysis, background adjustment and normalization [179]. Controlling the effects of systematic error while taking care of the biological variation are platform-specific and difficult to automate [166, 180]. *Image analysis* is the basis for data analysis, by converting the pixel intensities in the scanned image into intensity values per probe [135, 90]. Parts of the measured probe-level intensities do not come from gene expression, but rather from non-specific bindings and noise in the optical detection system and need to be assesed by *background adjustment*. The most critical step of a pre-processing analysis is a platform-adapted *normalization* method in order to remove any non-biological variation [166, 180, 25, 145]. Starting from here, biological questions can be adressed by bioinformatic strategies like SAM (significance analysis of microarrays) [167], PAM (predictive analysis of microarrays) [164] or GSEA (gene set enrichment analysis [156]) .

1.3.4 Mass spectrometry profiling

"Is Proteomics the New Genomics?" Jürgen Cox and Matthias Mann raised this question in 2007 [52]. They looked back to a period, starting in the mid of 1970s, where two-dimensional gel electrophoresis proteomics coupled with high-throughput tandem mass spectrometry (MS) revolutionized proteomics [122]. These have become the most popular and versatile methods to seperate and identify complex mixtures of peptides and proteins [3, 182, 34, 162].

> *"Proteins are central to our understanding of cellular function and disease processes, and without a concerted effort in proteomics, the fruit is of genomics will go unrealized." (Ian Humphery- Smith, University of Utrecht, one of HUPO's founder members)*

Especially the advances of mass spectrometry made biological molecules readable and John B. Fenn, Koichi Tanaka and Kurt Wüthrich have been awarded the nobel prize

in 2002 for their contribution in this area of research [178, 160, 59]. The Human Proteome Organisation (HUPO) was founded in 2001, an international proteomic initiatives to better understand human diseases. Now that the human genome sequence has been published, the HUPO turned their attention to identify the functions and expression patterns of proteins encoded by the genes. It could be argued that measuring the proteome already addresses the desired end point, which is the protein level of a gene of interest.

When we speak of the proteome, we mean the set of all proteins in a tissue of a living organism in a cell or cell compartment at a specific time point under exactly defined conditions [168]. It reflects the biochemical activity of a cell. Conceptually, this is similar to the transcriptomics technologies discussed in Chapter 1.3.3 on page 19.

Due to the more diverse chemical properties of proteins as compared to RNA, the field has a different and diverse set of methods. The analysis of the proteome delivers additional information which would not have been gained by studying the transcriptome alone, because a single gene can have one or more splice variants. Genes are of great complexity, and one gene can produce one or more different proteins with different functions, e.g. by addition of chemical groups (e.g methylation, phosphorylation) [141, 29]. More than 200 different types of post-transcriptional modifications are known, and it is predicted that on average three different modified proteins with different functions are produced from each human gene [74, 14, 54].

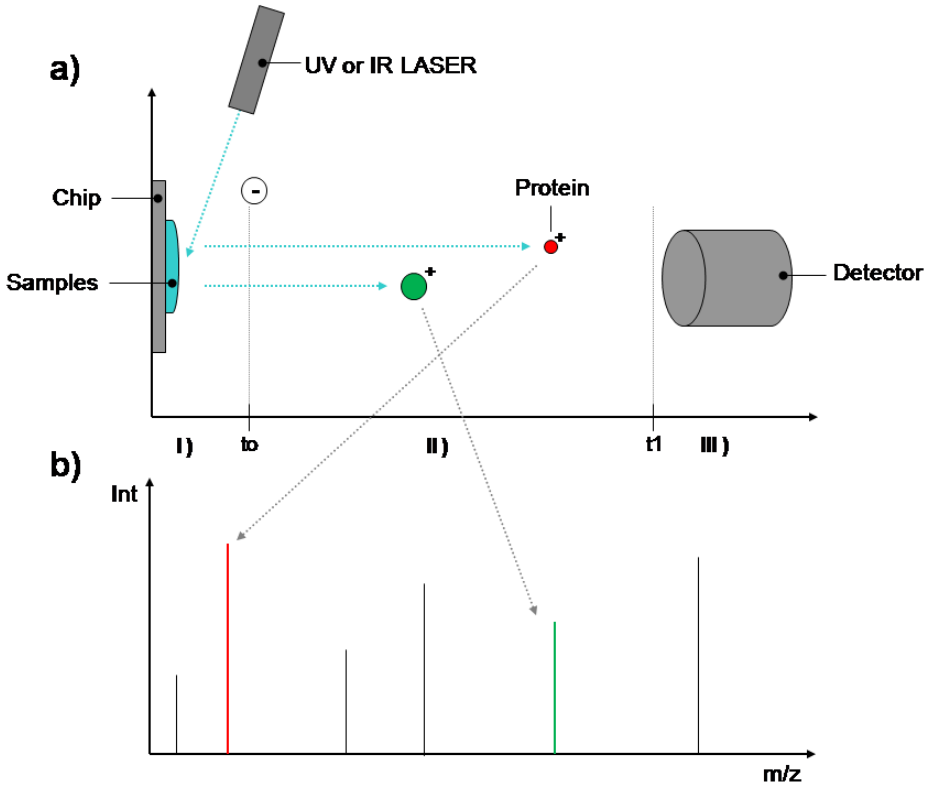


Figure 1.3: Desorption/ionization time-of-flight mass spectrometry. **a)** General setup of a mass spectrometer for MALDI and SELDI. *I) Ionization and Acceleration.* In both MALDI and SELDI, a biological sample of interest is applied to a surface. It is incubated and subsequently co-crystallized with matrix material. A laser is then fired at the co-crystallised mixture and initiates ionization and evaporation of proteins,which are then accelerated by an electric field. The energy of the laser beam is transferred via the matrix to the analyte sample and causes ionization. *II) Drifting.* An electrical field causes the ionized material to fly through the TOF tube (going from t_o to t_1). Lower mass peptides (red ball) fly faster through the tube than higher mass peptides (green ball). *III) Detector.* The peptides with a lower mass arrive earlier than the high mass peptides at the detector which is placed at the end of the flight tube. **b)** Schematic image of a mass spectrum. Using a quadratic equation, the mass-to-charge ratio (m/z) of a peptide can be calculated and plotted as a so called mass spectrum, with the intensity on the y- and the m/z-ratio on the x-axis. The peak height correlates to the protein concentration.

Proteomic Profiling is a new aspect of mass spectrometry which is used to analyze complex protein mixtures from tissue or body fluids, like blood. Typically, biological samples from different patients or different conditions are compared. A major goal, is to find a set of differentially expressed proteins. Proteomic profiling is often employed to identify biomarkers that can be used for diagnosis, prognosis or treatment. MALDI and SELDI coupled to Time-Of-Flight (TOF) discriminators are popular techniques widely used for proteome screening.

Also, matrix-based laser desorption/ionization (MALDI) and surface enhanced laser desorption/ionization (SELDI) have extended the application of mass spectrometry for the quantification of complex protein mixtures from e.g. body fluids like blood, sera or even from whole cells [60, 148, 67].

MALDI-TOF-MS stands for matrix-based laser desorption/ionization time-of-flight mass spectrometry. It is one of the best established ionization methods for mass spectrometric analysis, especially for the investigation of large molecules like proteins [107]. Thus, MALDI-MS has gained a crucial importance for protein analysis [160]. A chemical matrix, consisting of small organic molecules, plays a key role in the mass spectrometry technique by absorbing the laser light energy and causing a small part of the target substrate to vaporize in ionized form (Fig. **??** on page ??) [49]. The analysis by MALDI-MS can be divided into serveral steps. The first step involves the enrichment of proteins by magnetic beads with functionalized surface. A washing step removes unbound proteins followed by a elution of bound proteins from the beads. Afterwards the protein solution is de-salted and the proteins are co-crystallized with a matrix on a metal surface, the so called "target". The last step of the MALDI process involves desorption of bulk portions of the solid sample by a short pulse of laser light. Matrix molecules as well as probe molecules are unleashed in this process and accelerated through an electrostatic field towards the mass analysator [9].

The mass analysator used in MALDI is a time-of-flight (TOF) analysator which enables to exactly determine the masses in high vacuum (Fig. 1.3 on the preceding page). The ions formed within the short laser impulse are accelerated in the source by the electrostatic field and traverse after leaving the source a field-free drift distance in which they are isolated depending on their m/z-ratio (mass over charge) [105]. The abbreviation m/z-ratio is used to denote the quantity formed by dividing the mass m of an ion by it is charge number z. Smaller molecules with lower weight fly faster than large and heavy ones. With known acceleration, voltage, and flight route of the ions in the field-free drift distance, the m/z-ratio can be determined by measuring the flying time. The calibration is made by reference substances with well-known masses [177].

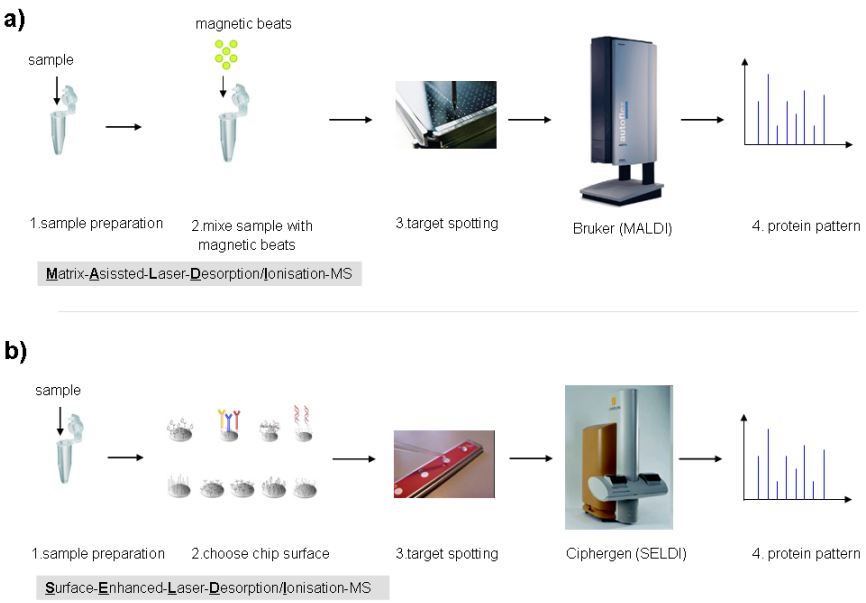


Figure 1.4: Workflow of proteomic profiling. a) First steps of a MALDI-TOF procedure include sample preperation of e.g. body fluids like serum. The sample is mixed with magnetic beads which catch only specific peptides. This target mixture is then spotted to a chip and is processed with an appropriate mass spectrometer.The resulting protein pattern displays the separated peptides in terms of their m/z-ratio. b) The SELDI-TOF workflow also includes sample preparation, target spotting and results in a protein pattern, but differs in utilizing a chip with a chromatographic surface instead of using magnetic beads.

SELDI-TOF-MS is surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. The underlying principles of mass spectrometry are closly related to the MALDI technique [12, 148, 174]. This technology essentially reverses the conventional MALDI sample preparation with magnetic beads as matrix by using a ProteinChip® array of addressable protein binding sites on a solid substrate which are generally chemical or biochemical affinity ligands (Fig. 1.4). Popular ProteinChip® ligands, also called surfaces, are reversed-phase, cation exchange, anion exchange and IMAC (immobilised metal affinity chromatography). Finally, a substance that absorbs laser energy, the SELDI equivalent of the MALDI matrix, is added to the chip array, and the chip array is subjected to "on-chip" laser desorption mass analysis to provide a molecular weight-based protein profile [93, 133, 177].

The main difference between MALDI and SELDI is, that SELDI normally uses a chip with a chromatographic surface, making the purification of the sample implicit. For MALDI, the purification needs to be done before application to the chip, by means of magnetic beads (Fig. 1.4).

Bioinformatic analysis of mass spectrometry data mostly strives for the goal of identifying a small number of features as meaningful diagnostic biomarkers for early diagnosis, prognosis, monitoring disease progression or response to treatment [124, 1, 175, 137]. A typical spectrum arising from SELDI or MALDI contains thousands of intensity measurements at a specific m/z-ratio which represent an unknown number of proteins. Algorithms for biomarker prediction start with the *raw data,* in most cases a list containing the measured intensities at a specific mass-value, the m/z-ratio [13]. *Baseline correction* avoides the displacement of the baseline function, which is a systematic error, often seen in mass spectrometry [82]. It is believed to be a part of the matrix molecules hitting the detector in the early part of the experiment, or to detector overload [108]. Not only the data resulting from such measurements are noisy but the variance between replicates of the same samples is also high. Adequate *normalization* methods, like the total ion count (TIC), addresses this measurement errors by reducing the effects of technical variance [33, 111]. Similar to microarray approaches, mass spectrometry data are very high dimensional and hence require *feature selection* methods to find promising biomarkers [56]. Different methods are used to reduce dimensionality by *peak detection* algorithms [47]. Popular methods compute the signal-to-noise (S/N) ratio and all local maxima in a spectrum that exceed a S/N-threshold are considered a peak [115]. After this preprocessing steps one gets for n spectra and p peaks a n x p matrix similar to gene expression microarrays. Once this matrix is obtained often machine learning methods like SVM (Support Vector Machines) coupled with *recursive feature elimination* procedures can be applied to discriminate disease states by differential protein patterns [100, 186].

1.4 Integrative bioinformatic analysis of omics-approaches

1.4.1 Correlation of chromosomal aberrations and gene expression

High-throughput technologies like aCGH enable the identification of DNA copy number aberrations (CNA) (Sec. 1.3.2). In the same way, gene expression microarrays allow for monitoring of thousands of genes and give new insights into underlying mechanisms of gene interaction (Sec. 1.3.3). However, numerous chromosomal alterations have been described, but molecular consequences remain unclear in most cases. To pinpoint genes that are directly affected by CNAs is a critical task. analyzing DNA copy number alterations and their effect on gene expression in parallel will enhance the knowledge about which genes are regulated, and are thus potential regulators in genetic processes and not just bystanders in alterd regions. These regulators may encode transcription factors or even signaling proteins which in turn activate hundreds of downstream genes. Unfortunately the regulator it iself may not be included in the genetic signatures. Several studies performed systematic analysis to discover wether CNAs are directly associated with changes in gene expression [87, 130, 38, 43, 94, 159, 155]. An adequate correlation between CNA and gene expression has been detected by J.Pollack *et al.* 2002. They showed that the overall patterns for amplified chromosomal regions and elevated gene expression in a subset of primary breast tumors and breast cancer cell lines is quite concordant [130]. Applying a linear regression model, they found 62% of high-level amplification to be associated with at least moderately increased gene expression. On average a 2-fold change in DNA copy number comes along with a about 1.5 fold-change in gene expression. Interestingly they noticed a significant shift of a histogram, generated from the correlation (going from -1 to 1) between CNA and expression values, in the positive direction from zero. From this they conclude a pervasive global influence of CNA on gene expression.

Hymen *et al.* 2002 analyzed the influence of genome wide CNAs on the expression of around 13.000 genes of 14 breast cancer cell lines in parallel [87]. For each gene they calculated the mean difference in gene expression between cell lines with and without amplification divided by standard deviations

$$w_g = \frac{m_{g1} - m_{g0}}{\sigma_{g1} + \sigma_{g0}}, \tag{1.1}$$

where m_g denotes the means and σ_g the standard deviation, 1 describes amplification and 0 no amplification. In doing so, their results illustrate that 44% of amplified genes (copy number ratio > 2.5) were up-regulated. This percentage decreased with lower level amplification.

Järvinnen *et al.* 2006 analyzed the correlation between CNA and gene expression of 20 samples of squamous cell carcinoma cell lines [94]. By using the same statistical method like J.Pollack *et al.* 2002, they found 39% of amplified regions to be upregulated and 14% of deleted regions to be downregulated. In total 739 genes were significantly

influenced by copy number increase. For these genes they calculated on average a Pearson correlation coefficient of 0.45 between DNA and RNA levels. In addition, they found 40 genes whose expression was systematically influenced by high DNA amplifications. Accordingly 502 down regulated genes were associated with deletions of corresponding chromosomal regions.

Serveral other studies also refer to the existence of correlation between changes in DNA copy number and their influence on gene expression [85, 114, 176, 188, 181, 165, 184] [114] [165] [176] [181] [184] [188].

However, no correlation between CNA and gene expression was reported by Björn Fritz *et al.* 2002. They analyzed alterations in DNA copy number and found no correlation with RQ-PCR expression of candidate genes for liposarcomas. Yao *et al.* 2006 confirmed these results. Their study of different subtypes of breast tumors by aCGH and Serial analysis of Gene Expression (SAGE) reveals no overall association between gene expression and amplification. They conclude that the correlation between CNA and gene expression is highly variable among tumors and conclude that different mechanisms of gene activation depend on the tumor subtype.

These studies demonstrate that the underlying effects of chromosomal aberrations on changes of gene expression are still not well understood. These studies only analyzed so called *cis*-effects, where CNAs are correlated with genes that are directly located at the same chromosomal position. Of much more interest are the interrelated alterations in DNA copy number, acting as *trans*-effect, on genes located on another chromosomal position. Soroceanu *et al.* 2007 observed in glioblastoma that a DNA loss of PTEN, which is located on chromosome 10, comes along with over-expression of IGF or EGFR. Both are not located on chromsome 10 but are potiential regulators in the formation of glioblastoma [149]. Other examples for *trans*-effects taking place in the interplay of structural changes of chromosomes on the expression of genes are given by Sweet-Cordero *et al.* and Huang *et al.* [158, 83].

Thus it has become clear that *trans*-effects can have a major effect on regulators of a gene signature. A promising method to analyze the existance of *cis*- and *trans-effects* is called SLAM (stepwise linkage analysis of microarray signatures) [2]. In order to identify candidate oncogene regulators in wound signatures, they link gene expression data to DNA copy number changes by a four-step method. First, they group the data into two classes based on absence or presence of a known gene expression pattern. Then they detect significant associations between chromosomal aberrations and gene expression signature by SAM [167]. In the next step, candidate regulators are identified by linkage analysis. Hereby, the existence of three neighboring amplified genes in only one class of the phenotyope is defined as a genetic linkage. Than the gene expression level of the potential regulator is compared with those of the genetic signature of interest. In the end they test whether the potential regulator mRNA level predicts the signatures in additional tumor samples. By applying this method, they find MYC and CSN5, both located on chromsome arm 8q, to be highly correlated with the wound-gene-signature they have identified previously [37]. Thus, the SLAM-method considers CNA and gene expression levels in an integrative manner, and the authors claim to offer new informa-

tion which could not be detected by just one method alone. However, SLAM fails to identify mechanisms through which wound-signature may be controlled by other regulators. Additionally this method does not answer how this regulators are associated with each other, e.g. in a conditional or combined manner.

Till now the most sophisticated approach to identify *cis*- and *trans*-effects in an integrated approach is presented by Lee *et al.* in 2008 [99]. They explore the underlying mechanism of CNA affecting gene expression by calculating the Pearson correlation which they store in a correlation matrix. Starting from here they searched for a set of CNAs and set of gene expression profiles that are highly correlated using a biclustering method called SAMBA [161]. The resulting modules of high correlation were analyzed for functional relevance by gene set enrichment analysis, coupled with hypergeometric statistics. The tested gene sets inlcude genes with specific biological functions, signaling pathways or cytoband locations. For the first time, their results based on the correlation matrix show that a large number of significant associations were derived from different cytobands. Among the top significant associations, 10 out of 515 combinations were found as potential *cis*-effects, and a number of 4386 out of 439151 were characterized as potential *trans*-effects. These results point out the strong association between chromosomal instability and gene expression related to different loci. Furthermore, by testing the enrichment of specific modules, they identified overrepresented gene sets which could not be verified when analyzing CNA and gene expression data on their own. Nevertheless this method does not face the fact of regulators acting in a combinatorial way and influencing other cytogenetic locations and genes in the big picture of an interacting network.

Our approach includes Bayesian networks (BN) and extends previous methods by identifying underlying *cis*- and *trans-effects*. BN are based on conditional probability relations and are therefore very useful to disclose the relationship between DNA copy number changes and gene expression. To our knowledge this thesis for the first time incorporates CNA and gene expression signatures in an integrative procedure via BN. A framework for a combined analysis is implemented, which took care of the joint probability distribution and results in a directed graph with nodes representing stochastic variables (chromosomal locations, genes), and edges account for directed dependencies among this variables. Principles of a BN will be introduced in section 1.4.2, and an example of an application will be given in section 2.1.2

1.4.2 Bayesian networks and computational biology

Bayesian networks are a representation of joint probability distributions (JPD). During the last 10 years they became increasingly important in the biological science. They were used to infer cellular networks [63], model protein signalling pathways **?**], data integration **?**], classification [27] and genetic datat analysis [16].

I used BN to gain new insights on how DNA copy number changes influence gene expression. It has been widely accepted that genes do not act as single players. They are rather merged as players in a network of interacting genes and can depend on copy number aberrations. These genes can be organised in pathways or biological functions. I tried to gave a contribution in understanding the mechanisms of genetic processes trig-

gered by alterations in DNA copy number and identifies potential regulators utilizing BN.

Again the overall question of this study is "do we see key players by analyzing CNAs and gene expression data in an integrative approach using BN." This idea, which also allows for additional information like clinical criteria including e.g. survival data or the prognostic index, is illustrated in Fig. 1.5. In this graphical representation, the vari-

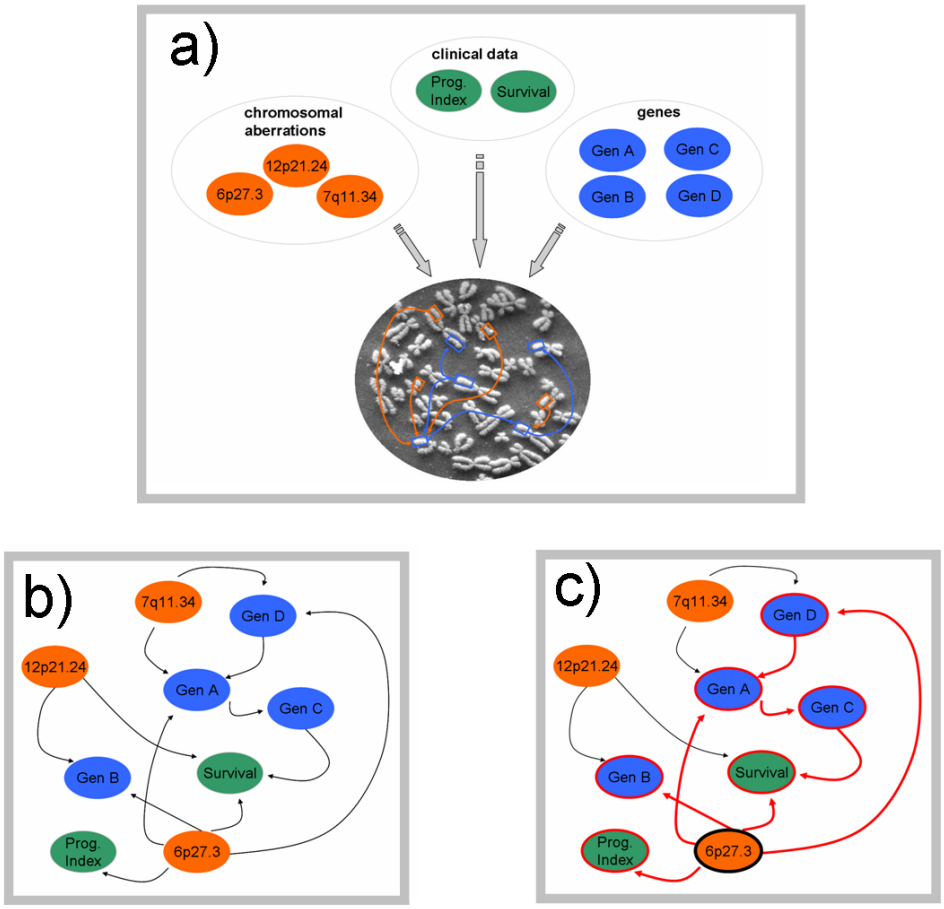


Figure 1.5: Basic idea for the identification of key players in molecular biological processes via Bayesian networks. Data are not real. a) Preselected nodes of interest origin from CNAs (orange), genes expression (blue) and clinical data (green) have to be chosen from every data type alone. It should be addressed, how these nodes interact and regulate each other depending on their chromosomal position. b) After infering BN this results in a directed acyclic graph with nodes and edges. c) Resulting dependencies allow to deduce key-players from that graph. In this example, a chromosomal aberration on 6p27.3 seems to have an important role on the clinical outcomes by influencing the expression of four genes.

ables (genes, clinical data or chromosomal aberrations) are represented by nodes that are connnected by edges. This edges represent relationships among the variables. The expression of each node is represented by one variable of the JPD which describes how

the variables are regulated by each other. A more detailed introduction to BN is given in Sect. 2.1.2.

Roughly speaking a BN is a tool that cna help to come to a decision in a specific situation. The situation can be seen as a model which is based on experiences someone made in his live. The experienced-based model affects the decision how to proceed in a given situation.

Such a model could give answers to almost any question, e.g., what is the chance to come into heavy rain when I leave the door in the morning. Another question could be, what is the lifetime risk to develop cancer. Considering the latter, a BN can be build up with nodes and edges. The nodes repesent variables, like being a smoker, age or other cofactors of interest, that might influence each other. The edges of a BN represent dependencies among these variables, e.g. being a smoker has an influence to suffer from cancer in the future. Once a model is made it is not irrevocable, instead we can change the estimation of a situation or even add new experience we made. For example if additional information is available like the level of alcohol intake, would certainly influence the estimate of getting cancer or not.

1.4.3 Separated and integrative analysis of BOS specific gene and protein expression

In order to expand the idea of this thesis to analyze and refresh the view on the biological dogma, we combined the expression of genes and proteins, related to Bronchiolitis Obliterans Syndrome (BOS), in an integrative step. This is done in accordance with the previous section **??**, where the work on analyzing the specific part of the biological dogma, where "DNA makes RNA" is figured out. In this section one step is made forward to "RNA makes proteins."

