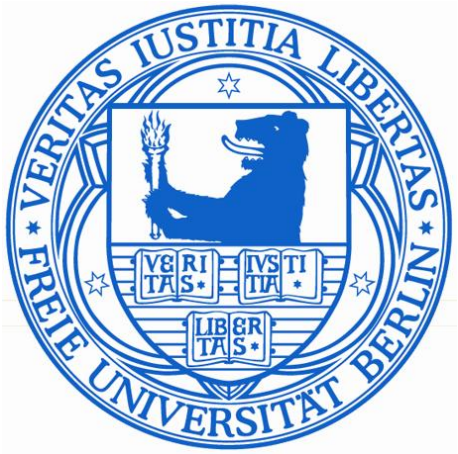
EXPLOITING PROTEOMICS DATA

by Chris Bauer, BSc, MSc



Department of Mathematics and Computer Science
Free University of Berlin

MicroDiscovery GmbH

for fulfillment of the requirements for the degree of
Doktor der Naturwissenschaften

Supervisor: Professor Dr. Knut Reinert

Second Supervisor: Professor Dr. Joachim Selbig

Supervisor MicroDiscovery: Dr. Johannes Schuchhardt

Disputation: 2013-05-07

August 2012

ABSTRACT

Proteomics plays a central role in understanding complex disease mechanisms, especially since it is well known that the effectors of biological functions are mostly proteins. Beside classical gel-based techniques especially Mass Spectrometry (MS) has emerged as the standard technique for proteomics experiments. MS-based proteomics has evolved into several different and partly complementary technologies. In this thesis we have analyzed data generated by the three complementary technologies: Matrix-Assisted Laser Desorption/Ionization (MALDI), Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) and 2D Difference Gel Electrophoresis (DIGE). The three technologies are applied to an obesity-induced mouse model in order to gain relevant knowledge on biological processes involved in diabetes. The primary goal of this thesis is to develop and implement specifically tailored data analysis methods for each technology with the aim to improve quality and reliability of the results compared to standard evaluation workflows.

The developed methods benefit from the fact that in proteomics a single protein is typically represented by several peptides showing more or less similar measurements. Combining this similarity information and advanced statistical testing, we are able to identify sets of potential biomarkers that may play an important role in diabetes disease mechanisms. The identified biomarkers are very well suited for building a classification engine to predict disease relations. However, peptides derived from the same protein may also show contradictory quantitations (e.g. a protein is two-fold up regulated and two-fold down regulated at the same time). This could be due to technical artifacts or biological properties (e.g. protein isoforms). We try to resolve these contradictions with PPINGUIN, a workflow developed for the reliable quantitation of iTRAQ experiments. Application of the developed methods leads to improved results compared to standard data evaluation methods.

The three technologies have a complementary character and therefore a direct comparison is difficult and shows only a small overlap. But a comparison based on the more abstract level of biochemical pathways shows a surprisingly good agreement of the results. In order to better understand the complex processes involved in diabetes a major challenge remains in integrating the results with other 'omics' technologies.

ZUSAMMENFASSUNG

Für das Verständnis von komplexen Krankheitsmechanismen ist Proteomics von zentraler Bedeutung, besonders da biologische Funktionen hauptsächlich von Proteinen ausgeführt werden. Neben klassischen Gel-basierten Verfahren hat sich vor allem Massenspektrometrie als Standardtechnologie für Proteomics-Experimente etabliert. Verschiedene, zum Teil komplementäre Technologien wurden entwickelt um Proteine zu untersuchen. In dieser Arbeit wurden drei verschiedene Technologien: MALDI, iTRAQ, und DIGE auf ein Maus-Modell angewandt. Das Ziel dabei ist wichtige Diabetis-bezogene biologische Prozesse zu analysieren. Das Hauptziel der vorliegenden Arbeit ist es, für jede der Technologien eine spezifische Datenanalysestrategie zu entwickeln, um die Qualität und die Reliabilität der Ergebnisse im Vergleich zu herkömmlichen Auswertungen zu verbessern.

Eine wichtige Eigenschaft von Proteomics Experimenten ist, dass ein einzelnes Protein oft durch eine Vielzahl von Peptiden charakterisiert wird. Die entwickelten Datenanalysestrategien machen sich diese Eigenschaft zu Nutze. Peptide, die vom selben Protein stammen, zeigen häufig ähnliche Messwerte. Diese Ähnlichkeit kombiniert mit statistischen Tests ermöglicht es, potentielle Biomarker zu identifizieren, die eine wichtige Rolle für Diabetis spielen. Die so identifizierten Biomarker sind sehr gut geeignet, um krankheitsrelevante Assoziationen zu prädizieren. Allerdings kommt es des Öfteren vor, dass Peptide, die vom selben Protein stammen ein widersprüchliches Signal aufweisen (z.B. Peptide die zweifach hoch- und andere die zweifach runterreguliert sind). Dieser Widerspruch kann entweder ein technisches Artefakt oder aber eine biologische Eigenart (z.B. Proteinisoformen) sein. Um diesen Widerspruch aufzulösen wurde PPINGUIN entwickelt, ein Workflow für die verlässliche Quantifikation von iTRAQ Daten. Im Vergleich zu herkömmlichen Auswertungen führen die entwickelten Verfahren zu verlässlicheren Ergebnissen.

Ein direkter Vergleich der drei Technologien wird durch den komplementären Charakter erschwert und führt auch nur zu wenigen Übereinstimmungen. Vergleicht man die Ergebnisse aber auf einem abstrakteren Level von molekularen Pathways, so ist der Überlap der unterschiedlichen Methoden erstaunlich hoch. Dennoch liegt die wohl größte Herausforderung um komplexe Krankheitsmechanismen in Zukunft besser zu verstehen in der Integration mit anderen 'Omics'-Technologien.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

C. Bauer, F. Kleinjung, D. Ruthishauser, C. Panse, A. Chadt, T. Dreja, H. Al-Hasani, K. Reinert, R. Schlapbach, and J. Schuchhardt. PPINGUIN: Peptide Profiling Guided Identification of Proteins Improves Quantitation of iTRAQ Ratios. *BMC Bioinformatics*, 13:34, Feb 2012

C. Bauer, F. Kleinjung, C. J. Smith, M. W. Towers, A. Tiss, A. Chadt, T. Dreja, D. Beule, H. Al-Hasani, K. Reinert, J. Schuchhardt, and R. Cramer. Biomarker discovery and redundancy reduction towards classification using a multi-factorial MALDI-TOF MS T2DM mouse model dataset. *BMC Bioinformatics*, 12:140, 2011

C. Bauer, R. Cramer, and J. Schuchhardt. Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol. Biol.*, 696:341–352, 2011

ACKNOWLEDGMENTS

It would not have been possible to perform this work and write this thesis without the support of MicroDiscovery GmbH and the patience and guidance of the kind people around me. It is to them that I want to express my deepest gratitude.

First and foremost, my special thanks goes to my supervisor at MicroDiscovery Dr. Johannes Schuchhardt, who was always a source of inspiration and motivation. I also want to thank my former colleague Dr. Frank Kleinjung for his help, advice and support especially in harder times throughout my work. I like to express further greatest thanks to my other colleagues and friends for scientific discussion, advice and continuous support.

I also want to express my special thanks to my first and second supervisor, Prof. Dr. Knut Reinert and Prof. Dr. Joachim Selbig.

I like to thank my family for the opportunity and encouragement to start and pursue a career in scientific research. I am particularly indebted to my parents for never-ending support especially to my father whose weekly question for the finishing date was a great source of motivation.

Last but not least, I am profoundly grateful to my wonderful wife, Josefine, for her patience, support and friendship through all these years. You are my sunshine every day and I am eternally grateful for your endless love.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

ACO	Ant Colony optimization
ANOVA	Analysis of Variance
AUC	Area Under Curve
B6	Black Six - C57BL/6J
CHF	Carbohydrate-Free Diet
CID	Collision-Induced Dissociation

CO	Combinatorial Optimization
CV	Coefficient of Variation
Cy2	3-(4-carboxymethyl)phenylmethyl-3′-ethyloxacarbocyanine halide
Cy3	1-(5-carboxypentyl)-1′-propylindocarbocyanine halide
Cy5	1-(5-carboxypentyl)-1′-methylindocarbocyanine halide
CWT	Continuous Wavelet Transform
Da	Dalton
DIfE	German Institute of Human Nutrition
DIGE	2D Difference Gel Electrophoresis
ESI	Electrospray Ionization
E-value	Expected Value
FDR	False Discovery Rate
FGCZ	Functional Genomics Center Zurich
GC-MS	Gas Chromatography Mass Spectrometry
GEO	Gene Expression Omnibus
GO	Gene Ontology
HF	High Fat
HPLC	High-performance liquid chromatography
ICAT	Isotope-Coded Affinity Tags
iTRAQ	Isobaric Tags for Relative and Absolute Quantitation
PTM	Post Translational Modification
MAD	Median Absolute Deviation
MALDI	Matrix-Assisted Laser Desorption/Ionization
MCMC	Monte Carlo Markov Chain
MPIMG	Max-Planck Institute for Molecular Genetics
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
m/z	mass/charge ratio
NA	Not Available
NFD	Number of False Discoveries
NZO	New Zealand Obese
PEG	Polyethylene Glycol

PPINGUIN	Peptide Profiling Guided Identification of Proteins
ppm	Part Per Million
RF	Random Forest
ROC	Receiver Operating Characteristic
RSS	Residual Sum of Squares
SD	Standard Diet
SILAC	Stable Isotope Labeling by Amino Acids in Cell Culture
SJL	Swiss Jim Lambert
SNR	Signal to Noise Ratio
SVM	Support Vector Machines
Sys-Prot	System-wide analysis and modelling of protein modification
T2DM	Type-2 Diabetes Mellitus
TOF	Time-of-Flight
UV	UltraViolet

1 INTRODUCTION

With the successful completion of the Human Genome Project[91], the foundation for a triumphant advance of transcriptomics and genomics was formed. Ever since a multitude of experiments addressing various kinds of diseases have been performed aiming at the characterization of disease-related transcriptomic and genomic alterations. Genomics was supposed to meet very high expectations:

> *"Within a decade, gene chips will offer a road map for prevention of illness throughout a lifetime".*[1]

Although reams of associated genetic loci, almost endless lists of differentially regulated transcripts and a plethora of associated functional annotations such pathways or Gene Ontology (GO) terms have been reported for many diseases, these approaches often did not meet the high expectations towards fully understanding the disease mechanisms. This is, however, not surprising keeping in mind that effectors of biological functions are almost exclusively proteins[19]. The level of protein expression not only depends on the transcript abundance but also on translational controls, degradation mechanisms or post-translational effects. Therefore more recently, the biologists' attention was expanded towards analysis of the proteome.

The shift from genomics to proteomics is clearly visible counting the occurrences of the terms 'genomics' and 'proteomics' in abstracts of all publications from 1996 till 2010 listed in PubMed (see Figure 1). While in the late 1990s genomics came into focus of scientific research with a exponential increase each year, proteomics started only at the beginning of the new millennium. Since then, both areas continuously increased in importance but however proteomics was trailing behind about three years. Since 2007 the number of publications for genomics seems to be rather constant while proteomics still increases in importance.

The term proteome was initially defined as the protein complement expressed by a genome[172]. The total number of biomolecules encompassed by the proteome is significantly larger compared to the genome. Currently, a total of ~ 56K human genes are annotated (*Ensembl[50] release 66 - February 2012*), ~ 22K of which are protein encoding. Especially due to alternative splicing (the estimated amount of alternatively spliced genes in human varies from 60%[26] to 94%[168]) the number of proteins is almost 5-6

1 President Bill Clinton's State Of The Union Address Part 2 - Jan. 27, 1998

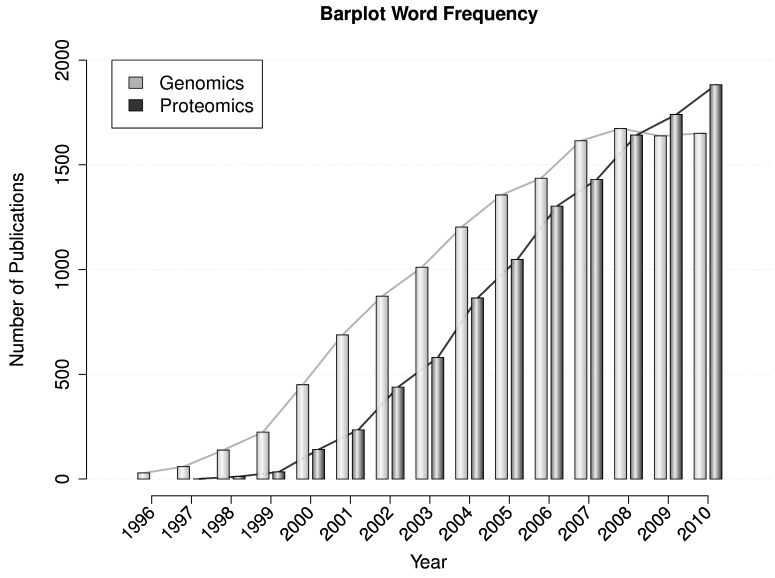


Figure 1: Frequencies of the words 'genomics' and 'proteomics' in PubMed publications from 1996 to 2010. Only occurrences in publication abstracts were counted.

times the number of known protein coding genes: UniProt Protein Knowledgebase currently comprises ~ 125K proteins (*UniProt[6] release 2012_02 - February 22, 2012*).

The term proteomics not only refers to the study of the proteome but also to the investigation of protein isoforms, Post Translational Modifications (PTMs) or protein interactions[162]. In particular PTMs are of big interest as they can determine activity state, localization, turnover or interactions with other proteins[102, 117].

Beside this huge amount of biomolecules, proteomics has to deal with more problems like limited and variable sample availability, a very high dynamic range as well as temporal and spatial specificity[162]. Especially the latter substantiates the dynamic nature of the proteome as it is not a static entity but highly variable in time and localization.

Within the last years proteomics has been a rapidly developing area and many advances were made especially towards quantitative proteomics. Due to these developments, in 2007 Cox and Mann[30] argued that proteomics could become the 'new genomics':

> *MS-based proteomics is finally ready for systems-wide measurement of protein expression levels. If so, many of the powerful systems-wide approaches previously restricted to the mRNA level could now be performed directly at the protein level.*

Even if proteomic approaches allow for high-accuracy protein quantification for several thousand proteins in complex proteomes[31], it is still a long way to go for measuring the complete proteome. The future will tell if proteomics could really become the 'new genomics'.

1.1 GOALS OF THIS THESIS

The principal goal of the thesis is an exhaustive statistical evaluation of three different complementary proteomics datasets (MALDI, iTRAQ and DIGE) created within the Sys-Prot project (EU sixth framework programme for research and technical development - Project Reference: 37457). In order to achieve this objective we developed specifically tailored methods for data evaluation. For each technology, state-of-the-art data analysis methods were inspected, extended and integrated into our analysis workflow in order to improve the quality and reliability of the results.

The experiments were performed to gain knowledge on disease mechanisms underlying Type-2 Diabetes Mellitus (T2DM). T2DM was chosen as the experimental objective since it is among the most common chronic diseases in nearly all countries and subject to intensive biomedical research.

For most of the proteomics technologies, the measurements are performed at the level of peptides which may later be assigned to proteins. Peptides are derived from proteins and a single protein is typically measured by several (more or less redundant) peptides. Peptides derived from the same protein are expected to show similar experimental measurements or a least a similar differential regulation. This property is a recurring element for the evaluation of the different datasets.

MALDI

For evaluation of MALDI data, we were aiming at developing an analysis workflow that is able to identify significant biomarkers (peaks) characteristic for the different mouse genotypes, the different diets or the growth of the mice. The combination of genotype and diet is of particular interest since one of the investigated mouse strains is resistant to obesity and diabetes, induced by high fat diet. Therefore, a special attention is paid to the identification of biomarkers which are characteristic for the combination of a specific mouse genotype and the applied diet.

Proteins are often measured by several peptides resulting in correlated peaks in the spectra. In order to group peptides that are potentially derived from the same protein we employ a correlation-based hierarchical clustering using experimental measurements. Both types of information about peak similar-

ity and peak significance are independent. The combination of both allows to identify sets of disease associated peaks that are potentially representing the same protein. The visualization of the combined information in a cluster dendrogram enables the interpretation of complex biological data in an intuitive manner. Furthermore, this method is very well suited for the purpose of feature selection for classification and prediction. The classification performance based on the selected set of features is similar to more complicated global optimization strategies.

Prior to statistical evaluation, a specialized preprocessing covering all important aspects such as baseline correction, peak picking and peak alignment guarantees that the methods can be applied and sample contaminations are removed.

iTRAQ

In the iTRAQ data we observed a considerable heterogeneity of peptides assigned to the same protein. This heterogeneity was often contradictory since the peptide spectra may indicate that the corresponding protein is (two-fold) up regulated and (two-fold) down regulated at the same time. With the aim to resolve these contradictions, we developed a workflow named Peptide Profiling Guided Identification of Proteins (PPINGUIN). This workflow employs a coarse-grained clustering of unidentified spectra (representing peptides) as an early step in data processing. Similar to the clustering used for MALDI data, this clustering was intended to group spectra (peptides) which are potentially derived from the same protein. Protein quantitation and identification was performed afterwards for each of the clusters independently.

We compared the results of PPINGUIN to state-of-the-art approaches using different aspects such as heterogeneity of peptides for the same protein, experimental reproducibility and accordance with prior knowledge. In result, the application of PPINGUIN led to more homogeneous peptide profiles without contradictions and to more reliable protein quantitation.

However, if a protein has differentially regulated isoforms (e.g. PTMs), they may result in contradictory quantitations. For instance, a phosphorylated protein might show an other differential regulation compared to the unphosphorylated protein. A protein with many peptides with contradictory quantitations, may be found in more than one cluster, which we see as a hint for potential protein isoforms. So application of PPINGUIN also allows the detection of potential novel protein isoforms.

DIGE

For the analysis of DIGE data, we wanted to develop a specialized pre-processing that benefits from the common reference pool

present on each gel. Furthermore we were investigating the existence of a protein specific dye effect that has been observed previously for DIGE data.

The experimental objective was similar to the evaluation of MALDI data. Our primary goal was to identify significant biomarkers (spots) characteristic for the different mouse genotypes or the different diets. Again, a special attention is paid to the identification of biomarkers, which are characteristic for the combination of a specific mouse genotype and the applied diet.

In DIGE data, different spots may represent the same or related proteins. We applied correlation-based hierarchical clustering in order to identify spots potentially representing the same protein. We found that spots in close proximity in the gel may represent the same protein or protein isoforms. The measurements of these spots are often correlated which may be an effect of PTMs or the labeling process.

1.2 STRUCTURE OF THIS THESIS

Chapter 2 gives insights into technical background and general statistical methods. The technical background of MALDI, iTRAQ and DIGE is described in Sections 2.1.1, 2.1.2 and 2.1.3. A detailed description of Sys-Prot project including the workflow of samples and data of all involved partners is given in Section 2.2.3. ANOVA as a popular method for multi-dimensional data analysis is briefly review in Section 2.3.

The main structure of this thesis is organized according to the three complementary proteomics approaches, which are referred to in different chapters of this thesis: MALDI - Chapter 3; iTRAQ - Chapter 4 and DIGE - Chapter 5. The three data analysis chapters are structured in a similar manner. Each chapter starts with an introduction followed by a brief description of the corresponding dataset and experimental design. State-of-the-art methods for data analysis are described in corresponding section of each chapter. Algorithms specifically developed in this thesis are referred to in the Methods section. Results of each data analysis are presented in the corresponding sections followed by a discussion and conclusion of the corresponding technology.

Subsequently, in the Chapter 6 the three different approaches are integrated and compared. Finally, Chapter 7 gives a conclusion and a brief outlook.

2

BACKGROUND

Chapter Contents

2.1 PROTEOMICS TECHNOLOGIES

Actually, the usage of high throughput technologies to analyze proteins has a quite long history. Beginning from the 1970s two-dimensional gel electrophoresis was engaged for proteomic research. But this technology never fully delivered on its promise of quantifying the proteome[30]. It was the development of Electrospray Ionization (ESI) and MALDI that focused the attention on MS as a versatile technique for proteomics research. The importance of this breakthrough was approved as John B. Fenn and Koichi Tanaka were awarded with the Nobel Prize in Chemistry for development of protein ionization methods in 2002. Ever since, MS has become the predominant technology used for proteomic research.

For some time MS has been restricted to qualitative analysis, but, in the last years MS turned towards quantitative investigations[118]. The capability for relative or even absolute protein quantification allows for investigation of protein concentrations under different conditions. Furthermore, the growing field of systems biology increasingly requires data including quantitative readouts as input for model construction and validation.

In principle, if absolute quantification is available relative readouts become redundant as they can be calculated easily. But absolute quantitation is difficult to achieve and more easily available relative readouts are often the alternative. Towards quantitative proteomics different mass spectral techniques have been developed and widely used encompassing labeling techniques such as iTRAQ[139], Isotope-Coded Affinity Tags (ICAT)[64] and Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)[119, 101] as well as some label free approaches [115]. The different quantitative techniques often play a complementary role in proteomics research since each technique has strengths and weaknesses.

2.1.1 *MALDI*

MALDI was first described in 1987 by Karas et al.[79]. In 1988 MALDI MS had been used for ionization of large biomolecules such as the bovine insulin (5733 Dalton (Da)), cytochrome-C (12384 Da) and carboxypeptidase-A (34472 Da) by Tanaka et al.[152].

Since then, MALDI-MS, particularly in combination with Time-of-Flight (TOF) instruments has become a promising tool for proteomics data acquisition[33] characterized by simplicity, good mass accuracy and high resolution[1]. It allows for processing a significant number of samples in a short time and therefore enables studies encompassing a multitude of samples[107, 54, 122]. MALDI-TOF MS profiling has been extensively used especially for investigating different types of cancer like breast cancer[165], lung cancer[54, 167], ovarian cancer[158] or colon cancer[37, 4], to name a few.

A modern MALDI-TOF instrument is made out of 3 main components:

1. A sample slide (containing matrix and analyst) and a pulsed laser (typically an UltraViolet (UV) laser). The matrix contains a large molar excess of chromaphores coupled with the laser frequency, causing essentially all laser radiation to be absorbed. Analyte molecules surrounded by matrix and salt ions are ejected, whereas the matrix molecules evaporate leaving the free ionized analyte.

2. A TOF analyzer accelerating ions by an electric field leading to a separation of the ions based on their mass/charge ratio (m/z). The essential principle is that if a cohort of ions moving in the same direction and having a constant kinetic energy but a distribution of masses, the resulting velocities are inversely proportional to $\sqrt{m/z}$ [62].

3. An ion detector usually involves photographic plates, faraday cylinders, or array detectors in conjunction with elec-

tron or photon multipliers to increase the intensity of the signal.

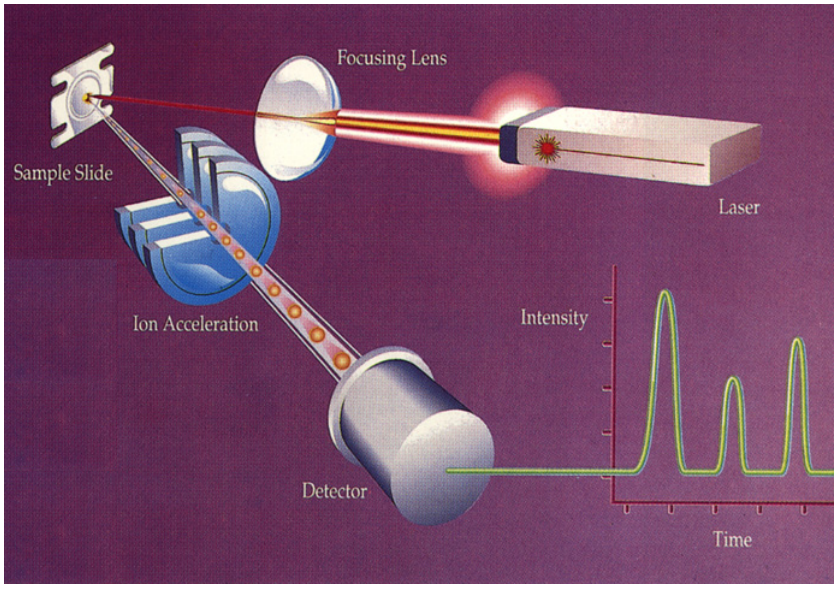


Figure 2: Scheme describing a modern MALDI-TOF mass spectrometer. Image is adopted from Finnigan[48].

The ions are traveling through an electric field (with potential V). When the ion is accelerated the ion's potential energy ($E_{pot} = zV$; z = ion charge) is converted to kinetic energy ($E_{kin} = 1/2 \cdot m \cdot v^2$; where m = mass and v = velocity). As the ion travels with constant velocity through the time-of-flight tube, the velocity can be described as the tube's length (d) divided by the travel time (t): $v = d/t$. Assuming potential energy is converted to kinetic energy and substituting velocity by $v = d/t$ leads to the following equation:

$$zV = \frac{1}{2} \cdot m \cdot \left(\frac{d}{t}\right)^2$$

that can rearranged:

$$t = \sqrt{\frac{m}{z}} \cdot \frac{d}{\sqrt{2V}}$$

As the length of the tube d and electric field V are given by the instrument settings the travel time of an ion through the time-of-flight tube is proportional to $\sqrt{m/z}$.

2.1.2 *iTRAQ*

Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) technology has initially been described in 2004 by Ross et al.[139].

They developed a set of reagents making derivatized peptides indistinguishable in MS, but exhibiting an intense low-mass Tandem Mass Spectrometry (MS/MS) signature. MS/MS is a sequential combination of two mass spectrometers. The first mass spectrometer (which can be similar to a MALDI-TOF - see section 2.1.1) is used to isolate a single precursor mass of a peptide. After a fragmentation step, the second mass spectrometer separates the fragments of the precursor ion and generates the corresponding MS/MS spectrum. The MS/MS spectrum is used for peptide identification by identification engines and protein databases.

iTRAQ isobaric reagents are placed at N termini and at the ϵ amino group of lysine side chains of peptides in a digested mixture. Isotopic labeling is constructed in the way that the resulting peptides are isobaric and chromatographically indistinguishable. After Collision-Induced Dissociation (CID) the different reagents have different signatures allowing for distinguishing individual members of the multiplexed reagent set.

The complete reagent molecules consist of a reporter group, a mass balance group and a peptide-reactive group (see figure 3). The overall mass of reporter group (mass range from m/z 114.1 - 117.1) and balance group (mass range 28 - 31 Da) are kept constant (145.1 Da). This is achieved by using differential isotopic enrichments including ^{13}C, ^{15}N and ^{18}O. Reacting with N termini or lysine side chains of peptides, the reagent forms an amine linkage similar to backbone peptide bonds ensuring similar fragment behavior (when subjected to CID). After fragmentation the balance group is lost leaving the 4 reporter groups appearing as distinct masses (114-117 Da) in the spectra. A demonstration using an exemplarily chosen spectrum is shown in figure 4. The complete spectrum is shown above and an excerpt of the spectrum zoomed to reporter region is depicted below. For each of the four reporters (m/z 114.1 - 117.1) typically corresponding to 4 different biological samples an individual intensity signal is obtained.

2.1.3 *DIGE*

Since first described in 1975 by O'Farrell[114] and Klose[85], for a long time two-dimensional gel electrophoresis was the only method available for simultaneously studying the abundance of thousands of proteins. Although more recently mass spectrometry has become the predominant technology used for proteomics, two-dimensional gel electrophoresis continued to be developed. In 1997 Unlu et al.[163] initially described the method - DIGE - allowing for multi-sample gel separation based on the two fluorescent dyes 1-(5-carboxypentyl)-1'-propylindocarbocyanine halide (Cy3) and 1-(5-carboxypentyl)-1'-methylindocarbocyanine

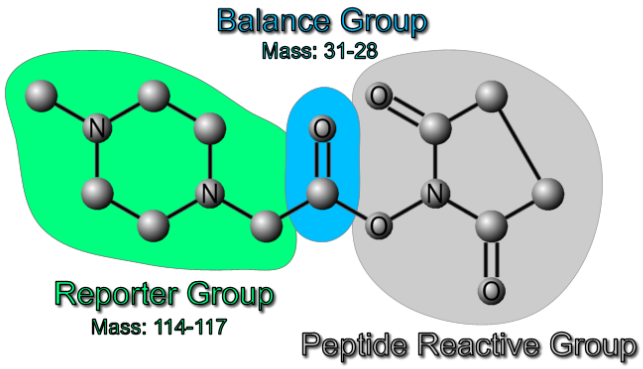


Figure 3: Chemical structure of the iTRAQ reagents. Complete molecules consist of a reporter group (left side - mass range from m/z 114.1 - 117.1), a mass balance group (middle part - mass range 28 - 31 Da) and a peptide-reactive group (right part). Image created following the images of Ross et al. and Boehm et al.[139, 17].

halide (Cy5). The dyes were developed according to 4 design rules: (*i*) dyes must react with the same amino acid (*ii*) the charge of the target must be preserved to maintain the isoelectric point (*iii*) the dyes must have similar molecular mass and (*iv*) to separate the dyes, they must have distinct fluorescent characteristics. Labeling of amino acid residues unavoidably affects molecular mass of the proteins. In order to avoid heterogeneity in protein populations, two labeling strategies are used: minimal labeling of lysine reactive dyes (labeling an estimated $< 5\%$ of total protein[161]) or alternatively saturation labeling of cysteine residues which shows improved sensitivity and dynamic range[144, 24].

The multiplexing capability leads to a substantial reduction of gel to gel variance raising confidence that observed fold changes can indeed hint to biological properties[164]. Extending this method by adding the dye 3-(4-carboxymethyl)phenylmethyl-3'-ethyloxacarbocyanine halide (Cy2) allows for multiplexing up to three samples on the same gel. But instead of multiplexing three samples Cy2 channel is often used to incorporate a pooled internal standard[3]. A common reference helps to eliminate between gel variation, as it improves a normalization and allows for statistical evaluation based on ratios instead of raw data values. Furthermore, the internal standard helps to map spots (proteins) between gels and therefore, increases comparability between different gels[87].

For DIGE, a protein mixture (for a given sample) is labeled with one of the three dyes which are spectrally distinct as well as charge and mass matched. Subsequently the labeled samples are

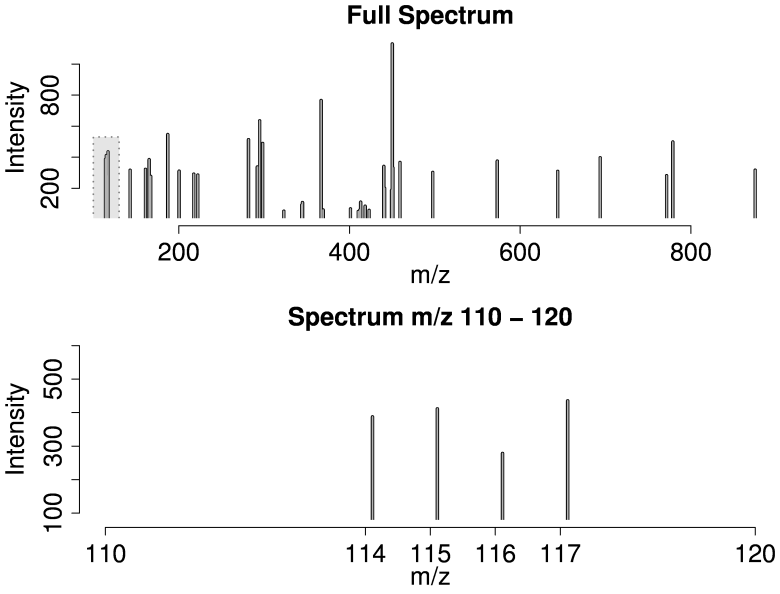


Figure 4: Exemplarily chosen spectrum. Full spectrum is drawn on upper side. Excerpt of the spectrum with zooming to reporter intensities (110-120 Da) - also marked in the full spectrum at upper part.

mixed and separated using the same gel. The fluorescent readouts are acquired by exciting each dye at the specific wavelength. Appropriate image analysis software and statistical analysis enables the selection of interesting features (e.g. differentially regulated spots). These spots can than be extracted and finally identified using MS/MS and database search. A schema of the complete process is depicted in Figure 5. Minden et al.[110] claimed that experimental design and statistical analysis were the most crucial aspects of performing informative DIGE -based proteomics experiments.

Several aspects have to be considered when performing DIGE experiments. Since DIGE is based on 2D gel electrophoresis the method is not suited for the detection of high or low molecular masses ($>$ 150kDa or $<$ 10kDa) or very basic or hydrophobic proteins. Furthermore, the labeling affects unavoidably the molecular mass leading to an electrophoretic separation of labeled and unlabeled proteins. However, for large proteins ($>$ 30kDa) the effect of a single dye molecule is negligible and even for small proteins ($<$ 30kDa) this shift is usually less than one half diameter of the protein spot[110]. A much bigger problem comes along with the assignment of protein species to a gel spot, since often multiple high confidence MS-based identifications are obtained. In this case usually the most abundant protein species is assumed to be the protein of interest which leads to a questionable protein identification [157].

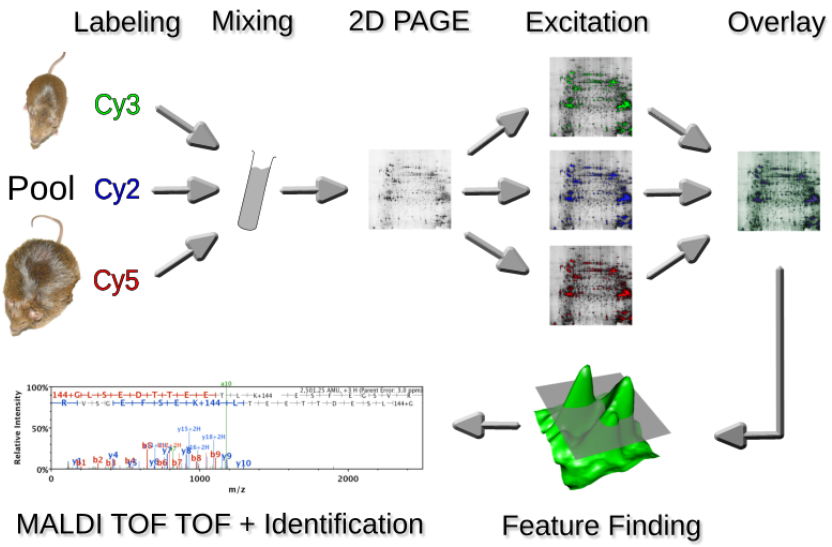


Figure 5: Scheme describing a typical DIGE experiment including labeling, image acquisition, feature finding and spot identification.

2.2 SYS-PROT PROJECT

The System-wide analysis and modelling of protein modification (Sys-Prot) project was funded by the European Commission, sixth framework programme for research and technical development (Project Reference: 37457) and involves five commercial and academic partners from three European countries. The declared aim of the project is:

> *the development of a new paradigm for the integration of proteomics data into systems biology. The goal is to gain relevant knowledge on biological processes that are important for human health and to use this knowledge for the purpose of disease modelling. The strategy to achieve this objective is an innovative, explorative systems biology approach both on the molecular and physiological level with a strong focus on protein function and modification* [151].

For establishing and developing technologies as well as novel and adequate data analysis strategies several complementary proteomics approaches (DIGE, MALDI and iTRAQ) are applied to an obesity-induced T2DM mouse model (see Figure 6). Diabetes mellitus was chosen as the experimental objective since it is among the most common chronic diseases in nearly all countries and subject to intensive biomedical research. The prevalence of diabetes will increase from 285 million in 2010 to 439 million in 2030[145]. Diabetes imposes an increasing economic burden on national health care systems world wide as 12% of the health expenditures are expected to be spent on diabetes in 2010. The global costs of treatment will raise from 418 billion USD in 2010 to 490 in 2030[182].

Figure 6: Obesity-induced T2DM mouse model (fat: NZO mouse, slim: SJL mouse).

Multiple studies have been performed assessing the diversity of the disease at the transcriptomic level revealing lists of candidate genes and associated pathways[156, 131]. At the proteomic level different techniques have been applied including gel-based[99] and MS-based quantitative approaches[130]. The majority of the studies follows a simple design and is restricted to the comparison of healthy versus diseased animal or human samples. No comprehensive proteomics study covering multiple experimental factors and comprising a multitude of samples has been published so far.

2.2.1 *Partners and Responsibilities*

The Sys-Prot project was coordinated by MicroDiscovery GmbH. The following partners are involved in the project (description of the partners are adapted from project homepage[151]):

1. **MicroDiscovery GmbH:** Based on the companies' experience in proteomics MicroDiscovery supported the partners in the analysis of the biomolecular data. A primary goal was the implementation of flexible data integration and quality control routines for different sources of complementary high-throughput proteomics data.

2. **German Institute of Human Nutrition (DIfE):** The DIfE was responsible for generation and characterization of mouse models for complex traits including polygenic obesity and T2DM.

3. **Functional Genomics Center Zurich (FGCZ):** The FGCZ's contribution to the project was the development and implementation of methods and tools based on iTRAQ technology.

4. **The BioCentre - University of Reading:** The Mass Spectrometry and Proteomics Unit at the BioCentre in Reading

uses its expertise in proteomic and mass spectrometric analysis for acquisition of MALDI and DIGE data.

5. **Max-Planck Institute for Molecular Genetics (MPIMG):** The MPIMG generated hybrid models based on model repository for signalling pathways, metabolic processes and gene regulatory pathways.

2.2.2 *Project Data and Sample Workflow*

The five project partners for Sys-Prot are scattered across three different european countries. The coordination of efforts from different project resources results in a work-flow of samples and data (see Figure 7). The German Institute of Human Nutrition (DIfE) was responsible for generation of mouse models and harvesting of the samples. After harvesting, the samples were stored in freezers. Complete batches of samples were then transported on dry ice to the BioCentre in Reading and to the Functional Genomics Center Zurich (FGCZ). While MALDI and DIGE experiments were performed by the BioCentre in Reading iTRAQ data was generated by the FGCZ. The measured data was sent to MicroDiscovery and MPIMG. MicroDiscovery was responsible for evaluation and integration of the different sources of data.