Understanding the molecular mechanisms of a disease like BOS is fundamental to the development of new therapies. The efforts of the last years in high-throughput methods like gene expression micorarrays and mass sepctrometry for protein profiling utilize a systems approach for biological procesess. Indeed, no single approach as a "stand alone" can fully unravel the complexity of fundamental biology. However, most integrative studies of mRNA and proteins searched for a correlation between this two levels of biology. Most popular are correlation analyzes of gene expression microarrays and 2-D gelelectrophoresis. The results are quite diverse likewise the studies related to the analysis of changes in DNA copy number and their effects on gene expression (Sec. 1.4.1 on page 26). For example, a study in yeast *Saccharomyces cerevisiae* found a correlation of 0.6 between the expression of 289 genes and their related proteins [89]. Others reported a correlation coefficients of -0.025 when analyzing the expression of 98 genes and proteins in lung adenocarcinomas in parallel [39]. Many reasons exists that might decouple the correlation between gene and protein expression measures. Many biological "sources of irritation" escort a mRNA on it is way through the biological dogma till it might eventually end in a protein. Example are mRNA degradation or alternative splicing (Sec. 1.3.4 on page 20). Also different post-transcriptional modifications influence the composition of a protein. These processes in a cell can not be measured with a gene expression microarray and thus lead to a worse correlation between mRNA and protein abundance.

1.5 Aims

The aim of this thesis was to develop a bioinformatic technique to gain new insights into how chromosomal aberrations affect gene expression and how genes act on protein expression. For this purpose a dataset of 81 patients suffering from neuroblastoma and a collective of 53 patients after lung transplantation was available. In particular, it had to be adressed whether cis- and trans-effects are underlying mechanism for the origin and progression of neuroblastoma. Furthermore, the effect of changes in gene expression on protein levels related to the Bronchiolitis Obliterans Syndrom had to be analysed. A meta-analysis approach of both mass spectrometry and aCGH data had to devised in order to discover new features that might not be found in a seperate analysis.

Chapter 2

Material and Methods

2.1 Integration of Neuroblastoma specific copy number changes and gene expression by BN

To analyze DNA copy number changes and their impact on gene expression current methods do not consider the underlying network like characteristics. This comes true especially for neuroblastoma. Widely applied methods analyze, so-called *cis-effects*, where the expression of those genes are monitored which lie within the same chromosomal region with lost or gained DNA. Our method consideres *cis*-effects as well but differs to other methods by tracking all possible state-combination between CNA (loss, balanced, gained) and gene expression (low, middle, high). Each state-combination got an assigned *consistency-score* and was used for cluster analysis

Furthermore, a new method to analyze *trans*-effects was developed. It was aimed to identify the underlying relationship in neuroblastoma between CNA and genes that are located on different chromosomal regions. This is done by computing a so called *equal-state-correlation-coefficient*, where we sum up equal states for each combination of CNA and gene expression.

With this *equal-state-correlation-coefficient* in hand a Bayesian network was applied to point out the network characteristics of genomic aberrations affecting gene expression. This method computed the probabilistic dependencies between CNA and gene expression and visualized the connections as a acyclic directed graph.

2.1.1 Data

In this study, paired aCGH profiles and gene expression data were used, comming from 81 patients suffering from neuroblastoma. For the application of aCGH data a previously published data set [152] was used. In this study whole genome aberrations were measured in neuroblastoma using a specifically designed high-resolution oligonucleotide 44 k aCGH microarrays (Agilent Technologies, Palo Alto, CA).The R package *GLAD* for detecting the breakpoints delimiting altered regions and assigning a status (normal, gained or lost) to each chromosomal region was utilized [86].

Gene expression data consisted of an already published data set in which gene expression profiles were generated as dye-flipped dual-color replicates using a customized 11k olignucleotide microarray [120]. The raw data were normalised by VSN (variance stabilization

normalization) [84].

2.1.2 BNtegrative. A comprehensive toolbox to screen genomic *cis*- and *trans*- effects

2.1.2.1 *k*-means discretization of gene expression values

In a typical DNA-micrarrray gene expression experiment, genes are labeled with a fluorophore (sec. 1.3.3). Such labeled transcripts are then hybridized to a microarray. The resulting fluorescence signal is detected by a scanner and is believed to be proportional to the relative abduance of the corresponding gene. The expression of all genes is then quantified by measuring the intensity via the scanner. Here, the gene expression values differ in the distribution compared to preprocessed aCGH data. As mentioned in section 1.3.2, the last preprocessing step during a aCGH experiment includes the assignment of states to each chromosomal position. These states are loss of DNA, balanced DNA content, and gain of DNA (-1,0,1). In order to analyze both data types in an integrative step, it is necessary to categorize gene expression data as well into comparable groups (down-regulation, no change, up-regulation of genes) (-1,0,1).

For each gene a k-means clustering approach was used to obtain up the three categories, mentioned above. Considering one gene, for each expression value X, over all amounts m of samples L the procedure started with randomly choosen $k = 3$ points as cluster centroids. The remaining gene expression values were assigned to the cluster centroid with the lowest Euclidian distance

$$dist_{eucl}\left(x_1, x_2\right) = \sqrt{\left(\sum_{i=1}^{m}\left(x_{1i} - x_{2i}\right)^2\right)}. \tag{2.1}$$

Then for each of the three clusters the centroid of n gene expression values is calculated, which is the arithmetic mean μ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.2}$$

Again, each gene expression value is assigned to the latest cluster centroids. These steps are repeated until the centroids are no longer moved. The method is illustrated in alg. 1.

Algorithm 1 k-means

for each gene
    randomly choose 3 centroids
    repeat
        for each expression value x
            1) assign x_i to centroid with the lowest Euclidian distance
            $dist_{euc}\left(x_1, x_2\right) = \sqrt{\left(\sum_{i=1}^{m}\left(x_{1i} - x_{2i}\right)^2\right)}$
            2) recalculate new centroids $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
    until the centroids no longer move

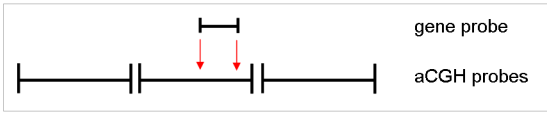
2.1.2.2 Matching gene probes with aCGH probes

Here we matched the probes represented at a gene expression microarray with the probes on an aCGH microarray. This was required when analyzing *cis*-effects of DNA copy number changes on the expression of genes, located on the same chromosomal position. In most cases, both data types are measured on different plattforms and therefore differ in the type of spotted probes on the microarray, like cDNA clones or short oligonucleotide sequences, see sect. 1.3.2 for further information.

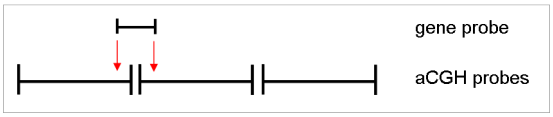
The algorithm required the chromosomal start and end points of the spotted probes, which enables to build up a link to a specific chromosomal region. For each gene, the algorithm searched for the aCGH probe on the same chromosome whose position matched most closely that on gene expressio microarray. If no perfect match was obtained, the method located the straight right or left neighboring aCGH probes. The matched gene to aCGH probes were saved as *cis*-effect connections, only if the neighboring aCGH probes had the same state {-1,0,1}. The method is shown in alg. 2.

Algorithm 2 Matching gene probes to aCGH probes

for each gene probe
do
search for the perfect match with aCGH probe



if no perfect match
search for the direct aCGH probe neighbor
if neighboring aCGH probes have the same state (-1,0,1)



save match as a *cis-effect*
else go to next gene probe

2.1.2.3 Calculation of patient-related *cis*-effects via consistency correlation

After the matching of gene probes to aCGH probes (sect. 2.1.2.2), here the algorithm calculated the correlation between DNA copy number changes and gene expression. The correlation was computed patient wise and results in a correlation value for each matched chromosomal *cis*-position.

Considering the matched *cis*-positions between aCGH probes and gene probes, the algorithm assigned for each possible state combination a correlation value. By state combination, the comparison of the actual individual state of each data type, at a specific

cis-position, for a single patient is ment. Again the states for the aCGH data were loss, balanced and gain. For the gene expression data the data were categorized into dow regulation ("loss"), no change ("balanced") and up regulation ("gain").

This resulted in a *consistency matrix* with columns for each single patient and rows as matched cis-*positions.* The data points of the matrix represented the *consistency-score* between CNA and the gene expression for that specific chromosomal position. The algorithm is schematically described in alg. 3 .

Algorithm 3 Build *consistency-matrix* of consistency score

for each patient
  for each chromosomal *cis*-position
    do
      assign a score value between gene state and aCGH state

aCGH	Gene Expression	assigned correlation value
loss	"loss"	3
loss	"balanced"	2
loss	"gain"	4
balanced	"loss"	1
balanced	"balanced"	0
balanced	"gain"	1
gain	"loss"	4
gain	"balanced"	2
gain	"gain"	3

save score in *consistency-matrix*

2.1.2.4 Hierachical clustering of patient-related *cis*-effects

At the end the patient-related *cis*-effects represented as a *consistency-matrix* were grouped together. This was done by hierachical clustering of the *consistency-matrix* in combination with the euclidean distance and the complete linkage algorithm.

2.1.2.5 Reduce dimensionality of aCGH data

I only considered frequently lost or gained chromosomal regions in order to reduce the dimensionality of the aCGH data. Therefore we left out chromosomal positions, where less than 20 % of all patients had an DNA aberration.

I build up *chromosomal location sets* (CLS) to further reduce the dimensionality of the aCGH data. CLS corresponds to each human chromosome and each cytogenetic band. In total we defined 426 CLS. I computed a mean-CLS-value for all aCGH probes that belong to the same CLS, by assigning the most frequent state (-1,0,1).

2.1.2.6 Identification of significant *cis*- and *trans*-effects over a whole set of patients

This part of the algorithm aimed to reduce the dimensionality of data by the identification of significant *cis*- and *trans*-effects over a hole set of patients. To avoid confusion, it should be kept in mind that from here on the algorithm did not consider patient-related cis-effects like provided in sec. 2.1.2.3.

Here the algorithm reached it is crucial part because the output served directly as an input to the BN analysis. The smaller the number of input variables for a BN, the shorter is the computation time and the more stable are the results.

A *similar-state-sum* was computed during the first step of this method. In detail, this step returns a measure of similarity for each gene probe γ with any other aCGH probe α. Consider that the two data types are repesented in two different matrices with patients in columns of the same order, and data type specific probes in rows. Starting with the first gene probe γ_1 as a vector, the sum of equal states over all patients compared with the vector of the first aCGH probe α_1, was computed. The individual vectors were of the length of the number n of patients. Only states which were unequal to balanced (0) for the aCGH data and no-change (0) for the gene expression data were considered. This sum of states, called *similar-state-sum* sss=$\{1, ..., n\}$ were computed for all combinations of gene probes with aCGH probes. At the end a *similar-state-sum-matrix* $sssm$ was build up with aCGH probes in rows and gene probes in columns. For a better understanding of how the $sssm$ was computed, it is schematically described in figure 2.1.

aCGH						
	A	**B**	**C**	**D**	**E**	**F**
a1	0	1	1	-1	0	1
a2	-1	1	0	0	-1	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

gene expression						
	A	**B**	**C**	**D**	**E**	**F**
g1	0	0	1	-1	-1	1
g2	-1	1	0	0	-1	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

sssm		
	a1	**a2**
g1	3	.
g2	.	4
⋮	.	.
⋮	.	.

Figure 2.1: Schematic computation of *similar-state-sum-matrix* . The aCGH matrix and gene expression matrix have the same structure. Rows refer to the data type-specific probes and columns for the patients in the same order. The discretized values for both matrices are -1,0 and 1. Green values identify equal states between aCGH and gene expression data. The similar-state-sum-matrix (sss) holds the sum of similar states between aCGH and gene expression data

Significance of *cis*- and *trans*-effects was tested by an empirical p-value. Therefore the labels of rows and columns of both data types were randomly relocated and again a *similar-state-sum-matrix* was computed. Then for each possible values of this permuted sss=$\{1, ..., n\}$, the p-value was computed by

$$p = \frac{\#sss > sss_i}{nrow_{sssm} * ncol_{sssm}} \tag{2.3}$$

where *nrow* was the number of rows and *ncol* the number of columns. This results in n p-values, that denote the significance threshold for the strength of a *cis*- or *trans*-effect.

2.1.2.7 Bayesian modeling of genome-wide *cis*- and *trans* -effects

BN are structures that represent probability distributions. For a set of variables $X = \{X_1, ..., X_n\}$, a BN consists of a network structure S. It is a *directed acyclic graph* with nodes as stochastic variables and edges as directed dependencies among these variables. If there is an edge from variable X_1 to X_2, then X_2 depends probabilistically on X_1. In this case X_1 is a parent of X_2, which is in turn the child of X_1. Nodes that do not have a parent are called unconditional variables [8]. *Local probabilty distributions* P are attached to each node in the network. They represent the strength of causal relationship between a variable and it is parents,

$$p(X_i|Pa_i) \tag{2.4}$$

where Pa_i are the parents of a variable X_i, and describe the behavior of that variable under every possible value assignment of it is parents.

The *joint probability distribution* of all conditional variables in a BN is the product of the local distribution,

$$p(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_i|Pa_i) \tag{2.5}$$

and can be seen as the probability of two or more events happening together.

Independence assumption is a key factor of BN and describes the task of breaking down the overall distribution of a BN into connected modules. The underlying rules to infer independence relations from the structure of a BN are given by *d-seperation*. These rules are similar to graph conectivity concepts and address the question whether a path is active in turns of creates dependency between end nodes. In the inactive situation a path is blocked by a node and dependency can not be created between the end nodes. For example, three random variables A, B and C (Fig. 2.2) are given. The variable A is d-separated from C given B if the path from A to C is blocked, given B

$$p(A, C|B) = p(A|B)p(C|B). \tag{2.6}$$

Blocked means that we have evidence e for B or in other words the value for B is known in the network and that implies that no information can flow between A and C.

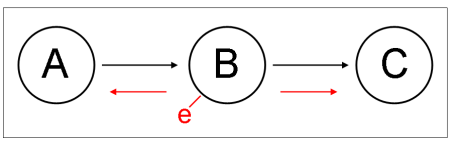


Figure 2.2: *d-seperation* of 3 nodes in a BN. A and C are *d-separated* given B because evidence e is given for B.

2.1.2.8 Structure learning of a BN

Structure learning in principle implies a set of conditional independence assumptions via *d-seperation* among the variables involved. There are two common algorithms for learning the structure of a BN. One approach aims to optimize a network score for a specific network by randomly changing the topology [48, 76, 42]. The second approach, a constrained-based method, deals with several independence tests between the variables of a network given other variables [126, 150].

Although the second method was used, also the first *score-based method* will be explained first because it is a straight forward way to generate a BN and helps to understand the *constrained-based method.*

Learning the structure of a BN from given data requires estimating the conditional probability distributions (parameters) and independence relations.

The *score-based method* assigns a score to each possible BN reflecting how well the BN describes the data set D. Assuming the structure S of the network, the score is

$$Score(S, D) = p(S|D) \tag{2.7}$$

in terms of posterior probabilities of S given the data D. Following the Bayesian theorem, this can be written as

$$Score(S, D) = \frac{p(D|S)p(S)}{p(D)} \tag{2.8}$$

where a score-base method attempts to maximize this score. Only the numerator needs to be maximized, since the denominator does not depend on S. On popular method to calculate the score of a network is the Bayesian Information Criterion (BIC score) [136]

$$BICscore(S, D) = \ln p(D|\hat{\Theta}, S) - \frac{d}{2} logN, \tag{2.9}$$

where $\hat{\Theta}$ is an estimate of the model parameters for the structure, d is the number of model parameters, and N is the size of the dataset. The BIC score is a measure of how well the model fit is the data. The problem of finding a structure of an optimal score of a BN is NP hard since the number of structures grows (super) exponential. Typical search methods implement greedy search strategies [41]. Starting with an initial network, edges are iteratively added, deleted or reversed until a local maximum of the score is found.

In this thesis a so called *constrained-based method* was used to learn the structure of a BN [150]. This kind of algorithm try to detect the dependencies and conditional independencies from data by statistical tests. The resulting dependencies and conditional independencies are then used to infer the structure of a BN. In order to use the results to reconstruct the structure, several assumptions have to be made: causal sufficiency assumption, causal Markov assumption, and faithfulness assumption.

> **Causal sufficiency assumption:** there exist no common unobserved variable in the domain that is a parent of one or more observed variables of the domain.

Causal Markov assumption: in a BN any variable is independent of all it is non-descendants given it is parents.

Faithfulness assumption: a BN structure S and a probability distribution P generated by S are faithful to one another if every conditional independence relationship is entailed by the causal Markov assumption in S.

En route the existence of an edge between two variables and the direction of an arc is discovered. Two straightforward constrained-based methods are the SGS (Spirtes, Glymour and Scheines [150]) algorithm and the PC algorithm [126].

To investigate the association of DNA copy number changes on gene expression, a constrained based method which is called *Growth-Shrink (GS) Markov Blanket (MB) Algorithm*[1] was used. The idea of a Markov Blanket of a variable is based on J. Pearl, 1997 (2nd Ed.) [125] and was improved by D. Margaritis, 2003 [109].

Markov Blanket (MB) is a minimal set of nodes which d-separates a node from all other nodes. The MB of a node X containes all parents, children and parents of children of that node. An example of a MB is given in figure 2.3.

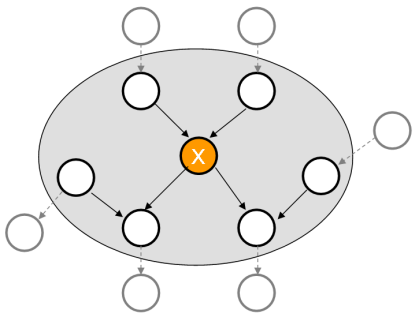


Figure 2.3: Markov blanket of a variable X. The members of the blanket are within the gray ellipse.

The MB of a variable X is computed by pairwise independent tests based on the mutual information (**MI**) criterion

$$\mathbf{MI}(X,Y) = \sum_{y \in Y} \sum_{x \in X} \log_2 \left(\frac{P(x,y)}{P(x)\,P(y)} \right). \tag{2.10}$$

It is a measure of strength of the association between the distributions of two variables X and Y. Significance is tested by means of a χ^2-distribution. The only parameter of this distribution is the degrees of freedom ν and is set to the number of state levels (-1,0,1) $df = 3$ of the input variables.

[1] `http://www.r-project.org`, *"bnlearn"* [109]

GS-Algorithm The algorithm consists of a growing and a shrinking phase of a **MB**. The growing phase starts with an empty set **S** and adds variables to **S** as long as they are dependent ($\sim \exists$) with X given the current contents of **S**. However, during this process variables are added to **S** that are infact outside of the Markov blanket. The shrinking phase accounts for this and removs all members of **S** as long as they are independent of X given the current **S**. The method is presented in algorithm 4. The symbols are explained in table 2.1. The algorithm is taken from Margaritas in 2003 [109] and is shown

Algorithm 4 Grow-Shrink Markov blanket algorithm

1) Start with empty **S**
  $\mathbf{S} \leftarrow \theta$
2) Growing phase
  While $\exists Y \in U - \{X\}$ such that $Y \pm X \mid S$
    Do $\mathbf{S} \leftarrow \mathbf{S} \cup \{Y\}$
3) Shrinking phase
  While $\exists Y \in \mathbf{S}$ such that $Y \perp X \mid \mathbf{S} - \{Y\}$
    Do $\mathbf{S} \leftarrow \mathbf{S} - \{Y\}$
4) $\mathbf{MB}(X) \leftarrow \mathbf{S}$

Table 2.1: Table of symbols

Symbol	Meaning
S	Set
U	Universe, set of variables variables in the domain: $\{X_1, ..., X_n, \}$
X, Y, Z	One-dimensional variables
$Y \pm X \mid S$	variables X and Y dependent upon conditioning on the variables in the set S
$Y \perp X \mid S$	variables X and Y are independent upon conditioning on the variables in the set S
$\mathbf{MB}\,(X)$	Markov blanket of variable X
$\mathbf{MI}\,(X, Y)$	mutual information of two variables X and Y
$\mathbf{N}\,(X)$	neighbors of variable X

in algorithm 5. It starts with the identification of the Markov blankets for each node, according to algorithm 4. Step 2 determines which members of the blanket of each node are actually direct neighbors **N**. This is done by computing pairwise independent tests, see above, between X and Y conditioned on all subsets of the smaller of $\mathbf{MB}(X) - Y$ and $\mathbf{MB}(Y) - X$. Step 3 represents the case where two variables (X, Y) have a common descendant (Z) and hence become dependend on each other, when conditioning on a set that includes any such descendant. It is possible that step 3 leads to directed cycles in the resulting graph which is not allowed in a BN. Therefore step 4 and 5 identify the minimum set of edges that need to be reversed for all cycles to disappear. Since not all directions can be determined during the last steps, this is resolved in step 6. Edges are orientated in a way such that they do not introduce a cycle, if the reverse direction

Algorithm 5 Learning the structure of a BN via GS-algorithm.

1. **Compute Markov Blankets**
   For all $X \in U$, compute the Markov blanket $\mathbf{MB}(X)$.
2. **Compute Graph Structure**
   For all $X \in U$ and $Y \in \mathbf{MB}(X)$, determine Y to be a direct neighbor of X if X and Y are dependent given $\mathbf{S}$ for all $\mathbf{S} \subseteq \mathbf{T}$, where $\mathbf{T}$ is the smaller of $\mathbf{MB}(X) - Y$ and $\mathbf{MB}(Y) - X$.
3. **Orient Edges**
   For all $X \in U$ and $Y \in \mathbf{N}(X)$, orient $Y \rightarrow X$ if there exists a variable $Z \in \mathbf{N}(X) - \mathbf{N}(Y) - \{Y\}$ such that Y and Z are dependent given $\mathbf{S} \cup \{X\}$ for all $\mathbf{S} \subseteq \mathbf{T}$, where $\mathbf{T}$ is the smaller of $\mathbf{MB}(Y) - \{X, Z\}$ and $\mathbf{MB}(Z) - \{X, Y\}$.
4. **Remove Cycles**
   Do the following while there exist cycles in the graph:
   - Compute the set of edges $\mathbf{C} = \{X \rightarrow Y \text{ such that } X \rightarrow Y \text{ is part of a cycle}\}$.
   - Remove from the current graph the edge in $\mathbf{C}$ that is part of the greatest number of cycles, and put it in $\mathbf{R}$.
5. **Reverse Edges**
   Insert each edge from $\mathbf{R}$ in the graph in reverse order of removal in Step 4, reversed.
6. **Propagate Directions**
   For all $X \in U$ and $Y \in \mathbf{N}(X)$ such that neither $Y \rightarrow X$ nor $X \rightarrow Y$, execute the following rule until it no longer applies: If there exists a directed path from X to Y, orient $X \rightarrow Y$.

necessarily did. If a direction of an edge could not be determined during the algorithm, each possible direction of each undirected edge is tested, and the one with the lowest p-value is accepted as the true direction for that edge.

Prior knowledge was integrated into the BN. It was required that arcs from gene nodes do not point to an aCGH node. Furthermore it was excluded that aCGH nodes could have a connection to other aCGH nodes. Although it is known that the expression of a specific gene can cause a chromosomal aberrations this was neglected with regard to the complexity of the model.

2.2 Integration of BOS specific gene and protein expression

Here the information derived from gene expression microarray experiments is combined with protein profiles of BOS in an integrative manner. This approach is based on translating the gene expression measures into "virtual protein spectra". This made both data types comparable. But first the gene expression and protein data were analyzed separately. The results from this isolated point of view were important for the understanding of the underlying mechanisms of BOS. Nevertheless the integrative approach gave the opportunity to obtain information that could not be interpreted by analyzing each data set on its own. Therefore the Wilcoxon rank test was applied to identify correleation of proteins expressed by their corresponding genes. It was considered that peaks in a protein spectrum could not directly be linked to a specific protein name, but rather coded by their m/z-value (Sect. 1.3.4 on page 20). The basic concept of this statistical test was based on comparing measured m/z-values of the protein profiles with the approximated m/z-values of the virtual gene-mass-spectra. Furthermore a meta-analysis approach integrated both data types and was adopted to gain new information which could not be achieved by analyzing both data sets on their own.

Under my supervision Mirjam Maier added during her diploma thesis functionality to carry out feature reduction during a classification step and performs data mining part presented in this section.

2.2.1 Data

Microarray data were obtained from Hannover Medical School (MHH). In total, 52 samples of patients and 10 control samples were used. The courses after lung transplantation was continuously monitored in periods of 9 to 24, 24 to 30, 30 to 36, 36 to 44 months (tab. 2.2). For gene expression analysis bronchial brush specimens were collected . For 23 out of 53 patient samples after lung transplantation (LT) and 6 out of 10 control samples, the gene expression profiling was performed (tab. 2.2). While some of the 23 patients were already affected by BOS, for the rest it was unclear whether they will develop this syndrome.

To obtain cells from the airway mucosa, a sheathed bronchial specimen brush[2] was pushed through the operating channel of the bronchoscope, positioned in a segment bronchus, and moved back and forth gently. After retracting the tip into the protective sheath the brush was removed (Fig. 2.4 on page 45). In order to harvest a sufficient number of cells, this procedure was repeated up to five times. The epithelial cells were gently removed from the brush by lightly shaking in saline solution, and were subsequently stored at -80°C. The extraction of RNA from the cells was performed according to the Trizol-method[3], followed by RNeasy Mini Kit[4]. The quality and integrity of the total-RNA was

[2]Boston REFNRI 1601
[3]Invitrogen, Karlsruhe, Germany
[4]Qiagen, Hilden, Germany

Table 2.2: Examinations at several time points after lung transplantation (LT), and the number of available patient samples and controls, respectively. The table is splitted into mass spectrometry analysis (MS), DNA-microarray gene expression analysis (microarray) and into overlapping patient and control cohorts for the integrative analysis of both data types .

No. examination	Months after LT	# patient (MS)	# control (MS)	# patient (microarray)	# control (microarray)	# matching patient (MS/microarray)	# matching control (genes/proteins)
1	9 to 24	52	10	23	6	23	6
2	24 to 30	27	10	-	-	-	-
3	30 to 36	12	10	-	-	-	-
4	36 to 44	1	10	-	-	-	-

determined using a Bioanalyzer[5]. Because of the low amounts of RNA it was necessary to amplify isolated mRNA from the sample done by a RNA amplification kit[6].