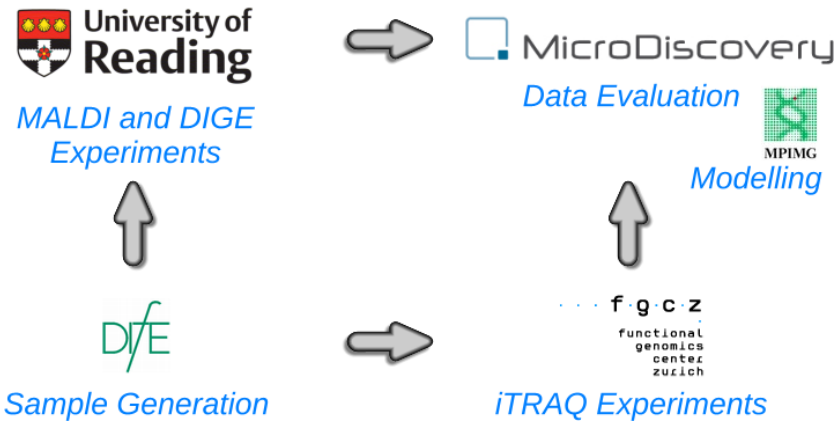


Figure 7: Schematic illustration of sample and data work-flow within the SysProt project. The biological samples are created by the DIfE. Harvested and frozen samples are sent to the BioCentre for MALDI and DIGE experiments and the FGCZ for iTRAQ experiments. Experimental data is further analyzed by MicroDiscovery and MPIMG.

2.2.3 *Sys-Prot Experimental Design*

The project is focused on the investigation of an established obesity-induced T2DM mouse model and comprises three experimental factors genotype, diet and time (see figure 8).

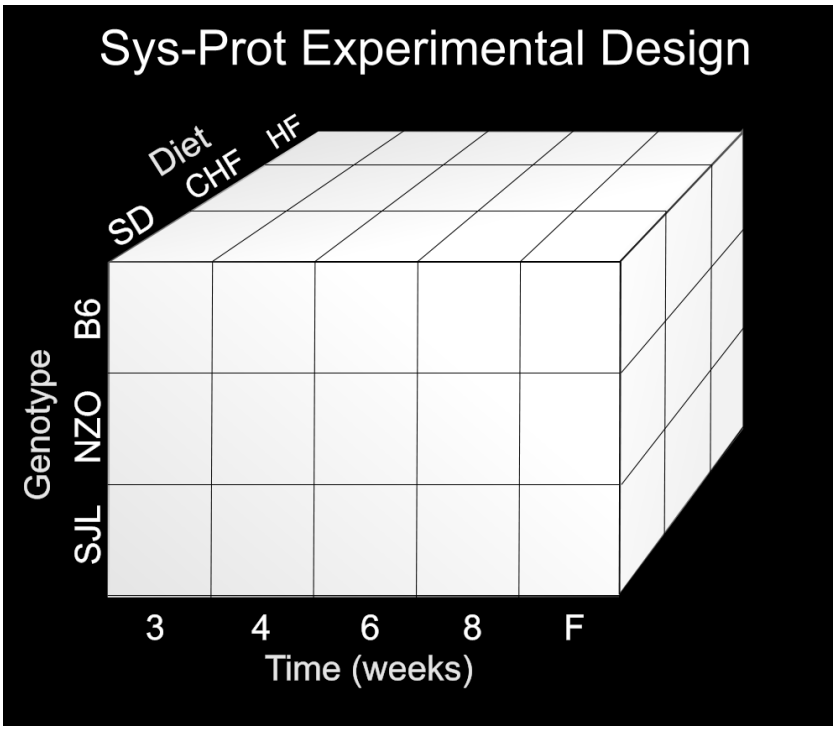


Figure 8: The three different experimental factors: genotype, diet and time used for Sys-Prot project are displayed as experimental cube. Each sub-cube (quadrant) represents a distinct combination of the three experimental factors.

Genotype

Three different mouse strains were examined: Black Six - C57BL/6J (B6), NZO and SJL. The NZO mouse strain exhibits polygenic obesity associated with hyperinsulinaemia and hyperglycaemia and presents additional features of a metabolic syndrome, including hypertension, and elevated levels of serum cholesterol and serum triglycerides[120]. NZO mice are highly susceptible to weight gain when fed a high-fat diet, resulting in the development of morbid obesity, with fat depots exceeding 40% of total body weight and the development of T2DM[75]. In contrast, the SJL mouse strain is lean and resistant to diet-induced obesity and diabetes[171]. B6 mice represent an intermediary phenotype between NZO and SJL at later age (> week 12) with respect to sensitivity to diet-induced obesity and diabetes. While the genetic and molecular basis for the different diabetes susceptibilities of polygenic mouse strains is largely unknown, Chadt et al. recently identified a naturally occurring loss-of-function mutation in the Tbc1d1 gene in SJL mice that increases lipid oxidation in skeletal muscle and as a result confers leanness and protects from diet-induced obesity and diabetes[23].

Diet

After weaning at week 3, male B6, NZO and SJL mice were raised on three different diets, a low fat diet (Standard Diet (SD); 8% calories from fat) and two different high fat diets, one containing carbohydrates (High Fat (HF); 35% calories from fat) the other one a carbohydrate-free (Carbohydrate-Free Diet (CHF); 72% calories from fat). It was shown previously that HF diet strongly induces insulin resistance and may lead to diabetes, whereas CHF equally induces peripheral insulin resistance but protects from diabetes [39, 76]. At week 8, mean body weight of SJL mice was 18.81 g (+/- 1.46 g) on SD, 20.04 g (+/- 0.99 g) on HF and 21.24 g (+/- 2.31 g) on CHF. In contrast, mean values for NZO mice were 31.94 g (+/- 1.36 g) on SD, 33.72 g (+/- 4.39 g) on HF and 36.6 g (+/- 4.83 g) on CHF, respectively. Mean values for B6 mice were 20.1 g (+/- 2.56 g) on SD, 20.54 g (+/- 0.78 g) on HF and 22.32 g (+/- 1.38 g) on CHF, respectively. The mice were sacrificed at 8 weeks of age and tissue and blood samples were processed and analyzed.

Time

Blood samples were collected at ages of 3, 4, 6 and 8 weeks from mouse tails. Final blood samples were taken directly from the heart at week 8. This sample was labeled F.

2.3 ANOVA

Analysis of Variance (ANOVA) describes a collection of approaches for statistical data evaluation. The main idea of ANOVA is to partition the variance into subcomponents with respect to one or more explanatory variables[73]. Similar to t-test, performing an ANOVA requires some assumptions to be fulfilled: first normally distributed data and normally distributed errors, second homogeneity of variance (homoscedasticity) and third independent measurements. The following three types of ANOVA can be distinguished[34]:

- One-way ANOVA
- Multi-way ANOVA
- ANOVA with mixed-effect model (nested ANOVA)

2.3.1 *One-way ANOVA*

One-way ANOVA is used to test for differences in one variable describing k (two or more) independent groups, e.g. multi-stage disease. For $k = 2$ one-way ANOVA is equivalent to t-test. Let μ_i denote the mean of the i^{th} group containing n_i elements then

ANOVA tests for the null hypothesis $\mu_1 = \mu_2 = ... = \mu_k$. If the null hypothesis is rejected than at least two of the means are not equal. The result does not provide any information about how many and which means differ.

Performing the corresponding $k \cdot (k-1)/2$ pairwise t-tests, leads to a loss in significance due to the required multiple testing corrections or if multiple testing corrections are neglected to an accumulation of type I error. Furthermore pairwise t-tests would not be independent of each other.

Standard t-test is a special case covered by one-way ANOVA. Assuming equal variances and equal sample sizes, the t-test t-value is calculated as:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(var_1 + var_2)}{n}}}$$

where n is the complete number of samples: $n = \sum_i n_i$.

The relation between t-value and f-value is $f = t^2$ and hence, the f value is given as:

$$f = \frac{n \cdot (\mu_1 - \mu_2)^2}{(var_1 + var_2)} \tag{2.1}$$

Building a factor based linear model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ where $i = 1, 2$ (number of different groups) and $j = 1..n$ (number of samples) the Residual Sum of Squares (RSS) of the model is calculated as:

$$RSS = \sum_{i,j} (x_{i,j} - \mu_i)^2 = (var_1 + var_2) \cdot (n-1) \tag{2.2}$$

The difference of means equals to the variance:

$$\begin{aligned}
&(\mu_1 - \mu_2)^2 \\
&= \mu_1^2 + \mu_2^2 - 2\mu_1\mu_2 \\
&= 1/2 \cdot ((\mu_1 - \mu_2)^2 + (\mu_2 - \mu_1)^2) \\
&= 1/2 \cdot \left(4 \cdot \left(\mu_1 - \frac{\mu_1 + \mu_2}{2}\right)^2 + 4 \cdot \left(\mu_2 - \frac{\mu_1 + \mu_2}{2}\right)\right) \\
&= 2 \cdot var(\mu_1, \mu_2)
\end{aligned} \tag{2.3}$$

Replacing numerator and denominator of 2.1 with 2.2 and 2.3 the f value can be calculated as:

$$f = \frac{2 \cdot n \cdot (n-1) \cdot var(\mu_1, \mu_2)}{RSS} \tag{2.4}$$

The numerator describes the between group variability whereas the denominator reflects the within-group variability. The formula can be extended to variables with more than two groups:

$$f = \frac{k \cdot n \cdot (n-1) \cdot \mathrm{var}(\mu_1..\mu_k)}{RSS} \tag{2.5}$$

This f value is distributed as: $f \sim F(k-1, N-k)$. The f distribution is the ratio of two χ^2 distributions.

2.3.2 *Multi-way ANOVA*

Multi-way ANOVA analyzes the effects of z (two or more) independent variables each with k_z (two or more) independent groups. Figure 9 visualizes the typical task for ANOVA with two different experimental factors (Genotype and Diet) each with two different values. The four groups are distinct and not nested (no group is subgroup of another group). The essential part is again the fitting of a linear model minimizing the RSS for the four groups without regarding non-linear effects due to combination of experimental factors. The coefficients of the multi dimensional linear model are calculated by projecting the values to each dimension. Hence the coefficients are identical with one-way-ANOVA. The resulting linear model is a z dimensional hyperplane - see upper part of Figure 9. In contrast to a one-way ANOVA, for the calculation of f-values and p-values the complete model (including all z variables) is used for calculation of RSS.

2.3.3 *ANOVA with mixed-effect model*

ANOVA with mixed-effect model assesses the effects of several (not necessarily independent) variables and also accounts for the effects due to combinations of variables, e.g. analyzing the effect of different genotypes for various diets. The starting point is very similar to multi-way ANOVA but with regarding combined effects. The calculation of the linear model is more difficult. As a first step, coefficient for the factor combination is calculated as the difference between the expected value due to the individual effects and the real value. In a second step the combination effect is subtracted from the individual effects in order to calculate the single parameter effects. The resulting model is a more complex surface - see lower part of Figure 9. For the calculation of f-values and p-values, the model with feature combinations (model 1 with p_1 parameters) is compared to the model without feature

combinations (model 2 with p_2 parameters). The first model is nested within the second one and the f-value is calculated as:

$$f = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)} \tag{2.6}$$

The f value is distributed as $f \sim F(p_2 - p_1, n - p_2)$.

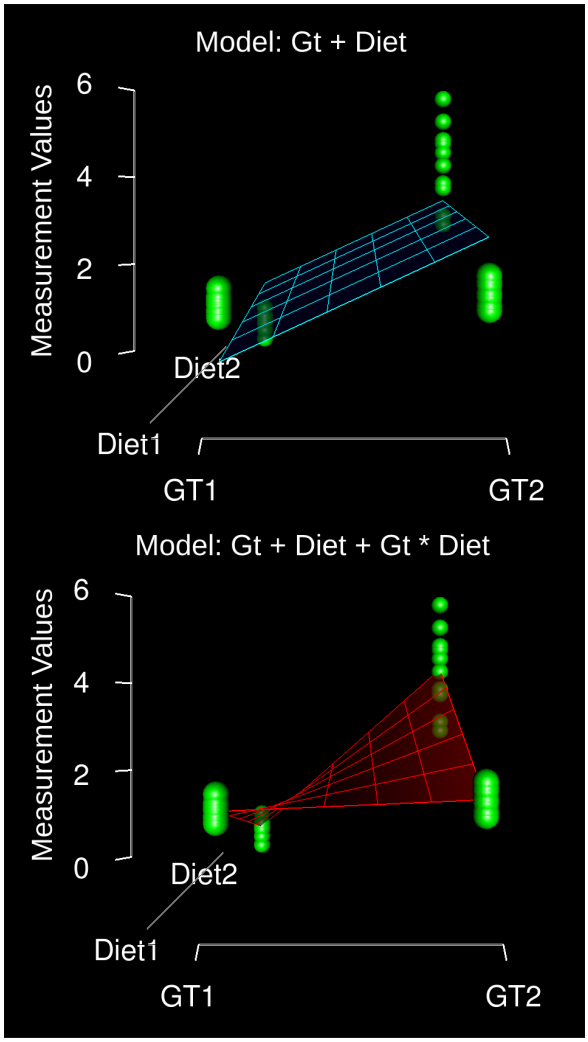


Figure 9: Demonstration of the linear model used for multi-way ANOVA (upper part) and ANOVA with mixed effects (lower part). Two variables (diet and genotype) each with two values together with the measurements span a 3 dimensional space. Multi-way ANOVA (upper part) considers each variable separately and in effect the model is a plane (blue plane). ANOVA with mixed effects (lower part) includes a combinatorial term of both variables and in effect the models is a more complex surface. The RSS of the linear model and the measurements is an essential part of an ANOVA analysis.

2.4 CLASSIFICATION AND FEATURE SELECTION

Besides biomarker identification, biological experiments are often intended for the purpose of classification, e.g. classification of tumor and control. Classification using a high dimensional dataset is always a challenging task as the problem of overfitting increases with the number of variables. Proper feature selection is essential for building a classifier that accomplishes good performance without overfitting.

Feature selection methods are distinguished into three groups: filter methods, embedded methods and wrapper methods[63]. Filter methods are independent of the classification and do not pay attention to the correlation of the features. Filter methods are typically statistical tests like t-test, χ^2-test or ANOVA. Embedded methods include the feature selection process in the construction of the classification system. Wrapper methods intend to find an optimal subset of features for the given classification task by wrapping a search algorithm around the classification model. For wrapper methods often non-linear global optimization strategies like genetic algorithms or swarm based intelligence approaches are used.

Wrapper methods often succeed in optimizing classification results but they also tend to overfitting and are computationally expensive as they scan through a huge search space. Embedded methods require complex algorithm adaptations for most classifiers and are therefore not applicable for many classification algorithms. Filter methods are straight forward[177], easy to implement and computationally fast but are often outperformed by the other methods in terms of classification performance. For a general bioinformatic review of the feature selection approaches currently applied see Saeys et al.[140]. A comparison of feature selection and classification especially with application to MALDI-MS data was published more recently by Liu et al.[98].

2.4.1 *Ant Colony Optimization*

ACO, introduced in early 1990s[28, 38], is a nature-inspired metaheuristic approach for the solution of hard Combinatorial Optimization (CO) problems. Generally, metaheuristic approaches such as simulated annealing, evolutionary algorithms or ACO are designed to obtain 'good enough' solutions to hard CO problems in a reasonable amount of computation time. A CO problem $P = (S, f)$ is an optimization problem to find a solution of minimal cost value for a given search space S (a finite set of solutions) and an objective function $f : S \mapsto \mathbb{R}^+$ that assigns a positive cost value to a given solution.

ACO is inspired by natural foraging behavior of real ants. When starting to search for food, ants initially explore their surrounding in a random manner. Having located a source of nutrition, they are carrying some of the food back to the nest. On the way back, they are depositing a trail of pheromone in order to guide the way for other ants of the colony. The level of pheromone depends on the quality and quantity of the found nutrition source. Using this indirect communication via pheromone trails, enables the ants to find shortest paths between the nest and high quality food sources.

The central part of the ACO algorithm is a pheromone model used to assign pheromone values τ_i to components of a solution S. Basically the algorithm consists of two iteratively applied steps:

1. Generation of a candidate solution in dependence of the pheromone model.
2. Updating the pheromone values using the candidate solution. The updating process intends to bias future solutions towards optimal solution.

In 2007 Ressom et al.[133] used ACO in combination with Support Vector Machines (SVM) for peak selection from MALDI-TOF data with 228 candidate peaks. Each of the 50 ants employed selected 5 peaks according to the probability function:

$$P_i(t) = \frac{(\tau_i(t))^\alpha \eta_i^\beta}{\sum_i (\tau_i(t))^\alpha \eta_i^\beta}$$

where τ_i is the pheromone value of feature i at time t, η_i represents prior information (uniform distribution or t-statistic) and α and β are used determine relative influence of pheromone values and prior information.

With the peaks selected each ant performs a classification using SVM and estimates accuracy of the classification using cross-validation. These candidate solutions are then used to update the pheromone values in such a way that high classification accuracy results in high amount of pheromone and vice versa:

$$\tau_i(t+1) = \rho \cdot \tau_i(t) + \Delta\tau_i(t)$$

where ρ is a constant between 0 and 1 representing decay of pheromones and $\Delta\tau_i$ is an amount proportional to classification accuracy.

This feature selection process is summarized schematically for 3 peaks in Figure 10. The goal is to identify features suited for classification (of dark gray and light gray). The first and third peak do not distinguish between the two groups, whereas the middle peak is well suited. In the first iteration every peak has an identical pheromone value and the ants choose all peaks with

similar probability. As deposition of pheromone depends on the accuracy of the classification result, only the pheromone level of the second peak increases. Due to the increased pheromone value, the ants choose the second peak more frequently and pheromone level further increases. In the third iteration all ant choose the middle peak due to high amount of pheromones.

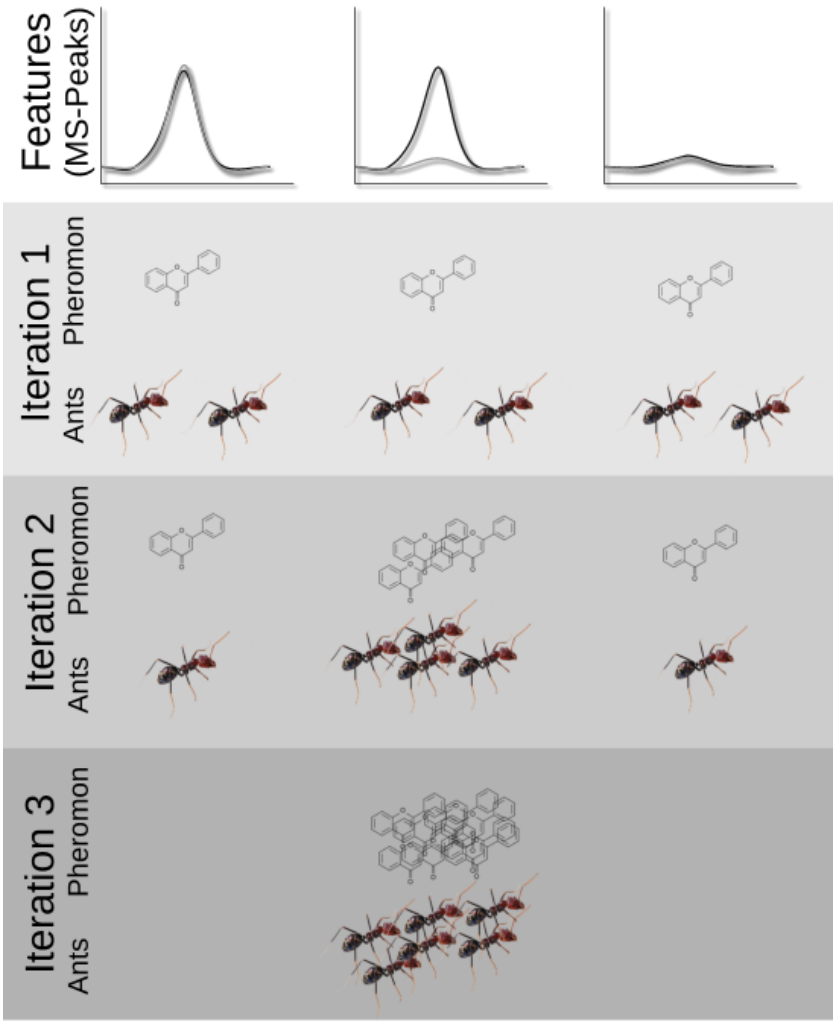


Figure 10: Schematic illustration of feature finding process using ant colony optimization. The first and third feature (MS peak) do not distinguish between the two groups (dark gray vs. light gray), whereas the middle feature is well suited for this task. A good classification performance of a peaks leads to an accumulation of pheromones.

3

MALDI: DATA ANALYSIS AND ALGORITHMS

Chapter Contents

Analysis of the Matrix-Assisted Laser Desorption/Ionization (MALDI) data aimed at the investigation of the mutual influence of different diets and mouse genotypes on composition of blood plasma proteins. The investigated diets and mouse genotypes are considered as an adequate model for Type-2 Diabetes Mellitus (T2DM). Blood plasma was selected since it is an attractive biological fluid, given its easy accessibility. Due to these properties identification of biomarkers in blood plasma is one of the 'Holy Grails' of proteomics[125]. MALDI is a promising tool for proteomics data acquisition and perfectly suited for our purpose as it allows the processing of a significant number of samples in a short time (c.f. Section 2.1.1).

For the evaluation of MALDI data, we developed a statistical method that exploits data correlation and integrates this method into a comprehensive work-flow designed for the analysis of multi-factorial MALDI-TOF MS data. Basically the work-flow for

evaluation of the data is divided into two different parts: preprocessing and biomarker discovery. The summarizing work-flow of the complete analysis is presented in Figure 11.

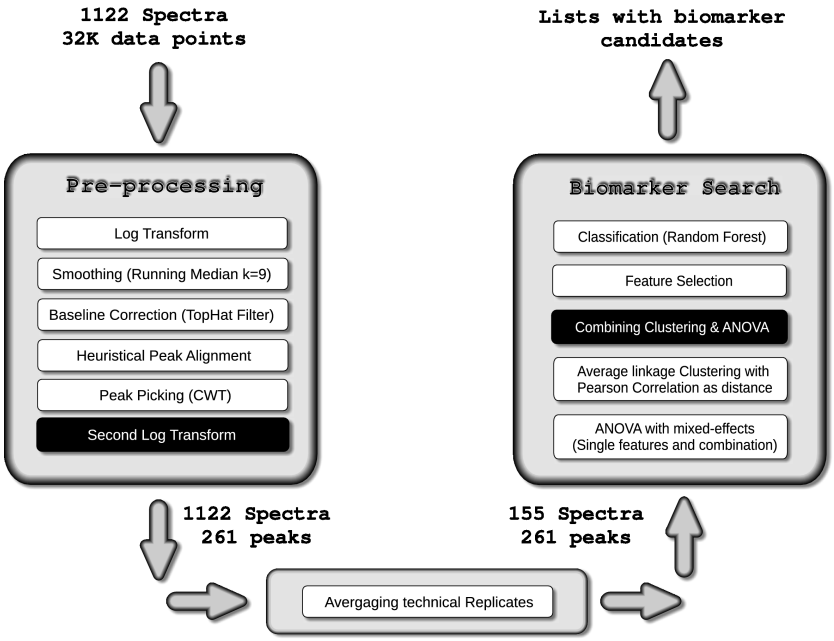


Figure 11: Schematic representation of the data analysis workflow developed for evaluation of MALDI data. Basically the workflow is divided in two different parts: preprocessing and biomarker discovery. In addition, the dimenstionality of the data is annotated prior to and after the main steps. Steps that have not been reported previously are highlighted.

This Chapter is organized in the following sections: After giving an introduction to the typical application of the technology, this chapter starts to deal with acquisition of data (Section 3.2) and the large number of preprocessing steps required (Section 3.3). Section 3.4 describes methods for preprocessing and data analysis particularly developed for this thesis. In the results section (Section 3.5) we present the effects of preprocessing, the results of biomarker discovery and results of feature selection and classification. Finally, in the last part of this chapter we will give a summary and conclusion.

3.1 INTRODUCTION

The multi-factorial study design encompassing the three different experimental factors: genotype, diet and time (Section 2.2), substantiates the need for a technology capable of processing a multitude of samples in a short time. More than 150 distinct biological samples resulting in more than 1100 spectra were measured and analyzed. MALDI is well suited for this task and is characterized

	week 3			week 4		
	SD	HF	CHF	SD	HF	CHF
B6	36/5	31/4	12/2	36/5	40/5	37/5
NZO	35/5	35/5	32/4	40/5	36/5	40/5
SJL	4/1	0/0	16/3	12/2	0/0	40/5

	week 6			week 8		
	SD	HF	CHF	SD	HF	CHF
B6	38/5	38/5	32/5	39/5	34/5	28/5
NZO	37/5	38/5	40/5	28/5	34/5	34/5
SJL	32/4	40/5	32/5	36/5	40/5	40/5

Table 1: Number of spectra and biological repeats for each factor combination. The first number indicates the number of spectra, the second states the number of biological replicates. In total there are 155 different biological samples from 31 different mouse individuals.

by simplicity, good mass accuracy and high resolution[1]. For technical details of MALDI technology see Section 2.1.1.

MALDI data requires a large amount of preprocessing prior to data analysis (smoothing, baseline correction, peak finding or peak matching - more detailed description is given below in Section 3.3.1). Various algorithms differing in principle, implementation and performance have been proposed to address different preprocessing steps [126, 112, 181]. For a comprehensive review of the most important preprocessing steps see also Yang et al. [178].

Typically the two main objectives of MALDI profiling studies are biomarker identification and classification. Various different methods have been applied addressing these two objectives ranging for classical t-test and ANOVA two more advanced genetic algorithm and swarm intelligence (see Section 3.3.2 for more details).

3.2 SAMPLE PREPARATION

Over 30 different mouse individuals where used to create over 150 distinct biological samples measured in more than 1100 MS spectra (see table 1).

Blood samples were obtained by cutting the tip of the mouse tail and collecting the blood from the dorsal and lateral tail veins into a Li-heparin-coated microcuvette. Immediately after blood collection each sample was centrifuged at 4°C for 5 min at 13,000 rpm. The blood plasma was then transferred into 200µL-microcentrifuge tubes, shipped on dry ice to the mass spectrome-

try laboratory and stored at -80°C prior to further sample preparation and MS analysis.

The amount of plasma obtained at each blood collection varied between 0 and 12 µl. Since 5 µl were needed for each sample preparation, it was possible to perform up to two sample preparations. In a few cases only one or no sample preparation could be performed. From each sample preparation 4 replicate MALDI MS profile spectra were acquired, resulting in a total of up to 8 technical replicates per sample. MALDI MS spectra were obtained using an Ultraflex MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). Spectra were acquired automatically for the m/z range of 700-10,000. MS profile peak identification was achieved similarly to the methods described by Tiss et al.[159] using a Q-TOF Premier mass spectrometer (Waters, Manchester, UK).

3.3 STATE-OF-THE-ART - MALDI

State-of-the-Art section describes commonly used methods for different steps of preprocessing as well as for biomarker identification and classification.

3.3.1 *Preprocessing*

The preprocessing workflow of MALDI data aims at transforming the large number of data points in raw spectral data (typically > 30.000) into a much smaller, statistically manageable set of peaks and at the same time addressing noise and technical bias. Comprising tens of thousands of data points in each spectrum, mass spectrometry data is inherently noisy. The main sources of noise are chemical in nature such as interference from matrix material and sample contamination or electrical noise which depends on the analytical set-up employed[88]. Furthermore, many factors like temperature or humidity may distort the machine's calibration, leading to shifts in m/z direction.

The widely accepted standard preprocessing steps are:

1. **Log transformation (Normalization):** Logarithmic transformation is typically performed in order to convert a multiplicative error behavior into an additive one. A additive error model with homogeneous error is required by many statistical tests such as t-test or ANOVA.

2. **Smoothing:** Smoothing mainly aims at removing high frequency noise. Beyond traditional signal processing techniques like Savitzky Golay filter [141], Mean/Median filter or Gaussian filters also wavelet based techniques are employed for data smoothing [29],[88].

3. **Baseline correction:** Baseline correction intends to remove low frequency noise and thus eliminates the correlation of nearby features. Typically methods like Top Hat filter[109] (see Section 3.3.1.1 for more details), Loess derivative filters [27], linear splines, polynomial fitting or convex hulls are applied to estimate the baseline.

4. **Peak picking:** The number of proposed methods for peak detection is immense. Most common algorithms make use of Signal to Noise Ratio (SNR), Continuous Wavelet Transform (CWT)[92, 40] or model functions like Gaussian function used as templates for peak detection.

5. **Peak alignment:** Peak alignment is employed to correct for shifts in m/z direction.

A multitude of software packages implementing the complete workflow is available. Commonly used public domain software tools are R[128] and Bioconductor[58] packages like `msProcess` or PROcess[134], Matlab packages like LIMPIC[103] or Cromwell[29] and the comprehensive C++ library OpenMS[149, 132].

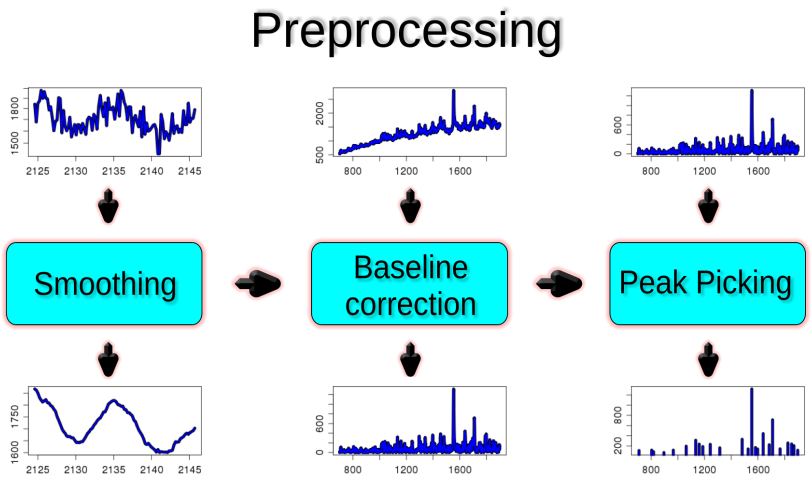


Figure 12: Schematic representation of main preprocessing steps: Smoothing removes high frequency noise; Baseline correction corrects for baseline drifts or low frequency noise; Peak picking converts continuous spectral data into a list describing position and intensity of identified features.

3.3.1.1 *Baseline Correction: Top Hat Filter*

Top Hat Filter is an effective way of baseline correction. The Top Hat Transform - first introduced in 1979 by Meyer[109] - is a mathematical morphology function allowing for extraction of peaks (in 1D or 2D) by removing local background levels[137]. Top Hat filter is based on a morphological operation called *opening*(O) that is defined as a successive application of *erosion*(E) followed by *dilatation*(D). Given a spectrum X and a structuring element B_x (with a reference point x), *erosion* describes all points x of X completely containing B_x and *dilatation* describes all points x of X touched by B_x:

$$E_{B_x}(X) = \{x | B_x \subseteq X\}$$

$$D_{B_x}(X) = \{x | B_x \cap X \neq \emptyset\}$$

$$O_{B_x}(X) = D_{B_x} \circ E_{B_x}(X)$$

The final baseline correction is done by subtracting the spectra from its opening.

Assuming the structuring element of the opening to be a line (with length (width) k) than the erosion operation for a spectrum

equals to a running minimum filter (with windows size k) and the *dilatation* equals to a running maximum filter (with windows size k). This is demonstrated in Figure 13.

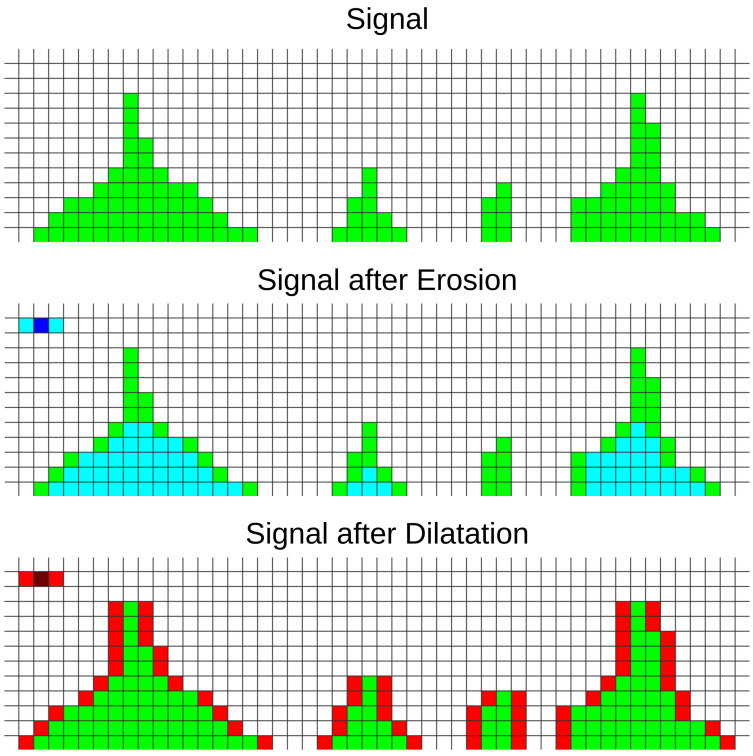


Figure 13: Schematic demonstration of erosion and dilatation. Upper part: Schematic presentation of a mass spectrum (green). Middle part: Signal after erosion (blue). The structuring element is a line with length 3. Erosion equals to a running minimum filter with length 3. Lower part: Signal after dilatation (red). The structuring element is again a line with length 3. The original spectrum is completely included in the dilatation. Dilatation equals to a running maximum filter with length 3.

3.3.1.2 *Peak Picking*

Peak picking aims at transforming a continuous m/z intensity signal into a list of peaks consisting of peak positions and peak intensities (typical peak heights or peak volumes). For investigation peak picking algorithms, we further analyze three common peak detection methods: Signal to Noise Ratio (SNR), Template based peak detection and Continuous Wavelet Transform (CWT). We have selected these three algorithms because they are very popular and widely-used.

Signal to noise ratio (SNR)

SNR is a very general approach. The essential part of this algorithm lies in the definition of noise. In statistics, noise is often defined as variance or Median Absolute Deviation (MAD) along different samples. In signal processing noise is often defined as an estimated background. For instance in the Bioconductor package `PROcess`[134] MAD of points within a window is used for noise estimation. Peaks are identified by searching a local maximum of points withing a certain neighborhood (e.g. about expected peak width) having a SNR bigger than a given threshold.

Template based peak detection

This algorithm assumes that the peaks to be detected are shaped like some model function e.g. a Gaussian function. Peak detection is performed by comparison with the template function using some similarity measure (e.g. correlation). The vector of spectrum intensities in transformed into a vector of correlation values. Peaks are identified by searching for high correlation values above a given threshold.

Continuous Wavelet Transform (CWT)

CWT[92, 40] is a more sophisticated approach that is used to split the signal into different frequency ranges. Regarding the m/z scale as generalized time scale, CWT constructs a time-frequency representation of the spectrum by mapping it from the time domain to the time-scale domain. The essential part of CWT is the mother wavelet whose translated and scaled versions are used to generate daughter wavelets.

3.3.1.3 *Peak Alignment*

To ensure comparability of different spectra, they must be aligned and especially artificially created shifts in m/z direction have to be removed. Compared to noise filtering or peak picking, an alignment of a multitude of spectra is more challenging and well established algorithms are missing. Furthermore this step is also very crucial as it paves the way for subsequent statistical analysis. Calculating the optimal alignment of a myriad of spectra is a computationally expensive and time consuming task and a variety of heuristic approaches have been applied[154, 129].

Beside heuristic algorithms different ideas for analytical approaches have been proposed. An attempt for analytically finding an optimal alignment are dynamic programming (DP) algorithms commonly used for sequence alignment (such as Smith-Waterman algorithm). For Gas Chromatography Mass Spectrometry (GC-MS) experiments employed for metabolic profiling,

Robinson et al.[135] applied a DP algorithm to find an optimal alignment. The peak scoring function ($P(i,j)$) for two spectra i and j takes care of both, peak distances and peak similarities:

$$P(i,j) = S(i,j) \cdot exp\left(-\frac{(t_i - t_j)^2}{2D^2}\right)$$

whereas $S(i,j)$ describes the similarity of the peaks (e.g. correlation), t_i, t_j are the two retention times (distances of the peaks) and D is some arbitrary parameter determining the importance of retention times.

As another approach for analytical peak alignment, Liu et al.[97] proposed an algorithm based on simple Monte Carlo Markov Chain (MCMC). They used a Bayesian approach assuming peak samples to be normally distributed around their true peak and directly addressed false negative and false positive peaks. However, application of the MCMC to an open-source ovarian cancer dataset (created by Wu et al.[174]) comprising 170 spectra (around 10% of the size of our dataset) took several days of computational time.

3.3.2 *Biomarker Identification and Classification*

MALDI profiling studies have typically two main objectives: biomarker identification or classification. Various different methods have been applied addressing these two objectives. As for biomarker discovery, commonly used methods range from classical t-test or Wilcoxon rank sum test[56] to more advanced techniques such as genetic algorithms and swarm based intelligence [133]. See also Section 2.4 for more details on feature selection and classification and Section 2.4.1 for more details for Ant Colony optimization (ACO).

As for classification virtually all of the common classification systems such as Bayes classification, decision trees, logistic regression, Random Forest (RF) or Support Vector Machines (SVM) have been used more or less successfully. Many papers and reviews were published considering classification task for MALDI data. Wu et al. [174] published a summary comparing different classification methods for ovarian cancer. In 2006, Zhang et al.[183] compared the performance of R-SVM and SVM-RFE. More recently, in 2009, Liu et al.[99] published an overview of additional feature selection and classification approaches, both using MALDI MS data sets.

3.3.3 *ROC Curves*

ROC curves are a widely used measure of performance of supervised classification rules. They plot false positive rate (or

1-specificity) versus true positive rate (or specificity). Perfect classifiers are characterized by false positive rate = 0 and high true positive rate = 1. Analysis of ROC curves is typically used compare different models and to select possibly optimal models. Furthermore, they are helpful to assess the trade-off between sensitivity and specificity. They are typically created by scanning through the parameter range of the classification system.

ROC curves are restricted to the case of two classes. For calculating a ROC curve for multi class classifier, two different approaches have been proposed:

1-vs-rest

Calculating the ROC curve for one class vs. all other classes[111]. For n classes n ROC curves could be created. The volume under the ROC surface could be approximated by projecting it down to two dimensional set of curves. This set of curves can now be averaged by weighting with the class probability.

1-vs-1

The ROC curves are calculated for each class combination. For n classes we create $n * (n - 1)$ ROC curves. This set of curves could be averaged unweightedly.[66]

3.4 METHODS AND ALGORITHMS

While section 3.3 describes the theoretical background of the required analysis steps, this section rather describes algorithms particularly developed in this thesis. For the preprocessing several existing algorithms like Top-Hat filter and different peak picking algorithms were adopted and implemented. Furthermore new algorithms were developed and implemented, e.g. for peak alignment or the detection of artificial peaks derived by Polyethylene Glycol (PEG). For statistical analysis a novel method was developed that combines significance information derived by ANOVA and redundancy information assessed by unsupervised clustering. All algorithms are implemented in statistical programming language (R[128] version 2.7.0 - 2008-04-22 and R version 2.12.1 - 2010-12-16).

3.4.1 *Preprocessing*

The preprocessing applied in this work is designed to address all typical preprocessing steps (see Algorithm 1). Figure 11) shows a complete workflow of the analysis of MALDI data including most parts of preprocessing.

Algorithm 1 General description of the preprocessing workflow for MALDI data. Preprocessing is basically a consecutive application of different steps.

for all Spectra **do**
  apply log transformation ($\log_2$)
  apply median filter ($k = 9$)
  apply adapted top hat filer (see section 3.3.1.1)
  apply peak picking (see section 3.3.1.2)
  apply peak alignment/matching (see section 3.3.1.3)
  apply additional data transformation (see section 3.4.1.5)
end for

3.4.1.1 *Peak Picking*

For investigation of peak picking algorithms, we chose to focus on three common peak detection methods: Signal to Noise Ratio (SNR), Template based peak detection and Continuous Wavelet Transform (CWT) (see 3.3.1.2). SNR and template-based approach were implemented in own R scripts while for CWT we used the R package `msProcess` - version 1.0.5 2009-01-20.

For SNR the noise was estimated as the background calculated using Top Hat filter(see section 3.3.1.1) with small window size. For template based peak detection, we scanned along the

spectra and calculated the correlation (Pearson correlation coefficient [136]) to a Gaussian function. For evaluation of CWT we used the second derivative of a Gaussian function (Mexican Hat Wavelet) as mother wavelet. For peak detection (using `R` package `msProcess`) the peak candidate has to be clearly distinguishable from the background (parameter: *snr.min*) and visible across at least 7 scale domains (parameter: *length.min*) excluding the first three high frequency wavelet scales (parameter: *scale.min*).

3.4.1.2 *Reference Peaks for evaluation of Peak Picking Algorithms*

For evaluation of the peak picking performance we defined a set of reference peaks. The peak picking algorithms are evaluated in terms of sensitivity (how many of the reference peaks are found) and specificity (how many of the found peaks are part of the reference set). The reference set was created in a semi-automatic process. To this end we initially picked peaks manually (in order not to favor any algorithm) and subsequently optimized peak positions automatically (to correct for manual inaccuracy). All in all the reference set contained a total of 381 peaks.

3.4.1.3 *Top Hat Filter*

The Top Hat filter is defined as a successive application of a running minimum followed by a running maximum. (see section 3.3.1.1). Following these assumptions we adapted the top hat filter into a successive application of a running 0.1 quantile filter, a running median filter and a running minimum filter (see Algorithm 2).

Algorithm 2 Baseline correction using an adapted Top Hat filter.

```
apply running 0.1 quantile (k=101)
apply running median (k=101)
apply running min (k=201)
```

3.4.1.4 *Peak Alignment*

Due to missing established algorithms for peak alignment and high computational costs of analytical approaches (especially considering the large amount of spectra in our dataset), we developed a heuristic approach suitable for our dataset. A multiple alignment of a huge amount of spectra is challenging when computing the aligning of all sequences to each other simultaneously. To avoid this complexity the alignment can be performed against a reference spectrum or alternatively, like used for multiple sequences alignment using a cluster based guide tree[97]. However, we prefer the first alternative and calculate the alignment to a

reference profile. The reference profile was defined as the average profile of all spectra.

Like in the algorithm presented by Randolph and Yasui[129], our approach starts with the identification of very prominent features (peaks) of the reference spectrum (RS). By applying the correlation based peak detection (see section 3.3.1.2) with high correlation threshold (0.8), we identified 43 reference peaks within the RS. For every spectrum to be aligned (s_i) and for every reference peak (rp_j), we apply peak picking algorithm for s_i within a given environment around rp_j. If a peak ($p_{i,j}$) is found we assume this peaks to be the same as rp_j and calculate the distance ($d_{i,j}$) of $p_{i,j}$ and rp_j. If the peak is missing in a spectrum the distance $d_{i,j}$ is set to Not Available (NA). Applying this method, for a certain spectrum k we compute a vector (of length 43) of distances ($d_{k,j}$) between identified peaks of s_k and rp_j.

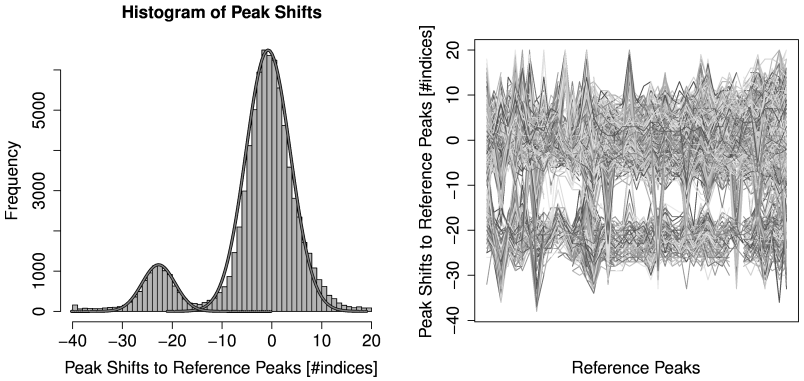


Figure 14: Distances of found peaks and reference peaks calculated during peak alignment process. Left hand side: Histogram of all distances. There are two groups of spectra showing different distribution of distances: one distributes around 0 and one around −22 indices. For both groups a fitted normal distribution is added to the histogram. Right hand side: Vector of distances for the spectra. Again the two groups of spectra are clearly visible. Furthermore for nearly all spectra the shift seems to be constant (on the index scale).

Regarding the vector of distances (right hand side of Figure 14) and the distribution of the distances (left hand side of Figure 14), leads to two observations: First, the vector of distances for a spectrum seems to be constant on the index scale, which corresponds to a quadratic distortion on m/z scale. Second: there are two groups of spectra showing different distributions of $d_{k,j}$. The majority of the spectra seems to be in good alignment with the reference spectrum, whereas a smaller group of spectra are more or less constantly shifted by −22 indices. These two groups are derived from different sample preparation batches.

Index scale refers to the consecutive ~ 32K data points without considering the m/z distances.

As observed above, distances to reference peaks seem to be constant for a particular spectra (see Figure 14), and hence the final displacement value (for a spectrum) is calculated by averaging the corresponding distance vector. A pseudo-code representation of the complete peak alignment is given in the Appendix see algorithm 5.

Having two groups of spectra (one with an index shift of ~ 0 and one with ~ -22) violates the assumption of Liu et al.[97] that peak samples are normally distributed around their true peak. This effect could also be a problem for our heuristic approach, because the mean spectrum contains each peak twice. For our dataset the group with an index shift of ~ 0 is much bigger than the other group and the mean spectrum is not affected badly. If both groups were equally sized the heuristic would fail to align the peaks.

3.4.1.5 *Additional Data Transform*

T-Test and ANOVA, both assume data to be normally distributed and variance to be homogeneous (c.f. section 2.3). However, even applying all preprocessing steps does not lead to a complete stabilization of the variance. Hence, in order to assure homoscedasticity additional steps were required.

Obviously, for peak intensities there is still a linear dependency between variance and intensity indicating a multiplicative error model. The standard treatment of data with a multiplicative error model is (another) log transformation. We added a pseudo-count of 0.1 to avoid the singularity at 0. Finally, we added an offset for shifting spectra to a positive scale.

3.4.2 *PEG Detection*

The polymer Polyethylene Glycol (PEG) has a broad range of applications, inter alia it is often used as an internal calibration compound in MS experiments or as stationary phase for gas chromatography. It is non-toxic, non-immunogenic, non-antigenic, highly soluble in water and FDA approved. Therefore, linking of one or more PEG molecules to a protein or peptides (known as PEGylation) is a common method especially in drug development because it prolongs residence in body, decreases degradation by metabolic enzymes and reduces or eliminates protein immunogenicity [166, 170].

PEG shows a typical fragmentation pattern, when subjected to MALDI MS experiment. PEG, having a mass of $44n + 62$ Da, results in a consecutive sequence of peaks with distances of 44 Da. In order to investigate whether a typical PEG fragmentation pattern can be found or not, we calculate all pairwise distances

of all n identified peaks ($n \cdot (n-1)/2$ distances). Figure 15 shows a histogram of the distribution of all pair-wise peak distances < 100 Da. The counts for the peak distances seem to be normally distributed with one prominent outlying bar with a distance of more than 5.5 times standard deviations from mean distance. This bar reflects peak distances between 43.94 and 44.34 Da which corresponds to PEG fragment distance indicating the presence of PEG in the experimental data.

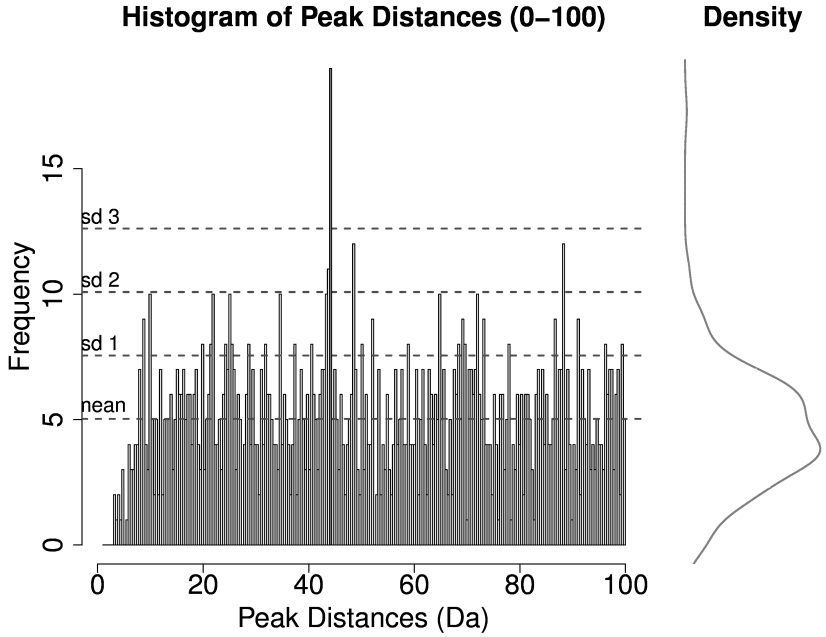


Figure 15: Histogram of the distribution of all pair wise peak distances < 100 Da. Horizontal dashed lines indicate the mean, and the first, second and third standard deviation. Curve on left hand side show the density of the counts. Counts seem to be normally distributed.

For further analysis we want to exclude peaks derived from PEG since our interest is focused on biologically relevant signals. Therefore, we developed an algorithm for identification of peaks derived from PEG rather than from biological biomolecules. The algorithm comprises several steps that mainly aim in identification of consecutive peak with a peak distance corresponding to PEG peak distance. The algorithm is presented on a high level description in algorithm 3.

3.4.3 *Handling of Technical Replicates*

As shown previously in Section 3.2 for a single sample up to 8 technical replications were performed. Technical replicates have an important effect on the data and on the statistical evaluation[82]. Thus an adequate handling of technical replicates is required. For

Algorithm 3 High level description of the algorithm for detection of peaks derived from PEG.

1: calculate all pair wise peak distances
2: identification of peaks with a distance of ~ 44 to another peak
3: reordering these peaks to lists of consecutive peaks
4: identifying lists of consecutive peaks with at least 3 peaks

the analysis is this dataset we use the standard approach of calculating the average of technical replicates. A detailed description of the effects of technical replicates on the statistical evaluation as well as alternative methods and the reasoning for choosing to average technical replicates is given in the Appendix.

3.4.4 *ANOVA with mixed effects*

The ANOVA model is designed to investigate effects of the three different experimental factors (genotype, diet and time) on blood proteins. A special emphasis is devoted to the mutual influence of diet and genotype. For the data evaluation presented in this chapter a straight forward approach is a mixed-effect ANOVA of the form:

$$Y \sim Genotype + Diet + Time + Genotype * Diet$$

This model simultaneously investigates effects derived from all three single experimental factors as well as the combination of genotype and diet - symbolized by the '*' (see Section 2.3 for more details of mixed effect ANOVA). The combination of genotype and diet is of particular interest since SJL mice are resistant to HF diet induced obesity and diabetes.

3.4.5 *Stratification and Clustering*

After preprocessing each peak should represent a peptide or peptide combination, respectively. The concentration of a peptide and hence the peak intensity varies in the diverse samples (dependent on experimental factors such as different diet-genotype combinations). Let an intensity profile be the list of intensities for a certain peak across all samples.

Due to fragmentation and degradation, each protein can split up into multiple peptides and therefore, lead to multiple peaks in the mass spectrum. These peaks are not independent and the corresponding intensity profiles are supposed to be correlated. A high correlation between intensity profiles can hint for related peptides such as multimer formations or PTMs. In order to benefit from this kind of correlation various methods have

been proposed [21]. To this end, we apply hierarchical clustering using average linkage[68] with $1 - \rho$ as distance measure, where ρ denotes the Pearson-correlation coefficient[136]. Each node in the cluster dendrogram represents several intensity profiles and similar intensity profiles are aggregated in close proximity.

It is useful to combine this similarity information with significance information by assigning p-values to the nodes. Using the ANOVA model described above (Section 3.4.4), we calculate a p-value for each peak and hence each leaf of the dendrogram. For a node representing several peaks, the p-value is calculated from the mean intensity profile of corresponding peaks. For technical and biological reasons intensity profiles can be different absolute scales. Therefore, prior to averaging intensity profiles, they are z-transformed (centered and scaled)[68]. So each leaf and each node of the dendrogram is annotated with p-values. The nodes not only aggregate similar peaks but also reflect the significance for an experimental factor. Merging both, similarity and significance information our approach allows for the interpretation of complex biological data in an intuitive manner.

3.5 RESULTS

In the following, we describe the effects of preprocessing and the results of subsequent data analysis. We will first investigate the distinct effects of different pre- processing steps such as Top Hat filter (section 3.5.1.1) or peak alignment (section 3.5.1.5). Special attention is paid to evaluation and comparison of different peak picking algorithms (Sections 3.5.1.3 and 3.5.1.4). Effects of preprocessing on a global level of the complete dataset are presented in Section 3.5.1.6.

Subsequently in Section 3.5.3 we present the results for identification of biomarker candidates for combination of experimental factors genotype and diet and for the single experimental factors, respectively. In the end of this section we present the result for classification and prediction (Section 3.5.4).

3.5.1 *Preprocessing*

3.5.1.1 *Top Hat Filter*

Effects of the Top Hat filter (for details see Section 3.4.1.3) for 100 randomly chosen spectra are demonstrated in Figure 16. Prior to baseline correction the different spectra show different baselines with shifts up to two orders of magnitude ($\log_2$). This is particularly visible in the zoomed spectra without baseline correction (upper right part of Figure 16). Top Hat filter successfully subtracted the baseline from each spectrum and all spectra are much more homogeneous. Especially at the level of the complete spectra (lower left side of Figure 16) the baseline shifts are not visible any more.

3.5.1.2 *Principles of peak picking algorithms*

In order to understand the three different peak picking algorithms: SNR, template based approach and CWT (see Section 3.4.1.1), we now illustrate their working principles. To illustrate the principles and assess the different algorithms we use the set of reference peaks (see Section 3.4.1.1).

Figure 17 gives a graphical impression of the underlying principles of the different algorithms. The first box shows the mean intensity spectrum of the complete data set in a mass window of m/z 1400-1800 Da. The noise level was defined as baseline calculated using Top Hat filter (see dashed line). The 33 peaks from the reference set within this mass window are indicated as vertical dashed lines.

The second part of Figure 17 shows the signal to noise ratio along the mass window of the mean spectrum. The SNR threshold used for peak identification was 1.75 (horizontal dashed line).

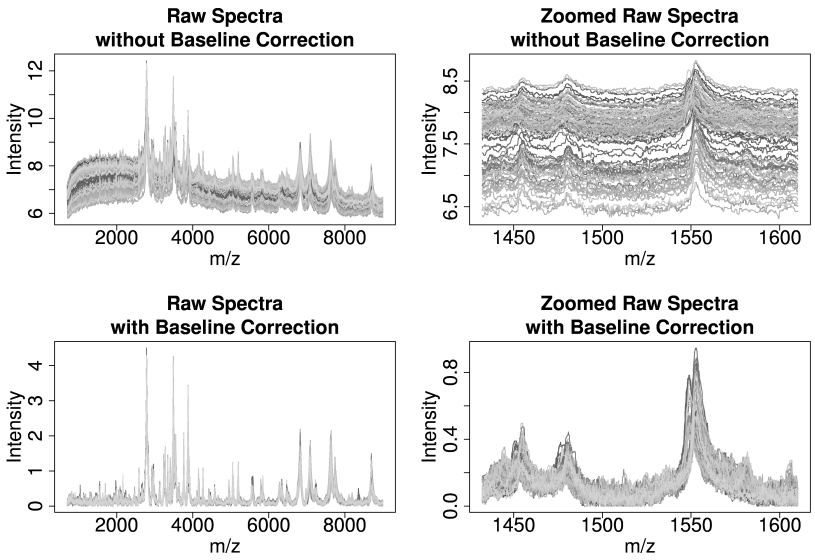


Figure 16: Effects of the adapted Top Hat filter for 100 spectra regarding the whole spectra (left hand side) and with a zoom (right hand side). The upper row reflects the spectra before baseline correction and the lower row after applying Top Hat filter.

Using SNR we identified 22 peaks in this mass range whereas we found 69% of our reference peaks (with the SNR threshold of 1.75). With this threshold we did not find any peak that was not part of the reference set.

The third box in Figure 17 visualizes the performance of template-based peak detection. The correlation coefficients along the spectrum are shown. The correlation threshold of 0.6 is shown as horizontal dashed line. All in all we found 31 of the 33 reference peaks (94%) plotted as dots above the peaks. We also found one peak that is not within the reference set (false positive) shown as asterisk above the peak.

The last part of Figure 17 demonstrates the peak picking using CWT. The first 7 daughter wavelets are depicted. Compared to the other two methods the peak picking is complicated by the fact that information from different time-scale domains has to be combined. The reference peaks again appear as vertical dashed lines and the picked peaks are marked above the peaks. Using CWT we identified 97% of the peaks but also got two false positive hits (marked with asterisks above the peaks).

3.5.1.3 *Evaluating peak picking algorithms*

We now assess the peak picking algorithms in terms of sensitivity (how many of the reference peaks are found) and specificity (how many of the found peaks are part of the reference set). An optimal algorithm has high sensitivity and high specificity. Sensitivity and specificity are used to generate the ROC curves in

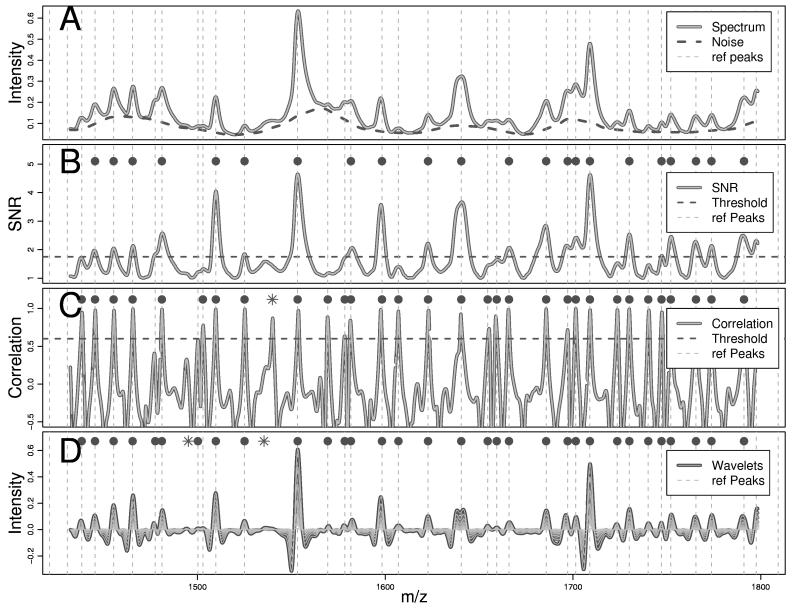


Figure 17: Comparison of the three different peak picking algorithms for the m/z-range 1400 - 1800. A: mean spectrum and noise/background (dashed line); B: SNR and threshold used for peak picking (horizontal dashed line); C: correlation coefficient and threshold (horizontal dashed line); D: first 7 wavelets. The vertical dashed lines are reference peaks. Marks above the plot indicate identified peaks (dot = contained in the reference set; asterisk = not in reference set - false positives).

Figure 18. For more details on ROC curves see Section 3.3.3. They are calculated by scanning the threshold values of the different algorithms e.g. changing the correlation threshold in the template based approach (for an illustration of the threshold operation see figure 17).

Furthermore, in order to evaluate the sensitivity to noise we added different quantities of high frequency noise (white noise). Since the observed error behavior for MS spectra indicates a multiplicative error behavior, we added a normally distributed noise with $mean = 0$ and an error of 2%, 4% and 10%. The performances of the three methods are affected to a very different degree (see Figure 18 for the ROC curves). The SNR is very sensitive to noise and the ROC curve worsens dramatically. The other two algorithms are much more robust. While on perfectly smoothed data the template correlation approach seems to be the method of choice, for noisy data the advantage of the template-based approach decreases and CWT shows the best performance. In conclusion the three presented peak picking algorithms show a different sensitivity to noise and therefore to the number of spectra and the choice of parameters for preprocessing steps.

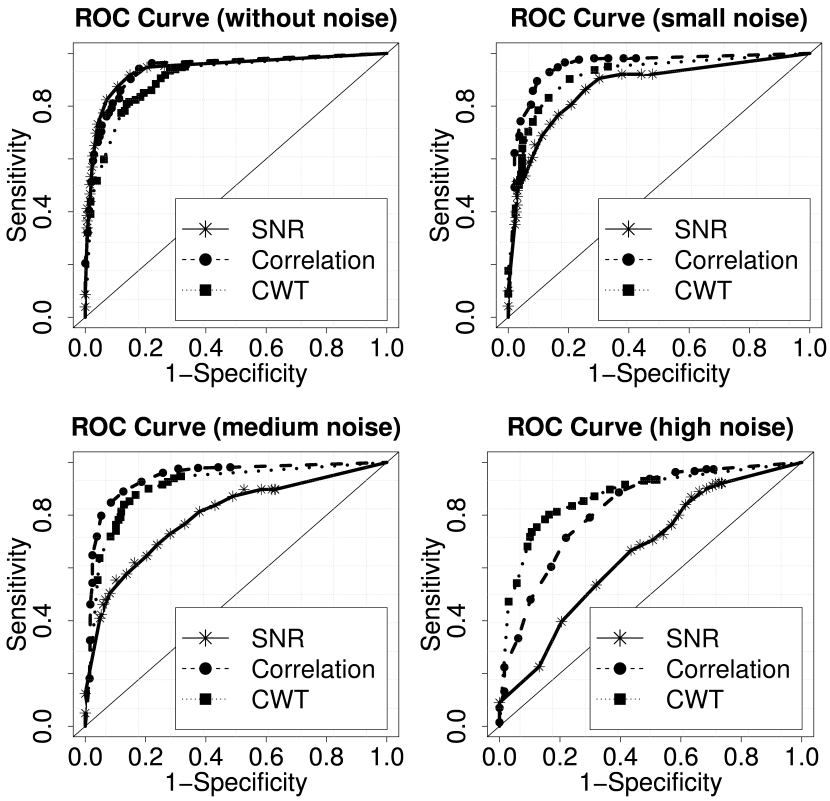


Figure 18: ROC curves for the three different peak picking algorithms on original and noisy Data: SNR (full lines), Correlations with Gaussian function (dashed lines) and CWT (pointed lines).

3.5.1.4 *Comparing peak picking algorithms*

The three different peak picking algorithms investigated here are distinct in terms of complexity, performance and stability. But all three methods have a common parameter: the estimated peak width. There are different ways to estimate the optimal peak width. For instance OpenMS[149, 132] offers the possibility to measure the peak width manually using graphical interface or the peak width can be estimated by the CWT algorithm itself.

For an overview of the advantages and disadvantages of the algorithms see table 2. Signal to Noise Ratio as a universally used signal processing technique is computationally fast, easy to implement and shows good performance on smoothed data. However, it is not very specific for this task as it ignores the shape of the peak. Since the noise is an integral part of the algorithm it is very sensitive to noise and therefore strongly depends on the quality of the data and on the performance of previously performed smoothing and baseline correction steps. This can be seen in lower right part of Figure 18.

The template-based approach is much more specific for the peak picking task assuming peaks to be shaped like a Gaussian

Method	PRO	CONTRA
SNR	• simple - easy to implement • fast performance • only few parameters	• depends on the definition of noise • unstable - very sensitive to noise • ignoring peak shape
Template Correlation	• simple - easy to implement • only few parameters • stable for small noise	• detection favors Gaussian shaped peaks • sensitive to high noise
CWT	• stable even for massive noise • internal data smoothing • flexible - tunable	• complicated algorithm • slow performance • difficult to tune - high number of parameters

Table 2: Summary of advantages and disadvantages of the three presented peak picking algorithms.

function. This assumption, however, might often not be exactly applicable because peaks may show a considerable asymmetry. Depending on the experimental parameters, particularly laser energy, significant deviation from a Gaussian peak shape can be obtained. Although this method has only a few parameters, it appears rather robust for lower levels of noise. However for high levels of noise the performance decreases.

CWT is like SNR a very universal signal processing technique used for many different tasks. In comparison to SNR and template-based approach CWT is more complex and computationally expensive. The large number of parameters allows for tuning CWT to be very specific for this task taking into account the shape of the peak. As smoothing is an intrinsic part of the algorithm CWT is very robust even to substantial amounts of noise. On the other hand tuning of the algorithm is difficult due to the large set of parameters and may result in over-specific solutions.

For perfectly smoothed data all three methods show good performances but CWT seems to be little worse than the other two. For data including a substantial amount of noise CWT clearly outperforms the other methods in terms of sensitivity and specificity.

3.5.1.5 *Peak Alignment*

The effect of the peak alignment is demonstrated using two exemplary chosen peaks at m/z 709 and 818 both of which are part of the 43 reference peaks used for alignment process. For

further details about peak alignment see Section 3.4.1.4. Figure 19 shows the two peaks before and after peak alignment. Prior to peak alignment (left hand side) some of the spectra are shifted: Two groups of spectra are visible each a different distribution of shifting positions: one distributes around 0 and one around −22 indices (c.f. Section 3.4.1.4). These groups are derived from different sample preparation batches. Without alignment statistical analysis would be biased. After peak alignment all peaks positions are homogeneous (right hand side of Figure 19).

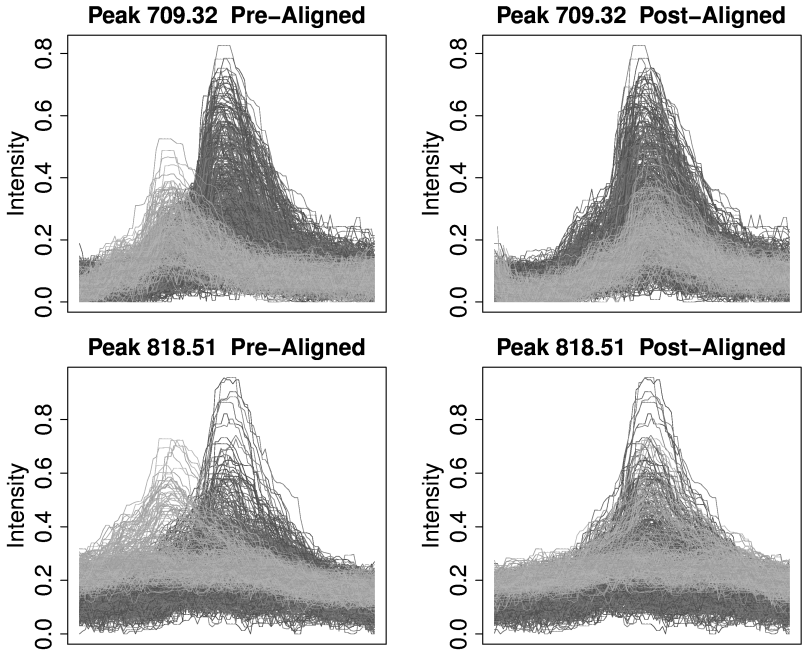


Figure 19: Two exemplarily chosen peaks 709.32 (upper row) and 818.51(lower row) before applying peak alignment (left hand side) and after peak alignment (right hand side). Before peak alignment two groups of spectra are visible: one group of spectra shifted by -22 indices (colored in light gray) and the other group in the center of each image (colored in dark gray). After peak alignment peak positions of all spectra are homogeneous and the groups are not separated by the index shift any more.

3.5.1.6 *Global Effects of Preprocessing*

The effects of log transformation, baseline correction and peak matching at the level of complete spectra as well as the corresponding error plots are depicted in Figure 20.

Typically preprocessing aims at reducing technically created bias and hence, assuring that the assumption required by statistical evaluation approaches such as ANOVA are fulfilled. For MALDI MS, raw data has a very strong correlation between signal intensity and variance (upper part of figure 20). After applying

logarithmic transformation to the spectra the correlation between variance and intensity is still strong (middle row of figure 20). However even the combination of log transformation, baseline correction and peak alignment does not lead to a complete stabilization of the variance which is necessary for applying our statistical analysis methods (lower part of figure 20).

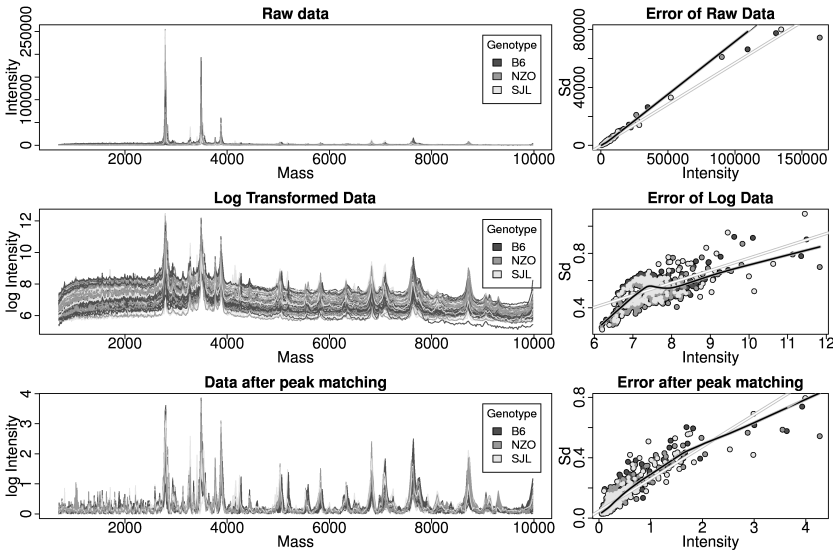


Figure 20: Effects of preprocessing: Raw data (top), log data (middle), after baseline correction and peak alignment (buttom). The left column shows the effect on the spectra while the right column shows the corresponding standard error plots including linear fit (white line) and lowess fit (black line). The different colors reflect different genotypes (dark gray: B6, gray: NZO, light gray: SJL). Non-aligned spectra could not be seen on this global scale.

3.5.1.7 *Additional data transform*

To assure homoscedasticity a second log transformation of the data was performed (see Section 3.4.1.5 for details). After additional data transformation steps data are homoscedastic (see Figure 21). Thus, the assumptions for the application of ANOVA are fulfilled.

3.5.1.8 *PEG*

Prior to statistical analysis we want to exclude peaks derived from PEG since our interest is focused on biologically relevant signals rather than artificial contaminations. Therefore, we applied the algorithm developed for identification of peaks derived from PEG (see Section 3.4.2). Using this algorithm, the following five consecutive PEG peaks lists were identified:

1. 774.76, 818.90, 861.72, 905.09, 948.54

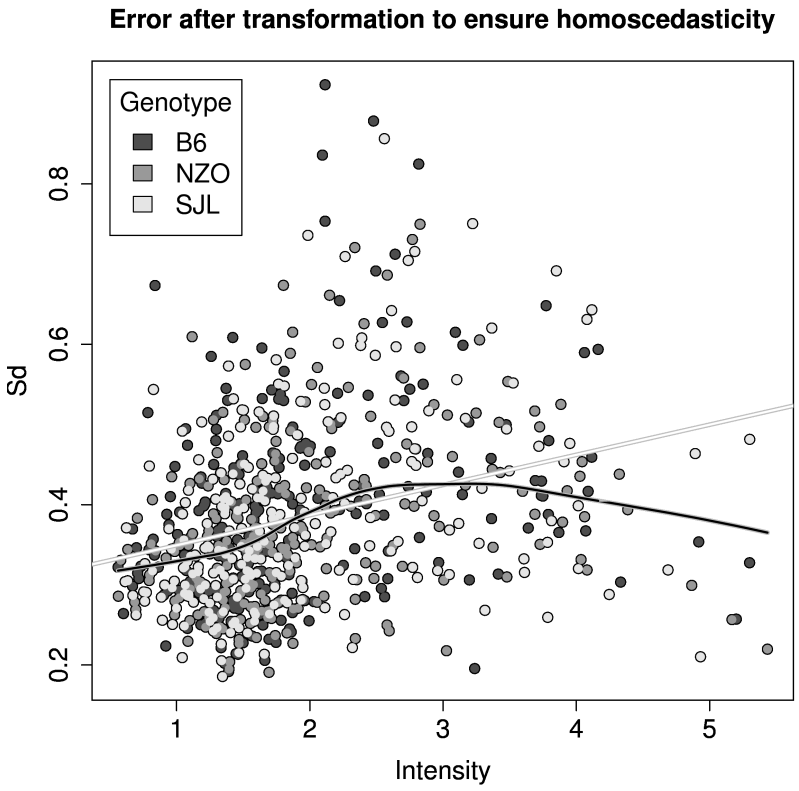


Figure 21: Error plot after log transformation to ensure homoscedasticity including linear fit (orange line) and lowess fit (black line). The different colors reflect different genotypes (dark gray: B6, gray: NZO, light gray: SJL).

2. 2272.72, 2316.89, 2360.60, 2404.50
3. 1112.53, 1157.08, 1201.41, 1245.28, 1289.61, 1333.39, 1377.73, 1421.76, 1465.97, 1509.97, 1553.89, 1598.09, 1686.09, 1730.41, 1774.35, 1818.25
4. 1578.81, 1622.99, 1666.11, 1709.41, 1752.70
5. 2068.82, 2112.86, 2156.51, 2200.61

all of which are discarded from further analysis.

3.5.2 *Average Linkage Clustering*

In parallel to ANOVA an average linkage clustering was performed as described in Section 3.4.5. The cluster dendrogram combining correlated peptides and ANOVA p-values is shown in Figure 22. The experimental factors have different impact on the data. The most significant p-values are obtained for genotype (up to 10^{-91}). The different mouse types can be easily distinguished using the profile data. Diet and the combination of genotype and diet seem to have much smaller but still substantial effect on the data (p-values up to 10^{-14}) whereas time has an even

greater effect (p-values up to 10^{-23}). The highly significant peaks are also an effect of the high number of samples. Nearly one third of all peaks - the whole right part of the dendrogram - is associated with the experimental factor time. On this global level the dendrogram gives an intuitive overview of the complete data set as both, similarity and significance information are shown in a unified representation.

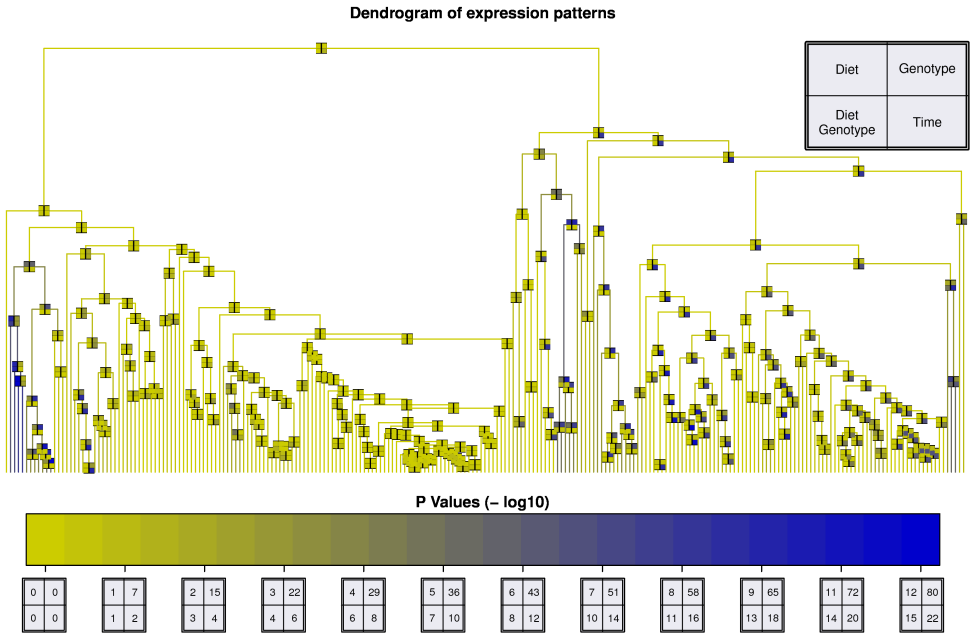


Figure 22: Dendrogram of all peaks identified in this dataset, see methods. Every node is characterized by four ANOVA p-values shown as a color-coded box with four fields: Diet (upper left), Genotype (upper right), Time (lower right) and combination of diet and genotype (lower left). The different $-\log_{10}$ p-value colorscales for the four factors are shown at the buttom.