Microarray analysis was performed according to standard protocols using the human cDNA chips of the Stanford Functional Genomics Faculty[7] [143]. The chip architecture was built by the *Resgen clone set* with more than 43,000 spots and is intended to cover the entire human transcriptome. An amount of 1.5 μg each of amplified RNA was labeled during reverse transcription with fluorochromes Cy3 (control RNA = a pool from six samples from healthy persons) or Cy5 (probe = one of 23 samples obtained after lung transplantation). Hybridization was performed for 14 to 18 hours in a hybridization chamber at 65 °C. After washing the slides, the fluorescence intensities of Cy5 and Cy3 were measured on a GenePix 4000 scanner[8] and analyzed using GenePix Pro 4.1 software[9]. This software package allowed the extraction of sample intensities or ratios at each printed cDNA location in the given microarray scan [170]. Areas of the microarray or spots that exhibited obvious damages were excluded from subsequent analyzes (Sect. 2.2.3 on page 48).

[5]Agilent Technologies 2100, Waldbronn
[6]MessageAmp aRNA kit, Ambion, Huntington, UK
[7]Stanford Functional Genomics Facility, Stanford, CA, USA
[8]Axon Instruments, Foster City, CA, USA
[9]Axon Instruments

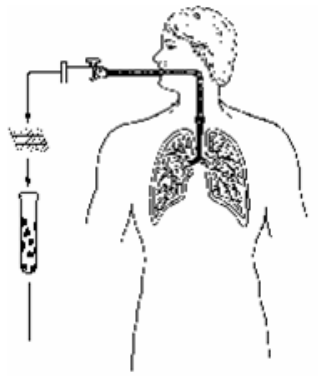


Figure 2.4: Bronchial brush specimen.

The Bronchoalveolar Lavage Fluid (BALF) from 52 patients after LT, and from 10 healthy controls, were analyzed by mass spectrometry (tab. 2.2). Bronchoscopy describes the process of filling saline solution in the lung for lavage. By means of a fiberoptic bronchoscope, the BALF was extracted out of the airways (Fig. 2.5) [106]. For this

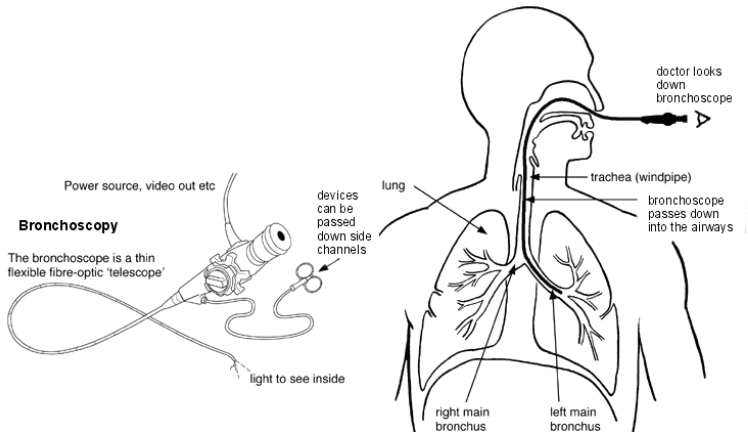


Figure 2.5: Bronchoscopy. A flexible bronchoscope is inserted through either the nose or mouth to the trachea and further down into the bronchus. Each area the bronchoscope passes can be examined. Specimen of lung tissues or lavages can be taken.

project, BALF was obtained by the Department of Pneumology at MHH. The BALF samples were collected during a routine clinical investigation after transplantation and directly delivered on ice after bronchoscopy and immediately processed in the laboratory.

Mass spectrometry was performed by means of an Ultraflex MALDI-TOF/ TOF-mass spectrometer. The analyzed samples were extracted from cells of the alveolar and bronchial airways. Superparamagnetic microparticles functionalized with C1 and C8 hydrophobic coating (MB-HIC 1 and MB-HIC 8 Beads) were used to enrich different subsets of proteins. The measurement was done in different mass (m/z) windows with a range from 1,000 to 10,000 Dalton and a second time with a range from 8,000 to 20,000 Da. In subsequent sections of this thesis, these different measurements will be denoted

as "1-10kDa" and "8-20kDa".

2.2.2 Transcriptome analysis

2.2.2.1 Preprocessing of gene expression data

Loess quantile normalization followed by a between-slide normalization was applied to the gene expression data [180]. The loess normalization used a robust scatterplot smoother (*loess*) to find a non-linear regression line through the center of the cloud of points in a two-dimensional scatterplot. By removing the calculated effects, a linear cloud of points was obtained that was centered on the diagonal of the scatterplot. The between-slide normalization step addressed the comparability of the distributions of log intensities between arrays. This was achieved by setting quantiles to identical values. Loess normalization first divided the whole chip into different sectors and then normalized each sector.

2.2.2.2 Individual significance analysis of gene expression data

Gene expression data were tested for differential expression by "Significance Analysis of Microarrays" (SAM) [167]. Furthermore Support Vector Machines (SVM) combined with Recursive Feature Elimination (RFE) was applied . Significance analyzes by SAMis based on a modified t-test statistic. it is an alternative way to detect differentially expressed genes. The approach performed in this thesis was established as Significance Analysis of Microarrays (SAM) which has been adapted specifically for microarrays [167].

SAM identifies genes with statistically significant changes in expression by conducting a set of gene-specific t-tests. A gene expression data matrix and the labels of that matrix (phenotype affiliation) serves as input for SAM. For each gene i a score d_i is assigned on the basis of it is gene expression change relative to the standard deviation. For comparison, the same statistic is calculated for every gene according to several random permutations. These results are denoted by d_{E_i}. Then a ranking of the d_i values is calculated by $d_{(1)} \leq d_{(2)} \leq d_{(n)}$ noted as $d_{(i)}$. Analogous rankings for the d_{E_i} are computed.

$$d_i = \frac{r_i}{s_i + s_0}, i = 1, 2, ..., p \tag{2.11}$$

with p number of genes, r_i differences of means and s_i the standard deviation. The variable s_0 is a small constant, which corrected the d-statistic of genes with small standard deviations to minimize the number of false positives. For calculating r_i and s_i for two groups C_1 and C_2 the following method is applied.

$$\bar{x}_{i1} = \sum_{j \epsilon C_1} \frac{x_{ij}}{n_1} \tag{2.12}$$

$$\bar{x}_{i2} = \sum_{j \epsilon C_2} \frac{x_{ij}}{n_2} \tag{2.13}$$

$$r_i = \bar{x}_{i2} - \bar{x}_{i1} \tag{2.14}$$

$$s_i = \left[\frac{(\frac{1}{n_1} + \frac{1}{n_2})\{\sum_{j \epsilon C_1}(x_{ij} - \bar{x}_{i1})^2 + \sum_{j \epsilon C_2}(x_{ij} - \bar{x}_{i2})^2\}}{n_1 + n_2 - 2} \right]^{1/2} \tag{2.15}$$

with n_k number of samples in C_k.

Genes with scores (difference between d_i and d_{E_i}) greater than a threshold Δ are considered potentially significant. The threshold Δ is adjusted to identify smaller or larger sets of genes, and *FDRs* are computed for each gene.

To find significant genes, a one-class SAM was applied on the loess quantile transformed data. A one-class SAM tests whether the mean gene expression differs from a user-specified mean [132].

Hierachical Clustering of significant genes is a powerful method to identify clusters of genes with similar gene expression patterns. A collection of objects is grouped into subsets or clusters, such that those within each cluster are more closely related to one another than objects assigned to different clusters. A central goal of cluster analysis is the notion of degree of similarity or dissimilarity between the individual objects being clustered [24]. Different clusters represent different classes of objects and often have variable size, shape and density.

Hierarchical clustering determines the hierarchy of clusters such that the clusters with minimal distance to each other are merged [73]. A dendrogram serves for visualizing the cluster analysis. This is a tree which represents the hierarchical distribution of the data set in major and minor subsets. The root of a dendrogram represents the whole data set as one big cluster. The leaves are single objects while the inner nodes represent the aggregation of all of their subtrees. Every branch between two clusters includs the distance between the represented objects. We used a bottom-up approach in combination with the *Canberra distance*.

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|} \tag{2.16}$$

This distance measure examines the sum of a series of fraction differences between coordinates of a pair of objects. Each term of a fraction difference has a value between 0 and 1. If one coordinate is zero, the term equals a unity regardless of the other value, thus the distance is not affected. This distance is very sensitive to a small change when both coordinates are close to zero.

2.2.2.3 Gene Ontology analysis

Gene Ontology (GO) categories, resulting from SAM, were tested for significance [17]. This GO categories were tested against the GO groups of all genes represented at the microarray. Fisher's exact test was performed to judge whether the observed difference is significant or not. For each GO term, a p-value was calculated representing the probability that the observed number of counts resulted by chance alone. We used the FDR to control the expected proportion of false positives.

2.2.3 Proteome analysis

2.2.3.1 Baseline correction

Baseline correction was performed to flatten the base profile of each spectrum by using an algorithm which attempt to remove the baseline slope and offset [134]. This was done by iteratively calculating the best fitting straight line through a set of estimated baseline points.

2.2.3.2 Interpolation

The peak resolution differed for each mass spectrum. Also the the range of the measured m/z-values varied and complicated the generation of a matrix with patients in columns and m/z-values in rows.

To address these two characteristics of mass spectra, a novel two-step-interpolation-method was implemented. The first step comprised an interpolation of spectra by approximating the missing data points such that the m/z intervals on the x-axis were given at equal resolution and the spectra were set to a common m/z range. For all spectra the m/z vector was interpolated to a common m/z vector using linear interpolation at the positions of the spectrum with the lowest resolution. The second step restricted the interpolation to the smallest common m/z range. This procedure is exemplary illustrated in Fig. 2.6.

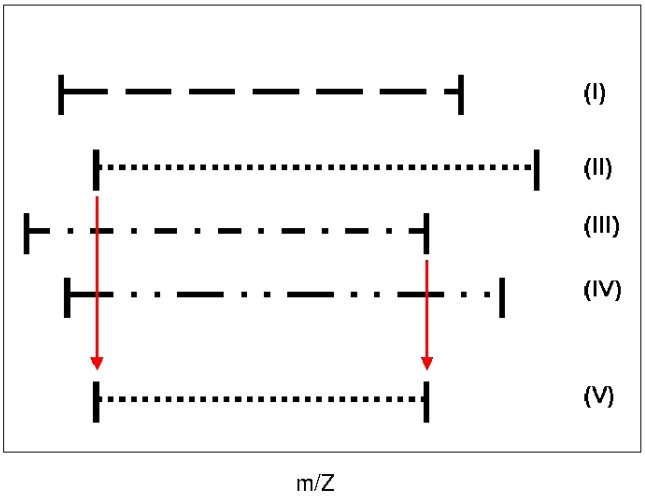


Figure 2.6: Resampling of mass spectra. Five mass spectra are shown exemplary as dashed lines. The spacing of the dashes represents the resolution of the respective spectra (I)-(IV). All spectra are interpolated to a common spectrum (V) with common m/z range and highest resolution. Therefore the largest starting point (here of spectrum (II)) and the smallest end point (here of spectrum (III)) of all spectra (I)-(IV) are chosen to be the master m/z range in (V). In addition, the highest resolution (here of spectrum (II)) is chosen as resolution for (V).

2.2.3.3 Alignment

Due to the error of measurement during a mass spectrometry experiment each sample hold the peaks at slightly different m/z-positions. These peak shifts caused a misalignment of proteins with similar molecular weight across all samples. The applied alignment procedure was based on an algorithm developed by Jeffries [91].

2.2.3.4 Normalization

There are different sources during a mass spectometry experiment that lead to a systematic variation between the spectra. A normalization method based on the total ion count was implemented and allowed for the comparison of the absolute peak intensities of different spectra [185].

2.2.3.5 Mean spectra

Multiple measurements of the same patient and control samples were performed to enhance the signal-to-noise ratio. The aim was to find the peaks which occur in all samples from one patient or control. If these spectra contained the same analytes with similar m/z values, redundant information could be compiled. Therefore a method was implemented to compute a mean spectrum for each patient. This procedure was inspired by the work of Hilario et al. 2006 [78].

2.2.3.6 Support Vector Machines

A SVM is a supervised learning method for classification. SVMs can deal with any data that can be represented as a vector in n dimensions and so can be classified by a hyperplane of $n-1$ dimensions. Special properties of SVMs are that they simultaneously minimize the empirical classification error and show a high accuracy with little bias towards overfitting [31]. Linear separation is used to assign a set of objects to their classes by inferring a hyperplane which best separates the two classes on the basis of training samples. The resulting hyperplane is the classifier. New unlabeled objects will be labeled depending on which side of the hyperplane are situated.

Formalization Let us consider data points of the form:

$$(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n) \tag{2.17}$$

where c_i is either 1 or -1 and denotes the class to which the point $\mathbf{x}_i$ belongs. In the case of $c_i = 1$, $\mathbf{x}_i$ belongs to the positive class and if $c_i = -1$, $\mathbf{x}_i$ belongs to the negative class. The data can be regarded as training data which denote the correct classification.

The aim of classification is to assign a label to a new unlabeled data point and correctly classify the new data point. SVMs approach this task by introducting a hyperplane between the positive and negative points (Fig. 2.7).

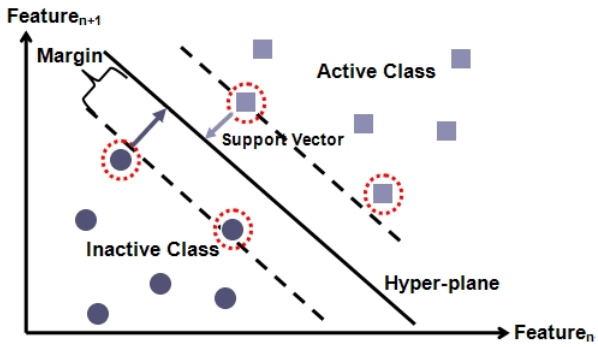


Figure 2.7: Separation of two classes (active and inactive) by a hyperplane computed by SVM in a n-dimensional feature space. The maximum margin hyperplane depends on the support vectors (red dotted circles).

There exist many hyperplanes (w, b) with $w\epsilon\Re^d$ and $b\epsilon\Re$ which can be defined by $< w, x > +b = 0$ with $< w, b >$ being the dot product between the vectors w and x. These hyperplanes satisfy

$$c_i(< w, x_i > +b) > 0, \forall i\epsilon 1, 2, ..., n. \tag{2.18}$$

The vector $\mathbf{w}$ points perpendicular to the separating hyperplane. Adding the offset parameter b allows to increase the margin. In its absence, the hyperplane is forced to pass through the origin, restricting the solution.
It is possible to choose an optimal maximum-margin hyperplane which is trained with samples from both classes. Samples along this hyperplane are called the 'support vectors'. These vectors all have the same distance to the hyperplane. The maximum-margin hyperplane is the solution of

$$\max_{w\epsilon\Re, b\epsilon\Re}(\min \|x - x_i\|), x\epsilon\Re \tag{2.19}$$

subject to $< \mathbf{w}, \mathbf{x} > +b = 0$, $i = 1, ..., n$.
The problem of maximizing the margin turns out to be a quadratic optimization problem and can be formulated as

$$\min(1/2)||\mathbf{w}||^2, \tag{2.20}$$

subject to $c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n$. The factor 1/2 is used for mathematical convenience.
The parameters of the maximum-margin hyperplane are derived by solving this optimization problem. There exist several well established algorithms from other fields for quickly solving the optimization problem that arises from SVMs, mostly reliant on heuristics for breaking the problem down into smaller, more-manageable chunks [140].
Writing the classification rule in its dual form reveals that classification is only a function of the support vectors, i.e. the training data that lie on the margin. The standard optimization technique for such problems is to formulate the Lagrangian and to solve the resulting dual problem:

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i,j} \alpha_i \alpha_j c_i c_j \mathbf{x}_i^T \mathbf{x}_j \tag{2.21}$$

subject to $\alpha_i \geq 0$, where α constitutes a dual representation of the weight vector in terms of the training set:

$$\mathbf{w} = \sum_i \alpha_i c_i \mathbf{x}_i \tag{2.22}$$

It is important to note that the hyperplane only depends on the support vectors.

Soft-margin In many cases it is not possible to find a hyperplane which correctly separates two classes. Sometimes this problem will be complicated due to outliers, which are single observations far away from the rest of the data. This frequent phenomenon in classification might shift the hyperplane into a wrong direction. For this reason, a modified maximum margin idea has to be developed that allows mislabeled examples. The *soft-margin* method is an alternative to the already explained *hard-margin* method. The goal is to improve the generalization performance of the SVM, i.e. its performance on test samples different from the training set [50].

Kernel trick However, even the *soft-margin* classifier can not solve real-world problems because a linear separation is not always possible. The idea now is to theoretically transform the data into a non linear higher-dimensional space, the feature space [140]. This is the so-called *kernel trick* because it is not necessary to know what the feature space looks like and to really transform the data into the feature space. It is only necessary to know the distances between the data points, thus the kernel function K acts like a similarity measurement.

$$K_\phi(\overrightarrow{x_i}, \overrightarrow{x_j}) =< \phi(\overrightarrow{x_i}), \phi(\overrightarrow{x_j}) > \tag{2.23}$$

Examples for kernels to use are linear, polynomial, sigmoid or radial basis functions. The optimization problem is a non-linear problem and difficult to solve, but there exist solutions to manage this as described before.
For an excellence resource about SVMs refer to *Learning with Kernels* [140].
The applied SVM used a linear kernel and optimized the cost parameter in the inner *3-fold CV* step. Here, 2/3 of the data were used as training set and 1/3 as test set. Nine values for the cost parameter were set which range from 2^{-6} to 2^{-10}. For every cost value, the *3-fold CV* was done and the accuracies were averaged. The cost value with best accuracy out of the resulting nine accuracies, was then chosen as optimal value.

Stratification is a challenge for all classification and feature selection methods to handle small sample sizes of data. The number of samples does in general not allow to set aside independent test and training sets of samples as common in machine learning [75]. Stratification is often used to find a remedy and assess the accuracy of the classifier. As shown in [186], feature selection results may vary even with a single-case difference in the training set when sample size is small. The choice of suitable training and test sets is important and the same sets should be applied for all used classifiers to guarantee a common basis for comparing their accuracy. The correct proportion of classes in the test set and thus guarantees an equal distribution of the instances was maintained by stratification. To balance class distributions, sets were stratified prior to classification by SVM. This means, if class (1) had 9 samples and class (2) 90, 9 samples from (1) and a random subset of 9 samples from (2) were chosen without replacement. This experiment was repeated 10 times in each run.

2.2.3.7 SVM with RFE

Mass spectrometry produced a high amount of high-dimensional data. Due to the disproportion between the number of observations n and the number of variables p, $n << p$, SVMs could not directly be applied to the data. Although SVMs can deal with high dimensionality, dimension reduction could still improve the performance dramatically [186]. Therefore SVM in combination with recursive feature eliminaton was used to find potential biomarkers associated with BOS.

The RFE approach in combination with the SVM algorithm allowed the direct determination of significant proteins [88]. The weights used by the SVM classificator to choose the most significant features were also applied in the recursive feature elimination process. SVM coupled with RFE used this weighting to eliminate the features with the lowest weight. The SVM was used to compute a hyperplane which was able to separate the input classes. The features were weighted according to their contribution to the separating hyperplane. Then the features with lower weight were removed and a new hyperplane was computed. These steps were repeated in a recursive way. The number of features that lead to the best performing classifier was then chosen to construct the final classifier.

The implementation of *SVM* in the *e1071* package[10] of R and was used in combination with *RFE* [6]. The integration of this algorithm in a *n-fold cross-validation-step* is shown in Fig. 2.8.

2.2.3.8 Alternative classification and feature elimination methods

In parallel to the RFE method the Hilbert-Schmidt Independence Criterion (BaHSIC) was used [98]. Similar to the RFE the BaHSIC method was combined with SVMs.

For comparison, SVMs were also applied without any feature selection. Alternative classification method, Prediction Analysis of Microarrays (PAM) was used [163]. In contrast to it is name, the PAM method can also be used for the analysis of proteomic data.

In total, 52 patient spectra (stage 01), 27 patient spectra (stage 02), 12 patient spectra (stage 02) and 10 control spectra (nt) remained for sample classification (tab. 2.3).

[10] `http://cran.r-project.org/src/contrib/Descriptions/e1071.html`

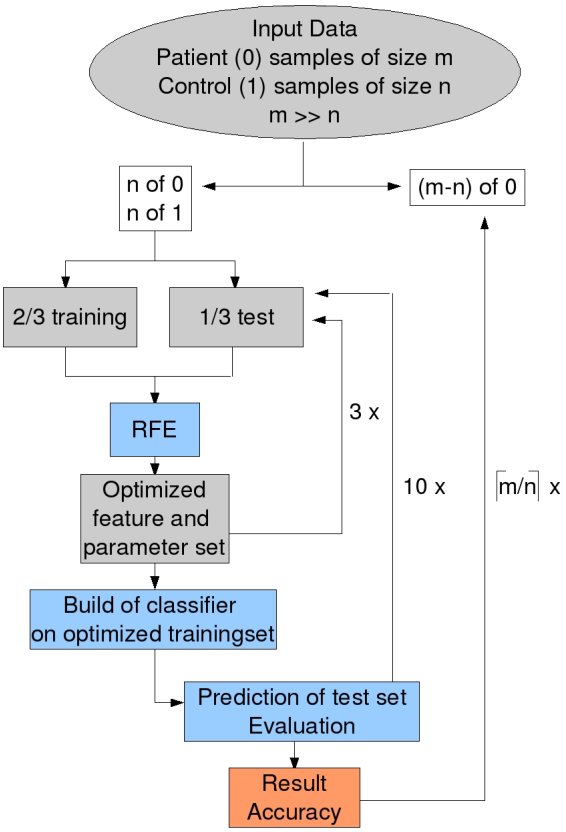


Figure 2.8: Cross validation setup including SVM coupled to RFE.

Table 2.3: Tested combinations of patients in different stages respectively control patients and their available sample numbers.

Combination	Number
All patients (01, 02, 03) vs nt	91 vs 10
Stage 01 vs nt	52 vs 10
Stage 02 vs nt	27 vs 10
Stage 03 vs nt	12 vs 10
Stage 01 vs 02	52 vs 27
Stage 01 vs 03	52 vs 12
Stage 02 vs 03	27 vs 12

2.2.4 Integrative analysis of gene and protein expression

2.2.4.1 Virtual proteomic mass spectra of gene expression levels

Every spot of a gene expression microarray contains a specific DNA fragment that is linked to different attributes, like *clone id*, *intensity* or *gene symbol*. The attribute *gene symbol* was used for further analyzes. However, it was not unique for every spot, thus some information is lost.

The integration was performed at the level of molecular masses of proteins which were available as attributes of the mass spectrometry data. Hence, it was necessary to map the genes to proteins and to determine their molecular weights. The *Compute pI/Mw tool*[11] from *ExPasy* was used. As input *UniProt*[12] identifiers are required which were not available as attributes from the microarray experiments were required. Thus, *gene symbols* were converted to *UniProt* identifiers. For the *gene symbol* conversion the *Gene ID Conversion Tool* from *DAVID*[13] was used. *Gene symbols* were extracted as a list from the microarray experiments and uploaded to the conversion tool. The output was a list of converted *UniProt* identifiers which served as input for the next step.

Computation of protein masses was done with *Compute pI/Mw tool*. The theoretical isoelectric point (pI) and molecular weight (Mw) of proteins was computed. *UniProt* sequences were processed to their mature forms. The resulting chains or peptides were used to infer the pI and Mw values. It was crucial for further analyzes that the proteins were mapped from mRNA to their mature form before calculating the pI value because this is the form in which proteins finally occur in the living organism after post-translational modifications. One major drawback was that protein phosphorylation, acetylation or glycosylation is not covered by *UniProt*. In some cases only fragments of a protein were available from the database. In such a case, no result was returned because pI and Mw cannot be computed accurately [65]. This lead to a shrunken set of proteins with appropriate masses. The output file was a list with *UniProt* identifiers, related theoretical isoelectric point and molecular weight.

Mapping of *gene symbols* to *uniprot* identifier was done by the mass information of the proteins. The *gene symbol* linked directly to the gene expression entry in the GPR file. The *UniProt* identifiers were linked to the molecular mass of the protein. These two parameters were mapped onto each other to directly infer the masses and their belonging gene expression. In the following, the mapped gene symbols were mentioned as *Gene2Prot*. For further analysis, only corresponding data from the same patients or controls were used. Transcriptomic data from 23 patients and 6 controls were available in addition to mass spectrometry data from 55 patients and 10 controls. The proteomic dataset covered all samples of the transcriptomic dataset. Hence, for further analyzes, the subset of these 23 patients and 6 controls has been selected.

[11] http://expasy.org/tools/pi_tool.html
[12] Universal Protein Resource, `http://www.expasy.uniprot.org/`
[13] `http://david.abcc.ncifcrf.gov/home.jsp`

Gene expressin data were composed of data from two channels: one corresponding to mRNA from patients' samples, the other one to mRNA from a pool of control samples (sec. 2.2.1). In order to combine them with mass spectrometry data, only data from one channel related to patients was used. This was done because it was assumed to be proportional to mRNA and hence protein abundance. If the ratio between two channels had been taken, this proportionality would have been lost. To be able to compare the results of integrative analysis with those of the separate analysis, the loess quantile normalized data were used for further steps.

Discretization of the gene expression values and the intensities of the protein masses was done to make the values suitable for numerical evaluation and comparison. There exist different discretization techniques like the division of values in specific quantiles or percentiles. Here, data were discretized by division into ten percentiles.

To convert gene expression data to virtual mass spectra a new matrix was computed. The matrix hold the samples in columns and the masses which belong to a *gene symbol* in rows.

The Wilcoxon rank sum test was applied to test the coherence between potentially significant patterns of markers related to BOS on the basis of mapped gene expression and mass spectrometry data. The correlation between the gene and protein expression of a cell was verified by using a Wilcoxon rank sum test [58].

The Wilcoxon rank sum test is an alternative to the *t-test* and assesses whether two samples of observations come from the same distribution. The two samples X and Y have to be independent and the observations have to be ordinal or continuous measurements. The null hypothesis $H_0 : x_{med} = y_{med}$ states that X and Y have the same mean value. Thus, the Wilcoxon rank test assumes that the values of X and Y are nearly equally distributed if the null hypothesis H_0 is valid. The test statistic T_w was built on the ranks of all observations $X_1, ..., X_n, Y_1, ..., Y_m$ (so called "pooled sample") so that $rg(X_1), ..., rg(Y_m)$ are obtained. This test statistic is defined as

$$T_w = \sum_{i=1}^{n} rg(X_i) = \sum_{i=1}^{n+m} iV_i \tag{2.24}$$

with

$$V_i = \begin{cases} 1 & \text{i-th observation of the pooled sample is X variable,} \\ 0 & \text{else.} \end{cases} \tag{2.25}$$

Different hypotheses were tested to examine the dependencies between the samples.

$$(a) H_0 : x_{med} = y_{med} \qquad H_1 : x_{med} \neq y_{med} \text{ (two-sided)} \tag{2.26}$$

$$(b) H_0 : x_{med} \geq y_{med} \qquad H_1 : x_{med} < y_{med} \text{ (one-sided)} \tag{2.27}$$

$$(c) H_0 : x_{med} \leq y_{med} \qquad H_1 : x_{med} > y_{med} \text{ (one-sided)} \tag{2.28}$$

Meta-analysis integrates gene and protein level by analyzing wether a combined approach outruns an isolated processing of both data types. One of the first approaches for meta-analysis was developed by Choi *et al.* in 2003 to compare microarray data of different platforms in order to find differentially expressed genes [44]. The R-package *GeneMeta* implemented the method described by Choi *et al.*, which considered the combination of two different sets of microarray data. This method was applied to the *Gene2Prot* and *MS* data from 23 patients and 6 healthy controls obtained from microarray and mass spectrometry data. The aim was to identify significant patterns associated with BOS which were not identified by analyzing individual studies alone.

The effect size model was assigned. To measure the true effect, it was important to eliminate the within-study variability and to calculate the between-study variability. μ denoted the parameter of interest (the average measure of difference) and y_i the observed effect size for independent studies $i = 1, 2, ...k$. The general model is given hierarchically as

$$y_i = \Theta_i + \varepsilon_i, \text{ with } \varepsilon_i \sim \mathcal{N}(0, s_i^2) \tag{2.29}$$

$$\Theta_i = \mu + \delta_i, \text{ with } \delta_i \sim \mathcal{N}(0, \tau^2), \tag{2.30}$$

where τ^2 representes the between-study variability and s_i^2 the within-study variability of study i [44].

Different models exist depending on whether or not between-study variability is non-vanishing. The fixed-effects model (FEM) assums $\tau^2 = 0$ which implies that the differences of observed effect sizes are from random sampling error alone and consequently $y_i \sim \mathcal{N}(\mu, s_i^2)$. The random-effects model (REM) explicitly accounts for differences between the studies with a study specific mean Θ_i and variance s_i^2. Furthermore, each δ_i is assumed to be drawn from some superpopulation with the overall mean μ and variance τ^2, thus $y_i \sim \mathcal{N}(\delta_i, s_i^2)$ and $\delta_i \sim \mathcal{N}(\mu, \tau^2)$. The homogeneity of study effects is tested to find out which model is appropriate for the data. This is equivalent to the hypothesis that τ^2 is actually zero [44]. The test of homogeneity was based on *Cochran's Q statistic* [45]:

$$Q = \sum w_i(y_i - \mu') \text{ with } w_i = s_i^{-2} \text{ and } \mu' = \frac{\sum w_i y_i}{\sum w_i}. \tag{2.31}$$

Under the hypothesis of homogeneity, *Cochran's Q statistic* follows a χ^2_{k-1} distribution. A large value of the *Q statistic* indicats a rejection of the hypothesis of homogeneity and the use of the REM model. This can be visualized in a *quantile-quantile (qq) plot* where a deviation from the diagonal indicats the use of a REM model.

Statistical significance of the meta-analysis was estimated by an algorithm similar to SAM [167] which was based on the concept of the false discovery rate (FDR). The comparison of FDRs of each of the two studies alone and the combined data set gave information about the significance of the combination.

Chapter 3

Results

3.1 Impact of DNA copy number changes on gene expression in neuroblastoma

In this study two already published neuroblastoma data sets were analyzed. In total 81 matching samples from NB patients from whom both aCGH data and gene expression data were investigated (sec. 2.1.1).

After several preprocessing steps of the gene expression as well as of the aCGH data, both data types were analyzed in an integrative step (sec. 2.1). A *consistency-matrix* was generated which reflected a correlation measure between the estimated DNA copy number of every chromosomal position and the coresponding gene expression value in *cis*-position for every patient . The following step resulted in a *similar-state-sum-matrix* which was tested for significance and served as an input to a BN approach based on Markov blankets. From here it was possible to identify *cis*- and *trans*-effects which took place in neuroblastoma.

3.1.1 Distribution of gene expression data after discretization

Discretization was used in order to get comparable distributions for the gene expression data and the aCGH data. The gene expression data were categorised into three categories by k-means discretization. These categories were: down-regulation (-1), no change (0) or up-regulation (1) of a gene, see section 2.1.2.1.

The resulting distribution of these discretized gene expression levels into one of the three categories is shown in table 3.1. About half of the genes (47.7 %) are assigned to the state "no change" whereas 29.5 % of the genes are in the state "down-regulation" and 22.8 % are assigned "up-regulation", respectively. .

Table 3.1: Distribution of gene states in percent after k-means discretization.

down-regulated (-1)	no change (0)	up-regulated (1)
29.5 %	47.7 %	22.8 %

3.1.2 Chromosome aberrations in neuroblastoma

Aiming to identify recurrent aberrations that are linked to neuroblastoma, the frequency of aberrations over all 81 patients was analyzed. Overall, several recurrent chromosome abberations previously described (Spitz et al., 2006) characteristic for neuroblastoma were detected. Frequent DNA losses were detected at 1p (32.1%), 8.p21 (45.8%), 9.q34 (40.7%), 11q (56.8%), 14.q32.31 (45.7%), 18.q21.33 (45.6). Gains were found 2.p24.3 (49.4%), entire chromosome 7 (39% - 53%), 11.q23.3 (50.7%) and 17.q (86.4%)

The losses and gains concerning the neuroblatoma data set were visualized as a frequency plot (Figure 3.1). Losses are highlighted in green and gains in red. This related

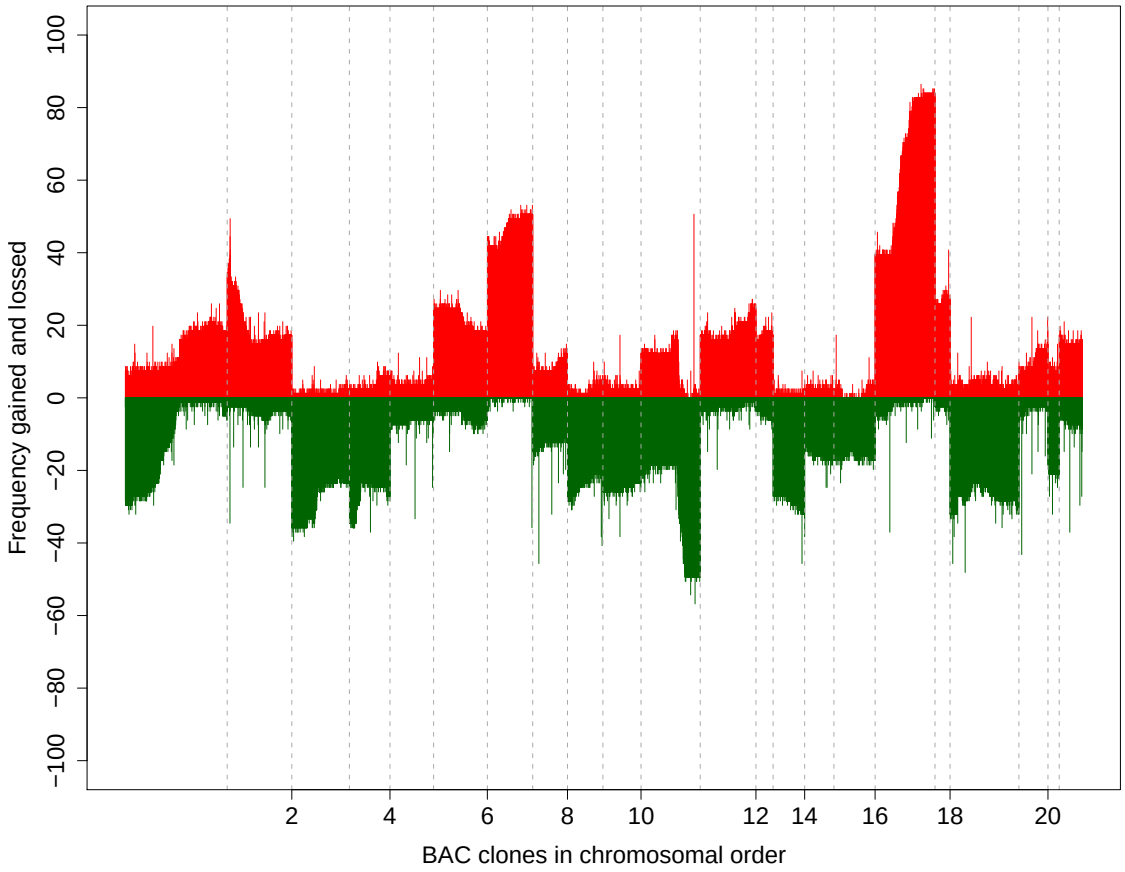


Figure 3.1: Frequency Plot of 81 neuroblastoma patients. Losses are displayed in green and gains in red. Chromosome boundaries are indicated by dashed lines.

to the following integrative step with the gene expression data, where typically a high gene expression level is coded in red and a low level in green.

3.1.3 Patient related cis-effects

The effects of chromosome aberrations on genes in *cis*-position related to neuroblastoma were studied. Matched probes represented on the customized 11k gene-expression-olignucleotide-microarrays to the probes of the high-resolution 44k oligonucleotide-aCGH-microarrays (sect. 2.1.2.2) were identified. A number of 1928 matching positions / genes were presented on both platforms.

The *consistency-matrix* was computed which contained for each patient and matching position the *consistency-score* (section 2.1.2.3). The *consistency-score* is a measure for the comparibility of chromosome aberrations with gene expression. It was computed patient-wise for each of the 1928 matching positions/genes. The results are represented in a matrix, the so called *consistency-matrix,* with positions/genes in rows and patients in columns.

We identified groups of patients with similar *consistency-score*s by one dimensional hierachical clustering of the *consistency-matrix* by using the euclidean distance and the complete linkage method. Positions/genes in rows were in order and only the patients were clustered (Fig. 3.2). The color coding is explained in Tab. 3.2.

The colored bars at the top of the colored map in Fig. 3.2 denote the values of the clinical variables: NB Status, MYCN and Stage. NB Status was subdivided into: darkblue - deceased, blue - alive without event, lightblue - alive with relapse/primary tumor; MYCN into white - not available (NA), gray - not amplified, black - amplified and stage into lightred - Stage 4S, darkred - Stage 4, purple - Stage 3, orange - Stage 2B, yellow - Stage 2A, blue Stage 2 and black - Stage 1.

5 different colors represent the *consistency-score*s: -4 in darkblue (aCGH loss, GE up); -3 in darkgreen (aCGH loss, GE down); 3 in red (aCGH gain, GE up); 4 (aCGH gain, GE down) in gray and -2 (aCGH down, GE no change), -1 (aCGH balanced, GE down), 0 (aCGH balanced, GE no change), 1 (aCGH balanced, GE up), 2 (aCGH up, GE no change) in white.

As can ben seen from Fig. 3.2 there was a group of patients that were characterized by a loss of DNA at chromosome 1 and also a down-regulation of genes in *cis*-position. The same is true for chromosome 3, 4, 9 to 11, 14 and 19. In contrast, chromosome 7 and 17 tend to hold regions were a gain of DNA corresponds to an up-regulation of genes in *cis*-position.

Patients with fatal outcome (dark blue, NB Status) seemed to suffer from the combined occurence of *cis*-effects on chromosome 1 and 17. The same is true for patients with an amplification of MYCN (black). Nearly all patients in Stage 4 hold distinct *cis*-effects at chromosome 7 and 17 (aCGH gain, GE up) and chromosome 11 (aCGH loss, GE down).

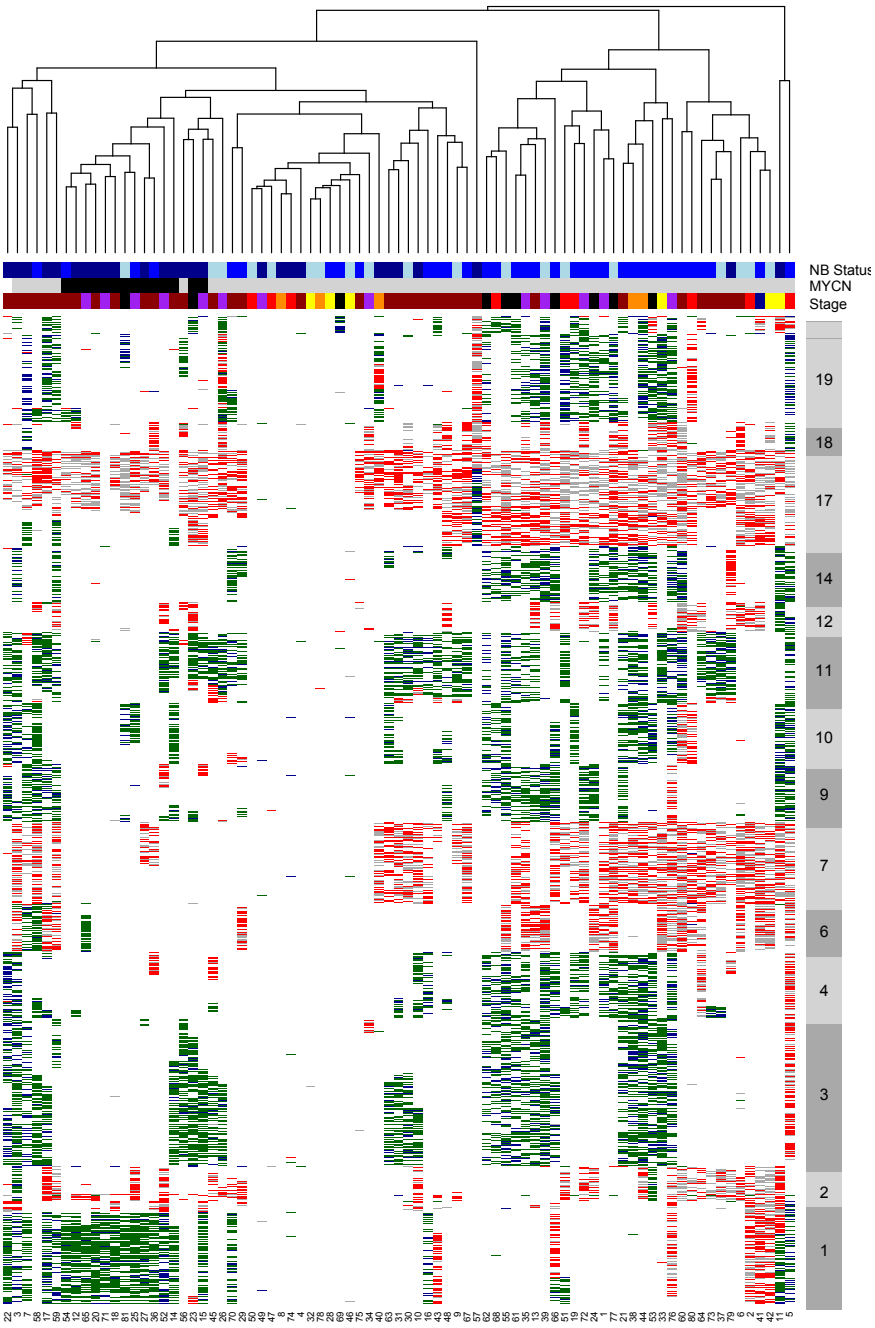
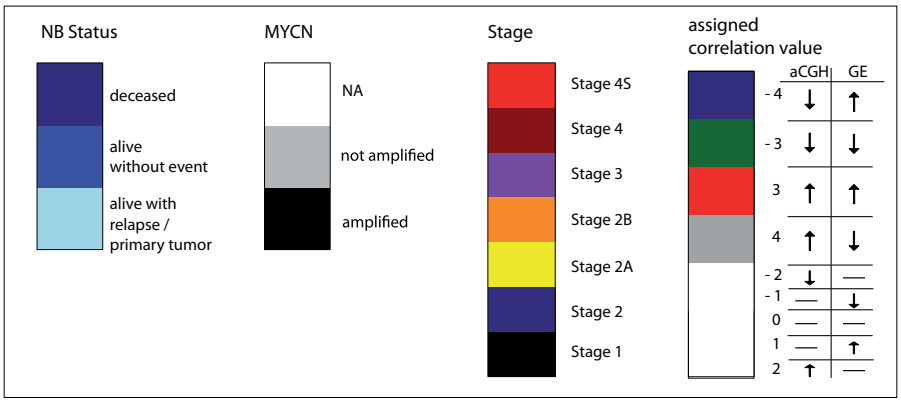


Figure 3.2: Heatmap of the *consistency-scores* in neuroblastoma . The colored bars at the top of the figure denote the values of the clinical variables: NB Status, MYCN and the Stage. Colors of the clinical variables as well as the color coded correlation values are explained in tab. 3.2. Chromosome boundaries are indicated by alternating light and dark gray bars at the right side.

Table 3.2: Color coding of the hierarchically clustered *consistency-scores*. NB Status, MYCN and Stage represent clinical variables. The assigned correlation values illustrate the color coded *consistency-scores* and are schematically explained with arrows on the right-hand side. An arrow pointing upwards denotes gain of a chromosomal region or up-regulation of gene expression, respctively. Arrows pointing downwards have analogous meaning. A horizontal line characterizes no change.



3.1.4 Identification of genomewide cis- and trans-effects via Bayesian Modeling

In sect. 3.1.3 *cis*-effects were computed patient-wise based on a *consistency-score.* This value describes a measure for every patient and estimates how well changes of DNA material correspond to gene expression of genes in *cis*-position.

By using BN, I sought to reveal additional trans-effects. This was done on the basis of all patients, i.e. effects got higher weights when they appeared in more patients. Especially *trans*-effects might have a role as regulators of many genes, see sect 1.4.1. Often they stay undiscovered in the background because it is hard to conclude which change of chromosome material affects changes of a gene expression level. In order to get more insight into this aspect, the following steps were applied to the paired neuroblastoma data set.

The dimensionality of the aCGH data was decreased in a two-step approach, (Sect. 2.1.2.5). First regions that showed no gained or lost chromosomal material in the genome of less than 20% of 81 patients were excluded. The second step compressed the overall amount of represented chromosomal locations to 462 *chromosomal location sets* (CLS).

As a measure of similarity between chromosome aberrations and changes in gene expression the s*imilar-state-sum ssm* was computed. This yielded a *similar-state-sum-matrix sssm* with aCGH probes in rows and gene probes in columns (Fig. 2.1). The *ssm* was tested for significance by computing an empirical p-value, with a threshold of $p < 0.01$.

A BN approach was used to analyze DNA copy number changes and their impact on gene expression. This method was based on Markov blankets which are a mininal set of nodes *d-separating* a node from all other (Sec. 2.1.2.8. The GS-algorithm (growth-shrinkage) iteratively computed the structure of the BN including computation of the Markov blankets, computation of the graph structure, orientation of edges, removing of cycles, reversing of edges and propagation of edge direction (Alg. 5).

At the end the structure was illustrated as a network (Fig. 3.3). Triangles denote chromosome aberrations and circles refer to genes. The colors reflect the characteristics of the nodes. Red color means that most patients ($> 50\%$) had a gain and up-regulation of that specific gene, and analogously for green color.

Two prominent changes in chromosome DNA influenced the topology of the network. Loss of genetic material at 11.q (highlighted by a green polygon) and gain at 17.q (highlighted by a red polygon) and were the main effectors. Biological relevance of the BN is discussed in section 4.

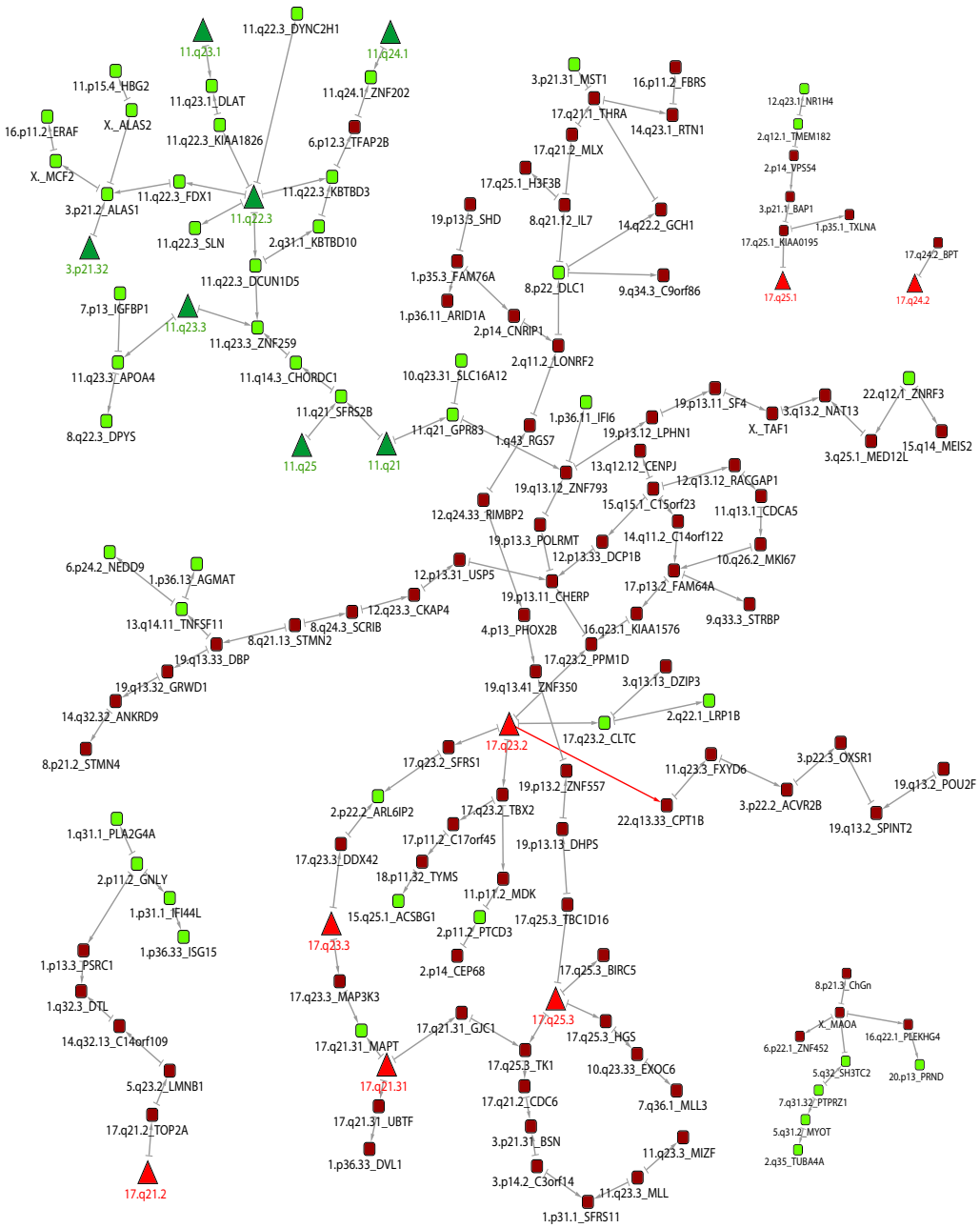


Figure 3.3: Bayesian network of *cis*- and *trans*-effects. Triangles represent chromosomal abberations and circles represent genes. The colors indicate the gene expression level respectively a gain or loss of chromosomal material (red = high/gain; green = low/loss).

3.2 Meta-analysis of genes and proteins identifies potential biomarker for BOS

3.2.1 Differently expressed genes and functional domains

SAM was used to detect significant changes in gene expression [167]. A one-class was performed on the loess-quantile-transformed data. This resulted in 1,306 significant genes.

A two-dimensional hierarchical cluster analysis with the 1,306 significant genes for the 23 patient samples and the 6 controls was performed (Fig. 3.4). The clustering was calculated using the Canberra distance (Eq. 2.16).

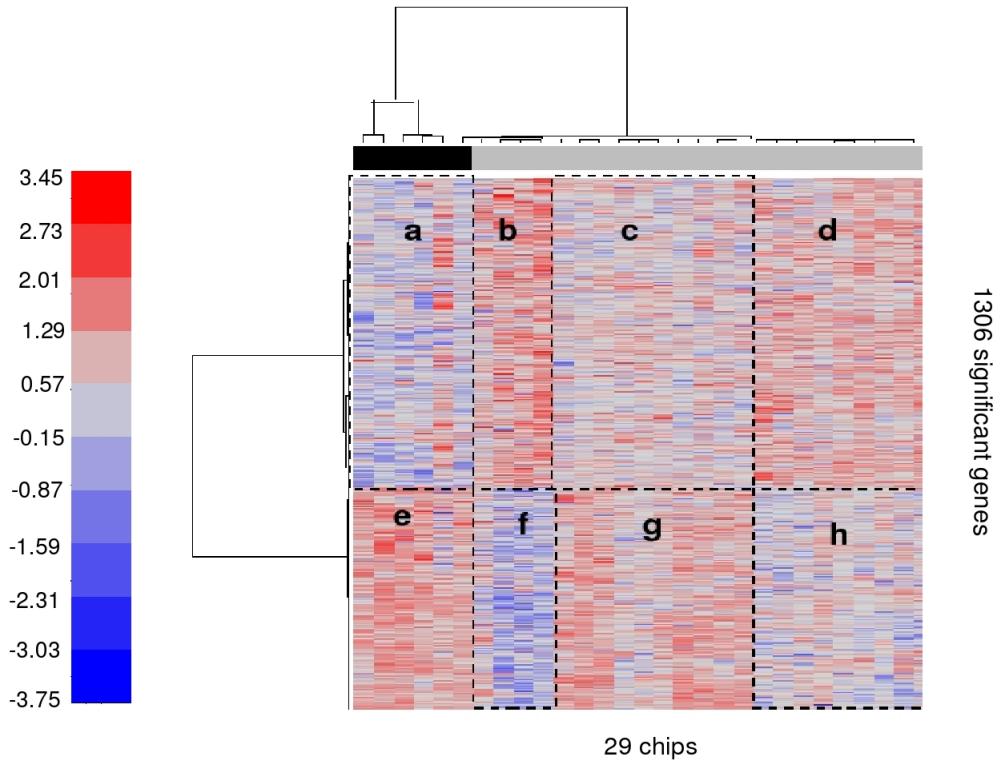


Figure 3.4: Heatmap representing the two-dimensional clustering of the 1,306 significant genes. Patients (gray) and controls (black) are shown in the bar. A clear separation between patients and controls exists. The blue color in the heatmap refers to downregulated genes and the red color to upregulated genes. The black dashed boxes a to h indicate that the clusters a and e have a similarity to clusters c and g, and clusters b and f are similar to d and h.