3.5.3 *Biomarker Identification*

3.5.3.1 *Factor Combination - Diet and Genotype*

Table 3 provides an overview of the two most significant clusters of peaks for the combination of experimental factors genotype and diet. Cluster *Combi:1* comprises three peaks with a mean correlation coefficient of 0.81. It contains the peak m/z 4075 with the most significant p-value for the combination of diet and genotype (10^{-14}) but also the most significant result for solely factor diet (p-value 10^{-14}).The second cluster comprises the peaks m/z 6116, 4041 and 8300 with a p-value of 10^{-8} for combination of genotype and diet and significant p-values for single factor genotype (10^{-24}) and time 10^{-19}.

A detailed illustration of the intensity profile for peak m/z 4075 as representative for cluster *Combi:1* can be seen in Figure 23. This peak shows high intensities for the combination of SJL-genotype and CHF-diet whereas it is almost constantly low for all other factor combinations. This effect is also visible for diet or genotype only. Looking solely at the factor diet we would conclude that peak m/z 4075 is correlated with diabetes-protective CHF diet[39, 76]. An extended analysis of the factor combination, however, shows that this correlation with the CHF diet is only given in SJL genotype, which is not visible in single factor analysis.

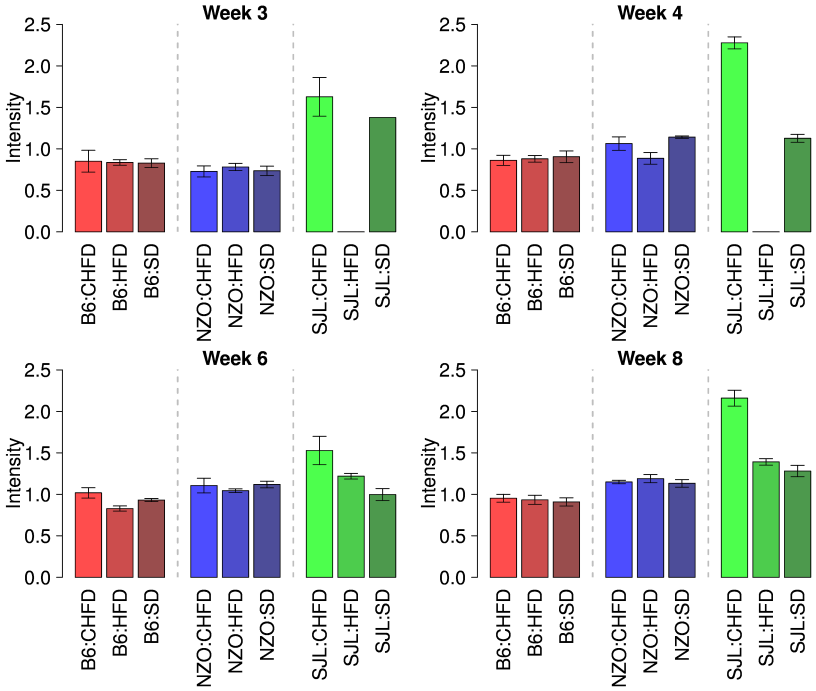


Figure 23: Normalized peak intensities for peak m/z 4075 exemplary for cluster *Combi:1*. Peak intensities for all 3 experimental factors are drawn as bar plots with error of mean error bars. Genotype and diet are given below the bars for each week. The ANOVA method is not sensitivly affected by the missing values due to sample harvesting problems occuring in SJL-HF in week 3 and 4. Higher peak intensities are found only for the combination of SJL genotype and HF diet.

3.5.3.2 *Single Factor - Diet*

The most significant cluster for experimental factor diet is shown in Table 4. Compared to the other experimental factors, diet has the smallest impact on the data. The most significant hit for diet is identical with the most significant hit found for the combination of diet and genotype (see table 3). A single factor analysis also identifies this peak but, however, only multi-factorial analysis offers detailed insights towards biological interpretation.

Cluster	mean-cor	Peak	p values							
			Diet		Genotype		Time		Diet ∗ Genotype	
Combi:1	0.81	2262	1.3e-10	0.00015	1.9e-18	3.8e-11	0.0018	0.97	4.6e-14	1.2e-05
		3618		6.7e-07		2e-08		0.004		8.8e-13
		4075		2e-14		1.5e-29		3e-08		7.9e-14
Combi:2	0.74	6116	0.0035	0.00026	9.2e-24	2.6e-05	1.8e-19	2.6e-13	2.6e-08	0.0091
		4041		1.9e-06		2e-30		3.1e-06		3.2e-05
		8300		1.7e-06		2e-19		3.9e-14		1.2e-07

Table 3: Most significant cluster of peaks for combination of experimental factors diet and genotype. For each cluster, peaks aggregated within this cluster, the average correlation of the peaks and the ANOVA p-values for the three different experimental factors and the factor combination of diet and genotype are given. P-values are given for each peak and for the complete cluster. Cluster *Combi:1* containing the tree peaks at m/z 2262, 3618 and 4075 has the most significant effect for combination of diet and genotype (p-value 10^{-14}) but also significant effect for genotype (10^{-18}) and diet (10^{-10}) alone. The p-value of the cluster for the factor combination is even sightly smaller that the p-values of the corresponding peaks.

Cluster	mean-cor	Peak	p values							
			Diet		Genotype		Time		Diet ∗ Genotype	
Diet:1	0.81	2262	1.3e-10	0.00015	1.9e-18	3.8e-11	0.0018	0.97	4.6e-14	1.2e-05
		3618		6.7e-07		2e-08		0.004		8.8e-13
		4075		2e-14		1.5e-29		3e-08		7.9e-14

Table 4: Most significant cluster of peaks for experimental factor diet. For each cluster, peaks aggregated within this cluster, the average correlation of the peaks and the ANOVA p-values of the corresponding ANOVA model are reported. P-values are stated for every peak and for the complete cluster. The most significant result for diet is the same as received for the combination of diet and genotype (see Table 3).

3.5.3.3 *Single Factor - Genotype*

The experimental factor genotype has the strongest effects on the data with p-values up to 10^{-91}. Table 5 presents the two most significant clusters of peaks for genotype, both with a p-value of 10^{-74}. Cluster *GT:1* comprises 4 peaks m/z 5822, 6329, 4237 and 5029, all of which show high levels for NZO and SJL genotype while the level for B6 genotype is very low (see bar plot of signal intensity for peak 5029 in figure 24 as a representative for this cluster). This peak perfectly distinguishes the B6 genotype from the other two.

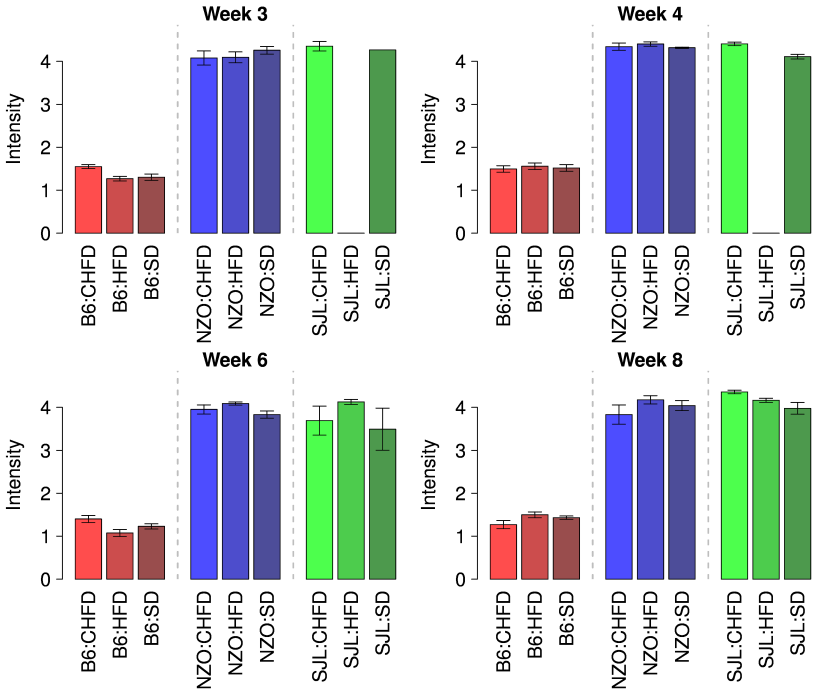


Figure 24: Normalized peak intensities for peak 5029 exemplary for cluster *GT:1*. Peak intensities for all combinations of 3 experimental factors are drawn as bar plots with error of mean error bars. Genotype and diet are given below the bars for each week. The ANOVA method is not sensitively affected by the missing values due to sample harvesting problems occuring in SJL-HF in week 3 and 4. The measured intensities for peak m/z 5029 are low for B6 mice but much higher for NZO and SJL mice. The high difference together with rather small errors and high sample size lead to very significant p-value of 10^{-91}.

Cluster *GT:2* containing the peaks m/z 3556, 3575, 2037, 2488, 3388 has very similar p-value like cluster *GT:1* but shows high signal intensities only for SJL genotype and lower intensities for B6 and NZO mice. A bar plot of peak intensities for peak 3388 as representative for *GT:2* is shown in figure 25. Peaks aggregated in cluster *GT:2* distinguish SJL mice from the other two mouse strains.

Beside the clearly visible biological effect the very low p-values are also an effect of the high number of samples.

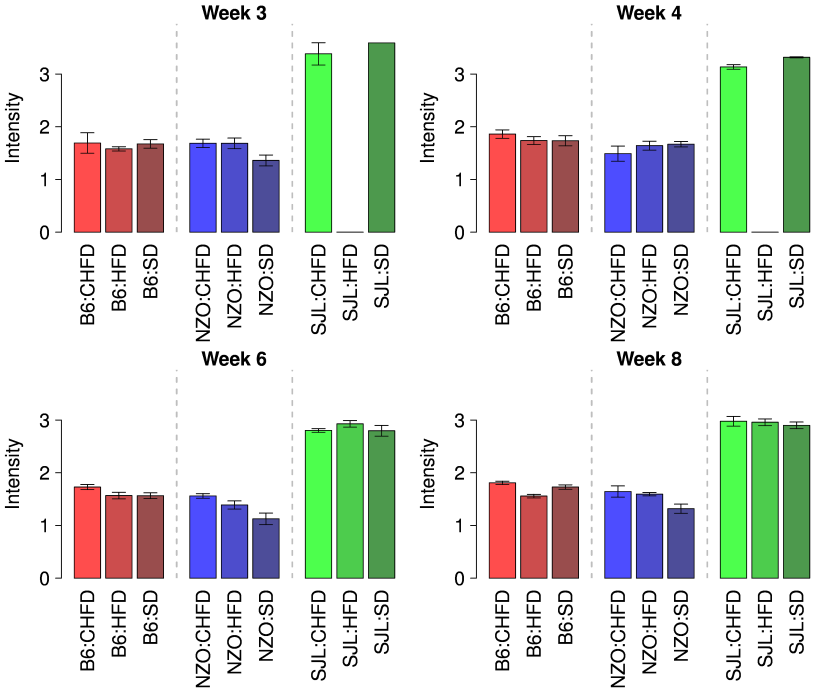


Figure 25: Normalized peak intensities for peak 3388 exemplary for cluster *GT:2*. Peak intensities for all combinations of experimental factors are drawn as bar plots with error of mean error bars. Genotype and diet are given below the bars for each week. The ANOVA method is not sensitively affected by the missing values due to sample harvesting problems occuring in SJL-HF in week 3 and 4. Peak m/z 3388 show high intensities for SJL genotype while for NZO and B6 the intensities are lower. The intensity difference together with rather small errors and high sample size lead to very significant p-value of 10^{-74} for genotype.

The combination of both peaks (3388 and 5029) allows for a perfect separation of all three genotypes. Figure 26 shows a scatter plot of intensity values for both peaks colored by different genotypes. For both peaks and every genotype value, an estimated normal distribution is drawn on top and right hand side. The three genotypes are perfectly distinguishable using these two peaks.

3.5.3.4 *Single Factor - Time*

The experimental design included the investigation of mice at different developmental stages and in effect many peaks are expected to be related to biological changes in growing up of the mice. This is most probably the reason why nearly one third of all peaks is associated with the experimental factor time (see figure 22 on page 50). The best two clusters are shown in table

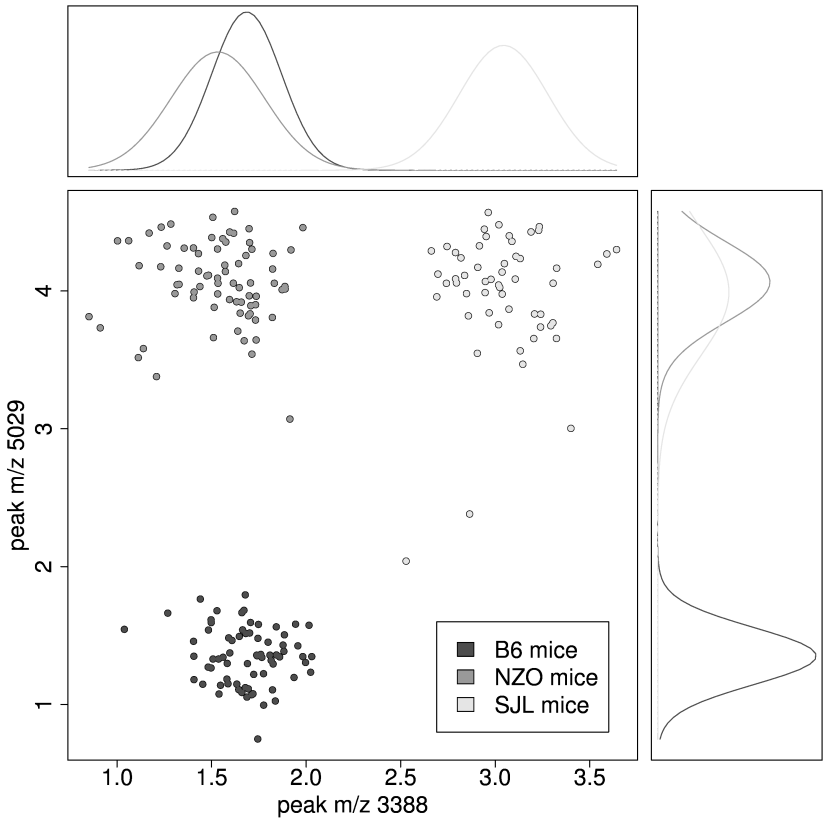


Figure 26: Scatter plot of peak intensity values for peaks 3388 and 5029. On top and the right hand side are fitted normal distributions for every genotype. Peak 3388 separates SJL genotype from the other two and peak 5029 distinguishes B6 from SJL and NZO. Using both peaks all three genotype can be easily separated.

6. The most significant peak (m/z 6569) has a p-value of 10^{-23}. This peak is part of cluster *Time:1* together with peaks m/z 9061 and 31634 having an average correlation coefficient of 0.93. The second cluster (*Time:2*) comprises two peaks m/z 3132 and 7058 with p-values of 10^{-20}. The cluster has a slightly more significant p value of 10^{-21}. The high p-values are caused by intensity differences together with high number of samples and rather low variances.

All p-values are given without multiple testing correction. Applying rigid Bonferroni multiple testing correction for 261 tests, the p-value threshold of 0.05 changes to $0.05/261 = 0.0002$. Hence all p-values discussed above remain significant.

3.5.3.5 *Redundancy - Hemoglobin Peaks*

Protein composition of blood is typically dominated by highly abundant proteins such as albumin and hemoglobin. Albumin and hemoglobin are large proteins represented by a multitude

Cluster	mean-cor	Peak	p values							
			Diet		Genotype		Time		Diet ∗ Genotype	
GT:1	0.96	5822	0.0012	0.0023	1.1e-74	5.7e-81	0.34	2.3e-07	0.18	0.82
		6329		0.022		6.7e-50		0.24		0.19
		4237		9.5e-05		1.1e-70		0.61		0.022
		5029		0.00014		1.3e-91		0.82		0.14
GT:2	0.82	3556	9.1e-06	3.8e-07	1.6e-74	2e-51	0.0058	0.023	0.0081	0.038
		3575		5.9e-05		1e-60		0.0042		0.0041
		2037		0.028		3.3e-35		0.051		0.017
		2488		0.014		8.4e-40		0.29		0.95
		3388		2.5e-12		1.8e-74		6e-04		0.0093

Table 5: Most significant clusters of peaks for experimental factor genotype. For each cluster, peaks aggregated within this cluster, the average correlation of the peaks and the ANOVA p-values for the corresponding ANOVA model are given. P-values are reported for every peak and for the complete cluster. Genotype has a very strong effect on the data with p-values up to $10^{-}74$ for cluster *GT:1* and *GT:2*. The high significance is caused by the high number of samples and by the strong differences (see Figure 25 for a visualization of peak intensities).

Cluster	mean-cor	Peak	p values							
			Diet		Genotype		Week		Diet ∗ Genotype	
Time:1	0.93	3163	0.33	0.4	2.1e-12	2.2e-08	1.4e-22	1.3e-19	0.36	0.48
		6569		0.32		6.9e-15		6.7e-23		0.38
		9061		0.21		6.6e-12		2.7e-19		0.24
Time:2	0.92	3132	0.029	0.059	0.31	0.34	5e-21	7e-20	0.74	0.92
		7058		0.013		0.0058		8.7e-20		0.33

Table 6: Most significant clusters of peaks for experimental factor time. For each cluster, peaks aggregated within this cluster, the average correlation of the peaks and the ANOVA p-values are reported. P-values are given for every peak and for the complete cluster. The most significant cluster for time (`Time:1`) shows p-values up to 10^{-22}. The peaks of cluster `Time:2` have p-values for experimental factor time of 10^{-20} while the cluster is slightly more significant (p-value: 10^{-21}).

of peptides and thus should be presented by multiple peaks in our dataset. Assuming that these peptides have the same originating protein, the intensity profile of the spots are supposed to be correlated. In effect, peptides derived from hemoglobin or albumin should be located in close proximity in the dendrogram. MS-based profile peak identification revealed one albumin and three hemoglobin peptides. Mapping the three hemoglobin peptide peaks into the dendrogram shows that they are indeed in close proximity (see Figure 27) supporting our assumption. In the middle of the dendrodram (see Figure 22) there is a cluster with a big number of peaks not associated to any experimental factor. The peak identified as albumin is located in that big cluster. Maybe the whole cluster presents peptides derived from albumin.

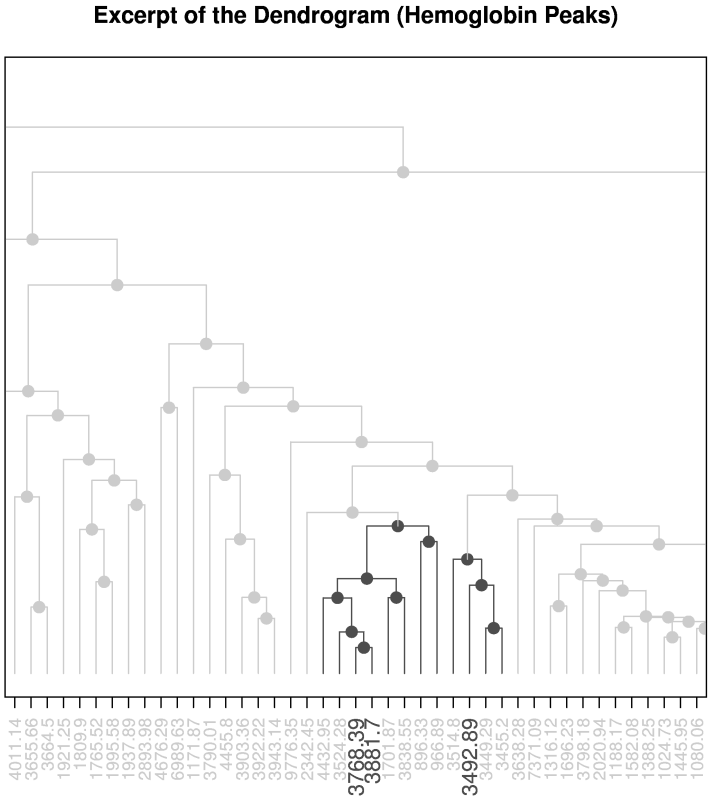


Figure 27: Excerpt of dendrogram (c.f. figure 22) with tree peaks identified as hemoglobin (colored dark gray on x-axis).

3.5.4 *Feature Selection and Classification*

The combination of ANOVA and similarity-based clustering established in the previous sections, is also well-suited for feature selection towards reliable and precise classifications and predictions. In the following, this is demonstrated in an exemplary manner for experimental factor diet. The other two experimen-

tal factors are not suited for purpose of demonstration because genotype classification is more or less trivial (c.f. figure 26) and time is not a factorial variable and has more distinct values. The classification performance is evaluated using cross validation error and ROC curves.

Using the combination of ANOVA and clustering described above (see Section 3.4.5) for feature selection we avoid the shortcoming of typical filter methods as clustering incorporates information about similarity and orthogonality (see Section 2.4). It is sufficient to use one representative feature from the cluster obtained from the statistical analysis to achieve classification performance comparable to wrapper methods. In order to demonstrate the advantages of cluster-based ANOVA we built a classification system with a decision tree ensemble based classifier (random Forest)[18] for the experimental factor diet.

Since the size of the optimal feature set for classification strongly depends on the classifier and on feature-label distribution[70], we performed classification with different feature set sizes: 3, 5 and 8. Feature selection was done by selecting top features from:

1. ANOVA analysis without clustering: Selection of peaks with the most significant p-values for experimental factor diet (Peaks m/z: 1883, 3267, 3407, 4075, 4237, 5176, 5536, 8332).

2. Ant Colony optimization strategy: Using an ACO strategy (see 2.4.1), we identified a set of features with optimized classification results in a similar way to Ressom et al.[133] (Peaks m/z: 3267, 3437, 3575, 4041, 4237, 4965, 6569, 7058).

3. ANOVA analysis including clustering: Selection of clusters or peaks with most significant p-values for experimental factor diet. For every cluster selection of the peak with the most significant p-value as representant for the cluster (Peaks m/z: 1883, 3267, 3407, 3556, 3943, 4075, 5176, 8332).

For ACO we used the in house implementation (see section 2.4.1) with the following parameter set: nAnts=75, nIter=100, nFeatures=$1-5$ and a 10-fold cross validation for evaluation of classification result. The feature selection algorithm took $\sim$ 5h with both CPUs on a Intel Core2 Duo CPU (2.66GHz). Figure 28 shows the course of the pheromone values. Most of the peaks are considered useless for classification, and therefore the level of pheromone is quickly trickling away. On the other hand, some features are well suited for purpose of classification represented by high pheromone values.

To compare the classification results using the three different feature selection methods, confusion matrices of a 10-fold cross validated classification for experimental factor diet are shown in table 7. Beside the classification error, we calculated a p-value

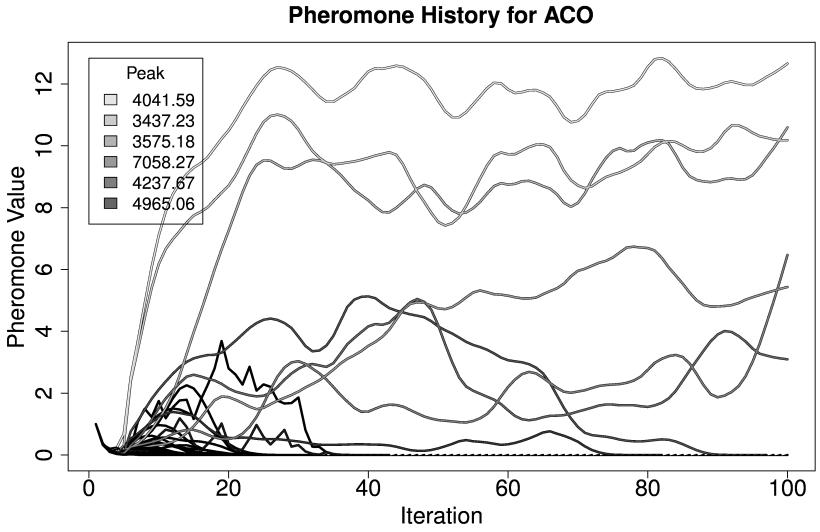


Figure 28: History of pheromone values during feature selection process of ACO for experimental factor diet. A line represents one peak and the x-axis reflects the current iteration. Lines are smoothed using lowess. For the majority of the peaks the level of pheromone rapidly drops close to zero. After 100 iterations only six peaks remain with a substantial pheromone level.

for the classification result by comparing the performance of the selected set of features with the performance of randomly selected sets. Since diet has three distinct values we expect the cross validation error to be normally distributed around 66% for randomly selected feature sets. Estimating this normal distribution by 1000 randomly selected feature sets (estimated: mean = 0.65, sd = 0.05), we were able to assign a p-value to each cross validation error.

Using ANOVA without clustering for feature selection leads to a 10-fold cross validation error of 53% for 3 features (p-value: 0.0028), 52% for 5 features (p-value: 0.006) and 42% for 8 features (p-value: $1 \cdot 10^{-06}$). As expected the ant colony feature selection clearly outperforms the simple filter method with a cross validation error of 40% for 3 features (p-value: $1 \cdot 10^{-08}$), 37% for 5 features (p-value: $2 \cdot 10^{-08}$) and 39% for 8 features (p-value: $5 \cdot 10^{-07}$). However, our improved feature selection technique leads to performances comparable to wrapper method in terms of cross validation errors (44%, 40%, 38% for 3, 5 and 8 features) as well as p-values ($1 \cdot 10^{-6}$, $7 \cdot 10^{-7}$, $3 \cdot 10^{-7}$ for 3, 5 and 8 features).

3.5.4.1 *ROC Curves*

ROC curves are typically used to compare different models and to select possibly optimal ones. Since we have three classes (HF,SD

nFeat	Method	CHF			HF			SD			Error	P-Value
		CHF	HF	SD	CHF	HF	SD	CHF	HF	SD		
3	ANOVA	33	14	18	17	24	24	16	16	36	0.53	0.0028
	ACO	45	4	16	10	33	22	11	16	41	0.4	1e-08
	Cl. ANOVA	40	12	13	15	28	22	9	16	43	0.44	1e-06
5	ANOVA	36	13	16	18	22	25	20	11	37	0.52	0.006
	ACO	48	3	14	12	33	20	10	15	43	0.37	2.7e-08
	Cl. ANOVA	40	13	12	15	38	12	4	24	40	0.4	6.7e-07
8	ANOVA	41	12	12	16	34	15	6	22	40	0.42	9e-06
	ACO	45	5	15	12	30	23	4	18	46	0.39	5.5e-07
	Cl. ANOVA	43	10	12	14	35	16	5	19	44	0.38	3.3e-07

Table 7: Confusion matrices for 10 fold cross validation for experimental factor diet using random forest classifier. Each lines corresponds to one confusion matrix. The feature selection was done by three different methods: ANOVA, ACO and cluster based ANOVA. The feature selection was performed three times with different number of features: 3, 5 and 8. Cells colored in light gray reflect true positives. ACO and Cluster ANOVA show much lower classification errors compared to ANOVA without clustering especially for 3 and 5 features.

and HF) we use the 1-vs-rest method to calculate a ROC curve for each diet (see Section 3.3.3 for more details). In total nine ROC curves are calculated: three diet curves and three feature selection approaches. The ROC curves are presented in Figure 29. For generation of the ROC curves 10-fold cross validation was repeated ten times in order to add error bars for sensitivity and specificity.

Quality of a ROC curve is typically assessed by calculating the Area Under Curve (AUC). The AUC values can range between 0.5 (random classifier) and 1 (perfect classifier). The AUC values for the nine ROC curves are shown in the legend in Figure 29 and presented in an overview in table 8. The classification of Standard Diet is very similar for all three methods as the curves are rather similar and the AUC is 0.87 for all three methods. High Fat diet is the hardest classification task for all three selected feature sets, whereas the ANOVA based feature selection shows the lowest values (0.75) compared to the other approaches (0.8 for cluster ANOVA and 0.86 for ACO). The best performances are received for classification of CHF with AUC values up to 0.9.

Cross validation errors and p-values as well as the ROC curves and AUC values demonstrate the usefulness of the cluster-based ANOVA for feature selection. Cluster-based ANOVA performs similar well as ACO for feature selection.

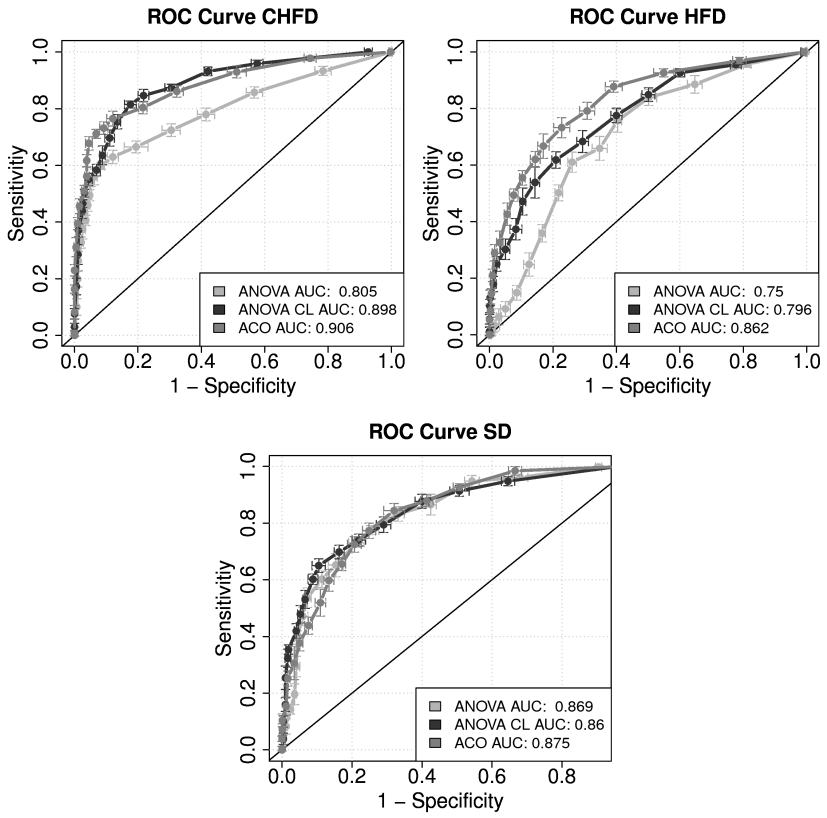


Figure 29: ROC curves for the classification of experimental factor diet. The three different diets are given in separate plots (left hand side: CHF, middle: HF and right hand side: SD). For every diet we compare the ROC curves for the three different feature selection methods.

	CHF	HF	SD
ANOVA	0.8	0.75	0.87
ACO	0.9	0.86	0.87
Cluster ANOVA	0.9	0.8	0.86

Table 8: Area Under Curve values for the three distinct diets and the three different feature selection approaches.

3.6 SUMMARY MALDI MS ANALYSIS

The ANOVA model applied analyzes the effects of single experimental factors as well as the combination of diet and genotype. Before applying ANOVA we ensured that all required assumptions are fulfilled (e.g. homoscedasticity and normal distribution). Hence, ANOVA is the perfect candidate for the statistical analysis and preferable to non-parametric Kruskal-Wallis test since is has better power.

Analyzing combinations of experimental features helps to biologically interpret experimental data. For instance, looking solely at the factor diet we would conclude that peak m/z 4075 is correlated with diabetes-protective CHF diet [39, 76] with a p-value of 10^{-14} (see Table 4). However, peak m/z 4075 was also the most significant result found for the combination of genotype and diet (see Table 3). An analysis of the factor combination, shows that the increased intensity of CHF diet is only visible for SJL genotype (see Figure 23). So peak m/z 4075 is correlated with diabetes-protective CHF diet only in SJL genotype. This biological interpretation is not possible using single factor analysis.

Our method might also help to see weak signals because a cluster comprising several peaks could result in more significant p-values than single peaks. This property of our approach is visible for experimental factor time in *Combi:2* of Table 3. The p-value for the cluster is 10^{-19} while the most significant peak in the cluster has a p-value of 10^{-14}.

One possible reason for a cluster having so many very good correlated peaks like the cluster in the middle of the dendrogram (see figure 22) is a huge common protein where the peaks are derived from. A perfect candidate for this role could be albumin as it consists of 608 amino acids. This hypothesis is supported by the fact that one of the peaks was indeed identified as albumin.

Table 9 shows distinctive properties of our approach compared to other methods. Standard t-test is often the method of choice for statistical testing and the selection of suitable features for classification and prediction. However, standard t-test is not adequate for multi-dimensional datasets since it investigates only one variable with exact two independent groups at the same time. F-test allows for testing multi-dimensional datasets and ANOVA enables to investigate factor combinations. Similarity of features is not considered by any of the statistical tests. Swarm intelligence or genetic algorithms are a different group of algorithms aiming at biomarker identification. Although they are applicable to multi dimensional datasets and take data redundancy into account they often fail in producing deterministic results and p-values. Our work is designed to retain all capabilities of statistical testing while considering feature similarities at the same time.

Method	Deterministic	Feature Selection	p-Values	Multi Dimensional	Combinations	Redundancy
t-Test	✓	✓	✓	×	×	×
F-Test	✓	✓	✓	✓	×	×
ANOVA	✓	✓	✓	✓	✓	×
Swarm Intelligence	×	✓	×	✓	✓	✓
GA	×	✓	×	✓	✓	✓
This Work	✓	✓	✓	✓	✓	✓

Table 9: Comparison of different methods for biomarker identification and feature selection.

Another advantage of our approach is the possibility to use only one, representative peak from a cluster for further analysis. We have seen that the peaks identified as hemoglobin are in close proximity in the dendrogram. Hence, we can assume that many of the surrounding peaks are also most likely derived from hemoglobin. Nonetheless, it has to be kept in mind that many peptides originating from the same parent protein will often behave differently. Our approach aims at identifying co-occuring peptides and hence leads to a reasonable reduction of the data. More complex interaction (e.h. high abundance of a protein causes low abundance of another peptide) would require other processing methods.

3.7 CONCLUSION MALDI MS ANALYSIS

We have introduced a method combining ANOVA and clustering-based redundancy reduction that is suitable for biomarker identification in multi-factorial MALDI-TOF MS profiling studies given an appropriate preprocessing. Applying this method to our data set we were able to identify peaks that are characteristic for the combination of two factors as well as peaks that are significant for single experimental factors. These results are significant even when applying rigid multiple testing corrections. It is shown that ANOVA is an adequate approach for the identification of biologically interesting biomarkers from MS profiling data based on multi-dimensional experimental design. Furthermore, classifications based on features selected with our approach perform similarly well to those generated with more complex global optimization methods.

4

ITRAQ: DATA ANALYSIS AND ALGORITHMS

Chapter Contents

Like the two other techniques, Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) experiments are performed with the aim to gain insights into biological mechanisms underlying T2DM. Compared to MALDI (see Chapter 3), iTRAQ is restricted to a low number of samples due to high costs and experimental complexity. Nevertheless, iTRAQ is well suited for our purpose as it allows for simultaneous quantitation of a large number of proteins while single MS MALDI is often restricted to qualitative results and the number of quantified proteins is limited.

In the first part of this chapter we give a brief introduction to the problems of iTRAQ technology and motivate the necessity for a novel analysis workflow (Section 4.1). Sample preparation and dataset description is referred to in Section 4.2. While Section 4.3 describes State-of-the-Art methods for protein identification and quantitation, Section 4.4 describes the methods and the workflow specifically developed for this thesis. In the first part of the results section (4.5) general results of iTRAQ data analysis such as normalization, parameters for protein identification and search for PTMs are presented. In the second part we describe the results obtained for applying our novel analysis workflow.

Typically, the first step in the evaluation of iTRAQ data is protein identification and quantitation. Data mining techniques such as clustering are typically applied afterwards (see left hand side of Figure 30). In this thesis, we will introduce a statistical analysis workflow for iTRAQ data employing a clustering approach as a very early step in data processing with the aim to reduce peptide heterogeneity (right hand side of Figure 30).