Identification of significant functional domains was based on gene ontology (GO) annotation. The biological functions of the 1,306 significant genes are visualized in Fig. 3.5. Most of the significant genes were involved in the induction of apoptosis and it is positive regulation.

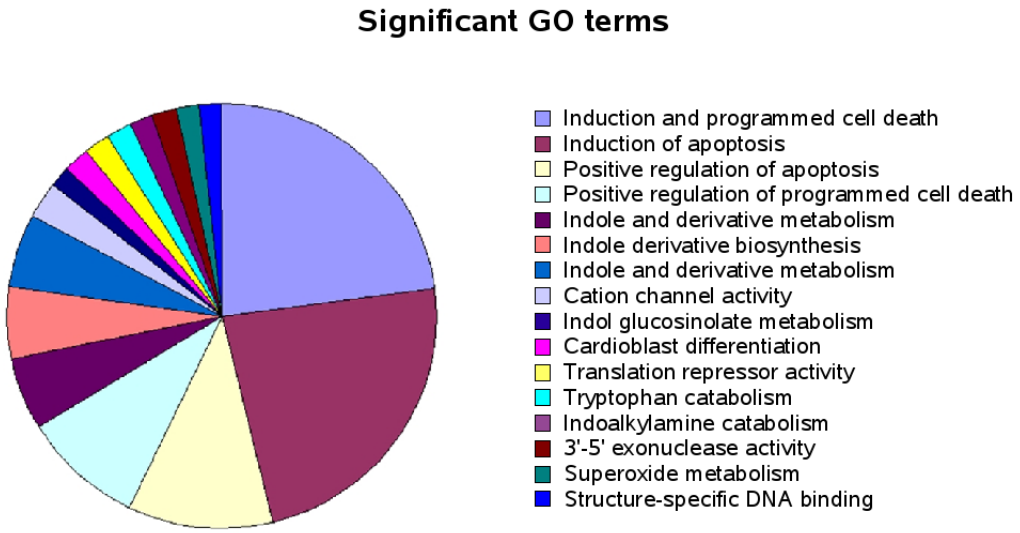


Figure 3.5: Pie chart showing the proportion of molecular functions of 1,306 significant genes.

3.2.2 Detection of significant proteomic patterns

For the detection of significant peaks, all different possible split is between pairs of classes were calculated to find proteomic patterns (Tab. 2.3). MS data of 52 patients (stage 01, including time series data of stage 02 and 03) and 10 controls (nt) were used for sample classification.

After preprocessing of the mass spectra a SVM coupled to RFE in a n-fold cross validation step was applied (Fig. 2.8). In parallel, analysis by SVM without feature selection, SVM with RFE, SVM with BaHSIC and PAM was performed and as well as their accuracy, sensitivity and specificity calculated. The most balanced classification results based on accuracy, sensitivity and specificity, were detected for SVM coupled to RFE on the data at 1-10k Da (Tab. 3.3).

Table 3.3: Accuracy, sensitivity, and specificity for classification by SVM coupled to RFE on data at 1-10kDa.

Combination (1-10kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.83	0.90	0.85
Stage 01 vs nt	0.73	0.73	0.73
Stage 02 vs nt	0.75	0.76	0.73
Stage 03 vs nt	0.68	0.71	0.66
Stage 01 vs 02	0.52	0.52	0.52
Stage 01 vs 03	0.41	0.42	0.39
Stage 02 vs 03	0.49	0.52	0.48

Table 3.4: Accuracy, sensitivity, and specificity for classification by SVM coupled to RFE on data at 8-20kDa.

Combination (8-20kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.57	0.60	0.53
Stage 01 vs nt	0.58	0.57	0.61
Stage 02 vs nt	0.61	0.62	0.60
Stage 03 vs nt	0.58	0.59	0.57
Stage 01 vs 02	0.46	0.47	0.45
Stage 01 vs 03	0.57	0.61	0.54
Stage 02 vs 03	0.47	0.47	0.47

SVM without RFE performed considerably worse (Tab. 3.5 and 3.6).

Table 3.5: Accuracy, sensitivity, and specificity for classification by SVM without feature selection on data at 1-10kDa.

Combination (1-10kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.55	0.55	0.55
Stage 01 vs nt	0.68	0.64	0.73
Stage 02 vs nt	0.66	0.60	0.71
Stage 03 vs nt	0.53	0.53	0.53
Stage 01 vs 02	0.53	0.54	0.53
Stage 01 vs 03	0.52	0.52	0.52
Stage 02 vs 03	0.58	0.52	0.64

Table 3.6: Accuracy, sensitivity, and specificity for classification by SVM without feature selection on data at 8-20kDa.

Combination (8-20kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.66	0.75	0.60
Stage 01 vs nt	0.68	0.70	0.68
Stage 02 vs nt	0.66	0.70	0.66
Stage 03 vs nt	0.51	0.55	0.49
Stage 01 vs 02	0.40	0.42	0.37
Stage 01 vs 03	0.41	0.00	0.46
Stage 02 vs 03	0.49	0.53	0.51

The analyzed range from 8-20 kDa did not show any discriminative pattern (Tab. 3.4).

SVM with BaHSIC resulted in similar values for accuracy, sensitivity and specificity as compared to SVM with RFE (Tab. 3.7 and Tab. 3.8). However, the estimated significant peaks by BaHSIC were an m/z range where no peak could be visually confirmed.

Table 3.7: Accuracy, sensitivity, and specificity for classification by SVM coupled to BaHSIC on data at 1-10kDa.

Combination (1-10kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.81	1.00	0.74
Stage 01 vs nt	0.83	0.97	0.77
Stage 02 vs nt	0.85	1.00	0.77
Stage 03 vs nt	0.74	0.77	0.70
Stage 01 vs 02	0.83	0.87	0.80
Stage 01 vs 03	0.50	0.50	0.50
Stage 02 vs 03	0.51	0.52	0.50

Table 3.8: Accuracy, sensitivity, and specificity for classification by SVM coupled to BaHSIC on data at 8-20kDa.

Combination (8-20kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.66	0.70	0.61
Stage 01 vs nt	0.67	0.72	0.60
Stage 02 vs nt	0.72	0.79	0.63
Stage 03 vs nt	0.78	0.83	0.72
Stage 01 vs 02	0.76	0.80	0.72
Stage 01 vs 03	0.51	0.51	0.51
Stage 02 vs 03	0.62	0.59	0.65

PAM had accuracy and sensitivity similar to SVM with RFE. But could not achieve as good results for the sensitivity and specificity (Tab. **??** and Tab. 3.10).

Table 3.9: Accuracy, sensitivity, and specificity for classification by PAM on data at 1-10kDa.

Combination (1-10kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.97	1.00	0.70
Stage 01 vs nt	0.88	0.96	0.50
Stage 02 vs nt	0.83	0.96	0.80
Stage 03 vs nt	0.86	0.92	0.00
Stage 01 vs 02	0.65	0.98	0.00
Stage 01 vs 03	0.81	1.00	0.00
Stage 02 vs 03	0.64	0.93	0.00

Table 3.10: Accuracy, sensitivity, and specificity for classification by PAM on data at 8-20kDa.

Combination (8-20kDa)	Accuracy	Sensitivity	Specificity
All patients (01, 02, 03) vs nt	0.79	0.98	0.00
Stage 01 vs nt	0.84	1.00	0.00
Stage 02 vs nt	0.65	0.91	0.00
Stage 03 vs nt	0.55	0.64	0.44
Stage 01 vs 02	0.66	1.00	0.00
Stage 01 vs 03	0.81	1.00	0.00
Stage 02 vs 03	0.68	1.00	0.09

The intersection of the most significant peaks for each method, SVM with RFE, SVM with BaHSIC and PAM were formed. This resulted in 7 highly significant peaks which are listed in Tab. 3.11. Two prominent peaks were detected at 1170 Da (galanin-like peptide precursor; Fig. 3.6) and 2160 Da (actin-related protein fragment; Fig. 3.7)

Table 3.11: The seven most significant peaks. Two proteins have been identified (1. and 2.), the other six are in the process of identification.

1. Peak at 1170 Da (exactly detected at Mw 1169.75 Da in lab). This peak has been identified as galanin-like peptide precursor. Identified by SVM-RFE (02 vs nt), PAM (02 vs nt), SVM-RFE (03 vs nt) and PAM (03 vs nt) (Fig. 3.6).

2. Peak at 2160 Da (exactly detected at Mw 2159.07 Da in lab). This peak has been identified as actin-related protein fragment in human. Identified by SVM-RFE (01 vs nt) (Fig. 3.7).

3. Peak at 3487 Da. A protein with matching Mw is the Neutrophile alpha-Defensine 3/human neutrophil peptide (HNP) 3. It has a possible participation in inflammation processes in chronical repulsion of transplants and has already been identified in BALF proteomes of patients [118]. Identified by SVM-RFE (patients vs nt), PAM (patients vs nt), SVM-RFE (03 vs nt) and PAM (03 vs nt).

4. Peak at 4135 Da. The corresponding protein to this peak has not yet been identified, but Zhang et al. [187] also detected this peak which was correlated with chronic lung transplant rejection. Identified by SVM-RFE (patients vs nt), PAM (patients vs nt), SVM-RFE (03 vs nt) and PAM (03 vs nt).

5. Peak at 4965 Da. The corresponding protein has not yet been identified. Zhang et al. [187] also detected this peak which appears in control samples and disappears over time in samples of patients who had a lung transplantation. Identified by SVM-RFE (patients vs nt), PAM (patients vs nt), SVM-RFE (03 vs nt) and PAM (03 vs nt).

6. Peak at 10803 Da. A protein with matching Mw is Calgranulin A/MRP-8, a macrophage-cytokines which is upregulated in chronical inflammation processes [4]. Calgranulin have been observed in conjuction with HNP in other body fluids with associated infections. Both are part of the immune response system [70]. Identified by SVM-RFE (02 vs nt) and PAM (02 vs nt).

7. Peak at 13792 Da. A protein with matching Mw is Transhyretine which is an anti-acute-phase protein [187]. Identified by SVM-BaHSIC (03 vs nt) and PAM (03 vs nt).

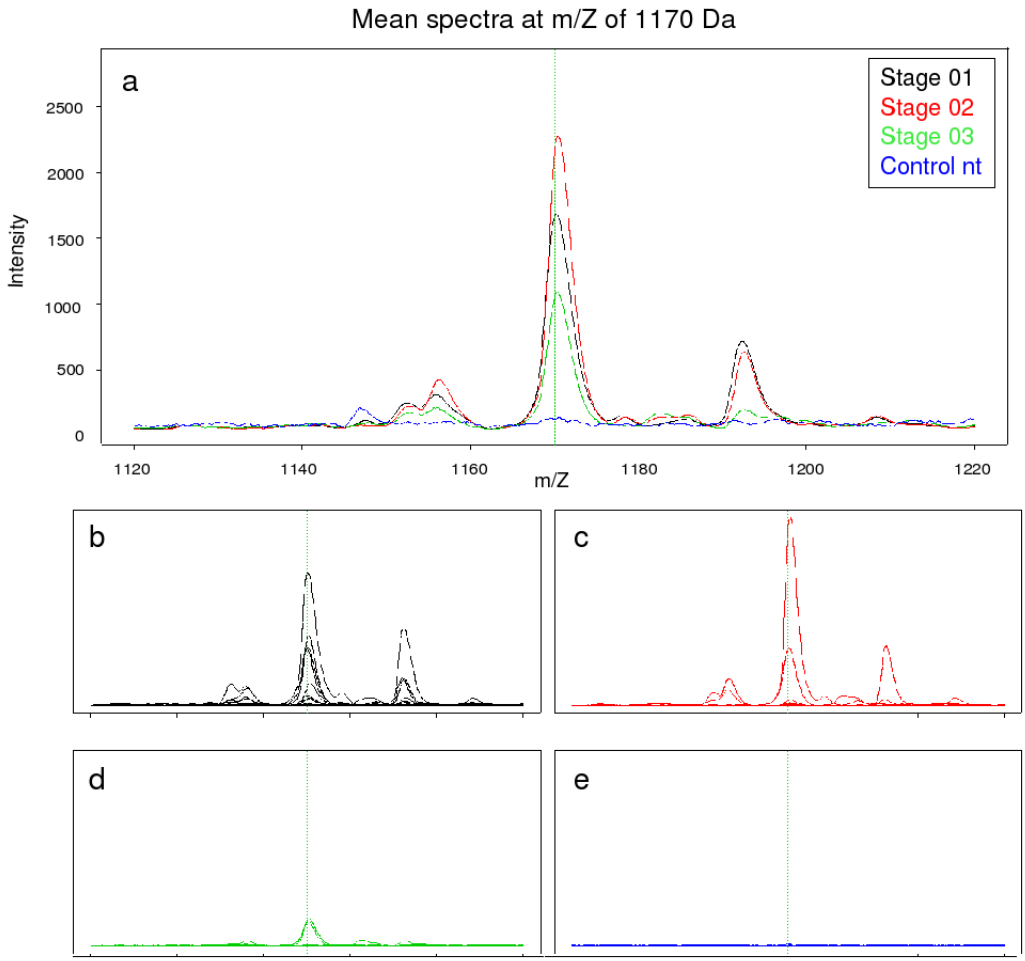


Figure 3.6: Peak at 1169 Dalton. Identified by SVM-RFE (02 vs nt), PAM (02 vs nt), SVM (03 vs nt), and PAM (03 vs nt). The upper plot represents the mean spectrum of the respective patient stage / control. The lower plot presents all spectra of patients in that specific stage. Black stage 1; red stage 2; green stage 3; blue stage nt (all controls). The plot on top (a) shows the four mean spectra of every stage. Mean spectra are composed of the spectra at the bottom: (b) indicates all the spectra of patients at stage 01 which result as mean spectrum (black) in the top image (a); (c) consists of the spectra of patients at stage 02 which contribute to the red mean spectrum in (a); (d) contains the spectra of patients at stage 03, the mean spectrum is shown in green in (a), (e) covers the spectra of controls (nt), which are presented as the blue mean spectrum in (a).

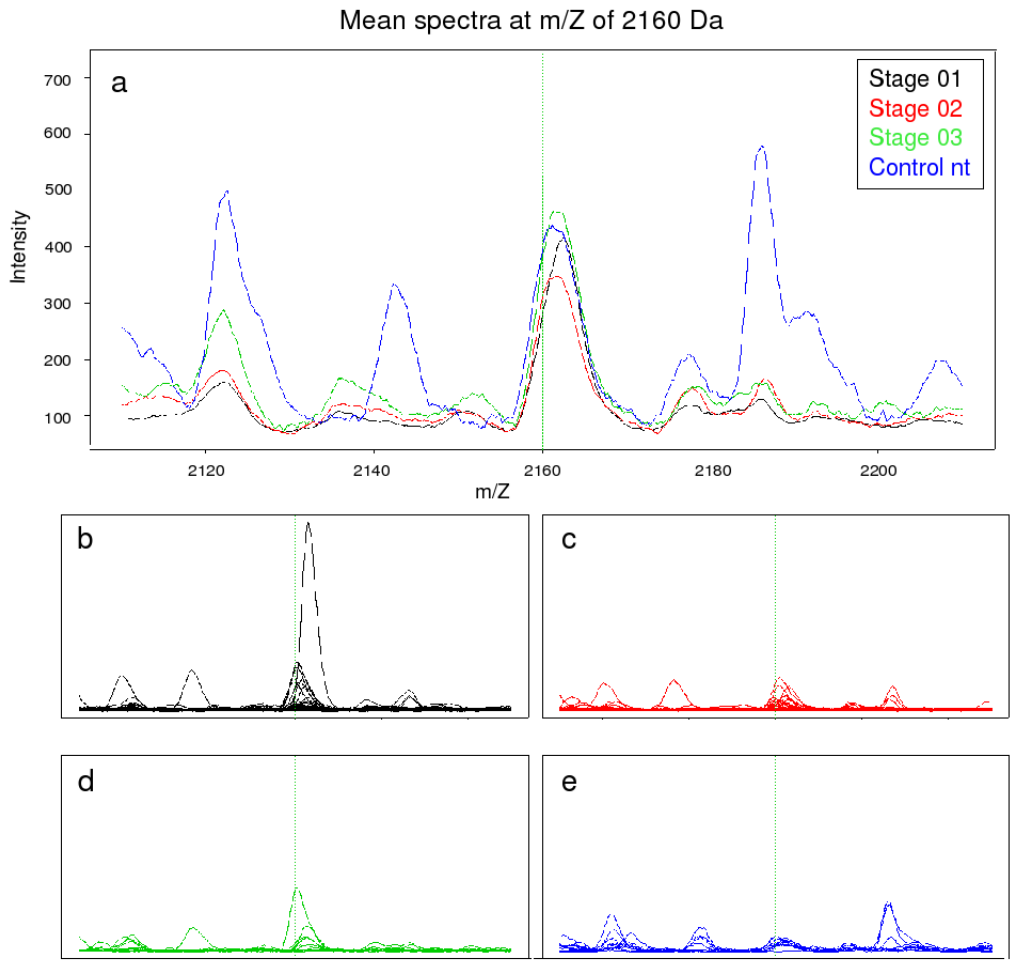


Figure 3.7: Peak at 2160 Dalton. Identified in in SVM-RFE (01 vs nt). See legend of Fig. 3.6.

3.2.3 Integrative analysis results in novel peaks

All genes of the microarray experiment, with *gene symbol* as identifier, were *translated* into their corresponding proteins by using the *Gene ID Conversion Tool*. These proteins all had *UniProt* identifiers, so the corresponding theoretical isoelectric point (pI) and molecular weight (Mw) values could be computed by the *Compute pI/Mw tool*. The molecular weights which were derived ranged from 443 to 869,000 Da whereas most of these weights ranged from 50,000 to 150,000 Da.

For the existent mass spectrometry profiles of 1-10 kDa and 8-20 kDa, 112 *Gene2Prot* Mw values mapped to the 1-10kDa scale, and 324 values to the 8-20 kDa scale. For these matchings, the charge in the mass-to-charge ratio (m/z) was assumed to be 1, which is usually the case with laser-assisted ionisation applied here. The subsets of Mw values served as input for further analyzes where only the red channel of the gene expression experiments was used in both cases. The loess-quantile-normalized data were discretized in ten equal-sized quantiles and averaged across either the patient or control group.

3.2.3.1 Virtual spectrum

The matrix had 23 patient and 6 control samples in columns and 112 mass values in rows for the data at 1-10kDa. For data at 8-20 kDa, the matrix had 324 mass values in rows. Depending on whether the matrix referred to gene expression or mass spectrometric values, it was denoted as 'Gene2Prot1-10', 'Gene2Prot8-20', 'MS1-10' or 'MS8-20' in subsequent sections.

The different matrices of patients or controls were mapped onto each other and plotted as virtual spectra and corresponding mass spectrometric data (Fig. 3.8to 3.11).

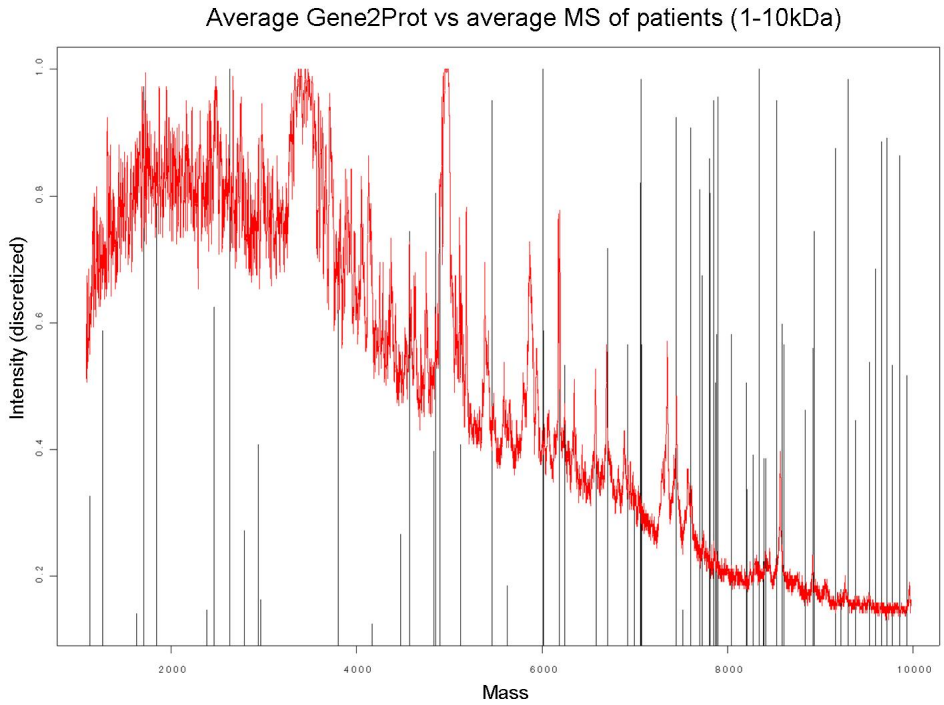


Figure 3.8: Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of patients for data at 1-10kDa.

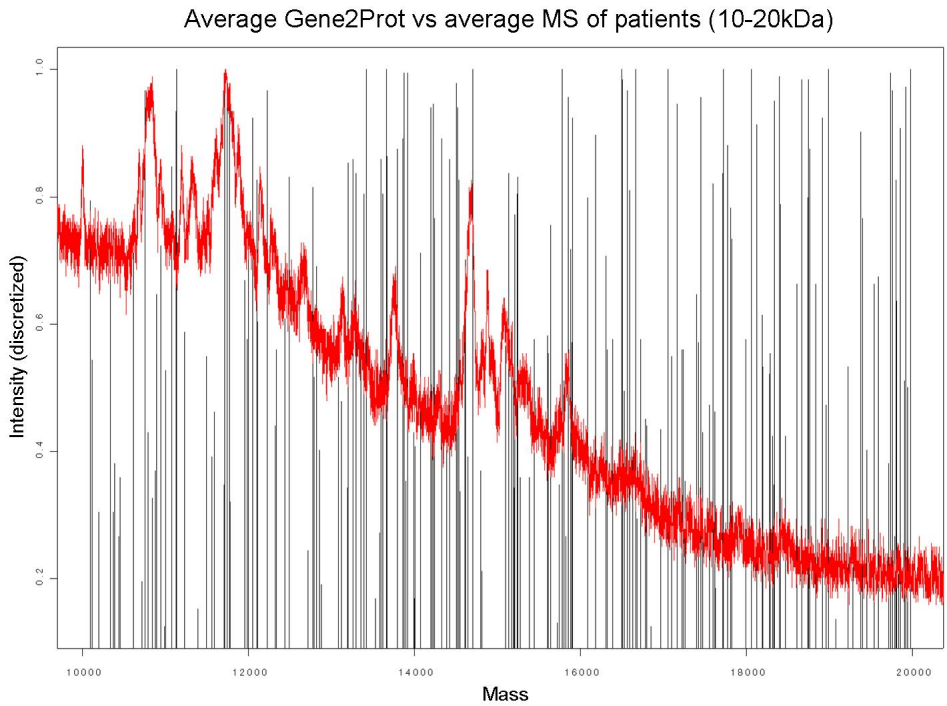


Figure 3.9: Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of patients for data at 8-20kDa.

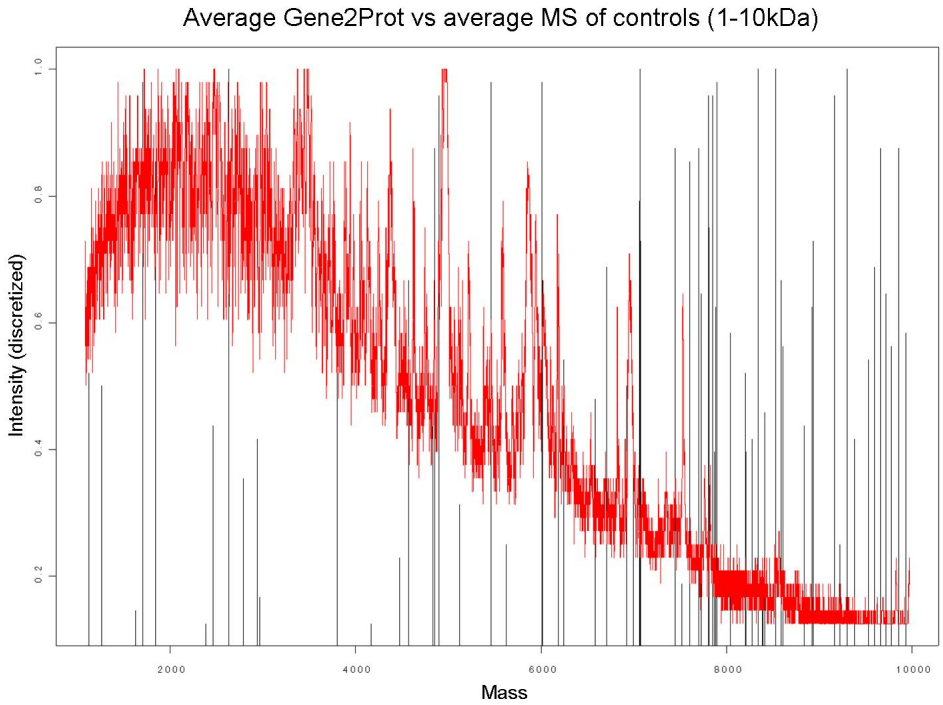


Figure 3.10: Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of controls for data at 1-10kDa.