4.1 INTRODUCTION

Quantitative proteomics is becoming increasingly important and over the last years many efforts have been made to develop and improve methods allowing for protein quantification. Besides gel based approaches[93, 47], mass spectral techniques encompassing labeling techniques such as iTRAQ[139], ICAT[64] and SILAC[119, 101] as well as label free approaches are widely-used for quantitative proteomics. iTRAQ has become a very popular technique for protein quantitation. The continuing popularity of iTRAQ requires the evaluation of the technique in terms of accuracy and precision[121]. Accuracy assesses the closeness to the true quantification value. Precision in this context refers to reproducibility of experiments. Since accuracy is difficult to evaluate, precision is the most frequently applied measure for experimental quality[20, 108]. Gan et al.[55] tried to assess the precision of iTRAQ data by analyzing technical (different channels of the same MS run), experimental (same channel but different runs)

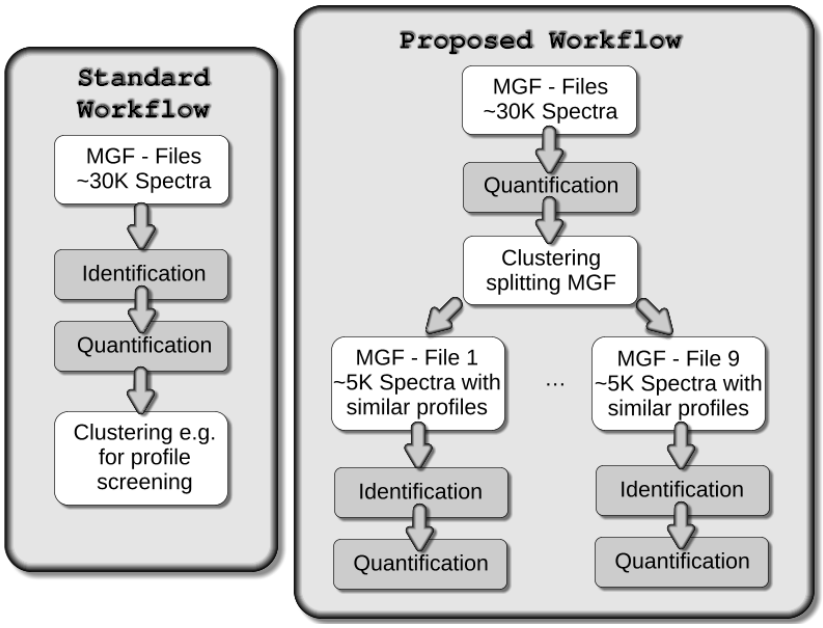


Figure 30: Schematic representation of the data analysis workflow employed for evaluation of iTRAQ data and comparison with standard workflows.

and biological variations (different biological samples). They designed different iTRAQ experiments covering the different types of replications. They found technical variation to be small (11%) whereas experimental and biological variations where more than twice as high. Therefore they underlined the necessity to include a sufficient number of biological replicates in iTRAQ experiments.

In 2008, Lacerda et al.[89] compared the two software packages MASCOT and Peaks (Bioinformatics Solutions Inc., Waterloo, ON, Canada)[100] using a six-protein mixture as well as a complex protein sample. They revealed significant differences in the two packages: For a complex protein mixture, only 26% of the proteins agreed within 20% error of quantitation ratios. These ambiguities are only due to algorithmic differences since both packages were applied to the same experimental data set. This implies that beside biological and technical variations caused sample preparation and mass spectrometer, there is also a considerable variation due to the software package used for evaluation. The highest fold change measured with iTRAQ differs widely among laboratories but rarely seems to exceed ten-fold, which was reported by Casado-Vela et al.[22] in a technical survey examining more than 200 articles. These are low fold-changes compared to microarray transcriptome profiling where a differentially expression of more than 32-fold (log2 fold of 5) is observed frequently.

For iTRAQ - like for the majority of MS based quantitation approaches - quantitation measurements are performed at the peptide level. Since often multiple peptides are measured for the

same protein, the need for some kind of summarizing strategy is obvious. Different ideas regarding the calculation of protein quantitation from multiple peptides have been applied including mean or median calculation[25, 17] and error weighted means[95]. Because of the fixed stoichiometric ratio, quantitation measurements for peptides uniquely assigned to the same protein should be strictly correlated[69]. But often this presumption is not fulfilled and the quantitation values exhibit a substantial heterogeneity. The heterogeneity is also observed for quantitation ratios and is not due to different ionization or fragmentation efficiency. This is illustrated in Figure 31 presenting the quantitation ratios of peptides for an exemplary chosen protein *40S ribosomal protein S30*. The 117/116 log-ratio (NZO genotype high fat diet vs. SJL genotype standard diet) varies from -0.5 (1.4 fold down-regulation) to 1 (2 fold up-regulation).

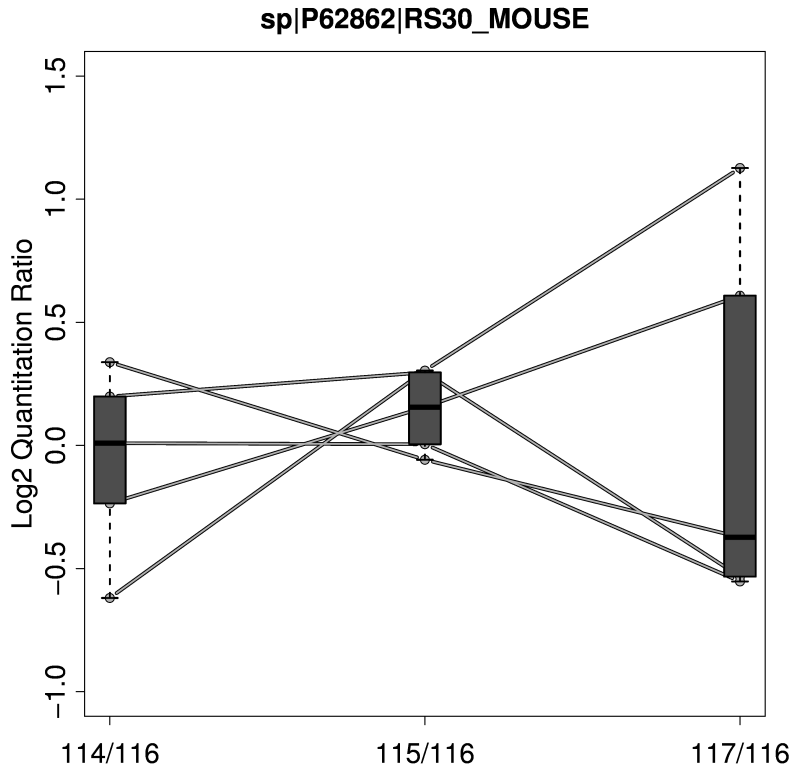


Figure 31: Demonstration of peptide heterogeneity for *40S ribosomal protein S30 (RS_30)*. Every line represents a unique peptide profile (peptide-spectrum-match) identified as originating from the RS_30 protein. iTRAQ ratios are calculated using 116 channel (SJL mouse with standard diet) as reference. For every ratio a box plot giving the lower quartile, median and upper quartile is drawn. Especially for the 117/116 ratio (NZO mouse with high fat diet) the quantitation ratios are heterogeneous ranging from -0.5 to $+1$ (corresponding to a 1.4 fold down-regulation or 2 fold up-regulation).

	NZO_SD	NZO_HF	SJL_SD	SJL_HF
Exp 1	mouse:1 channel:114	mouse:4 channel:117	mouse:7 channel:116	mouse:10 channel:115
Exp 2	mouse:2 channel:115	mouse:5 channel:114	mouse:8 channel:116	mouse:11 channel:117
Exp 3	mouse:3 channel:116	mouse:6 channel:115	mouse:9 channel:117	mouse:12 channel:114

Table 10: Experimental design and iTRAQ labeling (114 - 117) for three experimental replications (Exp 1, Exp2 and Exp 3). The experimental design combines technical and biological replications.

4.2 SAMPLE PREPARATION

Although iTRAQ is established as a reliable high throughput technology for quantitative proteomics, it is restricted to a small number of samples. However, the use of biological replicates is strongly recommended for statistical evaluation[55]. Considering these two facts, the study design that can be analyzed by iTRAQ experiments must be limited to the most relevant biological questions to keep the number of samples small. The dataset of the Sys-Prot project is restricted to the two different diets - Standard Diet (SD) and High Fat (HF) diet as well as the two mouse strains: New Zealand Obese (NZO) and Swiss Jim Lambert (SJL). This leads to a total of four distinct combinations of experimental factors. Each of the four distinct combinations is covered by three biological repeats (Exp 1, Exp 2 and Exp 3). The four distinct combinations of experimental factors together with three biological replications for each factor combination, lead to a total of $4 \cdot 3 = 12$ measurements. The experimental design and iTRAQ labeling strategy is shown in Table 10. Due to this experimental design both, technical variance (reflected by permutation of iTRAQ channels) and biological variance (due to different mouse individuals) are combined. Instead of trying to separate technical and biological effects we are more interested in effective variance.

The small number of samples is particularly evident compared to MALDI where more 30 different mouse individuals were used to create over 150 distinct biological samples measured in more than 1100 MS spectra (see Section 3.2). The complete experimental design of Sys-Prot project is described in Section 2.2.3.

4.3 STATE-OF-THE-ART

For evaluation of iTRAQ data, two basic steps are necessary: peptide/protein identification and peptide/protein quantitation. These two aspects are more or less independent and require different solutions.

4.3.1 *Protein Identification*

Protein identification is a complex process which aims to assign mass spectra to specific peptide sequences and thus to protein ids. Beside MS/MS peaks, a spectrum is characterized by retention time derived from High-performance liquid chromatography (HPLC) and precursor mass derived from the first MS (see also Section 2.1.2). Protein identification adopts the two-stage process of MS/MS. The first step for protein identification is to find peptide candidates in the protein sequence database whose mass matches the precursor mass within a given precursor mass tolerance (first MS). In the second step, theoretical spectra are generated for each peptide candidate (simulating MS/MS). The theoretical MS/MS spectra are then compared with the measured MS/MS spectra and peaks are matched if their masses are identical within a given fragment mass tolerance. This general paradigm of peptide identification is used (in different implementations) by the common protein identification tools such as Mascot, X!Tandem[32], Sequest[179] or Open Mass Spectrometry Search Algorithm (OMSSA)[57]. The process of peptide and protein identification is described in more detail for X!Tandem.

X!Tandem calculates a matching score as the sum of intensities of measured spectra using the matched y and b ions of the theoretical spectrum. This score is than multiplied by the factorial for the number of assigned b and y ions to calculate the X!Tandem HyperScore[46]:

$$\mathrm{HyperScore} = \left(\sum_{i=0}^{n} I_i \cdot P_i \right) \cdot N_b! \cdot N_y!$$

where I is the vector of spectrum intensities (normalized to a maximum of 100), P is a boolean vector indicating whether the spectra peaks are matched to theoretical peaks and N_b and N_y are the number of matched b and y ions. The motivation of the factorial values is given by Fenyo and Beavis[46] assuming an underlying hypergeometric distribution of the score ($\sum_{i=0}^{n} I_i \cdot P_i$). A score based on k y-ions compared to $k-1$ y-ions should be $\sim k$ times better. Fenyo and Beavis[46] argued that it is more intuitive and reasonable that a spectrum matching 10 peaks is 10! times better compared to a spectrum matching just a single peak. Without the additional term $N_b! \cdot N_y!$ the spectrum matching 10

peaks is only 10 times better compared to a spectrum matching just a single peak.

For every peptide candidate the HyperScore is calculated and the peptide candidate with the highest HyperScore is assumed to be the correct peptide. The decay of HyperScores for all peptide candidates for a spectrum is found to be exponentially distributed. Therefore the log distribution is assumed to be linear. A log-linear function representing the expected number of random matches is fitted to the right-hand tail of the HyperScores distribution. This function is used to extrapolate the Expected Value (E-value) for the best peptide. The X!Tandem E-value is a good measure of how good the best score is relative to the rest. A peptide is considered correct if the E-value is below peptide E-value threshold.

After calculation of peptide E-values, the next step is to infer protein identification. For a protein all best matching peptide sequences (best peptide E-value) are collected. If more than one spectrum is assigned to the same peptide sequence, then the peptide with the best E-value is kept and the rest is discarded. Based on the number of unique, high scoring peptides n of a protein and their respective scores e_i, a protein e-value is calculated:

$$e_{prot} = \frac{1}{sN^{n-1}} \cdot \binom{s}{n} \cdot p^n \cdot (1-p)^{s-n} \cdot \prod_{i=1}^{n} e_i$$

where s is the number of mass spectra in the dataset, N is the total number of tryptic peptides generated from the protein sequence, and p is N divided by the the number of tryptic peptides that were examined during the complete run.[44, 184] The last term is only the product of the underlying peptide E-values. The first terms describe the probability of random parent mass matches calculated by binomial distribution scaled by the number of spectra and the number of tryptic peptides for the protein to the power of number of high scoring peptides.

The whole process is completely different for other search engines. For example, Sequest calculates the XCorr score which is defined as the ratio of direct comparison of generated and observed spectra with the auto correlation background[59]. The direct comparison matches generated spectrum of the peptide candidate with the measured spectrum as the sum on overlapping peaks. For the autocorrelation background the measured spectrum is shifted backward and forward and the matching score is calculated.

Although Mascot is among the most widely used protein identification packages, many details of the Mascot search engine and probability-based Mowse scoring algorithm are not published.[116]

4.3.2 *Quantitation*

A wide range of quantification algorithms can be found in the literature. The most common algorithms are included in software packages such as MASCOT, ProQUANT, OpenMS[132, 149], i-TRACKER[143, 90], Multi-Q[180] or virtual expert mass spectrometrist (VEMS)[138]. In principle, iTRAQ quantitation of peptides refers to the extraction of the iTRAQ reporter mass intensities (see Chapter 2 - Figure 4). After extraction of quantitations an isotope correction has to be performed according to manufacturer's specifications (Applied Biosystems, Foster City, CA).

In a second step peptide quantitations have to be summarized in order to obtain protein quantitations. A typical problem in this process is that quantitation values of peptides assigned to the same protein often exhibit a substantial heterogeneity (c.f. Figure 31). To counteract this heterogeneity many approaches make use of outlier detection methods like Grubb's test[25] or Dixon's test[95] prior to averaging. But for several reasons outlier filters are problematic: First, outlier filtering can be applied only to proteins with a certain minimum number of peptides, a presumption often not fulfilled in iTRAQ datasets[83]. Second, if the heterogeneity is due to differentially regulated protein isoforms than the less frequent isoform is possibly regarded as an outlier and removed leading to a loss of information. Third, if outlier detection is applied after protein identification, peptides are removed that contributed to the protein identification score and hence the score is distorted a posteriori.

4.4 METHODS

4.4.1 *MASCOT*

Peptide identification and quantitation were performed using MASCOT search engine (version 2.2.04 Matrix Science, London). Peptides identified with a MASCOT score < 50 and a significance threshold of $p > 0.05$ were neglected. Parameters used for protein identification with MASCOT are shown left hand side of Table 11.

The database used was a SwissProt derived Functional Genomics Center Zurich (FGCZ) in-house mouse database from 2009 containing 43636 mouse protein sequences (OS=Mus musculus) and 259 additional FGCZ specific entries. All proteins are present in normal/forward sequences and decoy/reverse sequences. Randomized decoy database (reversed sequences) was used for controlling False Discovery Rate (FDR)[43, 77]. For calculation of FDR the list of proteins ordered by MASCOT *ProtScore* was cut if a given FDR level was reached. Because we intend to achieve reliable quantitation results instead of providing a comprehensive protein list, the false discovery rate was chosen restrictively: FDR=0.1%.

4.4.2 *X!Tandem and OpenMS*

Peptide identification was performed using X!Tandem software[32] version 2009.04.01.1. Parameter set used for protein identification with X!Tandem are shown right hand side of Table 11. All parameters where chosen to be similar to the MASCOT method in order to assure comparability of the results. Extraction of 4-plex iTRAQ quantitation data and isotope correction was performed using OpenMS[132, 149] svn revision 6265. The same decoy database as for MASCOT analysis was used and again false discovery rate was chosen restrictively: FDR=0.1%. For calculation of FDR the list of proteins ordered by X!Tandem protein identification score was cut if a given FDR level was reached.

4.4.3 *Peptide Profiling Guided Identification of Proteins*

In the following we will describe Peptide Profiling Guided Identification of Proteins, a novel workflow we developed for reliable and stable protein quantitation. PPINGUIN seizes on the presumption that quantitation profiles of peptides derived from the same protein are highly correlated as they have a common source. Pseudo-code representation for PPINGUIN is given in Algorithm 4.

We define an iTRAQ quantitation profile of a spectrum as the ordered list of the raw quantitation values, in our case the raw

	MASCOT	X!Tandem
Enzyme/Cleavage	Enzyme: Trypsin	cleavage site: '[RK]|P'
Missed Cleavages	2	2
Fixed Modifications	Methylthio (C), iTRAQ4plex (N-term), iTRAQ4plex (K)	Methylthio (C), iTRAQ4plex (N-term), iTRAQ4plex (K)
Variable Modifications	Oxidation (M), iTRAQ4plex (Y)	Oxidation (M), iTRAQ4plex (Y)
Peptide/Precursor Mass Tolerance	6 Part Per Million (ppm)	6 ppm
Fragment Mass Tolerance	0.1 Da	0.1 Da
Quantification	iTRAQ 4 plex with weighted protein ratio and median normalization of ratios	quantification with OpenMS and multi lowess normalization of ratios
Additional Parameters	mass values: monoisotopic; instrument type: ESI-FTICR; Isotope error mode: 0;	refinement of unanticipated cleavages
False Discovery Rate	0.1%	0.1%

Table 11: Direct comparison of the most important input parameter used for MASCOT and X!Tandem.

intensities of the four iTRAQ channels 114 to 117. As a first step and thus without regarding protein inference, iTRAQ quantitation profiles of the spectra are calculated by extracting the four quantitation values using OpenMS (I in Algorithm 4). Spectra with missing or incomplete quantitation profiles are discarded. The recommended isotope correction is performed according to manufacturer's specifications (Applied Biosystems, Foster City, CA) using OpenMS. Isotope correction aims at correcting for trace levels of isotopic impurities and is done by solving a system of equations. In addition, a complementary normalization of the four quantitation values is performed as described in Section 4.4.5 below (II in Algorithm 4).

Logarithmic quantitation profiles of the spectra are clustered in a coarse-grained manner using k-means algorithm[67, 148, 68] based on Euclidean distance and randomly selected starting points (IV in Algorithm 4). We use k-means clustering (k=5) as it is computationally fast and well suited to show the benefit of the pre-selection. The group size parameter k=5 was chosen according to two internal cluster validation measures (see Section 4.4.6). Since k-means clustering depends on the selected starting points a cluster stability analysis is performed (see Section 4.4.6.1 below).

Clustering intends to create groups of peptides with similar biological profiles (e.g. up-regulation for a certain combination of genotype and diet). As subsequent analysis is focused on relative iTRAQ ratios instead of absolute quantitation values, the profiles are centered prior to clustering (mean is set to zero) (III in Algorithm 4). In order to preserve differences between relative iTRAQ ratios no additional scaling was performed (standard deviation is preserved). This procedure is equivalent to a clustering using Euclidean distances on centered logarithmic quantitation profiles. With this procedure an explicit choice of a reference channel is not necessary.

Every spectrum is assigned to exactly one group and for every group the corresponding spectra show similar iTRAQ quantitation profiles. Quantitation and identification is now performed independently for each group with identical settings to X!Tandem and OpenMS approach (V in Algorithm 4 and see also workflow in Figure 30). Similar to the X!Tandem/OpenMS approach, the FDR was calculated by cutting the list of proteins ordered by X!Tandem protein identification score if a given FDR level was reached. The FDR is calibrated for each group individually and in effect, the X!Tandem threshold for protein identification differs in each group.

Finally, $\log_2$ ratio profiles are calculated using SJL genotype with Standard Diet (SD) as reference. Following the definition of quantitation profiles, ratio profiles are defined as the list of 3 possible iTRAQ ratios (e.g. for Exp 1: 114/116, 115/116 and 117/116).

All calculations (normalization and clustering) were performed using R statistical programming language (R[128] version 2.7.0 - 2008-04-22 and R version 2.12.1 - 2010-12-16). Protein inference and extraction of quantitation values was performed using X!Tandem and OpenMS as described previously.

4.4.4 *PPINGUIN with random clustering*

A main feature of PPINGUIN is the clustering based on quantitation profiles in order to group spectra with similar quantitation profiles. To assess the importance of the clustering, the whole procedure is performed again but k-means clustering is replaced by a random grouping. The rest of the workflow is performed exactly as described for PPINGUIN in Section 4.4.3

4.4.5 *Normalizing iTRAQ quantitations*

Additional normalization of the 4 quantitation values is required to correct for technical bias[121]. Karp et al.[83] observed a heterogeneity of variance for iTRAQ ratios where the width of the dis-

Algorithm 4 General description of PPINGUIN. Five main steps are labeled with Roman numerals. PPINGUIN uses k-means clustering with $k = 5$.

```
▷(I) Extract quantitation profiles for all spectra
(if available)
ListWithQPs ← NA
for all MS/MS Spectra (spec) do
    if AllQuantitationsAreAvailable(spec) then
        QP_spec ← ExtractQuantitationProfile(spec)
        IsotopeCorrection(QP_spec)
        AddToList(ListWithQPs,QP_spec)
    else
        Discard(spec)
    end if
end for
▷(II) Normalization and Logarithmic Transformation
for all Quantitation Profiles (QP_spec) in ListWithQPs do
    QP_spec ← Log2(QP_spec)
end for
MultiLowessNormalization(ListWithQPs)
▷(III) Centering
for all Quantitation Profiles (QP_spec) in ListWithQPs do
    QP_spec ← QP_spec−Mean(QP_spec)
end for
▷(IV) K-Means Clustering
ClusterList_{1..k,QP_{1..n_i}} ← KMeans(ListWithQPs,k)
▷(V) Apply Identification and Quantitation for each
cluster
Result ← NA
for i ← 1..k do
    ▷retrieve original MS/MS spectra
    OrigSpecs ← getOriginalMSMSSpectra(ClusterList_{i,})
    ▷apply identification and quantitation extration
    IDs ←X!Tandem(OrigSpecs)
    QPs_new ←ExtractQuantitationProfiles(OrigSpecs)
    ▷to store ratio profiles
    RPs ← NA
    for all QP_i in QPs_new do
        QP_i ← Log2(QP_i)
        RP_i ← getRatioProfile(QP_i,ref=(SD,SJL))
        Add(RPs,RP_i)
    end for
    Result_i ← Merge(IDs,RPs)
end for
```

tribution is significantly larger at low intensities. They proposed a variance stabilizing normalization based on VSN software[72]. We apply and compare three different normalization strategies: VSN, multi lowess algorithm - a multi dimensional extension of lowess normalization strategy[127] and simple median correction.

4.4.6 *Determining the Number of Clusters*

The number of clusters is an important parameter for clustering. The preferable number of cluster is determined using two different internal measures: gap statistic[155, 7] and Xie-Beni index[176].

The gap statistic compares the within-cluster sum of squares (W_k, k is the number of clusters) with its expected value a under null reference distribution. W_k is calculated as:

$$W_k = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ji} - c_j\right)^2$$

where c_j is the center of the j^{th} cluster. The null reference distribution reflects data without any groups. It is generated m times with a uniform feature distribution over the range of observed values. The expected value of the null reference distribution is calculated as the mean log within-cluster sum of squares of the m reference datasets. The Gap is calculated as:

$$Gap(k) = \frac{1}{m} \sum_{m=1}^{M} \log(W^*_{mk}) - \log(W_k)$$

Xie-Beni index is defined as the ratio of the compactness of the clusters and the separation between the clusters[176]. The separation between the clusters is defined as the minimum pairwise distance between all cluster centers.

Both measures were calculated for 25 replications.

4.4.6.1 *Cluster Stability Analysis*

The number of possible partitions is overwhelming and can be calculated with a Stirling number of the second kind (see formula in [148]). E.g. for 25 objects and 4 clusters, the number of non-empty partitions is $\sim 4.7 \cdot 10^{13}$. Finding the optimal partition with Euclidean sum-of-squares clustering is np-hard[5]. K-means aims at providing a good (hopefully optimal) partition in a reasonable amount of time. Depending on the starting points, k-means may get stuck at a local optimum resulting in different clusters for different starting points.

Cluster stability analysis investigates whether k-means finds (almost) the same partition for different starting points. Clustering is performed for 1000 replications each with different starting

points. Starting points are selected randomly from the set of elements to be clustered. We than calculate fraction of the number of points (quantitation profiles of the spectra) that are clustered in the same cluster for each of the 1000 replications. If the fraction is very high ($> 99\%$) than the clustering is very stable since the calculated partition is almost always the same.

4.4.7 *Differential Analysis*

Differentially expressed proteins are identified using a two-sided one sample t-test. The t-test analyzes whether the log ratio (e.g. NZO_HF/ SJL_SD) is zero (the null hypothesis: $\mu = 0$). P-values and fold changes are calculated protein-wise. Since a big number of tests is performed a multiple testing correction is indispensable. Perhaps the best known procedure in multiple testing correction is Bonferroni procedure[41]. Bonferroni correction rejects any hypothesis H_j with unadjusted p-value less than or equal to α/m (m = number of tests performed). The corresponding single-step Bonferroni adjusted p-values are given by $\tilde{p}_j = \min(p_j \cdot m, 1)$. Benjamini-Hochberg[15] as a more sophisticated multiple testing correction methods is monotone transformations and does not change the rank of the p-values. Since the ranking of p-values is used for further analysis, more simple Bonferroni correction is preferred.

The number of samples for t-test is restricted to 3. The average fold is a robust indicator for differentially regulated proteins. Top list of differentially expressed proteins is created by selecting proteins with mean absolute $\log_2$ fold changes > 0.5 ($\sqrt{2}$ fold change).

4.4.8 *Modification Search*

Searching simultaneously for a variety of protein modifications (almost 900 are currently annotated in UniMod[35]) is not possible with standard tools (such as MASCOT or X!Tandem). E.g. X!Tandem allows at most one modification for each amino acid residue. So we developed a modification search strategy aiming at the identification of relevant modifications. Protein identification with X!Tandem search engine (see Section 4.4.2) was repeatedly performed searching for variable modifications for every modification listed in UniMod (except iTRAQ4plex and Methylthio that are defined as fixed modifications). For every modification we count the number of peptides with good E-values identified from forward and reverse database. Relevant modifications are expected to be found frequently in the forward database but less frequently in the reverse database. Modifications are scored using the ratio between occurrences in forward and reverse database.

4.4.9 *Calculation of CV values for Peptide Homogeneity*

Let $y_{j,r}$ be the relative quantitation ratio for a peptide j and ratio $r \in R = \{$ NZO_SD/SJL_SD, NZO_HFD/SJL_SD and SJL_HFD/SJL_SD $\}$. To assess peptide homogeneity, we calculate the coefficient of variation of a protein p by using all unique peptides for proteins:

$$CV_p = \frac{1}{3 * n_p} \sum_{j \in p} \sum_{r \in R} \frac{\sigma_{j,r}}{\mu_{j,r}}$$

where n_p is the number of unique peptides for protein p and $\sigma_{j,r}$ and $\mu_{j,r}$ are the standard deviation and mean of relative quantitation ratios $y_{i,r}$ of all peptides uniquely assigned to protein p. The final coefficient of variation is calculated by averaging CV_p for all proteins.

4.4.10 *Calculation of CV values for Experimental Reproducibility*

Let $y_{e,i,r}$ be the relative quantitation ratio for experiment $e \in$ {Exp1, Exp2, Exp3}, protein $i \in I = 1..n$ and ratio $r \in R = \{$ NZO_SD/SJL_SD, NZO_HFD/SJL_SD and SJL_HFD/SJL_SD $\}$. In order to assess experimental reproducibility of r we calculate the average CV of all proteins occurring in all three experiments:

$$CV_r = \frac{1}{n} \cdot \sum_{i \in I} \left(\frac{\sigma_{i,r}}{\mu_{i,r}} \right)$$

where $\sigma_{i,r}$ and $\mu_{i,r}$ are the standard deviation and mean of relative quantitation ratios $y_{i,r}$ for protein i and ratio r for all three experiments:

$$\mu_{i,r} = \frac{1}{3} \sum_{e \in E} y_{e,i,r}$$

$$\sigma_{i,r} = \sqrt{\frac{1}{2} \sum_{e \in E} (y_{e,i,r} - \mu_{i,r})^2}$$

This value is reported together with mean standard deviation of $\log_2$ ratios:

$$StDev_r = \frac{1}{n} \cdot \sum_{i \in I} (\hat{\sigma}_{i,r})$$

where $\hat{\sigma}_{i,r}$ is the standard deviation of $\log_2$ ratios:

$$\hat{\mu}_{i,r} = \frac{1}{3} \sum_{e \in E} \log_2(y_{e,i,r})$$

$$\hat{\sigma}_{i,r} = \sqrt{\frac{1}{2} \sum_{e \in E} (\log_2(y_{e,i,r}) - \hat{\mu}_{i,r})^2}$$

4.5 RESULTS

We have developed Peptide Profiling Guided Identification of Proteins (PPINGUIN), a novel workflow for the quantitation of iTRAQ data. It is based on a peptide clustering using quantitation values followed by protein identification for each cluster independently (see Methods). In the first part of the result section, the effects of the different normalization strategies are presented (Section 4.5.1). Afterwards, we optimize the parameters for protein identification and demonstrated the results for identification of PTMs (Section 4.5.2 and 4.5.3). In section 4.5.5 we show the effects of the clustering used within PPINGUIN.

The quantitative results of PPINGUIN are compared with standard evaluation approaches using MASCOT and X!Tandem software. The quality is determined by three different criteria: (i) homogeneity of peptide profiles (ii) precision and (iii) accordance with prior knowledge (Section 4.5.7 - 4.5.9).

In section 4.5.10 we demonstrate the capability of PPINGUIN to detect potential protein isoforms. Finally, we compare the results obtained for proteomics with results of a selected transcriptomics microarray experiment.

4.5.1 *Normalization*

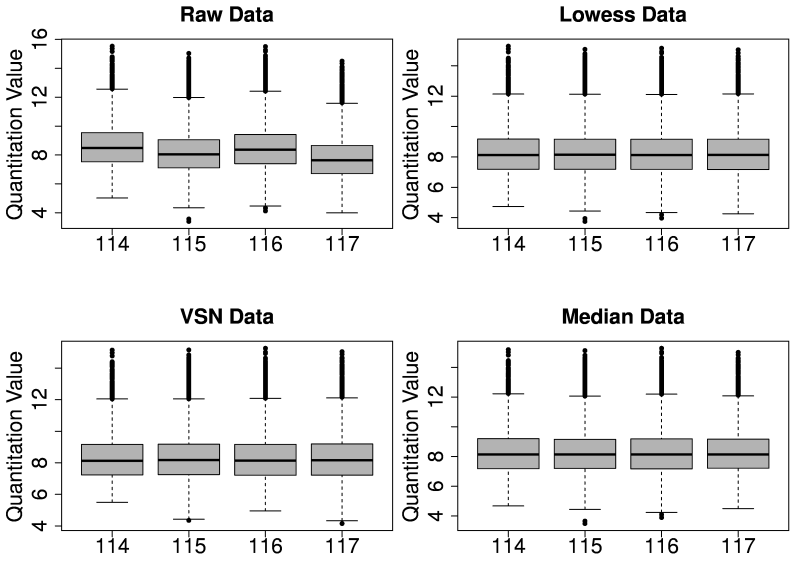


Figure 32: Box plot of raw iTRAQ data (upper left) and after application of three different normalization strategies: vsn (lower left), multi-lowess (upper right) and median correction (lower right). Prior to normalization there is a difference in median quantitation values that is removed by the normalization.

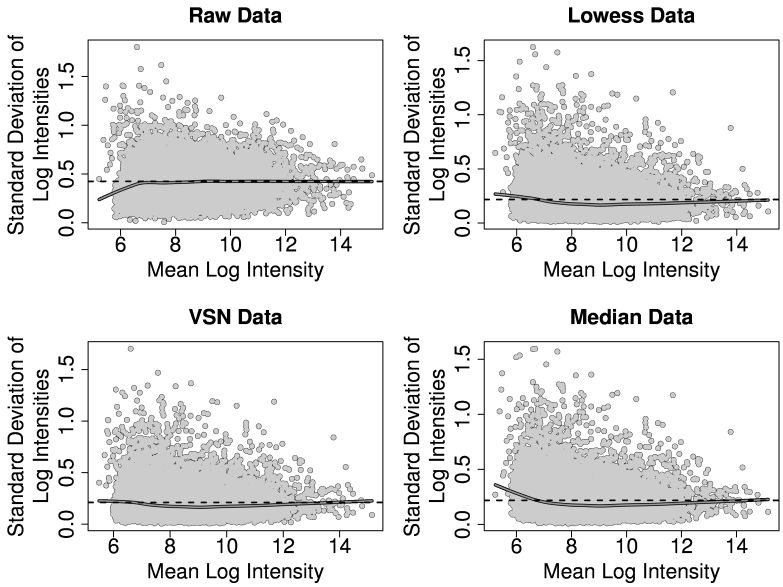


Figure 33: Standard error plot of raw data (upper left) and after application of three different normalization strategies: vsn (lower left), multi-lowess (upper right) and median correction (lower right). Bold line represents a lowess fit and black dashed line reflect median standard deviation.

Typically, raw iTRAQ data is biased by channel effects leading to differences in median quantitation values for each channel. This bias is visualized in the box plot of raw iTRAQ quantitation data exemplarily for the first experiment (upper left part of Figure 32). This bias is comparable to fluorescent dye bias for DIGE or multi-color microarray data. Without correction, these differences lead to a systematic bias in quantitation ratio. The medians of the iTRAQ channels span a range of 0.85 which would correspond to 1.8-fold differential expression if not corrected. A normalization strategy aims to remove this systematic effect. All three normalization algorithms applied: vsn, multi-lowess and median correction (see Section 4.4.5) successfully removed this bias and the differences in median quantitation value vanished (see Figure 32).

Another purpose of normalization is to assure homoscedasticity: homogeneity of variance (c.f. MALDI data evaluation Section 3.4.1.5). Homoscedasticity is a prerequisite for many statistical tests such as t-test or ANOVA. Figure 33 shows a standard error plot prior to and after application of the different normalization approaches. Prior to normalization the mean error of the 4 iTRAQ channels is 42% and the variance is lower for small intensities. After normalization the mean error of the 4 channels is only half as high: 22% for all normalization strategies applied. Median correction does not result in homoscedastic data since the error is higher for smaller quantitation ratios. This observation is in

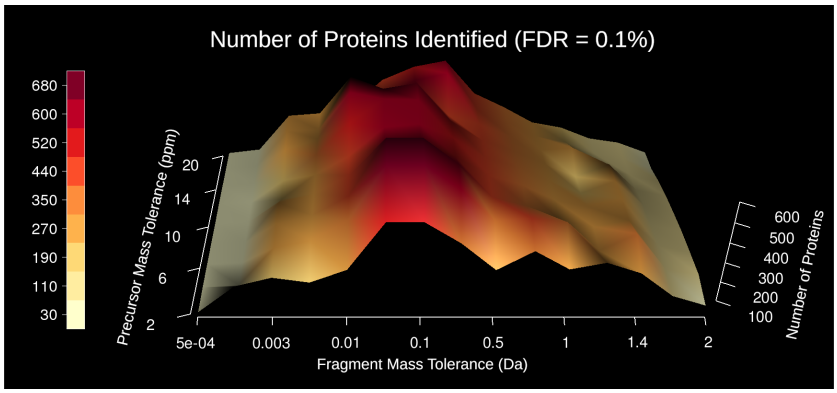


Figure 34: The number of identified proteins (z-axis) for different values of fragment mass tolerance (x-axis) and precursor mass tolerance (y-axis). Color code for the number of proteins is drawn left hand side.

accordance to the work of Karp et al. [83]. The other two normalization approaches lead to an almost constant variance and hence to homoscedasticity. Both, multi-lowess and VSN are very similar in terms of error and homogeneity of variance. Further analysis is based on multi-lowess normalized data.

4.5.2 *Finding Optimal Parameter for Protein Identification*

Protein identification (described in Section 4.3.1) strongly depends on the input parameters. Especially the number of (significantly) identified proteins depends on the choice of input parameter. Two of the most important parameters are precursor mass tolerance (or peptide mass tolerance in MASCOT) and fragment mass tolerance. Precursor mass tolerance changes the number of peptide candidates found for a spectrum. Fragment mass tolerance alters the mapping of theoretical and observed spectra and therefore changes the peptide HyperScore and E-value. So especially fragment mass tolerance is assumed to have an important influence on the number of identified proteins.

To investigate the influence of both parameters, we performed a screening with different parameter sets. The objective function was the number of proteins for a given FDR e.g. FDR $= 0.1\%$. The number of proteins as a function of both parameters is visualized in Figure 34. As expected, fragment mass tolerance strongly changes the number of peptides, whereas precursor mass tolerance has only minor influence (at least for the tested range). Choosing a very restrictive value for fragment mass tolerance (below mass accuracy of the instrument), the matching of theoretical and observed spectra is hindered. In effect a peptide candidate cannot show a high HyperScore or a low E-value respectively. On the other side choosing a permissive value for fragment mass

tolerance (high above mass accuracy of the instrument) leads to many peptide candidates whose theoretical spectra are matched almost perfectly to the observed spectra. In effect the exponential decay of HyperScores for peptide candidates is widened and the extrapolated E-value is less significant (see Section 4.3.1 for details on protein identification). This effect becomes clear keeping in mind that the E-value measures how good the best HyperScore is relative to the rest. If the rest of the peptide candidates results in higher HyperScores, than the distance to the best peptide shrinks and the E-value is less significant.

The number of proteins identified can be maximized by an optimal choice of the two input parameters. The relation between the two parameters and the number of proteins identified is very much the same also for different FDRs. The maximal number of proteins was found for a fragment mass tolerance of 0.1 Da and precursor mass tolerance of 6 ppm. We found a maximum of 680 proteins for a FDR = 0.1%, 840 proteins for FDR = 1% and 1010 proteins for FDR = 5%.

4.5.3 *Post-translational Protein Modifications*

The mechanism of Post Translational Modification (PTM) is known to play a key role in many biological processes. Examination of PTMs is critical for understanding mechanisms of these processes. Furthermore, some modifications are artifacts of sample processing. Several technical limitations hamper detection of PTM using MS/MS[2]: First, PTMs are often present at low concentrations and due to low sensitivity of the mass spectrometer and high dynamic range of proteins, the corresponding peptides may not be detected. To overcome this obstacle especially for detection of phosphorylations many enrichment strategies have been developed. Second, some modifications hinder enzymatic protein cleavage. This leads to long peptides which are difficult to detect due to the limited number of missed cleavages used by protein identification algorithms. Third, some modifications are known to induce unexpected fragmentation patterns which are difficult to interpret[2].

Database search tools such as MASCOT or X!Tandem can only screen a limited list of predefined PTMs. X!Tandem allows one modification for each amino acid residue. If a residue is listed multiple times, X!Tandem will use the last instance of the residue to set the modification (see X!Tandem API). The number of potential peptide candidates during database search grows exponentially with the number of modifications. E.g. considering variable phosphate modifications at any serine, threonine or tyrosine increases the effective search space ~15-fold[173]. Figure 35 shows the empirical CPU time needed for database search

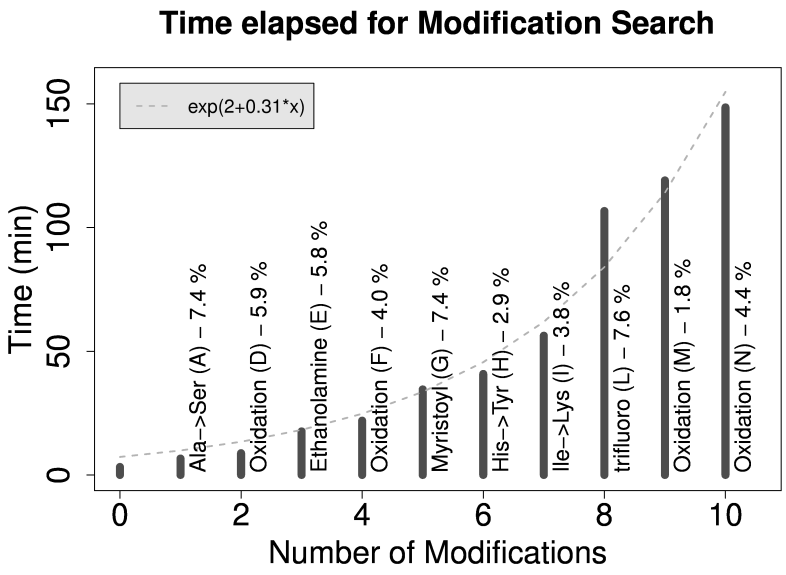


Figure 35: Time elapsed (in seconds) for database search including different number of variable modifications (from 0 - 10). Values are averaged over 5 replications (time variation was very small - 0.01%) An exponential function was fitted to the data (see legend). Next to every bar, the modification that was added including the frequency of the corresponding amino acid in vertebrates is stated [1].

with an increasing number of variable modifications (Intel(R) Core(TM)2 Duo CPU E8400 3.00GHz). As expected the time increases exponentially (see fitted exponential function) but the increase depends on the frequency of the corresponding amino acid. Searching for modification of a very frequent amino acid lysine (at position 8 in Figure 35) increases the time by almost 90% while adding less frequent amino acid methionine (at position 9) increases time by only 12%.

UniMod[35] - a comprehensive database of protein modification relevant to MS - currently lists almost 900 modifications. This includes biological enzymatic modifications, modifications due to sample preparation and modifications used for quantitation experiments such as iTRAQ. Searching for all of these modification is not possible with standard tools (such as MASCOT or X!Tandem) and also not necessary because the majority is neither biologically nor technically relevant. Due to this, we are aiming at the identification of relevant modifications see Section 4.4.8.

Figure 36 shows a histogram of modification score (ratios between peptides from forward and reverse database) for all modifications. The most frequent modification with the best score was oxidation of methionine which increased the number of peptides by almost 10%. Oxidation of methionine, whose impact on iTRAQ has been reported previously[153], can be caused by an enzymatic reaction but can also be due to sample preparation in the presence of reactive oxygen species. Apart from methionine, oxidation of two other amino acids: aspartic acid and asparagine are among the three most relevant modifications. Furthermore the deamination of glutamine and the substitution of glutamine to glutamic acid which are basically the same modification were found with good scores.

4.5.4 *Peptide E-Value distribution*

The Expected Value (E-value) for a certain peptide is the estimated probability that the peptide identification was incorrect (see Section 4.3.1 for E-value calculation). The decoy database (reverse database) allows for assessing peptide E-values since peptides from the decoy database are all random hits. Figure 37 shows the distribution of peptide E-values for the normal database (forward database) and the decoy database as a function of peptide length.

For small peptides (length 5-6) the number of peptides and the E-value distributions of forward and reverse databases are rather similar. Only a very small number of peptides show good E-values. With increasing peptide length we can see two main effects: First, the E-value distribution for the forward database is shifted to the left, while the reverse database distribution is unchanged. This implies that longer peptides show an increased

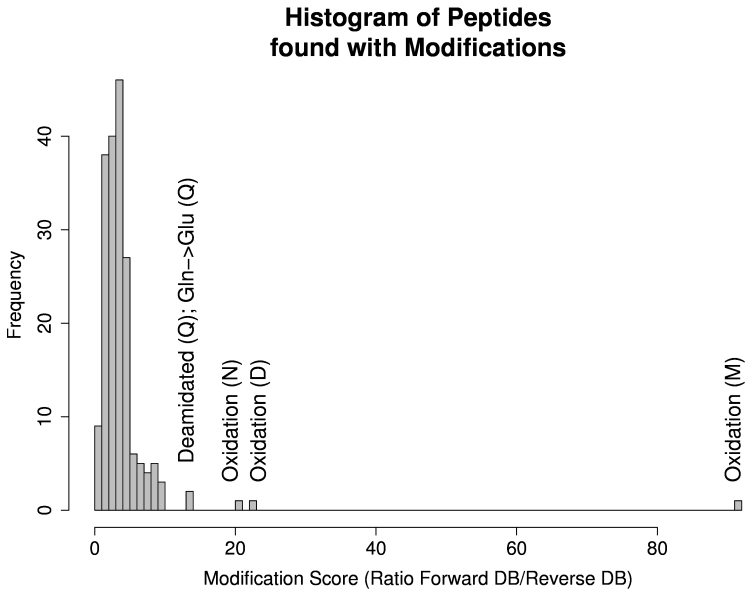


Figure 36: Histogram for modification score (ratios between peptides from forward and reverse database) for all modifications. The five modifications with the highest ratios are labeled.

proportion of significant E-values for the forward database. And second, the ratio between number of peptides from forward database and reverse database increases. The observed E-value distribution of the forward or reverse database does not change with an increasing number of modifications.

In effect, the E-value of X!Tandem allows a reliable assessment of randomness since peptides from the decoy database are characterized by low E-values. The E-value, however, does not fully comply with the assumed False Discovery Rate. Actually, an E-value of 0.1 equals a FDR of 10%. We found a total of 2752 peptides from decoy database, 56 of which have an E-value of < 0.1. This corresponds to a FDR of 2% instead of the expected value of 10%.

Especially short peptides (with length < 7) are often random hits. Only a small number (17%) of peptides from the forward database shows an E-value of < 0.1.

For MASCOT, the E-value distribution is different compared to X!Tandem especially for short peptides (see Figure 38). For small peptides (5 or 6 amino acids) 947 peptides from reverse database were identified with Mascot. 420 of these 947 peptides (44%) have an E-value < 0.1 which corresponds to a FDR of almost 50%. Apparently, the assessment of small peptides by Mascot leads to a high amount of false positive hits. The FDR decreases with increasing peptide length (16% for peptides with 7 or 8 residues and 5% for peptides with 9 or 10 residues).

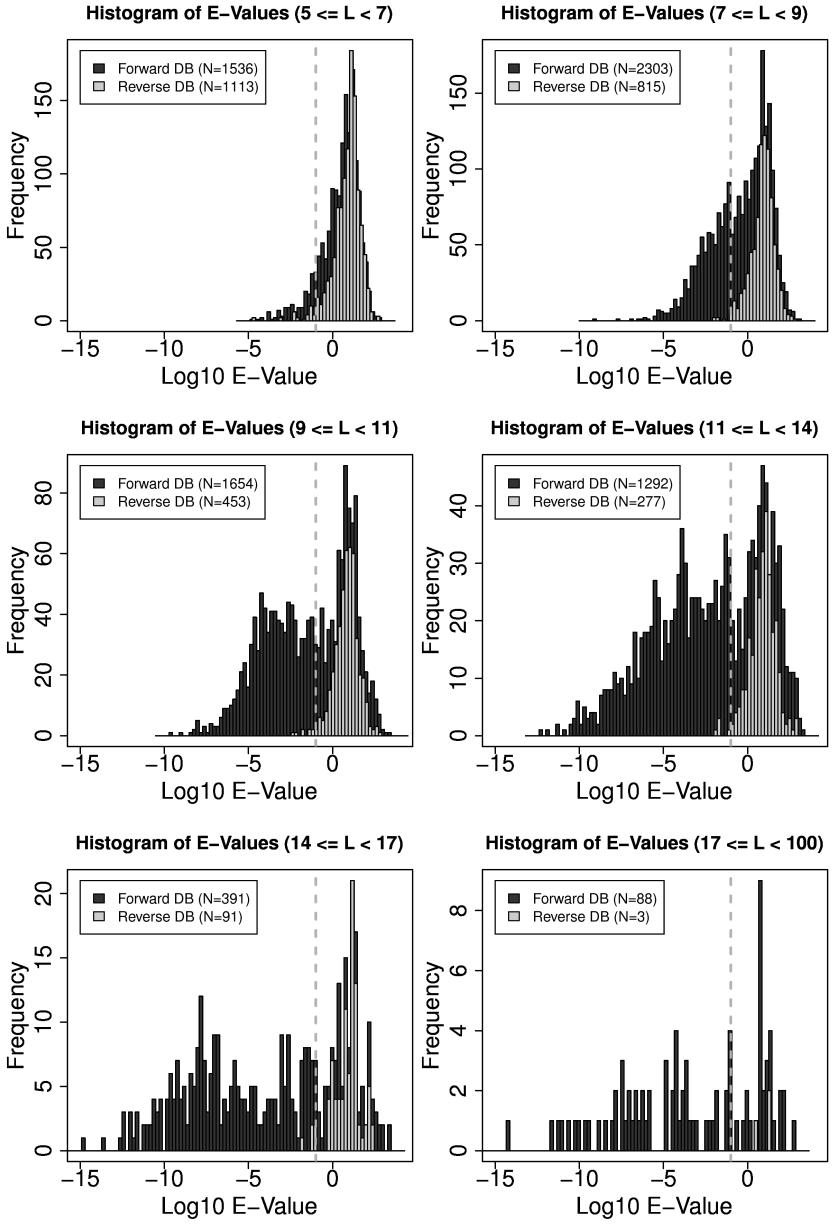


Figure 37: Histograms of E-values for peptides in the normal database (gray) and decoy database (red) identified with X!Tandem. Different panels refer to peptides with different peptide length. Vertical dashed line indicates the E-value threshold of 0.1.

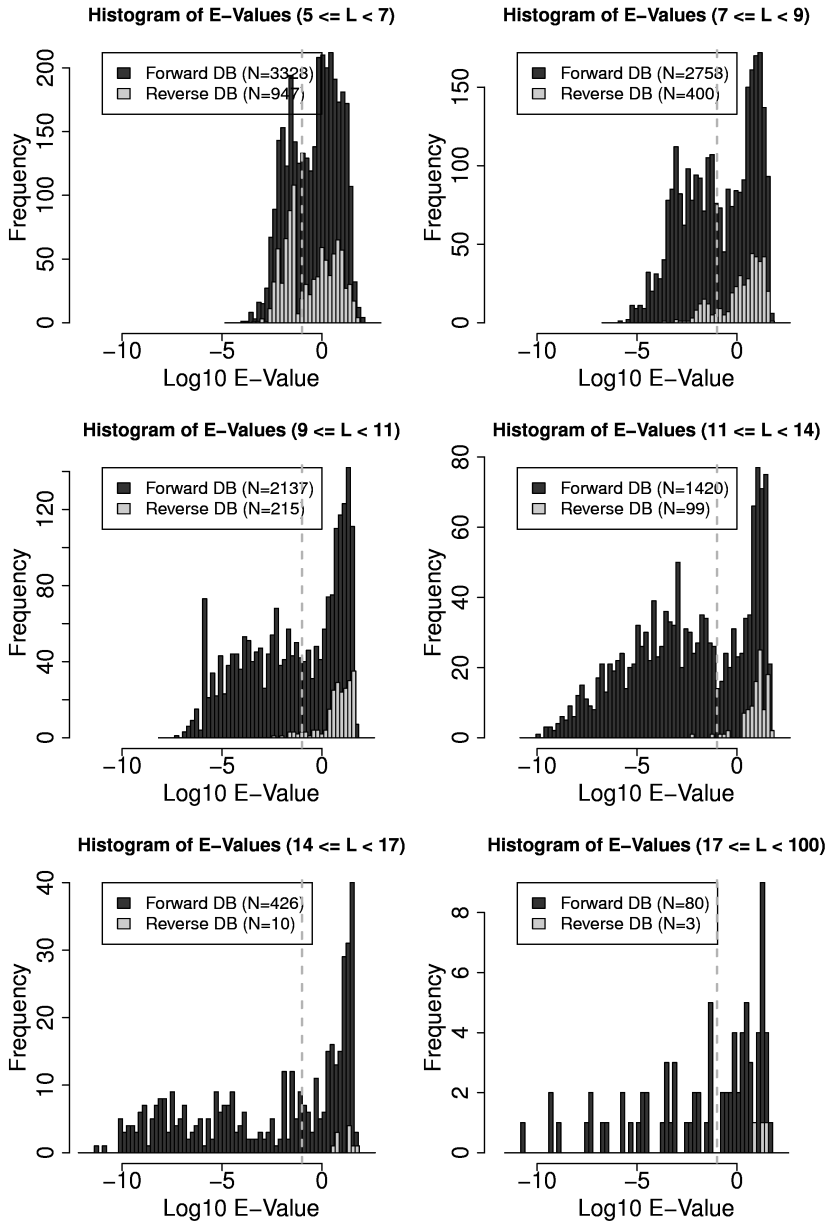


Figure 38: Histograms of E-values for peptides in the normal database (gray) and decoy database (red) identified with Mascot. Different panels refer to peptides with different peptide length. Vertical dashed line indicates the E-value threshold of 0.1.

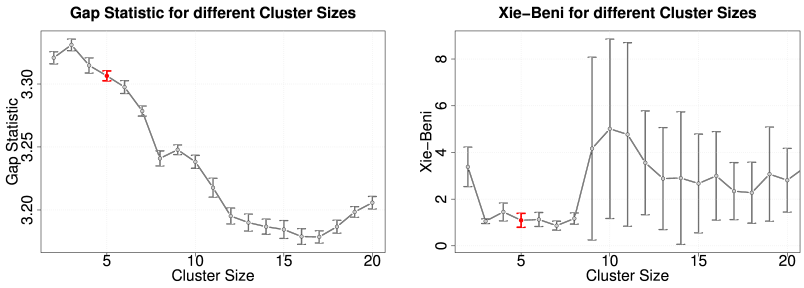


Figure 39: Evaluation of parameter k (number of clusters) with two different internal measures: Gap Statistics (left hand side) and Xie-Beni Index (right hand side). The preferred number of groups k = 5 is marked in red.

4.5.5 *Clustering*

4.5.5.1 *Number of Clusters*

For determining the preferred number of clusters we utilized Gap Statistics and Xie-Beni index as two internal cluster measures (see Section 4.4.6). The results for both measures are shown in Figure 39. High values for Gap Statistics and low values for Xie-Beni index are obtained for values of k between 3 and 7. Although the highest Gap Statistics was obtained for $k = 3$, we choose $k = 5$ as the preferred number of clusters. This is because we think that three clusters are not sufficient to to group the spectra according to their quantitation profile adequately. Five seems to be a reasonable number of clusters and still show high values for Gap Statistics and low values for Xie-Beni index.

4.5.5.2 *Cluster Stability Analysis*

Cluster stability analysis was done for 1000 cluster replications with randomly chosen starting points (see Section 4.4.6.1 for a description of cluster stability analysis). The clustering for $k = 5$ was found to be very stable since $> 99\%$ of the data points are assembled in the same group structure for all replications. This high stability was also found for values of $k = 3$ and $k = 4$. Stability decreases for bigger values of k.

4.5.5.3 *Clustering results*

In this section we demonstrate the result of the clustering which is part of PPINGUIN. This demonstration is restricted to the first of the three experimental replications (Exp 1). Figure 40 shows the ratio profiles of spectra assigned to the five clusters. Spectra aggregated within each cluster show similar ratio profiles. E.g. cluster 5 shows higher quantitation ratios for SJL_HF/SJL_SD ratio while cluster 3 contains spectra whose three ratios are ~ 0. Cluster

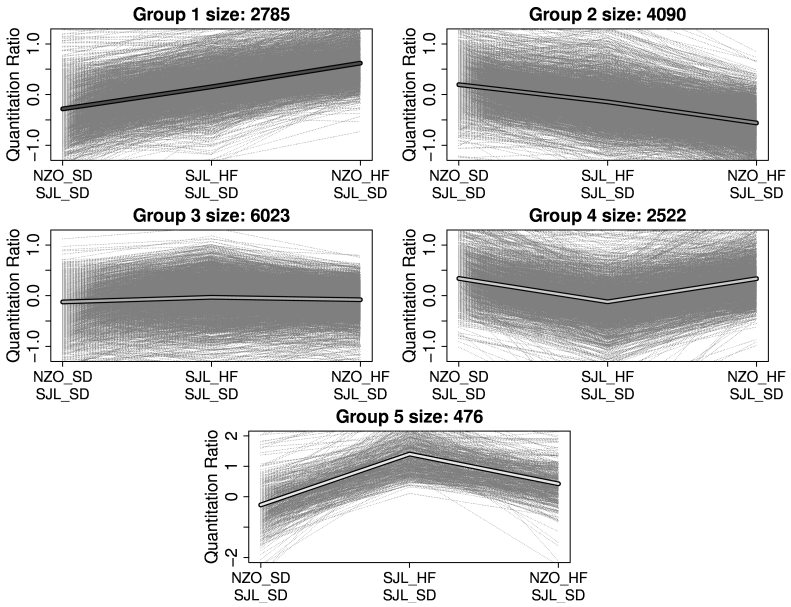


Figure 40: Clustering of the quantitation ratios of all spectra in Exp 1. Clustering was performed using k-means with k = 5. Mean profiles are drawn as bold lines. Number of spectra in each cluster are stated above the plot.

Cluster Index	1	2	3	4	5
#Spectra	2785	4090	6023	2522	476
#Peptides	88	732	429	255	7
#Proteins	6	318	152	181	38

Table 12: Sizes of the clusters created by PPINGUIN for the evaluation of Exp 1. For every cluster the number of spectra, peptides and proteins is stated.

sizes range between 476 spectra for cluster 5 and 6023 spectra for cluster 3 (see Table 12). The number of identified peptides varies from 7 for cluster 5 up to 732 for cluster 2.

The majority the proteins (77%) is identified uniquely in one cluster. 94% of all proteins are identified in one or maximal two clusters. Using PPINGUIN with random clustering (see Method section 4.4.4) less than 50% of the proteins are identified uniquely in one cluster. Clustering based on quantitation profiles employed in PPINGUIN preferably groups peptides that belong to the same protein.

4.5.6 *Proteins identified*

Proteins with stable identification (proteins identified in all three experimental replications) are of particular interest. The number of proteins identified in all three experiments with the same FDR

Method	#Proteins in Exps			Sum
	1	2	3	
MASCOT	152	133	236	521
X!Tandem/OpenMS	351	157	218	726
PPINGUIN (random clustering)	331	264	219	814
PPINGUIN	348	220	256	824

Table 13: Number of Proteins identified only in one, exactly two or all of the three experiments (Exp 1, Exp 2 and Exp 3). The last column gives to total number of proteins identified in at least one experiment (sum of the three columns). The three different methods (MASCOT, X!Tandem/OpenMS and PPINGIUN) are compared. Additionally PPINGUIN was performed a second time but with random clustering (see Method section 4.4.4).

differs for each method: 236 for MASCOT, 218 for X!Tandem and OpenMS and 256 for PPINGUIN (see 13). Ambiguous protein groups (e.g. H2B1B, H2B1C, H2B1F, ...), identified with exclusively non-unique peptides, were not counted here. The number of proteins found in all three experiments are more or less similar for all methods but in contrast the number of proteins found uniquely in a single experiment is much smaller for MASCOT. PPINGUIN shows the highest number of protein accessions found in all three experiments. Remarkably, the number of protein accessions for PPINGUIN with intensity profile clustering is much higher compared to random clustering (256 vs. 219). This hints for the practical benefit of the proposed clustering approach.

Most of the protein accessions received from X!Tandem were also detected using PPINGUIN (81% - see Venn diagram in Figure 3). The overlap between MASCOT and the other two approaches is good: 68% of the X!Tandem IDs and 58% of PPINGUIN IDs where found with MASCOT. All three methods have their set of unique proteins accessions: 54 for MASCOT, 17 for X!Tandem and 55 for PPINGUIN.

Most of these 54 unique MASCOT proteins are also found using X!Tandem but they remain below the significance threshold. This is mostly due to differences in the assessment of short peptides since MASCOT appears to include many small peptides for identification that are excluded by X!Tandem. Figure 42 shows a histogram of peptide length distribution of 54 proteins uniquely identified with Mascot, of peptides for the 55 proteins uniquely identified with PPINGUIN and of peptides for the 124 proteins found with all methods. Proteins uniquely identified with Mascot show an accumulation of short peptides (more than 25% of the peptides have only 5 residues) while proteins identified with all three methods preferably consist of longer peptides. This

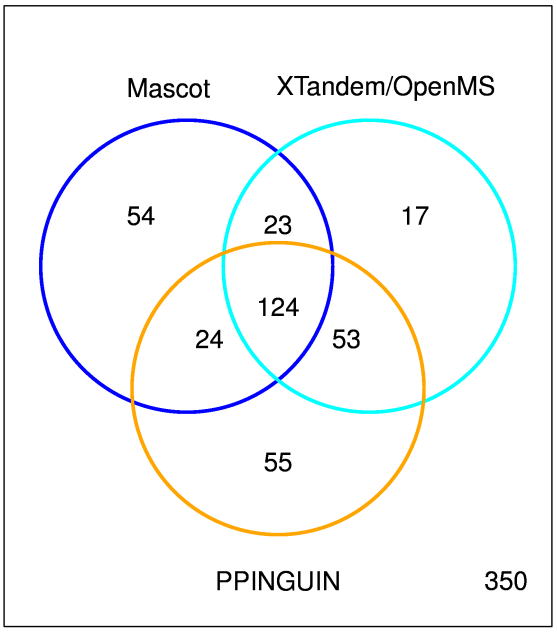


Figure 41: Overlap of protein identification of the three different approaches employed regarding proteins significantly identified within all three experiments.

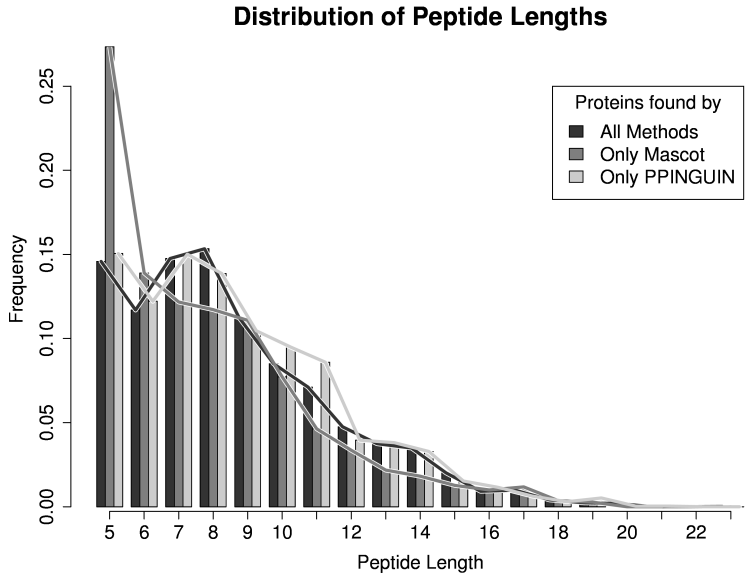


Figure 42: Histogram of peptide length distribution of 54 proteins uniquely identified with Mascot, 55 proteins uniquely found with PPINGUIN and 124 proteins found with all methods. Especially the proteins uniquely identified with Mascot show a clear accumulation of short peptides.

observation gives further support to the impression that Mascot is not reliable in assessing short peptides (see Section 4.5.4).

4.5.7 *Homogeneity of peptide profiles*

As described above, a protein represented by multiple unique peptides should result in strictly correlated quantitation ratios for these peptides.[69]. In practical situation, however, often heterogeneous ratio profiles are observed using MASCOT as well as X!Tandem, leading to difficulties in quantitative interpretation.

In the following, we compare the three methods (*i*) MASCOT, (*ii*) OpenMS and X!Tandem and (*iii*) PPINGUIN in terms of homogeneity of peptides assigned to the same protein. Homogeneity is assessed by calculating the Coefficient of Variation (CV) of the quantification ratios of all peptide belonging to a certain protein (see Section 4.4.9). For assessment of homogeneity, the first of the three experimental replications is used (cf. Table 10).

4.5.7.1 *MASCOT*

Evaluation based on MASCOT results in a CV for quantitation ratios of peptides for the same protein of 16% for 114/116 ratio, 16% for 115/116 ratio and 25% for 117/116 ratios (StDev=0.23, 0.22 and 0.36). Peptides belonging to the same protein frequently

	Ratio	MASCOT	X!Tandem OpenMS	PPINGUIN Random Cl	PPINGUIN
CV	114/116	0.16	0.16	0.14	0.13
	115/116	0.16	0.16	0.13	0.12
	117/116	0.25	0.23	0.2	0.16
$\log_2$ StDev	114/116	0.23	0.23	0.21	0.19
	115/116	0.22	0.23	0.2	0.18
	117/116	0.36	0.32	0.3	0.24

Table 14: Homogeneity of peptides averaged for all identified proteins: Variability of peptides for the same protein was calculated for the three approaches. For the three iTRAQ ratios (114/116, 115/116 and 117/116) the coefficient of variation and the standard deviation are stated. The three different methods (MASCOT, X!Tandem/OpenMS and PPINGIUN) are compared. Additionally PPINGUIN was performed a second time but with random clustering (see Method section 4.4.4).

show heterogeneous ratio profiles (see profiles for three exemplary chosen proteins in first row of 43). For example, peptides assigned to the protein *40S ribosomal protein S30 - RS30* (third column of Figure 43) show $\log_2$ quantitation ratios for NZO_HF vs. SJL_SD (117/116) ranging from -0.9 to $+1.3$, corresponding to 1.9-fold down-regulation and 2.5 up-regulation, respectively. This diversity of peptide quantitation is difficult to interpret.

4.5.7.2 *X!Tandem and OpenMS*

The second approach encompassing X!Tandem and OpenMS shows similar CV values of 16% for 114/116, 16% for 115/116 and 23% for 117/116 (StDev: 0.23, 0.23 and 0.32). Variability of peptides for the same protein is only slightly smaller compared to MASCOT (for 117/116 channel). Ratio profiles are still considerably heterogeneous and many profiles are still divergent (middle row of Figure 43). The protein *40S ribosomal protein S30 - RS30* (c.f. middle row and third column of 43) shows very similar peptide quantitations compared to MASCOT. Quantitative $\log_2$ ratios for NZO_HF vs. SJL_SD (117/116) range between -0.5 and $+1.1$ corresponding to $\sqrt{2}$ fold down-regulation and 2.3-fold up-regulation.

4.5.7.3 *PPINGUIN*

PPINGUIN shows considerably lower CV of peptides assigned to the same protein: 13% for 114/116, 12% for 115/116 and 16% for

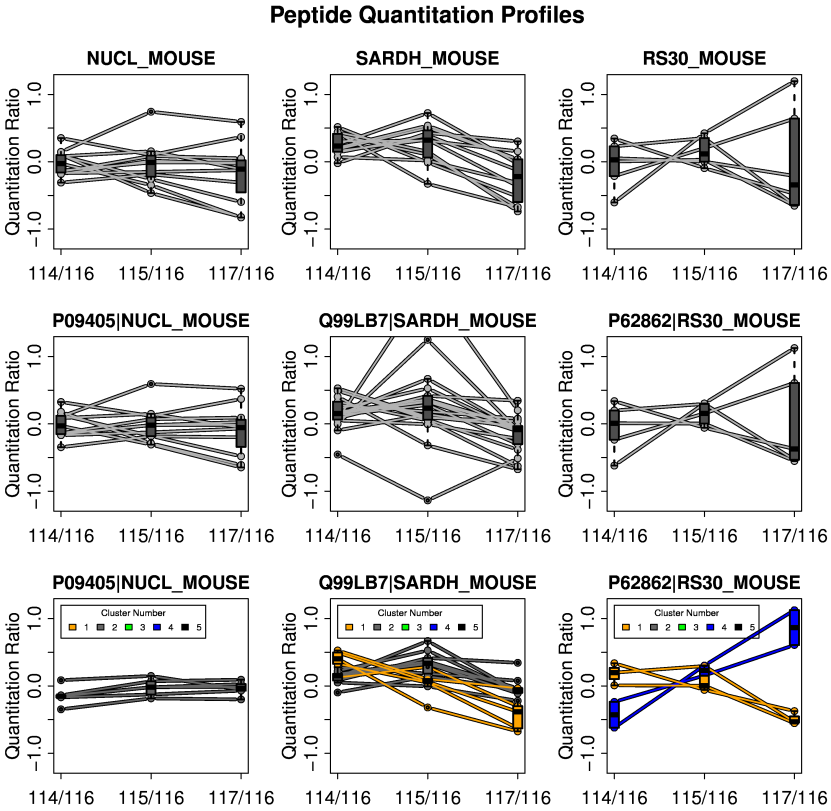


Figure 43: Visualization of peptide quantitation profiles for three different approaches employed (rows) demonstrated for 3 exemplary chosen proteins (columns). Each plot shows ratio profiles of peptides mapped to the corresponding protein. Rows correspond to the applied method: first row = MASCOT, second row = X! Tandem and OpenMS, last row = PPINGUIN. A unique feature of PPINGUIN is the separation of peptide into different groups with distinct quantitation profiles.

117/116 (StDev: 0.19, 0.18 and 0.24 - see 14). The reduced variance is an expected effect since the peptides within each cluster are similar by construction. Even for a random clustering the CV of peptides assigned to the same protein is lower compared to MASCOT and X!Tandem: 14% for 114/116, 13% for 115/116 and 20% for 117/116 (StDev: 0.21, 0.20 and 0.30 - see 14). PPINGUIN with (non-random) quantitation profiles based clustering shows lower CV of peptides assigned to the same protein compared to random clustering. The ratio profiles are more homogeneous and without outliers (bottom row 43). The resulting ratio profiles are in accordance with our assumption that peptides belonging to the same protein have similar relative quantitation measurements. A distinctive feature of PPINGUIN is demonstrated by *40S ribosomal protein S30 - RS30*: inconsistent quantitation profiles are resolved by splitting up in two groups each with homogeneous profiles. The protein is identified in two different clusters (1 and 4) with different peptide profiles. The peptides in cluster 1 show low relative concentration for NZO SD (114) and high relative concentration for NZO HF (117) while peptides in cluster 4 show the opposite behavior. This effect is described in more detail in Section 4.5.10.

For an independent assessment of our method we now proceed to investigate experimental reproducibility (precision) and accordance with prior knowledge.

4.5.8 *Precision - Experimental Reproducibility*

Experimental replicates are essential because biological systems are inherently variable. In order to test reproducibility we performed three independent experimental replications and three independent evaluations (see Dataset section: 4.2). We investigated two different mouse genotypes and two diets resulting in 4 distinct combinations. The 4 combinations define 3 ratios using SJL genotype with standard diet (SD) as reference. Quantitation ratios for a protein are calculated by averaging the log ratios of the corresponding peptides. To facilitate comparability we restrict the analysis to the set of proteins identified in all three experimental replications with each method.

We calculated CV and mean standard deviation of log quantitation ratios of all proteins (see Section 4.4.10) as popular measures for experimental reproducibility (see Table 1 for results). The analysis was performed separately for each of the 3 experimental ratios: NZO_SD/SJL_SD, NZO_HF/ SJL_SD and SJL_HF/SJL_SD.

Experimental variation of the MASCOT based evaluation is characterized by CV values ranging from 0.13 to 0.19 (see first column in Table 15). X!Tandem/OpenMS results in CV values ranging from 0.12 to 0.18 (second column in Table 15). Experimen-

	Ratio	MASCOT	X!Tandem OpenMS	PPINGUIN Random Cl	PPINGUIN
CV	NZO_SD/SJL_SD	0.13	0.12	0.12	0.10
	NZO_HFD/SJL_SD	0.16	0.16	0.13	0.13
	SJL_HFD/SJL_SD	0.19	0.18	0.14	0.14
$\log_2$ StDev	NZO_SD/SJL_SD	0.18	0.17	0.16	0.14
	NZO_HFD/SJL_SD	0.22	0.22	0.20	0.19
	SJL_HFD/SJL_SD	0.26	0.25	0.21	0.21

Table 15: Experimental reproducibility using the three different approaches. For the three experimental factors (NZO_SD/SJL_SD, NZO_HF/SJL_SD and SJL_HF/SJL_SD) the mean CV and the mean standard deviation for quantitation ratios of all proteins are stated. The three different methods (MASCOT, X!Tandem/OpenMS and PPINGIUN) are compared. Additionally PPINGUIN was performed a second time but with random clustering (see Method section 4.4.4).

tal variation is reduced using PPINGUIN with CV values ranging from 0.10 to 0.14 (fourth column in Table 15). The error of PPINGUIN is rather similar compared to random clustering. But the number of proteins found in all three experiments is 15% higher compared to random clustering (c.f. Table 13).

Different from the improved homogeneity in the previous section, the lower error of PPINGUIN is not a trivial effect since the complete analysis workflow is performed for every experiment independently. These results demonstrate that applying the proposed method for data evaluation leads to more stable quantitation values compared to Mascot and X!Tandem.

4.5.9 *Accordance with prior knowledge*

Typically, a primary goal of quantitative proteomics is the identification of differentially expressed proteins. In contrast to technical aspects in previous sections, we now identify differentially expressed biomarker candidates based on evaluation with the three different methods. To assess the results of the differential analysis, we use a reference set of gold standard genes identified in the context of T2DM[131]. This meta-analysis reports top genes candidates for mixture of genotypic and dietary effects based on different transcriptomics experiments. To assure comparability with the meta-analysis, differential analysis is performed comparing NZO mice with HF diet and SJL mouse with SD diet.

Identification of differentially expressed proteins is performed as described in Section 4.4.7. Due to the low number of replicates we use the fold instead of the p-value as criterion to judge differential expression. The fold change value of 0.5 was chosen

Protein ID	Description	log_2 Fold	P-Value	#Peptides	XTandem Score
O35490	betaine-homocysteine methyltransferase	-0.922	0.00552	33	105.1
P33267	cytochrome P450, family 2	-0.853	0.132	4	16.2
P97872	flavin containing monooxygenase 5	-0.71	0.0891	3	21
Q91V92	ATP citrate lyase	0.694	0.257	5	15.4
P01942	Hemoglobin subunit alpha	0.68	0.358	13	24.6
Q9Z2V4	phosphoenolpyruvate carboxykinase	-0.663	0.0824	4	10.3
Q8VCN5	cystathionase	-0.615	0.0121	3	8.5
Q9CPY7	leucine aminopeptidase 3	-0.6	0.0955	4	28.1
P70694	aldo-keto reductase	-0.589	0.0874	9	32.7
Q01279	epidermal growth factor receptor	-0.587	0.233	3	22.6
P10649	glutathione S-transferase	-0.581	0.125	9	20.7
Q5SWU9	acetyl-Coenzyme A carboxylase alpha	0.573	0.092	4	10.8

Table 16: Differentially expressed proteins comparing for NZO_HF/SJL_SD ratio are shown. Proteins highlighted in lightgrey color have previously been reported to be associated with obesity and T2DM.[131]

as threshold since it results in a reasonable number of differentially expressed proteins. For higher values the number of genes decreases fast and no gene has an absolute log_2 fold > 1 - see Table 16). For smaller threshold values the number of randomly detected genes increases (increase in false positive rate).

Evaluation based on Mascot identifies a total of 10 differentially regulated proteins of which 20% (2) are found in the reference. Using X!Tandem and Open MS identifies only 8 differential proteins of which 37% (3) are found in the reference set. PPINGUIN results in 12 differentially expressed proteins, of which 42% (5) are part of the reference set.

Of the three methods, PPINGUIN shows the highest agreement with the reference list. This remains true for changes of the threshold value (e.g. 0.4 or 0.7).

Figure 44 shows the volcano plot for NZO_HF/SJL_SD ratio evaluated using PPINGUIN. None of the proteins has a significant p-value (the Bonferroni corrected threshold of 0.0003 is below the scale) but some proteins show fold changes (up to 1 corresponding to a 2-fold differential expression). Table 16 presents the statistics of the differentially regulated proteins identified using PPINGUIN (proteins of the reference set are colored in lightgrey).

4.5.10 *Detecting Potential Protein Isoforms*

Protein isoforms and especially Post Translational Modification (PTM) play a key role in many biological processes. Examination of isoforms and PTMs is critical for understanding mechanisms of these processes. Different protein isoforms are often regulated separately and hence show distinct quantitation profiles.