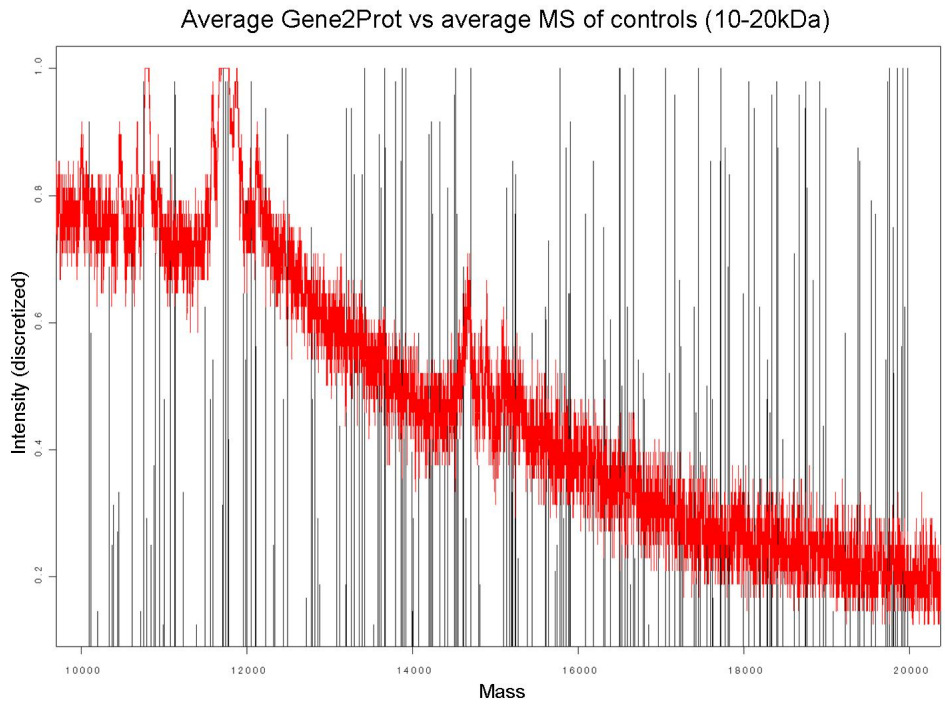


Figure 3.11: Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of controls for data at 8-20kDa.

3.2.3.2 Wilcoxon rank sum test

The *Wilcoxon rank sum test* was applied to Gene2Prot1-10 versus MS1-10 and to Gene2Prot8-20 versus MS8-20 in order to reveal dependencies between the data and formulate hypotheses on these data, like Gene2Prot (eq.3.1, eq. 3.2 and eq. 3.3). The outcome showed that is was not possible to determine similarities between the data.

$$(a) H_o : Gene2Prot_{med} = MS_{med} \quad H_1 : Gene2Prot_{med} \neq MS_{med} \tag{3.1}$$

$$(b) H_o : Gene2Prot_{med} \geq MS_{med} \quad H_1 : Gene2Prot_{med} < MS_{med} \tag{3.2}$$

$$(c) H_o : Gene2Prot_{med} \leq MS_{med} \quad H_1 : Gene2Prot_{med} > MS_{med} \tag{3.3}$$

3.2.3.3 Cross platform integration

Meta-analysis was applied to *Gene2Prot* and *MS* data. In order to decide whether a FEM or REM model is more appropriate for combining the data, *Cochran's Q statistic* was calculated for each mass value. Under the assumption that the differences in the effect sizes between studies was due to sampling error alone, the values for Q distributed according to a χ^2 distribution. The *qq-plot* for data at 1-10kDa for quantiles of the observed values of Q and the quantile of a χ^2 distribution are shown in Fig. 3.12. The deviation of the observed Q values from χ^2 distribution (diagonal) indicated to choose a REM model. The *qq-plot* for data at 8-20kDa looked similar (not shown). .

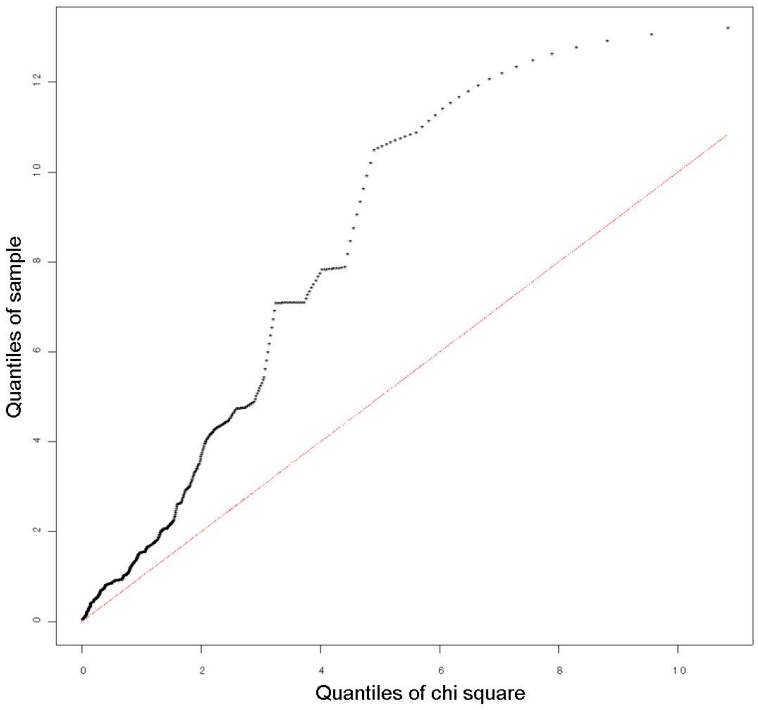


Figure 3.12: QQ Plot of data at 1-10kDa

Statistical significances of the meta-analysis were calculated for each of the two studies *Gene2Prot* and *MS* alone and for the combined data set (Fig. 3.13). The plot of data at 8-20kDa were not shown because the meta-analysis did not yield sufficient improvement compared to data at 1-10kDa. Throughout this analysis, 15 significant masses with a

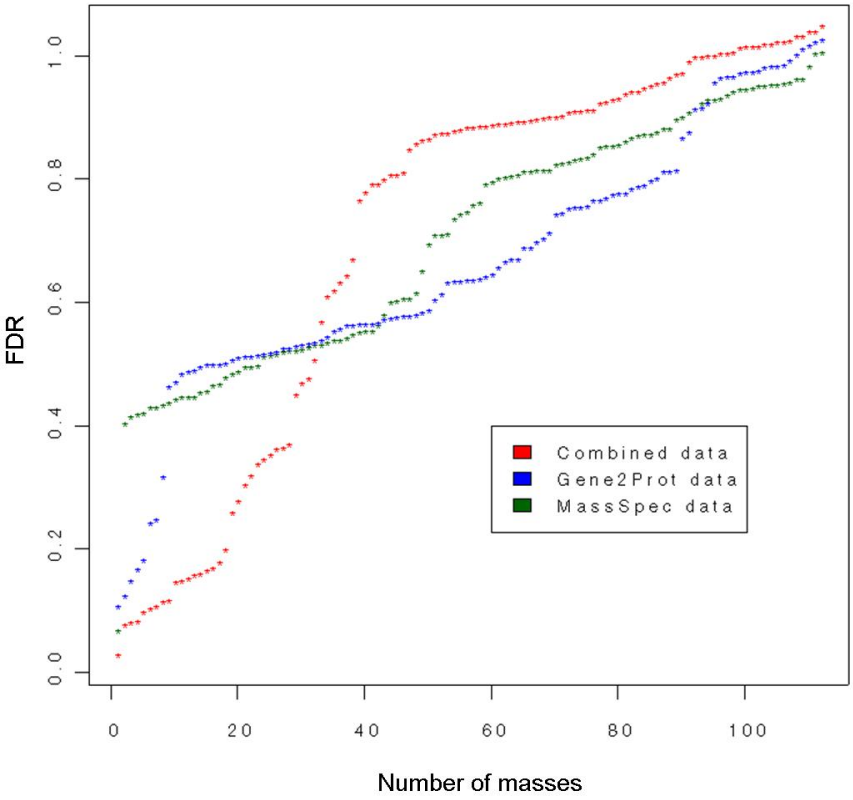


Figure 3.13: FDR Plot of data at 1-10kDa

FDR $\leq$0.2 were identified. These masses were found exclusively in the meta-analysis of the combined set. Tab. 3.12 lists these masses. The corresponding *geneSymbol*, FDR of *Gene2Prot*,v FDR of *MS*, and the FDR of combined set are listed, too. The meta-analysis on 8-20kDa data yielded no significant results (data not shown).

Table 3.12: Masses with gene symbol identified by meta-analysis of *Gene2Prot* and *MS* data at 1-10kDa with corresponding false discovery rate.

Mass	Gene Symbol	FDR(Gene2Prot)	FDR(MS)	FDR(Combined set)
2789 Da	INSL4	0.60	0.11	0.02
4572 Da	CPSF4	0.39	0.49	0.13
5123 Da	KIF14	0.43	0.60	0.10
5458 Da	CHD4	0.53	0.52	0.17
6010 Da	MGC18216	0.51	0.24	0.07
6013 Da	EVL	0.51	0.10	0.07
6707 Da	MPHOSPH6	0.51	0.56	0.15
6996 Da	LHX4	0.74	0.31	0.15
7446 Da	CCL15	0.42	0.53	0.09
7448 Da	FXYD3	0.44	0.50	0.14
7519 Da	FXYD2	0.54	0.66	0.16
7900 Da	LY6E	0.43	0.45	0.07
8343 Da	ACPP	0.42	0.57	0.14
8395 Da	CCL26	0.48	0.50	0.10
9598 Da	KLRC4	0.48	0.49	0.19

Chapter 4

Discussion

The main intention for writing this thesis was to contribute to the regulation on the flow of genetic information from DNA through mRNA to proteins [66]. Here two bioinformatic methods were presented which allow for an integrative analysis of genomic, transcriptomic and proteomic data. This task was split into two subparts. The first one included the analysis of gene expression patterns as a function of DNA copy number aberrations in neuroblastoma. A Bayesian approach gave insights into mechanisms of genetic processes triggered by alterations in DNA copy number. The second part focusses on improving the validity of gene and protein expression patterns related to the bronchiolitis-obliterans-syndrome by a meta-analysis approach. It was shown that an integrative analysis of both data types is superior to the results obtained by analyzing either data set individually.

4.1 Integrative analysis of genomic and transcriptomic data related to neuroblastoma

Two previously published neuroblastoma data sets including 81 patients sets were analyzed. The data were collected within aCGH and gene expression studies and analyzed in an integrative step. The gene expression data were discretized gene-wise into three categories (down-regulation, no change, up-regulation) by k-means clustering. This algorithm had weaknesses when all continuous values of a specific gene belong in principle to one category. In that case, the k-means algorithm returns still three categories for that gene. The k-means method is a data-driven method, which outperforms other methods like quantile or range discretization, where for each category an equal number of data values are mapped to equal-size bias. Other studies used self-organizing-maps in order to find the optimal number of categories, but that was no option in this study because a fixed number of 3 categories was choosen. Here k-means discretization was the option of choice because it was assumed that the gene expression data were normally distributed and hence all of the three categories were represented in the measured data for one gene. Limitations of this approach are the hard thresholding in discretization.

By analyzing the aCGH data, a number of 10 distinct aberrations of DNA copy numbers that took place in neuroblastoma were detected. Of those, the most frequently lost chromosomal region was deteted at 11q (56.8%). The most prominent gained region was 17q (86.4%). These results confirm the results that were obtained with a bigger set of neuroblastoma samples of which the samples used here were a subset [152]. Gain of chromosome 17 and loss of chromosome arm 11q are, besides loss of chromosome arm

1p, the most frequent abnormality detected in neuroblastoma [128, 151].

The impact of chromosome aberrations in neuroblastoma on genes located in *cis*-position was analyzed patient-wise. A *consistency-score* was computed for every sample. This value was a measure for the correlation between DNA copy number changes and the expression of genes in cis-position. The results of the one-dimensional hierachical clustering of the *consistency-matrix* revealed so far unknown interactions between chromosome aberrations and genes. For the first time it was demonstrated in a very high granularity how DNA copy number changes affect genes in *cis*-position. This method outperformes other methods, e.g. binning methods, were often a mean is taken over a large range of chromosomal regions. The results of clustering the *consistency-matrix* confirmed the results made by other groups. Known combinations of genetic changes, including 17q gain, and deletion of 1p and 11q are illustrated by the colored map resulting from clustering. Chromosome 1 was identified as a domain of lost chromosome material that came along with a down-regulation of genes at the same locus. That became in particular true for MYCN amplified patients. This group of patients was also characterized by gained chromosome material on chromosome 17 and a down-regulation of genes located in *cis*-position. These results hint towards a cross-relationship between chromosome aberrations at chromosome 17 and genes located at other genomic locations, especially chromosome 1. Other studies of neuroblastoma have revealed a high frequency of unbalanced translocations of chromosome 17. In consequence, genetic information on the partner chromosome can be lost. Prominent partner chromosomes are chromosome 1 and 11q. Especially patients in stage 4 seemed to suffer from a gain of chromosome 17 that came along with a loss of chromosome 11. Despite the aberrations the results demonstrated that some genes located at 11 are upregulated. This hints towards potential *trans*-effects that may arise from interactions between chromosome 11 and 17. See discussion below.

Trans-effects were identified by a *similar-state-sum* which was a measure of similarity for each gene expression probe to any other aCGH probe. That was achieved by summing up all equal states over all patients for the respective pair of analysis. An empirical p-value served as a selection criterion. By doing so, the amount of genes and chromosome aberrations was dramatically minimized and served as input variables for the BN approach. Here the GS-Algorithm was used to learn the structure of the BN. The main advantage of the algorithm comes through the use of Markov blankets to restrict the size of the conditioning sets. In order to determine the existence of an edge between two nodes, Markov blankets gave a measure for the association of the two distributions coming either from gene expression or aCGH data. The "direct neighbors" step, which does a number of dependence tests between X and Y and declare them direct neighbors only if all these tests have high confidence, helps to identify potential errors in the preceding Markov blanket phase. Integration of prior knowledge e.g. that an arc from gene nodes were not allowed to point to an aCGH node, helped to reduce the complexity of the model. Limitations are given by unobserved variables, e.g. miRNAs or other ncRNA that have an impact on the correlation results presented here.

The results of the BN are illustrated in Fig. 3.3. Focusing on highlighted in green, there are 13, in most cases down-regulated, genes which directly influenced by chromo-

somal losses at 11.q. These genes were either in *cis*-position or in direct chromosomal neighborhood (GPR83, SFRS2B, CHORDC1, ZNF259, APOA4, DCUN1D5, KBTBD3, ZNF202, SLN, FDX1, HBG2, KIAA1826, DLAT) . Interestingly the transcription factor TFAP2B (transcription factor AP-2 beta) which is located on 6.p.12.3 has an arc to another transcription factor ZNF202 (zinc finger protein 202) on 11.q.24.1. It is notable that TFAP2B is the only up-regulated gene connected to only down-regulated genes that characterized by loss of chromosomal material. Another interesting characteristic is that by a loss of 11.q.23.3 is connected the transcription factor ZNF259 and to APOA4 (apolipoprotein A-IV). APOA4 is known to bind to ZNF202, also located at 11.q, and is also known to be directly regulated by STAT3 (not in network), which is located at chromosome 17.q.21.2 [171, 142].

The chromosome aberrations at 17.q, highlighted in red, influenced a large part ot the network topology. It is also the carrier of the only real *trans*-effect that took place in the interaction of chromosome aberrations and the resulting gene expression. The Bayesian model predicted that a loss of 17.q.23.2 has a direct influence on the gene CPT1B (carnitine palmitoyltransferase 1B, 22.q13.33). CPT1B it iself points to the gene FXYD6 (11.q.23.3). It is reported to bind to TP53 (17.p13.1, not in network). FXYD6 was linked to OXSR1 bridged by ACVR2B. OXSR1 is known to directly regulate TP53 [153]. In the network topology OXSR1 is also connected to SPINT2 (serine peptidase inhibitor, Kunitz type 2, 19.q13.2). The SPINT2 protein is known to decrease the activation of human Erk protein which is known to increase phosphorylation of the p53 protein. TP53 protein mediates the activation of the human Erk protein [190].

A gain of 17.q.23.2 is connected to up-regulation of TBX2 (T-box 2) in *cis*-position. This genes encodes a transcription factor involved in the regulation of developmental processes. Expression studies indicated that this gene may have a potential role in tumorigenesis as an immortalizing agent. There is also an undirected *trans*-effect reaching from the gain 17.q.23.2, over TBX2 to MDK (midkine) on 11.p.11.2 (this chromosome region is lost). MDK is known to be involved in signal transduction and the development of the nervous system.

The gain of 17.q23.3 correlates with up-regulation of DDX42 in cis-position. This gene encodes a member of the Asp-Glu-Ala-Asp (DEAD) box protein family. Members of this protein family are putative RNA helicases, and are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Members of this family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division (provided by RefSeq).

Other cis-regulated genes are MAP3K3 (mitogen-activated protein kinase 3) and TOP2A (DNA topoisomerase II alpha). MAP3K3 is located inside the gained location 17.q.23.3, and the up-regulated gene TOP2A is located at 17.q21.2. TOP2A encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This nuclear enzyme is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication. It catalyzes the transient breaking and rejoining of

two strands of duplex DNA which allows the strands to pass through one another, thus altering the topology of DNA (provided by RefSeq).

4.2 Integrative analysis of transcriptomic and proteomic data related to BOS

Proteomic and transcriptomic data were analyzed in an integrative analysis. First gene expression and the mass spectrometry data were analyzed by their own. After preprocessing steps, 1,306 genes were detected as differently expressed by one-class SAM. Two-dimensional clustering resulted in two distinct subgroups and separated controls from patients. Different gene clusters hint towards BOS-related gene expression patterns. To get more insights into the functional role of these genes they was searched for significant functional gene ontology terms. The results suggested that especially apoptosis-inducing genes are involved in the development of LTX.

In addition the proteome of the controls and patients was analyzed. Different machine learning algorithms were used with and without feature selection methods. Support vector machines in combination with recursive feature elimination performed best and attained a classification accuracy of 83 % with 90 % sensitivity and 85 % specificity. When merging protein peaks with discriminative power over all used classification methods, 7 highly BOS-related peaks were identified. Galanin-like-peptide precursor and actin-related protein fragment were particulary noticeable.

To improve the analysis, an integrative step was carried out. The genes were converted to protein information comprising their isoelectric point and their molecular weights. This was deemed as virtual spectrum and compared with the corresponding mass spectrometry protein weights. A Wilcoxon rank test proved no differences in means between the virtual spectra and real spectra. It can be concluded that no dependencies are given between genes and their expressed proteins. However it must be considered that mass spectrometry also detects protein fragments which was not taken into consideration when building up the virtual spectra. Alternative splicing and post-transcriptional modification of proteins interfere with this integrative analysis. These events lead to complication with this method and seemd to have more explanatory power than the conclusion that the gene expression has no influence on protein expression.

The meta analysis approach integrated the virtual and real spectra. FDR was used to estimate the statistical significance. This was done for each spectrum type alone and also for the combined data set. The latter resulted in 15 masses which were found exclusively in the meta analysis of the combined set. Apparently meta-analysis increases the statistical power and thus generates more significant results in comparison to each data set alone.

In summary two integrative methods were presented that combined data derived from different levels of genetic information processing. These levels were: chromosome aberrations of DNA, gene expression and protein expression. The Bayesian approach called

BNtegrative offered new insights into the understanding of how chromsomal changes influence gene expression in neuroblastoma either in *cis*- or in *trans*-position. The second approach, based on meta-analysis of real and virtual spectra of BOS, resulted in outcomes that would not have been achieved when analyzing both data sets on their own. These results need to be validated experimentelly. Both methods, BNtegrative as well as meta analysis of virtual spectra are generic and can be used for any kind of tumor type or disease. Even similar array-based molecular-biological methods can be integrated, like methylation studies or two-dimensional gel electrophoresis.

Bibliography

[1] Bao-Ling Adam, Yinsheng Qu, John W Davis, Michael D Ward, Mary Ann Clements, Lisa H Cazares, O. John Semmes, Paul F Schellhammer, Yutaka Yasui, Ziding Feng, and George L Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*, 62(13):3609–3614, Jul 2002.

[2] Adam S Adler, Meihong Lin, Hugo Horlings, Dimitry S A Nuyten, Marc J van de Vijver, and Howard Y Chang. Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet*, 38(4):421–430, Apr 2006.

[3] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[4] Ali Ahmad, Darren L. Bayley, Shiping He, and Robert A. Stockley. Myeloid related protein-8/14 stimulates interleukin-8 production in airway epithelial cells. *Am J Respir Cell Mol Biol*, 29(4):523–530, Oct 2003.

[5] Iskander Al-Githmi, Nadia Batawil, Norihisa Shigemura, Michael Hsin, Tak Wai Lee, Gue-Wei He, and Anthony Yim. Bronchiolitis obliterans following lung transplantation. *Eur J Cardiothorac Surg*, 30(6):846–851, Dec 2006.

[6] Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10):6562–6566, May 2002.

[7] Balazs Antus, Janos Fillinger, Eszter Csiszer, Krisztina Czebe, and Ildiko Horvath. Bronchiolitis obliterans syndrome in lung transplant recipients. *Orv Hetil*, 146(19):953–958, May 2005.

[8] Ajith Abraham Arpad Kelemen and Yuehui Chen. *Computational Intelligence in Bioinformatics*. 2008.

[9] Alison E. Ashcroft. *An Introduction to Mass Spectrometry*. Astbury Centre for Structural Molecular Biology, Astbury Building, The University of Leeds., 2006.

[10] Edward F Attiyeh, Wendy B London, Yael P Mossé, Qun Wang, Cynthia Winter, Deepa Khazi, Patrick W McGrady, Robert C Seeger, A. Thomas Look, Hiroyuki Shimada, Garrett M Brodeur, Susan L Cohn, Katherine K Matthay, John M Maris, and Children's Oncology Group. Chromosome 1p and 11q deletions and outcome in neuroblastoma. *N Engl J Med*, 353(21):2243–2253, Nov 2005.

[11] T. K. Attwood. Genomics. The Babel of bioinformatics. *Science*, 290(5491):471–473, Oct 2000.

[12] Keith A Baggerly, Jeffrey S Morris, and Kevin R Coombes. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777–785, Mar 2004.

[13] Keith A Baggerly, Jeffrey S Morris, Jing Wang, David Gold, Lian-Chun Xiao, and Kevin R Coombes. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3(9):1667–1672, Sep 2003.

[14] R. E. Banks, M. J. Dunn, D. F. Hochstrasser, J. C. Sanchez, W. Blackstock, D. J. Pappin, and P. J. Selby. Proteomics: new perspectives, new biomedical opportunities. *Lancet*, 356(9243):1749–1756, Nov 2000.

[15] Michael T Barrett, Alicia Scheffer, Amir Ben-Dor, Nick Sampas, Doron Lipson, Robert Kincaid, Peter Tsang, Bo Curry, Kristin Baird, Paul S Meltzer, Zohar Yakhini, Laurakay Bruhn, and Stephen Laderman. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A*, 101(51):17765–17770, Dec 2004.

[16] Mark A Beaumont and Bruce Rannala. The bayesian revolution in genetics. *Nat Rev Genet*, 5(4):251–261, Apr 2004.

[17] Tim Beissbarth and Terence P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, Jun 2004.

[18] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Barbara A Rapp, and David L Wheeler. GenBank. *Nucleic Acids Res*, 30(1):17–20, Jan 2002.

[19] D. Benton. Bioinformatics–principles and potential of a new multidisciplinary tool. *Trends Biotechnol*, 14(8):261–272, Aug 1996.

[20] Frank Berthold, Joachim Boos, Stefan Burdach, Rudolf Erttmann, Günter Henze, Johann Hermann, Thomas Klingebiel, Bernhard Kremens, Freimut H Schilling, Martin Schrappe, Thorsten Simon, and Barbara Hero. Myeloablative megatherapy with autologous stem-cell rescue versus oral maintenance chemotherapy as consolidation treatment in patients with high-risk neuroblastoma: a randomised controlled trial. *Lancet Oncol*, 6(9):649–658, Sep 2005.

[21] Bernd Berwanger, Oliver Hartmann, Eckhard Bergmann, Sandra Bernard, Dirk Nielsen, Michael Krause, Ali Kartal, Daniel Flynn, Ruprecht Wiedemeyer, Manfred Schwab, Helmut Schäfer, Holger Christiansen, and Martin Eilers. Loss of a fyn-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. *Cancer Cell*, 2(5):377–386, Nov 2002.

[22] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, and Michel Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–370, Jan 2003.

[23] A. Boehler, S. Kesten, W. Weder, and R. Speich. Bronchiolitis obliterans after lung transplantation: a review. *Chest*, 114(5):1411–1426, Nov 1998.

[24] C. Boehm, K. Kailing, P. Kroeger, and H.-P. Kriegel. Immer groessere und komplexere datenmengen: Herausforderungen fuer clustering-algorithmen. *Datenbank-Spektrum*, 9, 2004.

[25] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[26] N. Bown, S. Cotterill, M. Lastowska, S. O'Neill, A. D. Pearson, D. Plantaz, M. Meddeb, G. Danglot, C. Brinkschmidt, H. Christiansen, G. Laureys, F. Speleman, J. Nicholson, A. Bernheim, D. R. Betts, J. Vandesompele, and N. Van Roy. Gain of chromosome arm 17q and adverse outcome in patients with neuroblastoma. *N Engl J Med*, 340(25):1954–1961, Jun 1999.

[27] James R Bradford, Chris J Needham, Andrew J Bulpitt, and David R Westhead. Insights into protein-protein interfaces using a bayesian network prediction method. *J Mol Biol*, 362(2):365–386, Sep 2006.

[28] Cameron Brennan, Yunyu Zhang, Christopher Leo, Bin Feng, Craig Cauwels, Andrew J Aguirre, Minjung Kim, Alexei Protopopov, and Lynda Chin. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res*, 64(14):4744–4748, Jul 2004.

[29] David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002.

[30] Garrett M Brodeur. Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer*, 3(3):203–216, Mar 2003.

[31] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[32] Declan Butler. Are you ready for the revolution? *Nature*, 409:758–760, 2001.

[33] David A Cairns, Douglas Thompson, David N Perkins, Anthea J Stanley, Peter J Selby, and Rosamonde E Banks. Proteomic profiling using mass spectrometry–does normalising by total ion current potentially mask some biological differences? *Proteomics*, 8(1):21–27, Jan 2008.

[34] Odile Carrette, Pierre R Burkhard, Jean-Charles Sanchez, and Denis F Hochstrasser. State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research. *Nat Protoc*, 1(2):812–823, 2006.

[35] B. Carvalho, E. Ouwerkerk, G. A. Meijer, and B. Ylstra. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol*, 57(6):644–646, Jun 2004.

[36] D. Chamberlain, J. Maurer, C. Chaparro, and L. Idolor. Evaluation of transbronchial lung biopsy specimens in the diagnosis of bronchiolitis obliterans after lung transplantation. *J Heart Lung Transplant*, 13(6):963–971, 1994.

[37] Howard Y Chang, Julie B Sneddon, Ash A Alizadeh, Ruchira Sood, Rob B West, Kelli Montgomery, Jen-Tsan Chi, Matt van de Rijn, David Botstein, and Patrick O Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol*, 2(2):E7, Feb 2004.

[38] J. Chaudhary and M. Schmidt. The impact of genomic alterations on the transcriptome: a prostate cancer cell line case study. *Chromosome Res*, 14(5):567–586, 2006.

[39] Guoan Chen, Tarek G Gharib, Chiang-Ching Huang, Jeremy M G Taylor, David E Misek, Sharon L R Kardia, Thomas J Giordano, Mark D Iannettoni, Mark B Orringer, Samir M Hanash, and David G Beer. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics*, 1(4):304–313, Apr 2002.

[40] Guoan Chen, Hong Wang, Tarek G Gharib, Chiang-Ching Huang, Dafydd G Thomas, Kerby A Shedden, Rork Kuick, Jeremy M G Taylor, Sharon L R Kardia, David E Misek, Thomas J Giordano, Mark D Iannettoni, Mark B Orringer, Samir M Hanash, and David G Beer. Overexpression of oncoprotein 18 correlates with poor differentiation in lung adenocarcinomas. *Mol Cell Proteomics*, 2(2):107–116, Feb 2003.

[41] D.M. Chickering. Learning bayesian networks is np-complete. *AI & STAT V*, 1996.

[42] D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554., 2002.

[43] Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, Fanqing Chen, Heidi Feiler, Taku Tokuyasu, Chris Kingsley, Shanaz Dairkee, Zhenhang Meng, Karen Chew, Daniel Pinkel, Ajay Jain, Britt Marie Ljung, Laura Esserman, Donna G Albertson, Frederic M Waldman, and Joe W Gray. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, Dec 2006.

[44] Jung Kyoon Choi, Ungsik Yu, Sangsoo Kim, and Ook Joon Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84–i90, 2003.

[45] B. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10:101–129, 1954.

[46] Francis S Collins, Michael Morgan, and Aristides Patrinos. The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617):286–290, Apr 2003.

[47] Kevin R Coombes, Spiridon Tsavachidis, Jeffrey S Morris, Keith A Baggerly, Mien-Chie Hung, and Henry M Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, Nov 2005.

[48] G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[49] Dale S Cornett, Michelle L Reyzer, Pierre Chaurand, and Richard M Caprioli. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nat Methods*, 4(10):828–833, Oct 2007.

[50] Corinna Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[51] Brian Cox, Thomas Kislinger, and Andrew Emili. Integrating gene and protein expression data: pattern analysis and profile mining. *Methods*, 35(3):303–314, Mar 2005.

[52] Jürgen Cox and Matthias Mann. Is proteomics the new genomics? *Cell*, 130(3):395–398, Aug 2007.

[53] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[54] L. Croft, S. Schandorff, F. Clark, K. Burrage, P. Arctander, and J. S. Mattick. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet*, 24(4):340–341, Apr 2000.

[55] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. Expression profiling using cDNA microarrays. *Nat Genet*, 21(1 Suppl):10–14, Jan 1999.

[56] Janusz Dutkowski and Anna Gambin. On consensus biomarker selection. *BMC Bioinformatics*, 8 Suppl 5:S5, 2007.

[57] Marc Estenne and Marshall I Hertz. Bronchiolitis obliterans after human lung transplantation. *Am J Respir Crit Care Med*, 166(4):440–444, Aug 2002.

[58] L. Fahrmeir, R. Kuenstler, and I. Pigeot. *Statistik*. Springer, Berlin, 2004.

[59] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 1989.

[60] P. Lee Ferguson and Richard D Smith. Proteome analysis by mass spectrometry. *Annu Rev Biophys Biomol Struct*, 32:399–424, 2003.

[61] S. Fischer, S. D. Cassivi, A. M. Xavier, J. A. Cardella, E. Cutz, V. Edwards, M. Liu, and S. Keshavjee. Cell death in human lung transplantation: apoptosis induction in human lungs during ischemia and after transplantation. *Ann Surg*, 231(3):424–431, Mar 2000.

[62] F. Forozan, R. Karhu, J. Kononen, A. Kallioniemi, and O. P. Kallioniemi. Genome screening by comparative genomic hybridization. *Trends Genet*, 13(10):405–409, Oct 1997.

[63] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004.

[64] Pengcheng Fu. Biomolecular computing: is it ready to take off? *Biotechnol J*, 2(1):91–101, Jan 2007.

[65] E. Gasteiger, M. R. Wilkins, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser. Protein identification and analysis tools in the expasy server. *Methods Mol Biol*, 112:531–552, 1999.

[66] W. Wayt Gibbs. The unseen genome: gems among the junk. *Sci Am*, 289(5):26–33, Nov 2003.

[67] Anne-Claude Gingras, Matthias Gstaiger, Brian Raught, and Ruedi Aebersold. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol*, 8(8):645–654, Aug 2007.

[68] G. Brian Golding. DNA and the revolutions of molecular evolution, computational biology, and bioinformatics. *Genome*, 46(6):930–935, Dec 2003.

[69] N. Goodman. Biological data becomes computer literate: new advances in bioinformatics. *Curr Opin Biotechnol*, 13(1):68–71, Feb 2002.

[70] Michael G Gravett, Miles J Novy, Ron G Rosenfeld, Ashok P Reddy, Thomas Jacob, Mark Turner, Ashley McCormack, Jodi A Lapidus, Jane Hitti, David A Eschenbach, Charles T Roberts, and Srinivasa R Nagalla. Diagnosis of intra-amniotic infection by proteomic profiling and identification of novel biomarkers. *JAMA*, 292(4):462–469, Jul 2004.

[71] C. Guo, P. S. White, M. J. Weiss, M. D. Hogarty, P. M. Thompson, D. O. Stram, R. Gerbing, K. K. Matthay, R. C. Seeger, G. M. Brodeur, and J. M. Maris. Allelic deletion at 11q23 is common in mycn single copy neuroblastomas. *Oncogene*, 18(35):4948–4957, Sep 1999.

[72] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, Mar 1999.

[73] Kriegel H.-P. and Pfeifle M. Hierarchical density-based clustering of uncertain data. In *Proc. 5th IEEE Int. Conf. on Data Mining (ICDM'05)*, pages 689–692, Houston, TX, 2005.

[74] J. Hanke, D. Brett, I. Zastrow, A. Aydin, S. Delbrück, G. Lehmann, F. Luft, J. Reich, and P. Bork. Alternative splicing of human genes: more the rule than the exception? *Trends Genet*, 15(10):389–390, Oct 1999.

[75] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Verl., New York, Berlin, Heidelberg, 2001.

[76] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243., 1995.

[77] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cdna microarray analysis. *BioTechniques*, 29, No. 3:548–562, Sept 2000.

[78] Melanie Hilario, Alexandros Kalousis, Christian Pellegrini, and Markus Mueller. Processing and classification of protein mass spectra. *Mass Spectrom Rev*, 25(3):409–449, 2006.

[79] J. F. Hocquette. Where are we in genomics? *J Physiol Pharmacol*, 56 Suppl 3:37–70, Jun 2005.

[80] http://genomesonline.org.

[81] http://www.ncbi.nlm.nih.gov/Genomes/index.html.

[82] Jianhua Hu, Kevin R Coombes, Jeffrey S Morris, and Keith A Baggerly. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic*, 3(4):322–331, Feb 2005.

[83] Erich Huang, Seiichi Ishida, Jennifer Pittman, Holly Dressman, Andrea Bild, Mark Kloos, Mark D'Amico, Richard G Pestell, Mike West, and Joseph R Nevins. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet*, 34(2):226–230, Jun 2003.

[84] Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[85] T. R. Hughes, C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer, D. Slade, J. Burchard, S. Dow, T. R. Ward, M. J. Kidd, S. H. Friend, and M. J. Marton. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet*, 25(3):333–337, Jul 2000.

[86] Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, Dec 2004.

[87] Elizabeth Hyman, Päivikki Kauraniemi, Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahloun, Olli-P. Kallioniemi, and Anne Kallioniemi. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–6245, Nov 2002.

[88] Guyon I., Weston J., Barnhill S., and Vapnik V. *Gene Selection for Cancer Classification using Support Vector Machines*, volume 46. Machine Learning, Springer Netherlands, 2002.

[89] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, May 2001.

[90] Yang Y.H.and Buckley M. J., Dudoit S., and Speed T. P. Comparison of methods for image analysis on cdna microarray data. *Journal of Computational & Graphical Statistics*, 11:108–136, 2002.

[91] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, Jul 2005.

[92] Kees Jong, Elena Marchiori, Gerrit Meijer, A. V D Vaart, and Bauke Ylstra. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20(18):3636–3637, Dec 2004.

[93] G. L. Wright Jr, L. H. Cazares, S-M. Leung, S. Nasim, B-L. Adam, T-T. Yip, P. F. Schellhammer, L. Gong, and A. Vlahou. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis*, 2(5/6):264–276, Dec 1999.

[94] A-K. Järvinen, R. Autio, S. Haapa-Paananen, M. Wolf, M. Saarela, R. Grénman, I. Leivo, O. Kallioniemi, A. A. Mäkitie, and O. Monni. Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses. *Oncogene*, 25(52):6997–7008, Nov 2006.

[95] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, Oct 1992.

[96] Minoru Kanehisa and Peer Bork. Bioinformatics in the post-sequence era. *Nat Genet*, 33 Suppl:305–310, Mar 2003.

[97] Sudhir Kumar and Joel Dudley. Bioinformatics software for biologists in the genomics era. *Bioinformatics*, 23(14):1713–1717, Jul 2007.

[98] Song Le, Bedo Justin, Borgwardt Karsten M., Gretton Arthur, and Smola Alex. The bahsic family of gene selection algorithms submitted. *submitted*, 2006.

[99] Hyunju Lee, Sek Won Kong, and Peter J Park. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–896, Apr 2008.

[100] Ilya Levner. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, 6:68, 2005.

[101] P. Lichter, K. Fischer, S. Joos, T. Fink, M. Baudis, R. K. Potkul, S. Ohl, S. Solinas-Toldo, R. Weber, S. Stilgenbauer, M. Bentz, and H. Döhner. Efficacy of current molecular cytogenetic protocols for the diagnosis of chromosome aberrations in tumor specimens. *Cytokines Mol Ther*, 2(3):163–169, Sep 1996.

[102] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.

[103] L. A. Loeb and F. C. Christians. Multiple mutations in human cancers. *Mutat Res*, 350(1):279–286, Feb 1996.

[104] L. A. Loeb, C. F. Springgate, and N. Battula. Errors in DNA replication as a basis of malignant changes. *Cancer Res*, 34(9):2311–2321, Sep 1974.

[105] Zorbas H. Lottspeich F. *Bioanalytik*. Spektrum Akademischer Verlag GmbH Heidelberg, Berlin, 1998.

[106] Dierich M., Skawran B., Steinmann D., Gottlieb J., von Neuhoff N., Szangolies J., Hohlfeld J., Schlegelberger B., Niedermeyer J., and Welte T. Compartment-specific gene- and protein-expression during early phases of bronchiolitis obliterans syndrome after lung transplantation. In *Clinical Research Unit Lung Transplantation KFO 123, Medizinische Hochschule Hannover*, 2006.

[107] Karas M., Bachmann D., and Bahr U. *Mass Spectrom, Ion Proc*, 78:53–68, 1987.

[108] Dariya I Malyarenko, William E Cooke, Bao-Ling Adam, Gunjan Malik, Haijian Chen, Eugene R Tracy, Michael W Trosset, Maciek Sasinowski, O. John Semmes, and Dennis M Manos. Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clin Chem*, 51(1):65–74, Jan 2005.

[109] D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon, University, Pittsburgh, 2003.

[110] J. M. Maris, C. Guo, P. S. White, M. D. Hogarty, P. M. Thompson, D. O. Stram, R. Gerbing, K. K. Matthay, R. C. Seeger, and G. M. Brodeur. Allelic deletion at chromosome bands 11q14-23 is common in neuroblastoma. *Med Pediatr Oncol*, 36(1):24–27, Jan 2001.

[111] Wouter Meuleman, Judith Ymn Engwegen, Marie-Christine W Gast, Jos H Beijnen, Marcel Jt Reinders, and Lodewyk Fa Wessels. Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics*, 9:88, 2008.

[112] Evi Michels, Jasmien Hoebeeck, Katleen De Preter, Alexander Schramm, Bénédicte Brichard, Anne De Paepe, Angelika Eggert, Geneviève Laureys, Jo Vandesompele, and Frank Speleman. Cadm1 is a strong neuroblastoma candidate gene that maps within a 3.72 mb critical region of loss on 11q23. *BMC Cancer*, 8:173, 2008.

[113] Robert Molidor, Alexander Sturn, Michael Maurer, and Zlatko Trajanoski. New trends in bioinformatics: from genome sequence to personalized medicine. *Exp Gerontol*, 38(10):1031–1036, Oct 2003.

[114] O. Monni, M. Barlund, S. Mousses, J. Kononen, G. Sauter, M. Heiskanen, P. Paavola, K. Avela, Y. Chen, M. L. Bittner, and A. Kallioniemi. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc Natl Acad Sci U S A*, 98(10):5711–5716, May 2001.

[115] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, May 2005.

[116] Yael P Mosse, Marci Laudenslager, Deepa Khazi, Alex J Carlisle, Cynthia L Winter, Eric Rappaport, and John M Maris. Germline phox2b mutation in hereditary neuroblastoma. *Am J Hum Genet*, 75(4):727–730, Oct 2004.

[117] A. Nakagawara, M. Arima-Nakagawara, N. J. Scavarda, C. G. Azar, A. B. Cantor, and G. M. Brodeur. Association between high levels of expression of the trk gene and favorable outcome in human neuroblastoma. *N Engl J Med*, 328(12):847–854, Mar 1993.

[118] Gary L Nelsestuen, Michael B Martinez, Marshall I Hertz, Kay Savik, and Christine H Wendt. Proteomic identification of human neutrophil alpha-defensins in chronic lung allograft rejection. *Proteomics*, 5(6):1705–1713, Apr 2005.

[119] H. J. Nickerson, K. K. Matthay, R. C. Seeger, G. M. Brodeur, H. Shimada, C. Perez, J. B. Atkinson, M. Selch, R. B. Gerbing, D. O. Stram, and J. Lukens. Favorable biology and outcome of stage iv-s neuroblastoma with supportive care or minimal therapy: a children's cancer group study. *J Clin Oncol*, 18(3):477–486, Feb 2000.

[120] André Oberthuer, Frank Berthold, Patrick Warnat, Barbara Hero, Yvonne Kahlert, Rüdiger Spitz, Karen Ernestus, Rainer König, Stefan Haas, Roland Eils, Manfred Schwab, Benedikt Brors, Frank Westermann, and Matthias Fischer. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol*, 24(31):5070–5078, Nov 2006.

[121] André Oberthuer, Barbara Hero, Rüdiger Spitz, Frank Berthold, and Matthias Fischer. The tumor-associated antigen prame is universally expressed in high-stage neuroblastoma and associated with poor outcome. *Clin Cancer Res*, 10(13):4307–4313, Jul 2004.

[122] P. H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*, 250(10):4007–4021, May 1975.

[123] Adam B Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004.

[124] C. P. Paweletz, B. Trock, M. Pennanen, T. Tsangaris, C. Magnant, L. A. Liotta, and E. F. Petricoin. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis Markers*, 17(4):301–307, 2001.

[125] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2nd Ed., 1997.

[126] J. Pearl and T. Verma. A theory of inferred causation, in j. allen, r. fikes, e. sandewall, principles of knowledge representation and reasoning. *Proceeding of the Second International Conference (Morgan Kaufmann, San Mateo, CA)*, 441-452, 1991.

[127] Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, and Jonathan D Wren. Evolving research trends in bioinformatics. *Brief Bioinform*, 8(2):88–95, Mar 2007.

[128] D. Plantaz, G. Mohapatra, K. K. Matthay, M. Pellarin, R. C. Seeger, and B. G. Feuerstein. Gain of chromosome 17 is the most frequent abnormality detected in neuroblastoma by comparative genomic hybridization. *Am J Pathol*, 150(1):81–89, Jan 1997.

[129] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, 23(1):41–46, Sep 1999.

[130] Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, Oct 2002.

[131] Manuela Pruess and Rolf Apweiler. Bioinformatics Resources for In Silico Proteome Analysis. *J Biomed Biotechnol*, 2003(4):231–236, 2003.

[132] Gentleman R., Carey V., Huber W., Irizarry R., and Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005.

[133] Guru Reddy and Enrique Dalmasso. SELDI ProteinChip(R) Array Technology: Protein-Based Predictive Medicine and Drug Discovery Applications. *J Biomed Biotechnol*, 2003(4):237–241, 2003.

[134] Anne C. Sauve and Terence P. Speed, editors. *Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data*. Proceedings Gensips, 2004.

[135] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*, Suppl 37:120–125, 2001.

[136] G. Schartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[137] Stefan Schaub, John Wilkins, Tracey Weiler, Kevin Sangster, David Rush, and Peter Nickerson. Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int*, 65(1):323–332, Jan 2004.

[138] M. Schena. Genome analysis with gene expression microarrays. *Bioessays*, 18(5):427–431, May 1996.

[139] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.

[140] Bernhard Schoelkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[141] P. A. Sharp. Split genes and RNA splicing. *Cell*, 77(6):805–815, Jun 1994.

[142] Ling Shen, Patrick Tso, Stephen C Woods, Randall R Sakai, W. Sean Davidson, and Min Liu. Hypothalamic apolipoprotein a-iv is regulated by leptin. *Endocrinology*, 148(6):2681–2689, Jun 2007.

[143] Radha Shyamsundar, Young H Kim, John P Higgins, Kelli Montgomery, Michelle Jorden, Anand Sethuraman, Matt van de Rijn, David Botstein, Patrick O Brown, and Jonathan R Pollack. A dna microarray survey of gene expression in normal human tissues. *Genome Biol*, 6(3):R22, 2005.

[144] S. Singh-Gasson, R. D. Green, Y. Yue, C. Nelson, F. Blattner, M. R. Sussman, and F. Cerrina. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol*, 17(10):974–978, Oct 1999.

[145] Gordon K Smyth and Terry Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, Dec 2003.

[146] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, 29(3):263–264, Nov 2001.

[147] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, and P. Lichter. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20(4):399–407, Dec 1997.

[148] Scot R Weinberger & Andreas Wiesner Sonja Vorderwülbecke, Steve Cleverley. Protein quantification by the seldi-tof-ms?based proteinchip® system. *Nature Methods*, 2:393 – 395, 2005.

[149] Liliana Soroceanu, Samir Kharbanda, Ruihuan Chen, Robert H Soriano, Ken Aldape, Anjan Misra, Jiping Zha, William F Forrest, Janice M Nigro, Zora Modrusan, Burt G Feuerstein, and Heidi S Phillips. Identification of IGF2 signaling through phosphoinositide-3-kinase regulatory subunit 3 as a growth-promoting axis in glioblastoma. *Proc Natl Acad Sci U S A*, 104(9):3466–3471, Feb 2007.

[150] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer-Verlag, New York, 1993.

[151] Ruediger Spitz, Barbara Hero, Karen Ernestus, and Frank Berthold. Deletions in chromosome arms 3p and 11q are new prognostic markers in localized and 4s neuroblastoma. *Clin Cancer Res*, 9(1):52–58, Jan 2003.

[152] Ruediger Spitz, Andre Oberthuer, Marc Zapatka, Benedikt Brors, Barbara Hero, Karen Ernestus, Joern Oestreich, Matthias Fischer, Thorsten Simon, and Frank Berthold. Oligonucleotide array-based comparative genomic hybridization (aCGH) of 90 neuroblastomas reveals aberration patterns closely associated with relapse pattern and outcome. *Genes Chromosomes Cancer*, 45(12):1130–1142, Dec 2006.

[153] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, Sep 2005.

[154] G. Stoesser, P. Sterk, M. A. Tuli, P. J. Stoehr, and G. N. Cameron. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, 25(1):7–14, Jan 1997.

[155] Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna de Grassi, Charles Lee, Chris Tyler-Smith, Nigel Carter, Stephen W Scherer, Simon Tavaré, Panagiotis Deloukas, Matthew E Hurles, and Emmanouil T Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb 2007.

[156] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[157] Jennifer Svetlecic, Agostino Molteni, Yayan Chen, Mohammad Al-Hamed, Tim Quinn, and Betty Herndon. Transplant-related bronchiolitis obliterans (bos) demonstrates unique cytokine profiles compared to toxicant-induced bos. *Exp Mol Pathol*, 79(3):198–205, Dec 2005.

[158] Alejandro Sweet-Cordero, Sayan Mukherjee, Aravind Subramanian, Han You, Jeffrey J Roix, Christine Ladd-Acosta, Jill Mesirov, Todd R Golub, and Tyler Jacks. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet*, 37(1):48–55, Jan 2005.

[159] Alejandro Sweet-Cordero, George C Tseng, Han You, Margaret Douglass, Bing Huey, Donna Albertson, and Tyler Jacks. Comparison of gene expression and DNA copy number changes in a murine model of lung cancer. *Genes Chromosomes Cancer*, 45(4):338–348, Apr 2006.

[160] Ido Y. Akita S. Yoshida Y. Tanaka K., Waki H. *Mass Spectrom*, 2:151 – 153, 1988.

[161] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*, 101(9):2981–2986, Mar 2004.

[162] Nilesh S Tannu and Scott E Hemby. Two-dimensional fluorescence difference gel electrophoresis for comparative proteomics profiling. *Nat Protoc*, 1(4):1732–1742, 2006.

[163] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002.

[164] Narasimhan B Chu G Tibshirani R, Hastie T. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 18(1):104–117, 2003.

[165] Dafna Tsafrir, Manny Bacolod, Zachariah Selvanayagam, Ilan Tsafrir, Jinru Shia, Zhaoshi Zeng, Hao Liu, Curtis Krier, Robert F Stengel, Francis Barany, William L Gerald, Philip B Paty, Eytan Domany, and Daniel A Notterman. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res*, 66(4):2129–2137, Feb 2006.

[166] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*, 29(12):2549–2557, Jun 2001.

[167] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, Apr 2001.

[168] R. M. Twyman, editor. *Principles of proteomics.* ISBN 1-85996-273-4. BIOS Scientific Publishers, New York, 2004.

[169] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang,

J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[170] von Neuhoff Nils. Ltx probe types. Technical report, Hannover Medical School (MHH), 2006.

[171] S. Wagner, M. A. Hess, P. Ormonde-Hanson, J. Malandro, H. Hu, M. Chen, R. Kehrer, M. Frodsham, C. Schumacher, M. Beluch, C. Honer, M. Skolnick, D. Ballinger, and B. R. Bowen. A broad role for the zinc finger protein znf202 in human lipid metabolism. *J Biol Chem*, 275(21):15685–15690, May 2000.

[172] T. Wahlers, A. Haverich, H. J. Schaefers, S. W. Hirt, H. G. Fieguth, M. Jurmann, C. Zink, and H. G. Borst. Chronic rejection following lung transplantation. incidence, time pattern and consequences. *Eur J Cardiothorac Surg*, 7(6):319–23; discussion 324, 1993.

[173] J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[174] Robert A. Weinberg. *The Biology of Cancer*. Garland Science, 2006.

[175] Axel Wellmann, Volker Wollscheid, Hong Lu, Zhan Lu Ma, Peter Albers, Karin Schütze, Volker Rohde, Peter Behrens, Stefan Dreschers, Yon Ko, and Nicolas Wernert. Analysis of microdissected prostate tissue with ProteinChip arrays–a

way to new insights into carcinogenesis and to diagnostic tools. *Int J Mol Med*, 9(4):341–347, Apr 2002.

[176] Maija Wolf, Spyro Mousses, Sampsa Hautaniemi, Ritva Karhu, Pia Huusko, Minna Allinen, Abdel Elkahloun, Outi Monni, Yidong Chen, Anne Kallioniemi, and Olli-P. Kallioniemi. High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia*, 6(3):240–247, 2004.

[177] Michael E Wright, David K Han, and Ruedi Aebersold. Mass spectrometry-based expression profiling of clinical prostate cancer. *Mol Cell Proteomics*, 4(4):545–554, Apr 2005.

[178] K. Wüthrich, G. Wider, G. Wagner, and W. Braun. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J Mol Biol*, 155(3):311–319, Mar 1982.

[179] Zhijin Wu and Rafael A Irizarry. Preprocessing of oligonucleotide array data. *Nat Biotechnol*, 22(6):656–8; author reply 658, Jun 2004.

[180] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, Feb 2002.

[181] Jun Yao, Stanislawa Weremowicz, Bin Feng, Robert C Gentleman, Jeffrey R Marks, Rebecca Gelman, Cameron Brennan, and Kornelia Polyak. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res*, 66(8):4065–4078, Apr 2006.

[182] John R Yates, Annalyn Gilchrist, Kathryn E Howell, and John J M Bergeron. Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol*, 6(9):702–714, Sep 2005.

[183] Sun Ying-Hao, Yang Qing, Wang Lin-Hui, Gao Li, Tang Rong, Ying Kang, Xu Chuan-Liang, Qian Song-Xi, Li Yao, Xie Yi, and Mao Yu-Ming. Monitoring gene expression profile changes in bladder transitional cell carcinoma using cdna microarray. *Urol Oncol*, 7(5):207–212, 2002.

[184] Claudia Zanazzi, Remko Hersmus, Imke M Veltman, Ad J M Gillis, Ellen van Drunen, H. Berna Beverloo, Joost P J J Hegmans, Marielle Verweij, Bart N Lambrecht, J. Wolter Oosterhuis, and Leendert H J Looijenga. Gene expression profiling and gene copy-number changes in malignant mesothelioma cell lines. *Genes Chromosomes Cancer*, 46(10):895–908, Oct 2007.