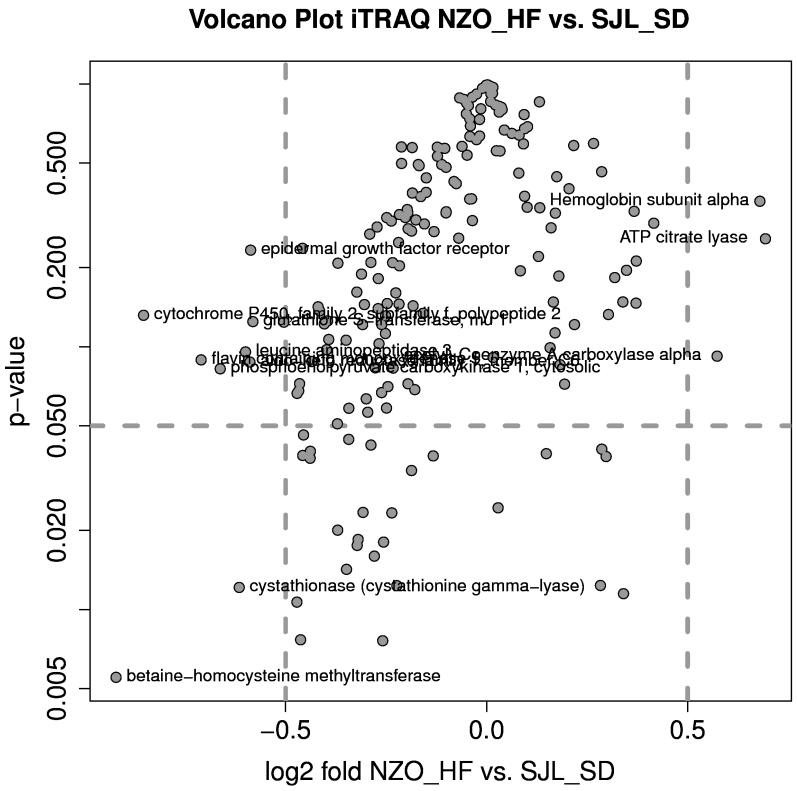


Figure 44: Volcano Plot for NZO_HF/SJL_SD. Horizontal dashed lines indicate the 0.05 significance threshold. Vertical dashed lines indicate fold threshold of 0.5 used for definition of differential expression (see table 16).

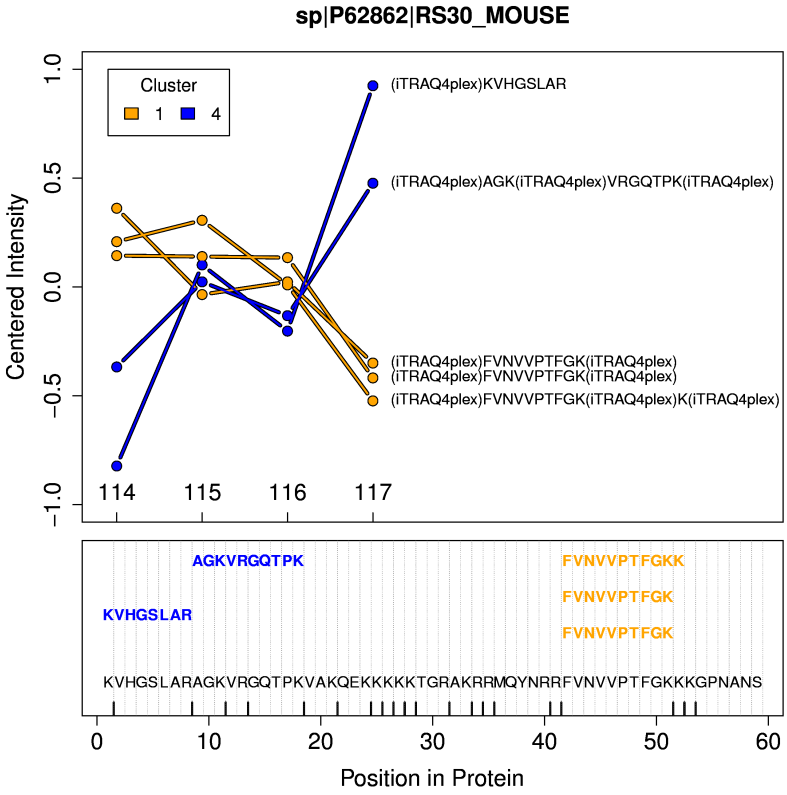


Figure 45: Upper part: Quantitation profile of the peptides assigned to the ribosomal protein RS30 detected in the first experiment. Labels are representing the samples: 114 - NZO_SD; 115 - SJL_HF; 116 - SJL_SD and 117 - NZO_HF. Colors orange and blue correspond to clusters 1 and 4 the peptides were identified in. Lower part: Protein sequence with positions of mapped peptides. 'Upward ticks' on the x-axis indicate predicted trypsin cleavage sites.

In the following we want to identify potential isoforms with PPINGUIN. A key feature of our approach is the separation of different peptide profiles for the same protein in multiple clusters. The peptides in each cluster exhibit distinct quantitation profiles which may correspond to protein isoforms. Potential protein isoforms can be detected by searching for proteins identified in multiple clusters. The majority of the proteins is found uniquely in a single cluster (77%) and does not show evidence for protein isoforms.

Figure 45 and 46 show quantitation profiles of two ribosomal proteins: 40S ribosomal proteins S30 and 60S ribosomal protein L29. Both proteins are identified in two clusters (1 and 4) with distinct quantitation profiles. These different profiles are probably due to protein isoforms. These isoforms are regulated differentially (~ 2-fold) for NZO mouse (channels 114 and 117). The similar behavior of two ribosomal proteins located on distinct

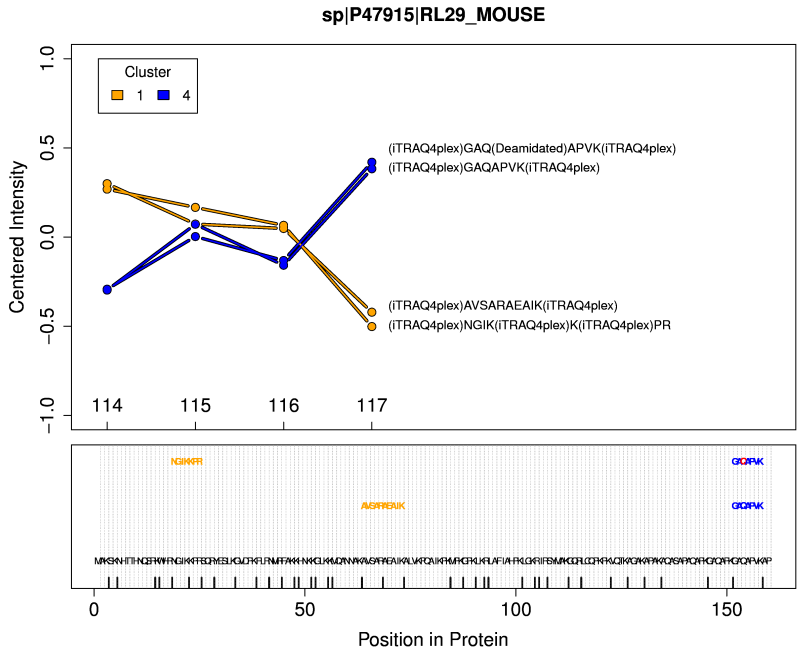


Figure 46: Upper part: Quantitation profile of the peptides assigned to the ribosomal protein RL29 detected in the first experiment. Labels are representing the samples: 114 - NZO_SD; 115 - SJL_HF; 116 - SJL_SD and 117 - NZO_HF. Colors orange and blue correspond to clusters 1 and 4 the peptides were identified in. Lower part: Protein sequence with positions of mapped peptides. 'Upward ticks' on the x-axis indicate predicted trypsin cleavage sites. Red colored residues indicate found modification.

sites suggests a possible biological mechanism and supports the rational of our procedure.

Typical reasons for isoforms are protein modifications, splice variants or degradation effects. For further investigation of PTMs as a source for protein isoforms, protein identification was performed anew, searching for the 20 most frequent modifications (the most frequent for each residue - see section 4.5.3). As for RS30 protein, we did not find further evidence for protein modification. But RL29 protein shows one peptide with a modification: deamidation of glutamine. This modification affects only one of the two peptides from cluster 4 but gives further support to the theory that protein modification is a possible reason for protein isoforms.

Investigating splice variants as a possible explanation for protein isoforms, we found that RS30 protein is transcribed from exon 4 and 5 of the FAU (Ensembl-ID: ENSMUSG00000038274) gene. The peptides from different clusters are located in different regions of the protein which also correspond to the different exons of the FAU gene, but there was no indication for differential splicing in the database. The positions of the peptides make both, degradation effects or splice variants possible explanations for the observed RS30 isoforms. However, the FAU gene may have two variants: the RS_30 protein with 59 amino acids and the completely transcribed protein with 133 amino acids. PPINGUIN finds two variants of the RS_30 gene. These two variants may correspond to the two potential variants, which of cause would require further investigation.

The RL29 protein is transcribed from Ensembl gene: ENSMUSG00000048758. Two different transcripts are annotated for this gene. Again, the peptides (of RL29 protein) from different clusters are located in different regions of the protein. For the RL29 isoforms all three phenomena: protein modification, splice variants and degradation effects could be possible reason for the different isoforms.

4.5.11 *Non-unique proteins*

In some cases peptides are not unique for a single protein but instead could possibly be derived from several proteins. Typically these peptides are removed because the originating protein can not be determined. The clustering used in PPINGUIN may help to resolve these ambiguities.

Figure 47 shows four exemplary cases of proteins with non-unique and unique peptides. The first row contains peptides assigned to proteins: `Q5SWU9-1`, `Q5SWU9-2` and `Q99MR8`. Three peptides are uniquely assigned to a single protein: one peptide to protein `Q5SWU9-1` in cluster 1 and two peptides to protein `Q5SWU9-2`

in cluster 2. Additionally to the unique peptides, one non-unique peptide (dashed grey line) was found that could be derived from three different proteins. The non-unique peptide shows a similar profile like the peptides derived from `Q5SWU9-2` and are found in the same cluster (cluster 2). The clustering suggests that the non-unique peptide belongs to protein `Q5SWU9-2`.

The second row (of Figure 47) shows peptides assigned to proteins `P10126` and `P62631`. Four peptides are found in cluster 3, two of which are uniquely identified as protein `P10126`. The other two non-unique peptides from cluster 3 are probably also derived from protein `P10126`. The other six non-unique peptides from cluster 2 show a different quantitation profile and may rather be derived from protein `P62631`.

The last two rows contain proteins with a single uniquely identified peptide. In addition, non-unique peptides with very similar profiles are found (in the same cluster). The similarity suggests that the non-unique peptides belong to the same protein as the uniquely identified peptide.

Of cause there are still many ambiguous non-unique peptides that can not be assigned to a single protein. At all, protein inference leads to a total of 88 groups of non-unique proteins. For more than 52% of these non-unique peptides PPINGUIN does not provide any further information regarding ambiguous identification. This is because these proteins are represented only by non-unique peptides and do not have any unique peptide that can be used to resolve ambiguities. However, PPINGUIN may help to resolve ambiguities for almost 50% of the non-unique proteins.

4.5.12 *Comparison with Genomics*

PPINGUIN is based on the assumption that peptides derived from the same proteins should be highly correlated. This assumption is even more relevant in genomics or transcriptomics with microarray technology. In both fields (transcriptomics and proteomics), measurements are performed at the level of features which belong to a common superior structure: In proteomics, quantitative measurements are typically on the scale of single peptides belonging to proteins while for transcriptomics microarrays measurements are performed on the scale of oligonucleotides which belong to genes. In order to quantify the superior structure (genes or proteins), a summarization strategy is required. The summarization strategy, always assumes the summarized features (peptides, oligonucleotides) show more or less similar quantitation values. In the following, we will investigate whether the assumption that biomolecules derived from a common source should be highly correlated holds true for microarray based experiments.

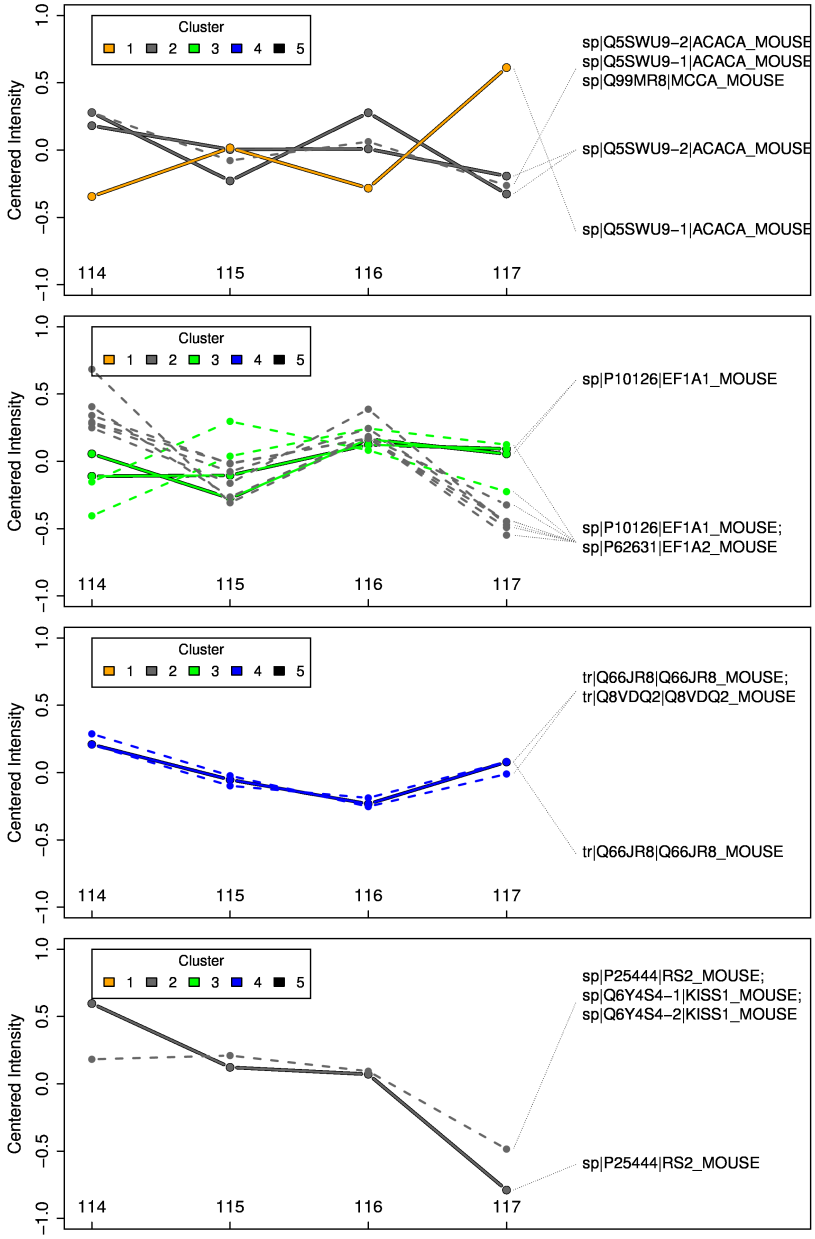


Figure 47: Four exemplary chosen proteins (rows) with non-unique peptides (dashed lines) as well as unique peptides (bold lines). Peptide identification is given right hand side of the plot. Non-unique peptides are characterized by multiple possible proteins assigned. Cluster assignment may help to unravel the ambiguities: E.g. the first row (proteins: Q5SWU9-1, Q5SWU9-2 and Q99MR8) has three unique peptides one assigned to protein Q5SWU9-1 (in cluster 1) and two assigned to Q5SWU9-2 (in cluster 3). The clustering suggests that the non-unique peptide (dashed grey line) belongs to protein Q5SWU9-2.

In order to optimize the comparability of the oligonucletide-based microarray analysis and iTRAQ data evaluation, we are interested in a microarray study with similar experimental design. The key properties of iTRAQ technology is relative quantification that requires a reference channel. Therefore, a good study candidate employs two color microarrays, with a common pooled reference design. We searched NCBI Gene Expression Omnibus (GEO) database[9] for a suitable study.

4.5.12.1 *Microarray Study*

In 2009 Tombol et al.[160] published an oligonucleotide microarray study investigating four groups of human adrenocortical tumors: normal cortex, inactive adenoma, cortisol secreting adenoma, adrenocortical cancer (GEO DataSet ID: `GSE14922`). The experiment was performed using a two-color microarray designed to compare individual samples to a uniform reference pool of all samples. The hybridization was done using Agilent Whole Human Genome 4x44 K microarray (GEO Platform ID: `GPL6480`). In total, the study by Tombol et al. comprises 16 microarrays - 4 biological replicates for each of the 4 experimental groups.

Thinking in terms of iTRAQ design, the microarry experiment corresponds to a hypothetical '5-plex iTRAQ' study resulting in 4 quantitative ratios. The microarray images where processed using Agilent Feature Extraction Software 8.5 and original array normalization was chosen as default normalization scenario for Agilent 4x44 K two-colour array platforms[160].

The normalized logarithmic ratio data was downloaded from GEO as *Series Matrix File*. The average ratio of all oligonucleotides on a chip is 0 for all chips (see box plot of all 16 chips in Figure 48). In order to validate the assumption that biomolecules derived from a common source should be highly correlated, we investigate the homogeneity of ratio profiles of oligonucleotides for the same gene in a similar way as done for homogeneity of peptides (section 4.5.7).

4.5.12.2 *Homogeneity of Oligonucleotide Profiles*

To evaluate homogeneity of oligonucleotides, we follow similar strategy as applied for homogeneity of peptides (section 4.5.7). The 44K Aligent probe IDs, are mapped to the corresponding ENSEMBL gene IDs using Bioconductor annotation package `hgug4112a.db`. If multiple Aligent IDs are mapped to the same ENSEMBL gene, we expect these probes to show correlated profiles as they represent the same ENSEMBL gene. For the 4 ENSEMBL genes with the highest number of Agilent IDs assigned, the ratio profiles are shown in figure 49. For calculation of ratio profiles the experimental replications were averaged. The ratio

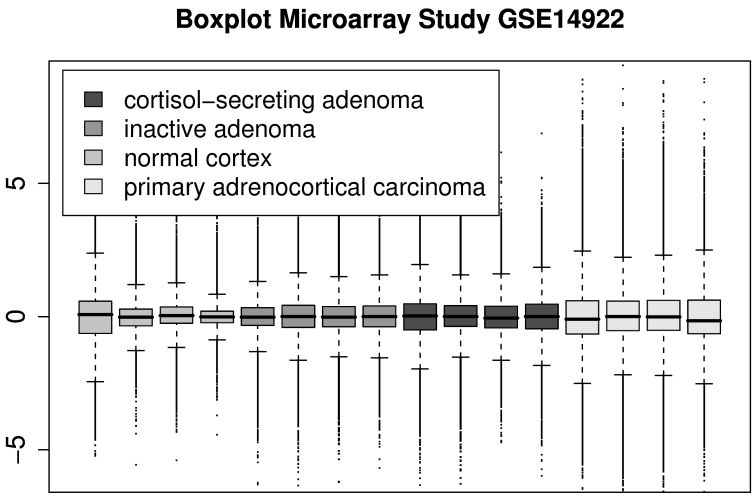


Figure 48: Box plot of all normalized signals from 16 microarrays of the GSE14922 study downloaded from GEO. The 4 different experimental groups are colored with shades of gray. All ratios are distributed around 0.

profiles are partly heterogeneous and do show only medium correlation (average correlation coefficient of all corresponding genes = 0.5). Overall CV of Agilent IDs mapped to the same ENSEMBL ID is 22% and therefore, comparable to homogeneity of peptides (see section 4.5.7).

The heterogeneity of oligos mapped to the same ENSEMBL ID complicates the assignment of quantitative values for single genes. The biological entity measured by oligo nucleotide microarrays is mRNA. A single oligo nucleotide typically represents a single mRNA whereas a single (ENSEMBL) gene may be transcribed to several different mRNAs due to alternative splicing. The number of alternatively spliced genes was conservatively estimated to be $40-60\%$[105, 26] but more recent estimates suggest that $> 92-94\%$ of human genes undergo alternative splicing[168]. The different mRNAs and resulting proteins produced by alternative processing often differ in structure, function, localization and regulation. Considering these facts, the question arises whether (ENSEMBL) gene IDs should be considered as the common sources for oligo nucleotides. Or if, actually, (ENSEMBL) transcripts IDs are a more adequate common source for the probe IDs.

Following this idea, all 44K Aligent gene IDs where mapped to the corresponding ENSEMBL transcripts (mapping tables are included in the dataset). Similar to previous analysis, for the 4 ENSEMBL transcripts with the highest number of Agilent

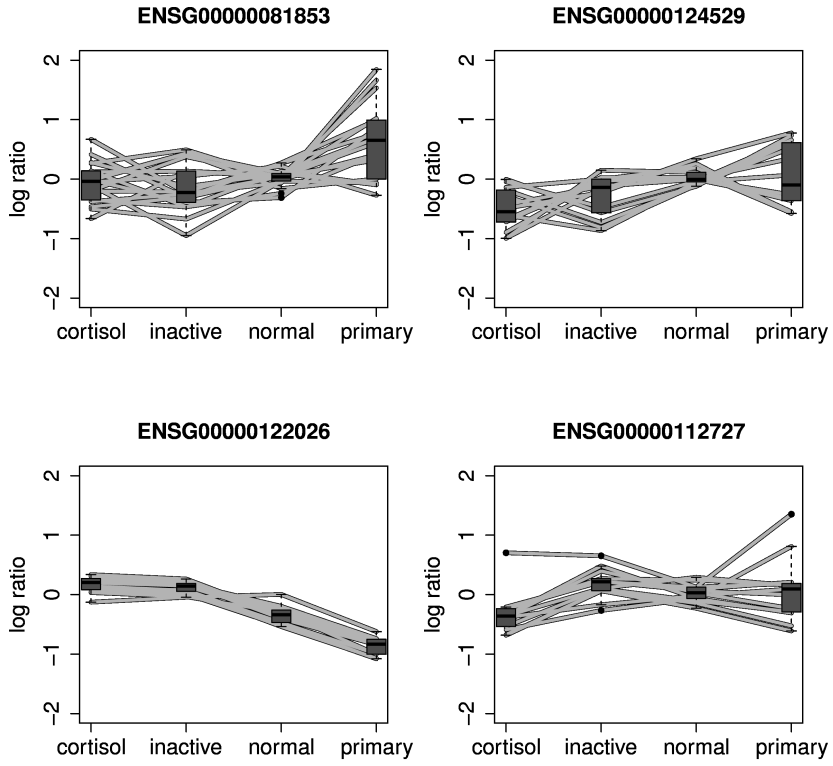


Figure 49: Ratio profiles of the 4 ENSEMBL gene with the highest number of mapped Agilent IDs. Every plot represents one ENSEMBL gene, and every line in a plot reflects one Agilent gene mapped to the corresponding ENSEMBL IDs.

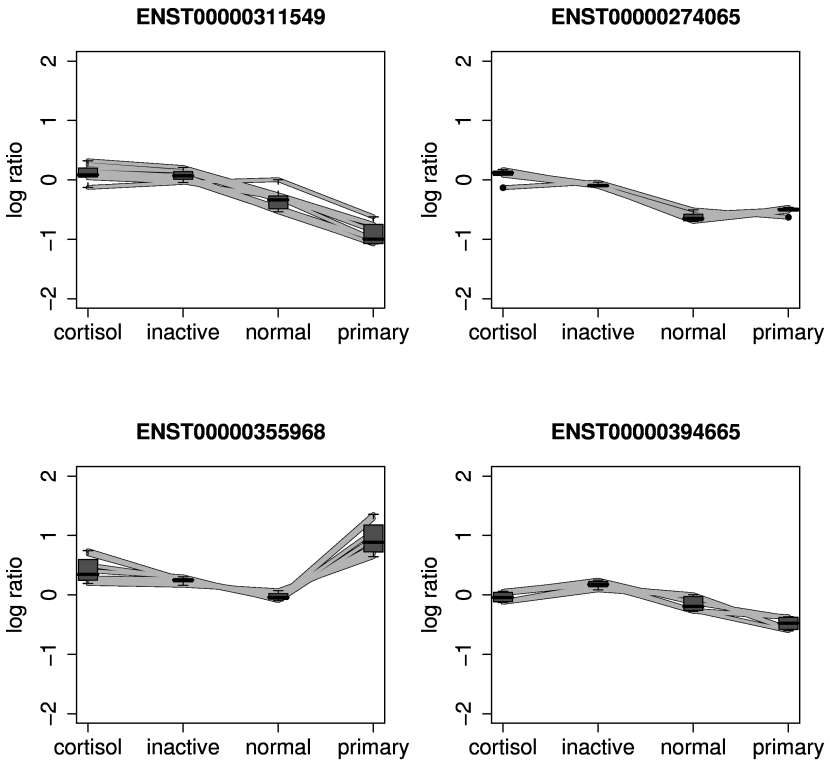


Figure 50: Ratio profiles of the 4 ENSEMBL transcripts with the highest number of mapped Agilent IDs. Every plot represents one ENSEMBL transcript, and every line in a plot reflects one Agilent gene mapped to the corresponding ENSEMBL IDs.

IDs, the ratio profiles are shown in figure 50. In comparison with figure 49, the profiles for a common transcript are more homogeneous which is also confirmed by an increased average correlation coefficient of all transcripts (0.67 vs. 0.5). The overall CV of all agilent IDs for the same ENSEMBL transcript is lower in comparison with ENSEMBL genes (18% vs. 22%).

The results presented suggest that biomolecules derived from a common source show similar profiles in the case of oligo nucleotides and transcripts.

4.5.12.3 *PPINGUIN - for Transcriptomics*

The motivation of PPINGUIN is that peptides (represented by spectra) belonging to the same protein show correlated quantitation profiles. For transcriptomics microarray technology we have demonstrated that mRNA probes belonging to the same transcript often show correlated quantitations. If the clustering employed in PPINGUIN indeed groups elements belonging to the same superstructure, than the mRNA probes belonging to the same transcript should be assigned to the same cluster. In con-

sequence if a gene is found in multiple clusters, the different clusters may correspond to different transcripts of the gene. We now investigate whether these hypotheses are true for transcriptomics data.

To maximize the comparability with the implementation of PPINGUIN, we do not use the normalized data provided by Tombol et al.. Instead we downloaded the raw data of the study from GEO (see Section 4.5.12.1). For performing the clustering we followed the implementation of PPINGUIN: The raw data was normalized using multi-Lowess normalization. For each gene, a 5-tupel intensity profile is calculated representing the average of the four samples and the reference. The intensity profile is centered (without scaling) and subsequently k-means clustering ($k = 5$) is applied to the intensity profiles using Euclidean distance.

For the 214 ENSEMBL transcript IDs that are represented by more than five agilent mRNA probe IDs, we check whether they are pooled in one cluster or spread across multiple clusters. 159 (75%) of these transcripts are indeed pooled in a single cluster. This proportion corresponds surprisingly well to the number of proteins found in a single cluster (which was 77% - see Section 4.5.5.3) If the hypothesis of correlated quantitation profiles is correct (which holds true for transcript mRNA oligos) than the clustering employed in PPINGUIN groups together what belongs together.

Figure 51 shows two exemplary genes ENSG00000050165 - DKK3 and ENSG00000108187 - PBLD that are found in two different clusters. Both genes are measured by mRNA oligos showing two different groups of 5-tupel intensity profiles. For some oligos a mapping to the corresponding transcript is not annotated although a gene id is given. The mRNA oligos belonging to the same group are assigned to the same transcript if the transcript is annotated. So the different groups reflect different isoforms of the corresponding gene (in this case different transcriptions). The hypothesis that oligo nucleotides which are found in different clusters reflect different isoforms, is true for these two exemplarily chosen genes. If the same is true for the analysis of iTRAQ data, than proteins which are identified by PPINGUIN in different clusters correspond to different protein isoforms (see section 4.5.10).

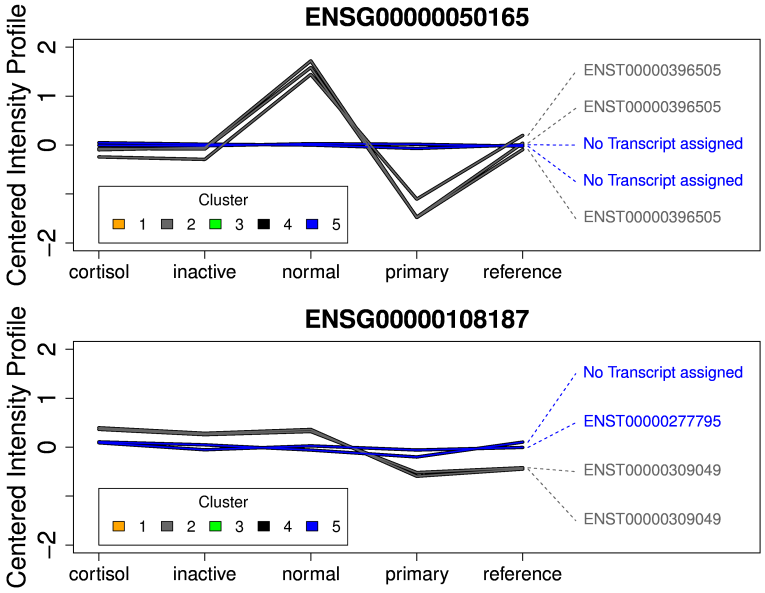


Figure 51: Two Genes (ENSG00000050165 - DKK3 and ENSG00000108187 - PBLD) that are split into two different clusters. For each gene, the centered 5-tupel intensity profile is (used for clustering) is shown. The different clusters are color coded. On the right hand side the corresponding ENSEMBL transcripts are given.

4.6 SUMMARY AND DISCUSSION: ITRAQ

During protein identification (with MASCOT or X!Tandem) often multiple peptides are assigned to the same protein. The assumption that these peptides are really derived from one originating protein is questionable, if the peptides show heterogeneous quantitation values. This is because a specific protein (or peptides of the same protein) can not show up- and down-regulation simultaneously. However, heterogeneous behavior of peptides could be derived from biological or technical reasons. Biologically, protein isoforms, PTMs or multimer formations could lead to heterogeneous peptide quantitation values. Technically, the heterogeneity can be caused by false positive database identifications. Whatever reason causes the heterogeneity, protein quantitations are distorted by averaging heterogeneous peptide quantitations. The effects of differentially regulated protein isoforms may neutralize each other resulting in log fold changes of ~ 0. PPINGUIN is specifically designed to resolve these ambiguities.

In contrast to the standard workflow, PPINGUIN employs clustering prior to protein identification as a very early step in data processing (see workflow comparison in Figure 30). Typically, data mining techniques are applied after protein identification and quantitation. Recently different approaches have been proposed to improve protein identification using peak intensities[69, 94]. In contrast to these works, our major goal is to improve quantitation itself based on the proven and tested identification tools.

Compared to MASCOT and X!Tandem/OpenMS, evaluation based on PPINGUIN shows improved results regarding all three quality control steps: First: heterogeneity of peptides for the same proteins is strongly reduced; Second: experimental reproducibility is elevated and Third: accordance with prior knowledge is also more conclusive.

Our approach is based on pre-selection of peptides by applying unsupervised clustering using k-means. We decided to use k-means as it is computationally fast and sufficient to demonstrate the benefit of the pre-selection. We have also tested affinity propagation[51] as a more sophisticated clustering approach but it turned out to be computationally very expensive while the main results (decreased heterogeneity and increased reproducibility) were similar to k-means. The value of k=5 was determined using two cluster validation scores: gap statistic and Xie-Beni index. Furthermore we found the clustering to be stable with respect to different starting points for the clustering. Also for different number of clusters within a reasonable range the results are not changed. Altogether we think that despite its simplicity k-means (with k=5) is well suited for our purpose.

Due to our experimental design different experiments refer to different biological samples combined with a permutation of iTRAQ channels leading to a mixture of technical and biological variations. Applying the multi-lowess normalization, the data is homoscedastic and technical variation and biological variation are similar to values reported by Gan et al.[55]. Applying only simple median correction the variance is higher for smaller quantitation ratios, which has been reported previously [83].

PPINGUIN can intrinsically identify protein isoforms if they are expressed differentially. PPINGUIN can detect potential novel splice variants and thus, it may help to improve protein or even nucleotide databases. As an example PPINGUIN finds two variants of the RS_30 gene (see Figure 45). The corresponding FAU gene may have two variants: the RS_30 protein with 59 amino acids and the completely transcribed protein with 133 amino acids (see Section 4.5.10). The two isoforms found by PPINGUIN probably correspond to the two protein variants. Verification of this hypotheses would require further in-depth investigation.

Beside detection of potential isoforms, PPINGUIN may also help to assign non-unique peptides to the true origination protein. E.g. if non-unique peptides for two potential proteins are clustered in two different groups each together with unique peptides, than the uniquely assigned peptide can hint for the true original protein (see Figure 47). Examples of this effect are given in Section 4.5.11.

Each of the three evaluated methods shows a distinct set of identified proteins. PPINGUIN has the highest number of proteins (256) identified in all three experiments. This is especially an effect of the biologically motivated clustering used by PPINGUIN since a random clustering results in less proteins identified in all three experiments (219 - see Table 13). A relatively large set of found proteins is found only by PPINGUIN and not by X!Tandem. This is due to two combined effects: First, exploiting quantitation profile information, our clustering leads to a relative enrichment of peptides belonging to the same protein in a cluster and second, by splitting spectra into groups, clustering decreases the total number of spectra in each identification process. The reduced number of spectra per cluster alters the identification threshold used for calibration of the false discovery rate and in effect new proteins are identified. MASCOT also results in a large set of uniquely identified proteins. Most of these unique MASCOT proteins are also found using X!Tandem but they remain below the significance threshold. This is mostly due to differences in the assessment of short peptides since MASCOT appears to include many small peptides for identification that are excluded by X!Tandem (see also Figure 42). In general MASCOT seems to have problems in assessing small peptides since we have found an increased proportion of short peptides with significant E-values

also from the decoy database (see Figure 38). This may be an overlapping effect of forward and reverse database, since Elias and Gygi[43] reported that for small peptides (length 5) almost all peptides are shared between forward and reverse database. However, testing this effect with our databases, from the 947 peptides from the reverse database with length 5-6, only 3% are shared by the forward database. The reason for the increased proportion of short peptides with significant E-values from MASCOT remains unclear.

The set of quantified protein accessions received by PPINGUIN is characterized by an increased experimental reproducibility compared to the other methods. This implies that using PPINGUIN for evaluation, one experimental result is a reliable predictor for the results of a similar experiment. Finally, the comparison with prior knowledge showed a surprisingly high agreement of our top proteins with the reference set, which we deem representative for diabetes and obesity. This hints for the practical benefit of our method.

4.7 CONCLUSION AND OUTLOOK

We proposed a novel method for evaluation of iTRAQ data motivated by the observation that relative concentrations of peptides derived from the same protein often show unexpectedly heterogeneous correlation patterns. Exploiting correlations of quantitation ratios achieves more consistent quantitation ratios than the standard approaches. This is demonstrated by an increased reproducibility of independent experiments. Besides leading to a more reliable quantitation, the method can reveal new isoform candidates.

We see our work as a promising step towards quantitation guided identification. In general, we recommend to use our method in case accurate quantitation is a major objective of research. Regarding the increasing importance of quantitative proteomics we think that this method will be useful in practical applications like model fitting or functional enrichment analysis.

We expect that our approach will be still more valuable with an increasing number of parallel quantified samples (e.g. 8-plex iTRAQ) since the importance of the clustering increases. The proposed approach can also be very useful for other quantitative proteomics technologies like, e.g., SILAC. A next step will be to extend the algorithms to include spectra with incomplete iTRAQ quantitations. Future versions of PPINGUIN will aim at further refinement of protein quantitation by incorporating the rapidly growing public knowledge on splice variants and protein isoforms.

5

DIGE: DATA ANALYSIS AND ALGORITHMS

Chapter Contents

5.1 INTRODUCTION

2D Difference Gel Electrophoresis (DIGE) was initially described in 1997 by Unlu et al.[163] as a multi-sample gel separation method based on the two fluorescent dyes Cy3 and Cy5. Extending this method by adding the dye Cy2 2D Difference Gel Electrophoresis (DIGE) theoretically allows for multiplexing three samples. In practice, DIGE is used for multiplexing only two samples and the Cy2 channel is often used to incorporate a pooled internal standard[3] (see section 2.1.3). In general, pooling biological samples is often discouraged because combining biological distinct subjects in a single pool makes estimation of individual biological variation more complicated or even impossible. Experiments that do not allow for estimation of biological variation should not be performed[113]. However, a common pooled reference leads to an increased comparability of the samples without introduction of an artificial (maybe out of context) reference. An internal standard improves the accuracy of relative quantitation by acting as a loading control and facilitating spot matching between gels and alleviates somewhat the inherent gel-to-gel variation[157].

5.2 DATASET

Analogous to iTRAQ experiments (see section 4.2), DIGE was performed to investigate effects of Standard Diet (SD) and High Fat (HF) diet on the two mouse strains New Zealand Obese (NZO) and Swiss Jim Lambert (SJL) (see section 2.2.3 for complete experimental design of the Sys-Prot project). A total number of 10 DIGE gels where performed, see table 17 for experimental design of DIGE experiment. Labeling of samples follows a randomized experimental design. This helps to avoid biases often observed in DIGE experiments leading to systematic errors in the data[80].

Gel ID	Cy2	Cy5		Cy3	
		Genotype	Diet	Genotype	Diet
869	pool	NZO	SD	NZO	HFD
870	pool	NZO	HFD	SJL	HFD
871	pool	SJL	SD	NZO	SD
578	pool	SJL	HFD	SJL	SD
873	pool	NZO	SD	SJL	SD
874	pool	SJL	HFD	NZO	HFD
875	pool	SJL	HFD	NZO	SD
876	pool	NZO	HFD	SJL	SD
877	pool	SJL	SD	SJL	HFD
878	pool	NZO	SD	NZO	HFD

Table 17: Description of experimental design. 10 different gels colored with 3 different dyes (Cy2, Cy3 and Cy5) were created. Labeling of samples follows a randomized experimental design. Cy2 channel was used as a pooled mixture of all samples.

5.3 STATE OF THE ART

Since the first introduction of DIGE in 1997[163], DIGE has been applied to a broad range of research areas including cell signaling[24, 146], neuroscience[164, 53] and cancer research[52, 84].

Given suitable dyes and a sensitive imaging system, Minden et al.[110] claimed that experimental design and statistical analysis are the most crucial aspects of performing informative DIGE experiments. Typically the whole workflow for DIGE data analysis is performed using commercial DeCyder software package (GE Healthcare). This includes gel matching and identification of matched spot groups, image analysis and feature quantitation as well as subsequent statistical analysis including normalization and detection of differential biomarker candidates.