[185] M. Zapatka, P. Bewerunge, B. Brors, and R. Eils. Classify patient samples using mass spectra, an improved variable selection procedure. In progress.

[186] Xuegong Zhang, Xin Lu, Qian Shi, Xiu-Qin Xu, Hon-Chiu E Leung, Lyndsay N Harris, James D Iglehart, Alexander Miron, Jun S Liu, and Wing H Wong. Recursive svm feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, 7:197, 2006.

[187] Yan Zhang, Matthew Wroblewski, Marshall I Hertz, Christine H Wendt, Tereza M Cervenka, and Gary L Nelsestuen. Analysis of chronic lung transplant rejection by maldi-tof profiles of bronchoalveolar lavage fluid. *Proteomics*, 6(3):1001–1010, Feb 2006.

[188] Hongjuan Zhao, Young Kim, Pei Wang, Jacques Lapointe, Rob Tibshirani, Jonathan R Pollack, and James D Brooks. Genome-wide characterization of gene expression variations and DNA copy number changes in prostate cancer cell lines. *Prostate*, 63(2):187–197, May 2005.

[189] Xiaojun Zhao, Cheng Li, J. Guillermo Paez, Koei Chin, Pasi A Jänne, Tzu-Hsiu Chen, Luc Girard, John Minna, David Christiani, Chris Leo, Joe W Gray, William R Sellers, and Matthew Meyerson. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, 64(9):3060–3071, May 2004.

[190] Shougang Zhuang and Rick G Schnellmann. A death-promoting role for extracellular signal-regulated kinase. *J Pharmacol Exp Ther*, 319(3):991–997, Dec 2006.

[191] E. Zubritsky. Spotting a microarray system. *Anal Chem*, 72(23):761A–767A, Dec 2000.

List of Figures

1.1 Example plot of a genomic profile. The y-axis depicts the copy number ratio of all measured chromosomal regions. The chromosomal positions are displayed at the x-axis. Gray vertical lines define the different chromosomes numbered in the same color. 17

1.2 Comparative genomic hybridisation. Fragments of normal (red) and tumor (green) DNA are differentially labeled with two different fluorophores. These fragments hybridise to metaphase chromosomes. A red signal indicates that only normal DNA is annealed, but no tumor DNA was present. This is when a loss of DNA in the tumor DNA has occured. In the case of a yellow signal, both normal and tumor DNA are bound in the same amount, i.e. the tumor DNA shows no chromosomal aberration. A gain of tumor DNA is indicated by a a green signal, which denotes that more tumor DNA is annealed than normal DNA. Figure taken from [174]. . . 18

1.3 Desorption/ionization time-of-flight mass spectrometry. **a)** General setup of a mass spectrometer for MALDI and SELDI. *I) Ionization and Acceleration.* In both MALDI and SELDI, a biological sample of interest is applied to a surface. It is incubated and subsequently co-crystallized with matrix material. A laser is then fired at the co-crystallised mixture and initiates ionization and evaporation of proteins,which are then accelerated by an electric field. The energy of the laser beam is transferred via the matrix to the analyte sample and causes ionization. *II) Drifting.* An electrical field causes the ionized material to fly through the TOF tube (going from t_o to t_1). Lower mass peptides (red ball) fly faster through the tube than higher mass peptides (green ball). *III) Detector.* The peptides with a lower mass arrive earlier than the high mass peptides at the detector which is placed at the end of the flight tube. **b)** Schematic image of a mass spectrum. Using a quadratic equation, the mass-to-charge ratio (m/z) of a peptide can be calculated and plotted as a so called mass spectrum, with the intensity on the y- and the m/z-ratio on the x-axis. The peak height correlates to the protein concentration. 22

1.4 Workflow of proteomic profiling. a) First steps of a MALDI-TOF procedure include sample preperation of e.g. body fluids like serum. The sample is mixed with magnetic beads which catch only specific peptides. This target mixture is then spotted to a chip and is processed with an appropriate mass spectrometer.The resulting protein pattern displays the separated peptides in terms of their m/z-ratio. b) The SELDI-TOF workflow also includes sample preparation, target spotting and results in a protein pattern, but differs in utilizing a chip with a chromatographic surface instead of using magnetic beads. 24

1.5 Basic idea for the identification of key players in molecular biological processes via Bayesian networks. Data are not real. a) Preselected nodes of interest origin from CNAs (orange), genes expression (blue) and clinical data (green) have to be chosen from every data type alone. It should be addressed, how these nodes interact and regulate each other depending on their chromosomal position. b) After infering BN this results in a directed acyclic graph with nodes and edges. c) Resulting dependencies allow to deduce key-players from that graph. In this example, a chromosomal aberration on 6p27.3 seems to have an important role on the clinical outcomes by influencing the expression of four genes. 29

2.1 Schematic computation of *similar-state-sum-matrix* . The aCGH matrix and gene expression matrix have the same structure. Rows refer to the data type-specific probes and columns for the patients in the same order. The discretized values for both matrices are -1,0 and 1. Green values identify equal states between aCGH and gene expression data. The similar-state-sum-matrix (sss) holds the sum of similar states between aCGH and gene expression data 37

2.2 *d-seperatio*n of 3 nodes in a BN. A and C are *d-separate*d given B because evidence e is given for B. 38

2.3 Markov blanket of a variable X. The members of the blanket are within the gray ellipse. 40

2.4 Bronchial brush specimen. 45

2.5 Bronchoscopy. A flexible bronchoscope is inserted through either the nose or mouth to the trachea and further down into the bronchus. Each area the bronchoscope passes can be examined. Specimen of lung tissues or lavages can be taken. 45

2.6 Resampling of mass spectra. Five mass spectra are shown exemplary as dashed lines. The spacing of the dashes represents the resolution of the respective spectra (I)-(IV). All spectra are interpolated to a common spectrum (V) with common m/z range and highest resolution. Therefore the largest starting point (here of spectrum (II)) and the smallest end point (here of spectrum (III)) of all spectra (I)-(IV) are chosen to be the master m/z range in (V). In addition, the highest resolution (here of spectrum (II)) is chosen as resolution for (V). 48

2.7 Separation of two classes (active and inactive) by a hyperplane computed by SVM in a n-dimensional feature space. The maximum margin hyperplane depends on the support vectors (red dotted circles). 50

2.8 Cross validation setup including SVM coupled to RFE. 53

3.1 Frequency Plot of 81 neuroblastoma patients. Losses are displayed in green and gains in red. Chromosome boundaries are indicated by dashed lines. 58

3.2 Heatmap of the *consistency-scores* in neuroblastoma . The colored bars at the top of the figure denote the values of the clinical variables: NB Status, MYCN and the Stage. Colors of the clinical variables as well as the color coded correlation values are explained in tab. 3.2. Chromosome boundaries are indicated by alternating light and dark gray bars at the right side. 60
3.3 Bayesian network of *cis*- and *trans*-effects. Triangles represent chromosomal abberations and circles represent genes. The colors indicate the gene expression level respectively a gain or loss of chromosomal material (red = high/gain; green = low/loss). 63
3.4 Heatmap representing the two-dimensional clustering of the 1,306 significant genes. Patients (gray) and controls (black) are shown in the bar. A clear separation between patients and controls exists. The blue color in the heatmap refers to downregulated genes and the red color to upregulated genes. The black dashed boxes a to h indicate that the clusters a and e have a similarity to clusters c and g, and clusters b and f are similar to d and h. 64
3.5 Pie chart showing the proportion of molecular functions of 1,306 significant genes. 65
3.6 Peak at 1169 Dalton. Identified by SVM-RFE (02 vs nt), PAM (02 vs nt), SVM (03 vs nt), and PAM (03 vs nt). The upper plot represents the mean spectrum of the respective patient stage / control. The lower plot presents all spectra of patients in that specific stage. Black stage 1; red stage 2; green stage 3; blue stage nt (all controls). The plot on top (a) shows the four mean spectra of every stage. Mean spectra are composed of the spectra at the bottom: (b) indicates all the spectra of patients at stage 01 which result as mean spectrum (black) in the top image (a); (c) consists of the spectra of patients at stage 02 which contribute to the red mean spectrum in (a); (d) contains the spectra of patients at stage 03, the mean spectrum is shown in green in (a), (e) covers the spectra of controls (nt), which are presented as the blue mean spectrum in (a). 72
3.7 Peak at 2160 Dalton. Identified in in SVM-RFE (01 vs nt). See legend of Fig. 3.6. 73
3.8 Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of patients for data at 1-10kDa. 75
3.9 Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of patients for data at 8-20kDa. 75
3.10 Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of controls for data at 1-10kDa. 76
3.11 Virtual average spectrum of Gene2Prot (black) and average spectrum of MS (red) of controls for data at 8-20kDa. 76
3.12 QQ Plot of data at 1-10kDa 77
3.13 FDR Plot of data at 1-10kDa 78

List of Tables

2.1 Table of symbols 41
2.2 Examinations at several time points after lung transplantation (LT), and the number of available patient samples and controls, respectively. The table is splitted into mass spectrometry analysis (MS), DNA-microarray gene expression analysis (microarray) and into overlapping patient and control cohorts for the integrative analysis of both data types 44
2.3 Tested combinations of patients in different stages respectively control patients and their available sample numbers. 53

3.1 Distribution of gene states in percent after k-means discretization. . . . 57
3.2 Color coding of the hierarchically clustered *consistency-scores.* NB Status, MYCN and Stage represent clinical variables. The assigned correlation values illustrate the color coded *consistency-scores* and are schematically explained with arrows on the right-hand side. An arrow pointing upwards denotes gain of a chromosomal region or up-regulation of gene expression, respctively. Arrows pointing downwards have analogous meaning. A horizontal line characterizes no change. 61
3.3 Accuracy, sensitivity, and specificity for classification by SVM coupled to RFE on data at 1-10kDa. 67
3.4 Accuracy, sensitivity, and specificity for classification by SVM coupled to RFE on data at 8-20kDa. 67
3.5 Accuracy, sensitivity, and specificity for classification by SVM without feature selection on data at 1-10kDa. 68
3.6 Accuracy, sensitivity, and specificity for classification by SVM without feature selection on data at 8-20kDa. 68
3.7 Accuracy, sensitivity, and specificity for classification by SVM coupled to BaHSIC on data at 1-10kDa. 69
3.8 Accuracy, sensitivity, and specificity for classification by SVM coupled to BaHSIC on data at 8-20kDa. 69
3.9 Accuracy, sensitivity, and specificity for classification by PAM on data at 1-10kDa. . . 70
3.10 Accuracy, sensitivity, and specificity for classification by PAM on data at 8-20kDa. . . 70
3.11 The seven most significant peaks. Two proteins have been identified (1. and 2.), the other six are in the process of identification. 71
3.12 Masses with gene symbol identified by meta-analysis of *Gene2Prot* and *MS* data at 1-10kDa with corresponding false discovery rate. 79

List of Algorithms

1 k-means 34
2 Matching gene probes to aCGH probes 35
3 *Build* consistency-matrix *of consistency score* 36
4 Grow-Shrink Markov blanket algorithm 41
5 Learning the structure of a BN via GS-algorithm. 42

Appendix

Tables

	Experiment	patient	SEX	stage
4	US22502540_251271410111	10447	male	3
5	US22502540_251271410007	10504	male	4S
8	US22502540_251271410147	11805	male	4
11	US22502540_251271410332	12246	female	4
13	US22502540_251271410068	13164	male	4S
14	US22502540_251271410025	13169	female	4
16	US22502540_251271410109	13264	female	4
25	US22502540_251271410500	13746	female	2B
26	US22502540_251271410150	13747	female	4
30	US22502540_251271410222	13947	male	4
34	US22502540_251271410329	14312	female	2A
35	US22502540_251271410136	14359	male	4
36	US22502540_251271410220	14360	female	4
39	US22502540_251271410085	14529	male	4
51	US22502540_251271410043	15239	female	3
52	US22502540_251271410050	15240	male	4
53	US22502540_251271410030	15259	female	4
54	US22502540_251271410324	15282	male	4
57	US22502540_251271410042	15303	male	4S
58	US22502540_251271410334	15316	female	4
59	US22502540_251271410184	15347	male	4
60	US22502540_251271410335	15377	male	4
61	US22502540_251271410631	15403	male	1
64	US22502540_251271410552	15675	male	1
65	US22502540_251271410126	15732	male	3
69	US22502540_251271410124	15800	female	3
72	US22502540_251271410002	15821	female	4
75	US22502540_251271410060	15865	male	2A
78	US22502540_251271410168	15983	male	4
79	US22502540_251271410294	15991	male	4
82	US22502540_251271410157	16261	male	4
83	US22502540_251271410570	16270	female	2A
85	US22502540_251271410442	16437	male	2A
86	US22502540_251271410525	16500	male	3
87	US22502540_251271410166	16543	female	3
89	US22502540_251271410046	16561	male	4
92	US22502540_251271410639	16656	female	4
94	US22502540_251271410539	16663	male	2B
95	US22502540_251271410532	16677	male	3
97	US22502540_251271410507	16797	female	2B
98	US22502540_251271410545	16885	female	2
102	US22502540_251271410546	16980	male	2A

Table A.1: Clinical information neuroblastoma patients

	Experiment	patient	SEX	stage
104	US22502540_251271410098	17001	female	4
109	US22502540_251271410540	17189	male	2B
110	US22502540_251271410092	17209	male	4
119	US22502540_251271410328	17315	male	2A
127	US22502540_251271410107	17663	male	4S
128	US22502540_251271410031	17665	male	4
133	US22502540_251271410602	17721	female	3
148	US22502540_251271410596	18004	female	4S
157	US22502540_251271410497	18154	female	4S
159	US22502540_251271410501	18173	male	3
166	US22502540_251271410115	1870	female	1
169	US22502540_251271410252	2000	male	4
171	US22502540_251271410502	2110	male	1
172	US22502540_251271410106	2117	female	4
173	US22502540_251271410331	226	female	4
176	US22502540_251271410054	2864	male	4
178	US22502540_251271410279	3103	male	4
179	US22502540_251271410277	312	male	4
183	US22502540_251271410047	325	male	1
184	US22502540_251271410167	327	male	1
185	US22502540_251271410090	3383	male	4
198	US22502540_251271410070	417	female	4
200	US22502540_251271410276	4188	male	3
203	US22502540_251271410104	4443	female	1
204	US22502540_251271410102	459	male	4
211	US22502540_251271410088	5043	male	4S
215	US22502540_251271410006	527	male	1
216	US22502540_251271410148	5643	male	4
217	US22502540_251271410503	5703	male	3
219	US22502540_251271410057	575	female	3
221	US22502540_251271410121	587	female	4
223	US22502540_251271410327	595	female	4S
232	US22502540_251271410295	629	male	4
239	US22502540_251271410026	6763	male	3
241	US22502540_251271410567	7363	male	1
247	US22502540_251271410065	9123	female	2B
248	US22502540_251271410178	9243	female	4
249	US22502540_251271410003	9323	male	4S
251	US22502540_251271410154	9923	male	1

Table A.1: Clinical information neuroblastoma patients

Code Documentation

BNtegrative

Functions for the integration of aCGH and DNA-microarray data.

These functions are written by the author and are part of the framework BNtegrative. All functions and a complete workflow with example data are provide at the enclosed CD.

---

averageReplicates

DESCRIPTION

Computes the median of replicates for each gene probe.

USAGE

```
averageReplicates(x, gene.list)
```

ARGUMENTS

x	Matrix of gene expression data. Matrix contains replicates for gene probe.
gene.list	Vector of unique gene identifiers.

VALUES

matrix	Matrix of gene expression data. Matrix contains the median value for the replicates of each gene probe.

checkGenomicEffects

DESCRIPTION

Computes the correlation between DNA copy number changes and gene expression. This can be done either for patient-related *cis*-effects in terms of a *consistency-score* or for the complete set of patients by computing the *trans*-effects which results in a *similar-state-sum-matrix*.

USAGE

```
checkGenomicEffects(genematrix, cghmatrix, mapping, effect, debug, B)
```

ARGUMENTS

genematrix	Matrix of gene expression data.
cghmatrix	Matrix of aCGH data.
mapping	Object returned by mapGeneIdToBacClone().
effect	String representing the algorithm. Either "*cis*" or "*trans*."
B	Integer of permutation steps. Use for simulation runs.

VALUES

matrix (if effect == "*cis*")	A matrix of the classified gene expression data. The a *consistency-score* represents the correlation between DNA copy number changes and gene expression for each patient. Possible values are -1, 0, 1.
matrix (if effect == "*trans*")	A matrix representing the *similar-state-sums* as a *similar-state-sum-matrix*. Possible values are between 0 and *n* amount of patients.

combineGeneSets

DESCRIPTION

Computes one expression value for genes that belongs to the same gene set.

USAGE

```
combineGeneSets(x, gene.sets )
```

ARGUMENTS

x	Matrix of gene expression data.
gene.sets	Vector of gene set identifiers.

VALUES

matrix	Matrix of gene expression data which are grouped into gene set.

computeBN

DESCRIPTION

Estimates a Bayesian network based on Markov Blankets. It is a wrapper function for the *gs* function of the R-package *bnlearn*.

USAGE

```
computeBN(matr,debug, strict, direction, blacklist,whitelist)
```

ARGUMENTS

matr	Matrix of *similar-state-sum*.
strict	Boolean. If TRUE conflicting results in the learning process generate an error; otherwise they res1ult in a warning.
direction	Boolean. If TRUE no attempt will be made to determine the orientation of the arcs; the returned (undirected) graph will represent the underlying structure of the Bayesian network.
whitelist	Data frame containing a set of arcs to be included in the graph.
blacklist	Data frame containing a set of arcs not to be included in the graph.

VALUES

object	Object of class *bn*.

excludeBalancedRegions

DESCRIPTION

Excludes aCGH probes that are balanced over a set of patients. Specified by a user defined threshold.

USAGE

```
excludeBalancedRegions(x, fraction)
```

ARGUMENTS

x	Matrix of aCGH data.
fraction	Float value to specify the percentage of frequency as a threshold.

VALUES

vector	Boolean vector that specifies which aCGH probe did not reach the threshold

excludeNodesWithoutArcs

DESCRIPTION

Excludes nodes without an arc from the graph.

USAGE

excludeNodesWithoutArcs(x)

ARGUMENTS

x	Object of class bn.

VALUES

object	Object of class *bn*.

findCisEffectsFromCloneNeighbors

DESCRIPTION

Identifies present *cis*-effects included in the *similar-state-sum.* Serves as input for generateWhiteList().

USAGE

findCisEffectsFromCloneNeighbors(x, rand.effects)

ARGUMENTS

x	Integer that specifies the *similar-state-sum*.
rand.effects	Matrix returned by checkGenomicEffects(effect == "*trans*") when used with a permuted data set.

VALUES

list	List which contains for each aCGH probe a vector of gene probes in *cis*-position.

frequencyPlot

DESCRIPTION

Plots a frequency plot for a aCGH dataset.

USAGE

```
frequencyPlot(data)
```

ARGUMENTS

data	Matrix representing a aCGH data matrix

VALUES

data.frame	Dataframe with percentage of loses and gains.

generateBlackList

DESCRIPTION

Prepares a set of arcs to be definitely not included in the Bayesian network. Serves as input parameter for computeBN().

USAGE

```
generateBlackList(x)
```

ARGUMENTS

x	List which contains for each aCGH probe a vector of gene probes in *cis*-position

VALUES

data.frame	Data frame with two columns, containing a set of arcs to be definitely not included in the Bayesian network.

generateWhiteList

DESCRIPTION

Prepares a set of arcs to be definitely included in the Bayesian network. Serves as input parameter for computeBN().

USAGE

generateWhiteList(x)

ARGUMENTS

x	List which contains for each aCGH probe a vector of gene probes in *cis*-position

VALUES

data.frame	Data frame with two columns, containing a set of arcs to be definitely included in the Bayesian network.

getRelatedEffects

DESCRIPTION

Filters out all *cis* and *trans*-effects that do not reach the user specified threshold. The effects are represented as a *similar-state-sum*.

USAGE

getRelatedEffects(x, threshold)

ARGUMENTS

x	Matrix returned by checkGenomicEffects().
threshold	Integer that specifies a user defined threshold for the *similar-state-sum*.

VALUES

list	Filtered list which contains for each aCGH probe a vector of matched gene probes.

kMeans

DESCRIPTION

Performes k-means clustering on a data matrix. The function classifies each row of the gene expression data matrix into three classes.

USAGE

```
kMeans(matr)
```

ARGUMENTS

matr	Matrix of gene expression data.

VALUES

matrix	Matrix of classified gene expression data. Values are -1, 0, 1.

mapGeneIdToBacClone

DESCRIPTION

Matches the gene expression probes with probes/BAC clones from an aCGH microarray. This function requires the chromosomal start and end points of the spotted probes. Matching gene to aCGH probes are saved as *cis*-effects.

USAGE

```
mapGeneIdToBacClone(id.gene,chr.gene, start.gene, end.gene,
id.bac,chr.bac, start.bac, end.bac,debug)
```

ARGUMENTS

id.gene	Vector of gene ids
chr.gene	Vector of chromosome information for each gene probe. Length of id.gene.
start.gene	Vector of start positions for each gene probe. Length of id.gene
end.gene	Vector of end positions for each gene probe. Length of id.gene
id.bac	Vector of aCHG probe ids
chr.bac	Vector of chromosome information for each aCGH probe. Length of id.bac.
start.bac	Vector of start positions for each aCGH probe. Length of id.bac
end.bac	Vector of end positions for each aCGH probe. Length of id.bac
debug	Boolean. If TRUE progress information will be printed out

VALUES

An object of type list.

bacsASlist	A list which contains for each aCGH probe a vector of matched gene probes.
midpoint	Vector of midpoints for each matched aCGH probe.
chromosome	Vector of chromosome information for each matched aCGH probe.

pFromRandomEffects

DESCRIPTION

Computes for each *similar-state-sum* a p-value.

USAGE

```
pFromRandomEffects(x, rand.effects)
```

ARGUMENTS

x	Integer that specifies the *similar-state-sum*.
rand.effects	Matrix returned by checkGenomicEffects() when used with a permuted data set.

VALUES

list	Filtered list which contains for each aCGH probe a vector of matched gene probes.

singleProfilePlot

DESCRIPTION

Plots a aCGH profile of a single patient.

USAGE

```
singleProfilePlot(profile)
```

ARGUMENTS

profle	Vector representing the aCGH data of a single patient.

VALUES

no return value

TBImass

Functions for the integration of mass spectrometry and DNA-microarray data.

These functions are written by the author and are part of the TBImass package. Functions that are written by Mirjam Maier are labeled by a *. All functions and a complete workflow with example data are provide at the enclosed CD.

---

alignSpecs

DESCRIPTION

Does a two-step-interpolation of mass spectrometry dataset. The first step approximates the missing data points such that the m/z intervals on the x-axis were given at equal resolution and the spectra were set to a common m/z range. The second step restricted the interpolation to the smallest common m/z range.

USAGE

```
alignSpecs(specs, specs.obj)
```

ARGUMENTS

specs	Object of class specs
specs.obj	Boolean. If TRUE the return type is of type *specs* otherwise a matrix will be returned.

VALUES

specs (if specs.obj == TRUE)	Object of class specs.
matrix (if specs.obj == FALSE)	Matrix of mass spectrometry data.

aveSpecs

DESCRIPTION

Estimates a mean spectrum for each class.

USAGE

```
aveSpecs(specs, align.specs,ave.all, ave.each,bsl.cor)
```

ARGUMENTS

specs	Object of class specs.
align.specs	Boolean. If TRUE alignSpecs will be called.
ave.all	Boolean. If TRUE a mean spectrum of all spectra will be computed.
ave.each	Boolean. If TRUE a mean spectrum for each class will be computed.
bsl.cor	Boolean. If TRUE bslnOff will be called.

VALUES

specs	Object of class specs

bslnOff

DESCRIPTION

Performs a base line correction of mass spectrometry profiles.

USAGE

```
bslnOff(raw)
```

ARGUMENTS

raw	Object of class specs.

VALUES

specs	Object of class specs

dataInput

DESCRIPTION

Reads mass spectrometry raw data from the file system.

USAGE

```
dataInput(dir.data, sep, skipnr, pattern, header)
```

ARGUMENTS

dir.data	String with directory that contains the raw data.
sep	String with delimeter.
skipnr	Integer that specifies the number of columns to skip..
pattern	String with the file extension.
header	Boolean. If TRUE files contain a header.

VALUES

specs	Object of class specs

normalize

DESCRIPTION

Normalizes mass spectrometry raw data.

USAGE

```
normalize(specs, norm.type, cutoff)
```

ARGUMENTS

specs	Object of class specs
norm.type	String that defines the normalization method.
	"Sum" = m/z values will be divided by the sum of all m/z values.
	"Median" = m/z values will be divided by the median of all m/z values.
	"Mean" = m/z values will be divided by the mean of all m/z values.
cutoff	Integer that defines the minimum m/z-value.

VALUES

specs	Object of class specs

pearson

DESCRIPTION

Computes the pearson correlation value for a mean spectrum and the mass spectrometry profiles.

USAGE

```
pearson(specs, method = "overall")
```

ARGUMENTS

specs	Object of class specs
method	String that defines which mean spectra to take. *"overall"* = mean spectrum of all spectra will be used. *"Median"* = mean spectrum of each class will be used. *"Mean"* = m/z values will be divided by the mean of all m/z values.

VALUES

specs	Object of class specs

plotMeanSpecs

DESCRIPTION

Plots the mean spectra for each class. The top part shows the mean spectra and the lower part displays all spectra split by their class

USAGE

```
plotMeanSpecs(specs, p.type=, see.raw, peak.oi , p.r=200)
```

ARGUMENTS

specs	Object of class specs
p.type	String that defines the design of the plot. *"mean"* = mean spectra of all spectra will be displayed at the upper part and the class specific spectra at the bottom. *"zoom.mean"* = zooms only the mean spectrum. *"zoom.mean.raw"* = zooms only the class specific spectra

VALUES

no return value

plotIntersectFeatures*

DESCRIPTION

Plots the intersection peaks that arise from different classification algorithms and feature selection methods.

USAGE

```
intersectFeatures(intersectFeatures, c1, c2, specs)
```

ARGUMENTS

intersectFeatures	List that holds for each classification algorithm a vector with peaks
c1	String that names the class
c2	String that names the class
specs	Object of class specs

VALUES

no return value