5.4 METHODS

5.4.1 *Pre-processing / Normalization*

The different dyes show distinct fluorescence properties (e.g. Cy5 has a higher extinction coefficient). This dye effect leads to a systematic shift in channel intensities (on log scale) or to an offset after ratio calculation, respectively. A similar effect was observed for the four different iTRAQ channels (c.f. Section 4.5.1). The normalization/calibration aims at removing intensity bias within each gel as well as bias between the gels. The easiest normalization strategy is a linear regression using a scaling factor and a background offset[81]. Kultima et al.[87] compared different normalization strategies for DIGE data. Beside removing within and between gel biases, they devoted a special emphasis to spacial bias within the gel. They found only 2D loess and 2D quantile normalization successfully removed both, intensity and spatial bias.

For calibration/normalization we use multi lowess algorithm - a multi dimensional extension of lowess normalization strategy[127]. The algorithm assumes the majority of the features to be expressed in similar manner regardless of expression level and as an effect the correlation of the samples to be high. The normalization is performed for each gel separately for two reasons: First, between gel correlation is much lower than within gel correlation and Second, the dye effect is different for each gel (see also Jung et al.[74]). This normalization approach entails that after normalization the gels must not be compared directly. Instead the comparison of different gels has to be performed using ratio data employing the common pooled channel.

5.4.2 *Protein specific dye effect*

A global normalization strategy aims at removing global dye effects. Krogh et al.[86] showed that protein-specific dye effect occurs in DIGE data that can not be removed by a global normalization strategy. They analyzed three different DIGE experiments and found $19 - 34\%$ of the proteins to show a statistically significant dye effect (0.001 significance level). They proposed an analysis tool (DIGEanalyzer) using linear mixed model to remove protein specific dye effects that is also implemented in Proteios[65]. Krogh et al.[86] argued that dye effects in DIGE may result from a combination of preferential dye binding or differences in gel migration and fluorescent properties.

In order to investigate the existence of a protein-specific dye effect in our dataset we used ANOVA (see chapter 2.3) with linear mixed model regarding all single effects:

$$Ratio \sim Diet + Genotype + Dye$$

Since multiple tests are performed a multiple testing correction has to be performed (c.f. section 4.4.7). Bonferroni correction[41] rejects any hypothesis H_j with unadjusted p-value less than or equal to α/m (m = number of tests performed).

5.4.3 *Differential Analysis*

Differential analysis investigates effects of the experimental factors genotype and diet analogous to MALDI and iTRAQ data analysis (see Chapter 3 and 4). Following the statistical analysis of MALDI data (Section 3.4.4) we simultaneously investigate effects derived from genotype and diet as well as the factor combination at the same time. A good candidate for that analysis is again ANOVA (Section 2.3). The ANOVA model employed comprises both experimental factors genotype and diet including mixed effects:

$$Ratio \sim Diet + Genotype + Diet * Genotype$$

The dye effect is not considered here, because we can not find evidence for a statistically significant dye effect (see Section 5.5.2).

Considering all factor combinations each combination is represented by 5 replications (c.f. Table 17). Multiple testing correction is performed according to Bonferroni.

5.4.4 *Spot Similarity*

The analysis of MALDI and iTRAQ data has demonstrated the benefit of clustering methods investigating the relation between peptides or proteins. Following these approaches, we investigate whether clustering is also helpful for DIGE data. Different kinds of clustering methods of spots/proteins have been performed previously on DIGE data[142, 96, 104]. These approaches are typically based on unsupervised learning algorithms such as hierarchical clustering or k-means using euclidean distance measure and showed several spots/proteins with high similarities referring to the experimental design.

For clustering we reuse the concept of quantitation profiles introduced for analysis of iTRAQ data - Section 4.4.3. A quantitation profile is the list of intensities for a certain spot across all (20) samples. High correlation between intensity profiles hints for common origin as e.g. multimer formations or PTMs or possibly similar biological functions. However, correlation of spots is often

dominated by sample effects (effects of genotype and diet). In order to investigate spot similarities the sample effects must be neglected. Therefore correlation is calculated for each genotype-diet combination separately and averaged afterwards. Average linkage hierarchical clustering is performed using $1 - \rho$ as distance measure where ρ denotes the average Pearson correlation across the samples of each genotype/diet combination.

Spots with similar intensity profiles are aggregated in close proximity in the dendrogram. For the identification of spots with similar intensity profiles we cut the tree at a given height level ($h = 0.15$). Subsequently we select all clusters with at least three spots.

5.5 RESULTS

5.5.1 *Pre-processing / Normalization*

Raw gel data (without any normalization) show medium reproducibility with between gel correlation coefficients of 0.76 - 0.9 (Pearson correlation[136]). In contrast, the within gel correlation (correlation of the three channels across all spots) is much higher with correlation coefficients: 0.95-0.98 (see right hand side of figure 52 for correlation plot). For unnormalized data, the systematic shift in channel intensities is clearly visible in the box plot (left hand side of figure 52). For the majority of the gels, the red channel (Cy5) shows the lowest mean intensity while the blue channel (Cy2) has the highest average intensity. The highest channel bias of 0.67 is observed for gel *871* which corresponds to a 1.6 fold differential expression if not normalized. This systematic shift substantiates the need for a normalization/calibration strategy.

Gel-based multi lowess normalization successfully removed the systematic shift of the different dyes of each gel. But a substantial offset between the different gels remains (see left hand side of figure 53). Consequently, the different gels must not be compared at the level of the normalized data but rather on the level of ratios.

The Cy2 channel represents a pooled reference of all samples, that is present on every gel. Therefore, the Cy2 channel is used as reference channel to calculate the Cy3/Cy2 and Cy5/Cy2 ratios for each gel. At the ratio level the different gels are comparable because the average ratios are distributed with mean 0 and similar variances. The correlation of the 20 ratios ordered by the experimental factors is shown in figure 54. Most of the correlation coefficients are close to zero showing that the majority of proteins is not differentially expressed. The genotype effects are clearly visible in the correlation matrix. This implies that the genotype effects are rather strong which corresponds to results obtained

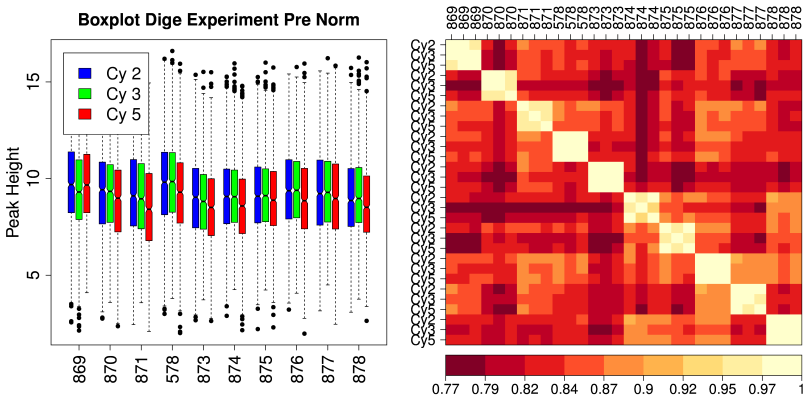


Figure 52: Data prior to normalization. Left side: Box plot of all channels of the 10 gels colored by the dye used (blue = Cy2, green = Cy3 and red = Cy5). Different dyes show different average intensities. Right hand side: Correlation plot (Pearson Correlation) of all channels of the 10 gels. The diagonal boxing is due to high within gel correlation.

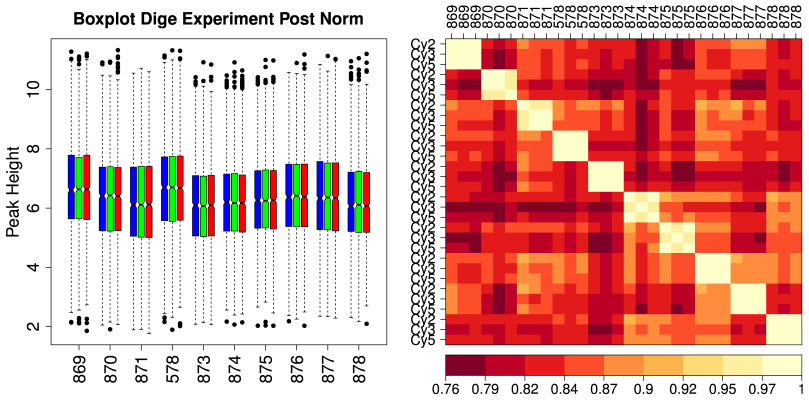


Figure 53: Data after application of normalization. Left side: Box plot of all channels of the 10 gels colored by the dye used (blue = Cy2, green = Cy3 and red = Cy5). After normalization different dyes are homogeneous within the same gel. But an offset between the different gels remains. Right hand side: Correlation plot (Pearson Correlation) of all channels of the 10 gels. The diagonal boxing demonstrates high within gel correlation.

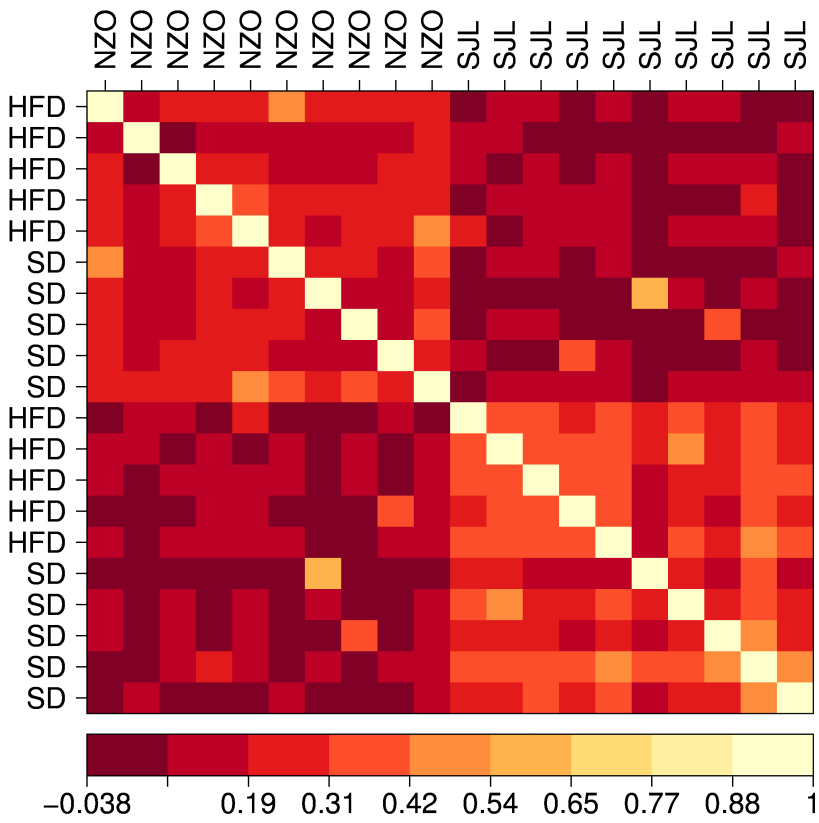


Figure 54: Correlation plot (Pearson Correlation) of the data after application of normalization and calculation of ratios using Cy2 channel as reference. The majority of correlation coefficients are 0 and within gel correlation vanished. Genotype has a rather strong effect on the data since the NZO and SJL genotypes samples are distinguishable while the diets are not.

from MALDI and iTRAQ data (see Chapter 3 and 4). Furthermore, the strong within gel correlation vanished.

5.5.2 *Protein specific dye effects*

We investigate the existence of a protein specific dye effect as described in Section 5.4.2. The volcano plot of effect strength and ANOVA p-values for protein specific dye effects is shown in Figure 55. Regarding significance threshold after the Bonferroni correction only 4 proteins (0.03%) show statistically significant dye effects (see right hand side of Figure 55 for a box plot of these four proteins). Especially spot number 1040 has a very significant dye effect with p-value of 10^{-10}. However, compared to the results of Krogh et al.[86] who found $19-34\%$ of the proteins to show statistically significant dye effects, the protein-specific dye effects in our dataset are rather marginal effects (2.7% of all spots at comparable significance level of 0.001).

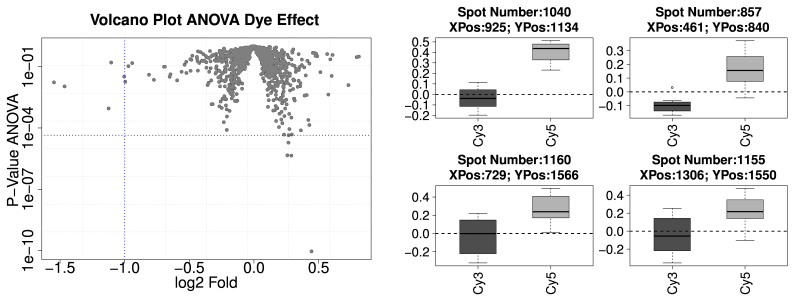


Figure 55: Protein specific dye effect after global normalization. Left side: Volcano Plot, blue vertical dotted lines reflect the $\log_2$ fold of -1 and 1, blue horizontal dotted line marks the Bonferroni Multiple Testing corrected p-value threshold of 0.05. Right hand side: Box plot of the intensity distribution separately for each genotype regarding the four spots with the most significant p-values.

Since the dye has only minor effects on our dataset and due to the randomized experimental sample design the protein specific dye effect is neglected in the subsequent analysis.

5.5.3 *Differential Analysis*

The effects due to differences in genotype are stronger than the effects due to diet (c.f. MALDI analysis in Section 3.5.3) which is already visible regarding the correlation plot of the DIGE ratios (figure 54). Statistical analysis is performed to investigate the effect due to single experimental factors as well as factor combination as described in Section 5.4.3.

5.5.3.1 *Genotype*

As already seen during the analysis of MALDI and iTRAQ data, NZO and SJL mice show distinct proteomic patterns. Left hand side of Figure 56 shows the volcano plot for experimental factor genotype. Some spots show differential expression with $\log_2$ folds up to ± 3 which corresponds to a differential expression of 8 fold. ANOVA p-values are significant up to 10^{-14}. A box plot of the two genotypes for the six most significant spots is presented at the left hand side of Figure 56.

The effect of genotype is indeed strong leading to the identification of various differently expressed spots. The 15 spots with the most significant p-values for experimental factor genotype are presented in Table 18 including protein identification if available.

Spot ID	Identification			Mean NZO	Mean SJL	$\log_2$ Fold	P-Value
	ID	Description	Score				
807	P48036	Annexin A5	429	0.492	-0.784	-1.03	4.69e-15
	Q64374	Regucalcin	265				
557				0.713	-2.25	-2.85	1.04e-14
556	Q63836	Selenium-binding protein 2	514	0.641	-2.26	-2.94	6.76e-14
407	P38647	Stress-70 protein, mitochondrial	1657	0.585	-1.34	-2.02	1.68e-12
	P63017	Heat shock cognate 71 kDa protein	680				
	P07724	Serum albumin	465				
	P38647	Stress-70 protein, mitochondrial	5099				
	P07724	Serum albumin	2540				
1025	Q02257	Junction plakoglobin	225	0.368	-0.399	-0.702	4.05e-12
	P70296	Phosphatidylethanolamine-binding protein 1	535				
928				-0.984	0.198	1.23	4.31e-12
943				-0.731	0.41	1.14	5.42e-12
330	P63017	Heat shock cognate 71 kDa protein	750	0.485	-1.09	-1.61	6.16e-11
367	P38647	Stress-70 protein, mitochondrial	175	-0.174	0.215	0.383	6.47e-11
1159	P12710	Fatty acid-binding protein, liver	503	0.914	-1.48	-2.38	1.34e-10
	P52760	Ribonuclease UK114	447				
	P52760	Ribonuclease UK114	599				
828				-0.489	0.321	0.752	7.31e-10
461				-0.559	0.346	1.02	2.68e-09
927				-0.866	0.309	1.12	2.99e-09
580	P32020	Non-specific lipid-transfer protein	1172	0.599	-0.608	-1.01	6.11e-09
	P56480	ATP synthase subunit beta, mitochondrial	1084				
925				-0.715	0.273	0.955	1.02e-08

Table 18: 15 spots with the most significant p-values for the experimental factor genotype. If available protein identification information is given.

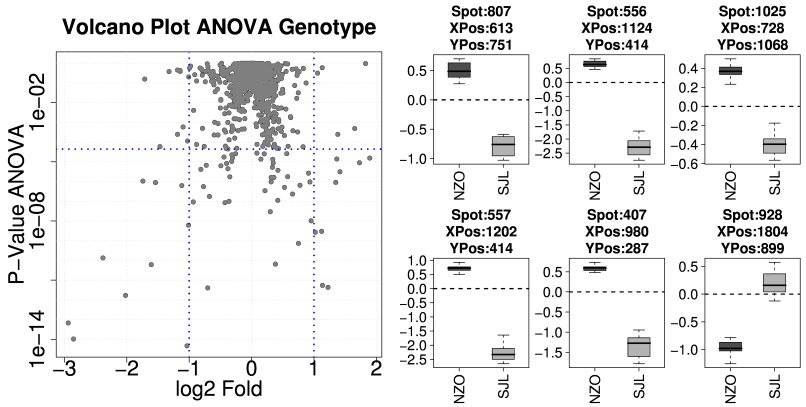


Figure 56: Effects of the experimental factor genotype. Left side: Volcano Plot, blue vertical dotted lines reflect the $\log_2$ fold of -1 and 1, blue horizontal dotted line marks the Bonferroni Multiple Testing corrected p-value threshold of 0.05. Right hand side: Box plot of the intensity distribution separately for each genotype regarding the six spots with the most significant p-values.

5.5.3.2 *Diet*

Compared to experimental factor genotype, the different diets have only poor effect on the data. The most significant p-value of ANOVA for diet is only 10^{-5}. Some folds are up to 1.5 but none of which is significantly changed. However, the best folds of the most significant spots are $\sim \pm 0.5$. For the volcano plot see left hand side of Figure 57. The box plots of the the six most significant spots are shown on the right hand side of Figure 57. Some spots are differentially expressed but these effects are poor compared to the genotype effects. The 15 most significant spots for the experimental factor diet are presented in Table 19.

5.5.3.3 *Factor Combination*

Analogous to the analysis of MALDI data we analyze the mutual effect of feature combination genotype and diet (see section 5.4.3 for ANOVA model). The volcano plot for factor combination is depicted left hand side of figure 58. None of the resulting p-values is significant regarding the multiple testing corrected threshold of 0.05. Furthermore the folds for the most significant spots are small ($\sim \pm 0.5$). Fold in the case of ANOVA reflects the coefficient of the linear model used within ANOVA which can be interpreted like an effect strength. The box plot of the six most significant spots for each distinct combination is shown on right hand side of figure 58. Spots 14, 61, 62 and 68 all show a similar pattern: HF/SJL is lower than the rest. Spot 482 shows higher signals only for the combination of SD/NZO. The 15 most

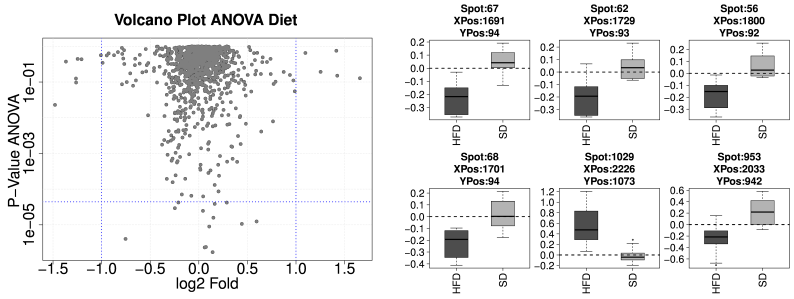


Figure 57: Effects of the experimental factor diet. Left side: Volcano Plot, blue vertical dotted lines reflect the $\log_2$ fold of -1 and 1, blue horizontal dotted line marks the Bonferroni Multiple Testing corrected p-value threshold of 0.05. Right hand side: Box plot of the intensity distribution separately for each diet regarding the six spots with the most significant p-values.

Spot	Identification			Mean	Mean	$\log_2$	P-Value
ID	ID	Description	Score	NZO	SJL	Fold	
67				-0.223	0.0504	0.142	1.71e-06
68				-0.225	0.0164	0.0604	2.32e-06
62				-0.198	0.0466	0.0735	2.39e-06
1029				0.54	-0.00314	-0.754	4.02e-06
56				-0.183	0.0658	0.108	8.94e-06
953				-0.249	0.221	0.443	1.06e-05
461				-0.345	0.132	0.595	1.24e-05
367	P38647	Stress-70 protein, mitochondrial	175	-0.0574	0.0987	0.15	1.58e-05
61				-0.193	0.0415	0.0428	2.46e-05
499				0.17	-0.111	-0.246	2.92e-05
89				-0.191	0.146	0.164	3.85e-05
73				-0.461	-0.067	0.288	4.19e-05
405				0.131	-0.163	-0.21	4.22e-05
14				-0.232	0.00169	0.0449	6.23e-05
794				0.0874	-0.225	-0.0767	6.5e-05

Table 19: 15 spots with the most significant p-values for the experimental factor diet. If available protein identification information is given.

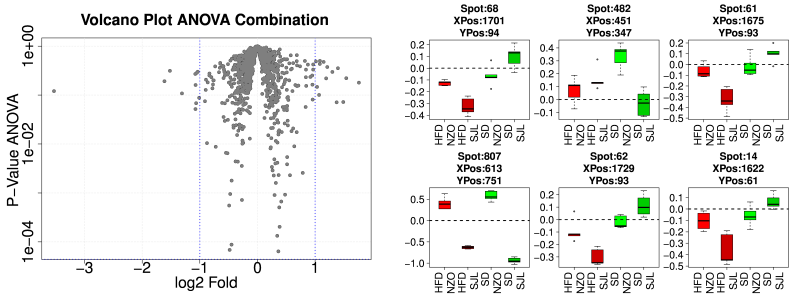


Figure 58: Effects of the combination of experimental factors diet and genotype. Left side: Volcano Plot, blue vertical dotted lines reflect the $\log_2$ fold of -1 and 1, blue horizontal dotted line marks the Bonferroni Multiple Testing corrected p-value threshold of 0.05. Right hand side: Box plot of the intensity distribution separately for each distinct combination of genotype and diet regarding the six spots with the most significant p-values.

significant spots for the factor combination are presented in table 20.

5.5.4 *Spot Similarity*

A heatmap[42] (clustering of samples and genes) of the best 50 differentially expressed spots for genotype is shown in Figure 59. Spots in close proximity in the dendrogram are characterized by correlated intensity profiles (similar color pattern of the rows). Mainly two different types of spots/proteins are discernible one with up-regulation in SJL and the other with down-regulation, respectively.

The dendrogram of the spots on left hand side of the heatmap is dominated by genotype effects. In order to investigate spot similarities the sample effect must be removed. Therefore correlation is calculated for each genotype-diet combination separately and averaged (see Section 5.4.4). Similar spots are detected by cutting the tree at a given height level ($h = 0.15$) - see Figure 60 for a visualization. Table 21 shows all identified clusters of similar spots including average and minimal correlation coefficients. Cutting the tree at a height of 0.15 corresponds to a minimal correlation coefficient of 0.85 when using minimal linkage clustering (see table 21).

The majority of the clusters contains spots in close spatial proximity on the gel (see x,y coordinates in table 21). Spots in close proximity on a gel might reflect identical proteins due to two reasons. First, labeling itself leads to an electrophoretic separation of labeled and unlabeled protein which is more evident in the lower mass range[60, 157] but usually less than one half diameter

Spot ID	Identification			Mean				Fold	P-Value
	ID	Description	Score	HF/NZO	HF/SJL	SD/NZO	SD/SJL		
68				-0.125	-0.325	-0.0649	0.0978	0.363	6.39e-05
807	P48036 Q64374	Annexin A5 Regucalcin	429 265	0.405	-0.629	0.578	-0.939	-0.483	6.85e-05
482				0.0713	0.157	0.335	-0.0286	-0.45	8.77e-05
62				-0.0942	-0.302	-0.0207	0.114	0.343	0.000133
61				-0.0593	-0.327	-0.0165	0.0994	0.384	0.000203
14				-0.105	-0.359	-0.0605	0.0639	0.378	0.000506
563				-0.276	-0.0816	-0.0699	-0.367	-0.491	0.000553
69				-0.168	-0.454	-0.155	-0.0274	0.413	0.000667
213				-0.207	-0.0467	-0.137	-0.245	-0.269	0.000717
258				-0.133	-0.44	-0.116	-0.127	0.296	0.000904
58				-0.00163	-0.218	0.0168	0.28	0.479	0.000922
794				-0.126	0.3	-0.202	-0.249	-0.472	0.000951
282				0.0145	-0.119	-0.0898	0.0958	0.319	0.0016
336				-0.127	-0.0575	0.0848	-0.294	-0.448	0.00179
56				-0.115	-0.251	-0.00748	0.139	0.282	0.00227

Table 20: 15 spots with the most significant p-values regarding the combination of experimental factor diet and genotype. If available protein identification information is given.

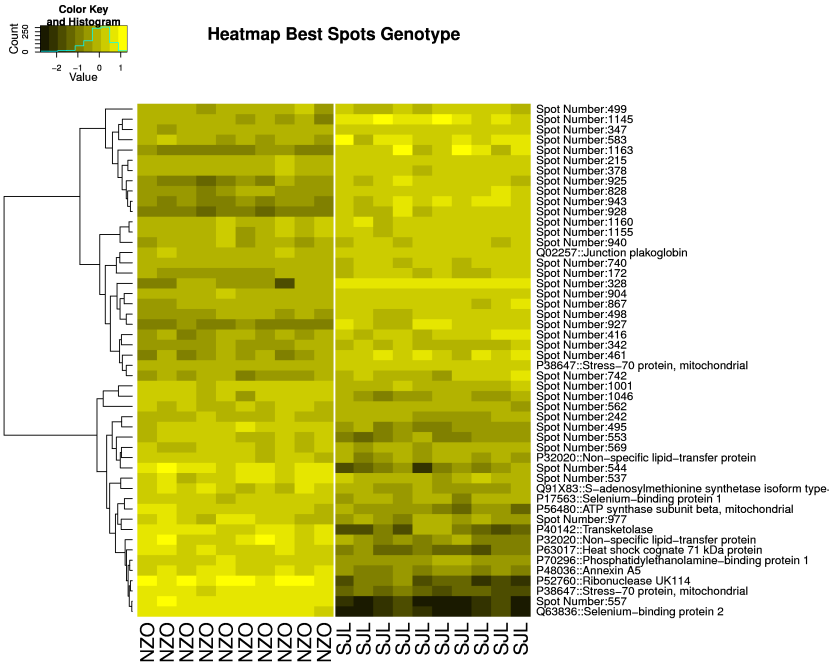


Figure 59: Heatmap of the best 50 differential spots (spots with best fold) for experimental factor genotype. The two genotypes are clearly visible. Clustering was performed using 1 - Correlation as distance measure.

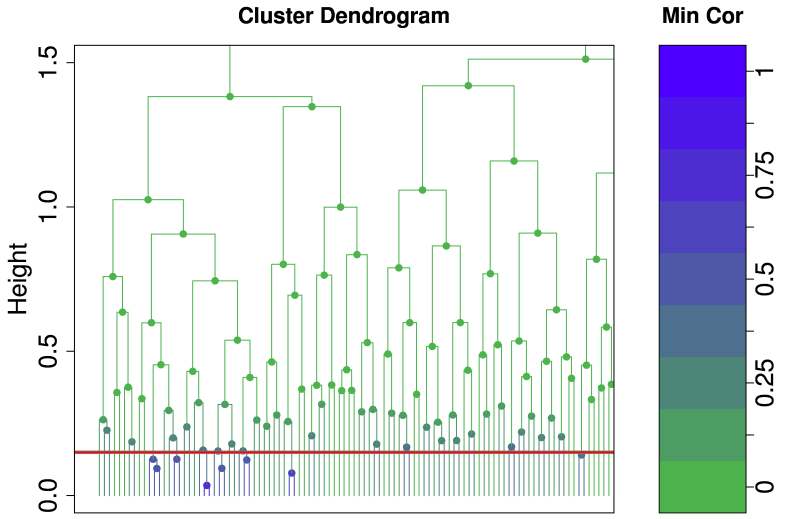


Figure 60: Excerpt of the cluster dendrogram. Every leave reflects one spot (Spot ID not annotated). The spots are clustered using 1 - Correlation as distance measurement. The clusters are colored according to the minimal correlation coefficient off all spots belonging to this cluster. The red vertical line indicate the height used to cut the tree.

of the protein spot[110]. Second, PTMs cannot only change the mass of a protein but they can also affect the charge of the protein which leads to a horizontal shift in the gel[45]. Hence proteins present with multiple isoforms might be found in spatial proximity on the gel. Different protein isoforms that are not differentially expressed should be highly correlated[45].

5.6 SUMMARY AND DISCUSSION

We have developed and applied a pre-processing workflow specifically tailored for DIGE data created within the Sys-Prot project. Statistical evaluation reveals several differentially expressed spots for different experimental factors. Especially forthe factor genotype we found several differentially expressed spots. A protein specific dye effect as seen by Krogh et al.[86] was not observed in our dataset.

Protein identification based on MS revealed a total number of 22 proteins belonging to 24 spots. Among the 24 spots identified, the spots: 330, 407, 556, 580, 807, 1025, 1159 are found differentially expressed for experimental factor genotype (see Section 5.5.3.1). All spots showed very similar intensity profiles: up-regulation for NZO and down-regulation for SJL - see left hand side of Figure 61.

Cluster No.	Spot IDs (X,Y)	Mean Cor	Min Cor
1	603 (817,474), 610 (816,481), 613 (803,483)	0.97	0.96
2	1155 (1306,1550), 1162 (639,1572), 1160 (729,1566)	0.93	0.91
3	506 (883,369), 634 (1412,505), 786 (1235,711)	0.9	0.85
4	246 (597,185), 213 (1104,166), 430 (451,303)	0.89	0.87
5	1170 (1598,1616), 1169 (1585,1616), 1172 (1630,1616)	0.89	0.86

Table 21: Spots with very similar intensity profiles. The cluster dendrogram of all spots was cut at height of 0.15. All clusters with at least 3 spot are presented here including average and minimal correlation coefficients.

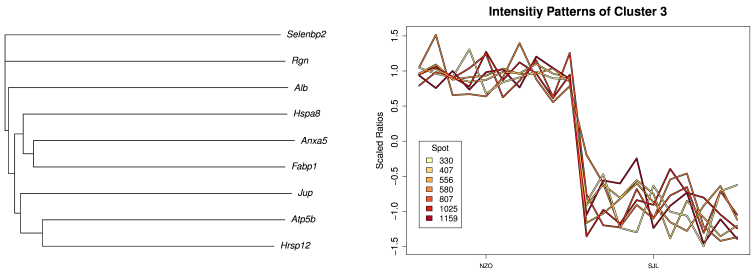


Figure 61: Left hand side: Phylogenetic tree of the protein identified within cluster 3. Sequences are obtained from UniProt and a multiple alignment was performed using ClustalW. Right hand side: Intensity patterns of the spots of cluster 3. The ratios have been scaled as the correlation intrinsically uses scaled data.

In contrast to the majority of identified groups (c.f. Table 21), this set of spots is neither in spatial proximity of the gel, nor are they derived different protein isoforms. They all show similar intensity patterns but they do not show these similarities on sequence level (see 61). Furthermore they do not seem to have a similar biological function as they are not annotated in similar KEGG or Reactome pathways. The reason for the similar profiles remains unclear. One possible explanation could be some common upstream regulatory mechanism.

Often multiple high confidence MS-based identifications were obtained for the same spot. In this case usually the most abundant protein species is assumed to be the protein of interest which is often not true[157].

5.7 CONCLUSION

A specialized pre-processing benefits from the common reference pool present on each gel and increases between-gel comparability. Towards biomarker identification ANOVA is well suited to analyze DIGE data. We were able to detect spots that are differentially expressed for single experimental factors and combination of genotype and diet. Spot correlation is very useful for DIGE data. Due to labeling process or PTMs, spots in close proximity in the gel may represent the same protein or protein isoforms and are therefore often correlated.

6 INTEGRATION OF RESULTS

6.1 COMPARING THE APPROACHES

Three proteomics technologies: MALDI, DIGE and iTRAQ were employed for analyzing T2DM mouse models. The three different technologies are distinct in motivation, methodological properties and results. Although all three technologies are employed for biomarker identification, the results are scarcely comparable. iTRAQ provides large numbers of protein quantitations including protein identifications, while for MALDI and DIGE the number of identified proteins is limited (typically < 20) and depends on follow-up experiments. Especially for the MALDI data analyzed in this thesis, none of the most significant differential peaks has been identified in an experimental follow-up yet.The missing identification hinders a direct comparison of MALDI with one of the other two technologies. Furthermore, the three analysis strategies developed in this work are suited for different tasks. MALDI and DIGE data evaluation primarily aimed at the identification of biomarker candidates and classification of experimental factors. In the iTRAQ data we observed a high degree of ambiguities and therefore we developed PPINGUIN, a method allowing for stable and reliable protein quantification. Compared to MALDI and DIGE, identification of biomarker candidates was not the primary goal of PPINGUIN.

Even with an increasing number of identified proteins the overlap is excepted to be rather small. This is because the different technologies select proteins according to different physicochemical properties.

6.1.1 *Common Properties of the three Approaches*

Despite these differences, all three approaches have several properties in common. First of all, the three proteomics experiments are performed to investigate the effects of T2DM. All data are based on similar experimental designs using the same mouse samples (see Section 2.2). Second, all developed approaches employ signal correlation (either protein-wise or peptide-wise) to improve results although the correlation is used for a different purpose: For MALDI clustering of quantitation profiles especially in combination with ANOVA was used for feature selection for classification. On the other hand, clustering of quantitation profiles of iTRAQ spectra improved reproducibility and reliability of

protein quantitations. Third, all three analyses commonly showed very clear results (identified peaks/spots show highly significant p-values even after rigid multiple testing correction). Fourth, in all three analyses the different mouse genotypes have the most significant effect on the data while diet has only minor effects: In our MALDI data we found differentially expressed peaks with p-values up to 10^{-91} for genotype (very low p-values are also an effect of the large number of samples), while the best p-values for other experimental factors was only 10^{-23} (see Section 3.5.3). Analyzing the DIGE data, spots with p-values up to 10^{-14} were identified for genotype while the most significant spot for diet had a p-value of 10^{-5} (see Section 5.5.3). For iTRAQ data the top differential protein had a $\log_2$ fold for genotype of 1.8 while the best $\log_2$ fold for diet was 0.6 (p-value is not a reliable criterion for our iTRAQ data because only three experimental repeats are available).

6.1.2 *Comparing biomarkers*

iTRAQ allows for the identification of a large number of proteins. For DIGE and MALDI identification of proteins requires additional follow up experiments. For the MALDI data analyzed in this thesis, none of the most significant differential peaks has been identified at the current stage. However, for DIGE some of the most differential spots have been identified. Hence, at least protein identifier obtained from the analyses of iTRAQ and DIGE can be compared. For the analysis of the DIGE experiments, protein identification was restricted almost exclusively to proteins with high genotype effects (see Table 18 in Chapter 5). To assure comparability of DIGE and iTRAQ data, the statistical evaluation of iTRAQ data as described in section 4.4.7 and 4.5.9 in Chapter 4 was repeated. The repeated evaluation aimed at the identification of genotype effects without dietary effects (comparing NZO_SD with SJL_SD).

The overlap between differential proteins found with DIGE and differential proteins found with iTRAQ (for the same experimental comparison) is low (see Venn diagram in Figure 62). From the 17 differentially expressed proteins from DIGE experiments only 3 (FABPL, NLTP and METK1) where also detected with X!Tandem in iTRAQ data. However, an overlap of 3 proteins a significant effect (fisher test p-value: 10^{-9}; using all available UniProt identifier as reference - which is also used later for pathway identification).

A small overlap of differentially expressed proteins identified by DIGE and iTRAQ is in accordance with other studies. E.g. Wu et al.[175] reported only a single protein identified with DIGE and iTRAQ investigating HCT-116 cell lysates. Wu et al. argued that the small overlap supports the hypothesis that both technologies

are complementary in nature. Joining iTRAQ and DIGE results a total of 39 differentially regulated protein ids were identified for experiment factor genotype.

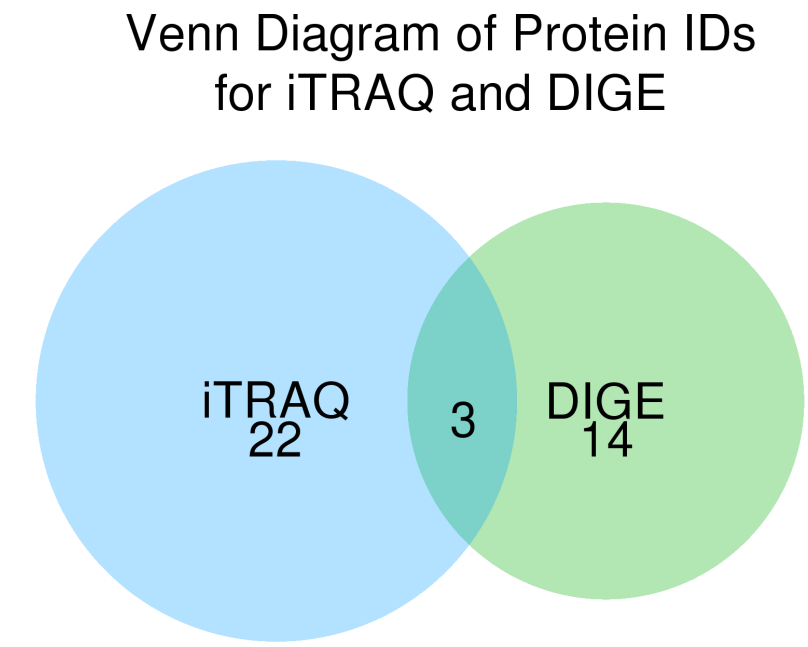


Figure 62: Venn diagram comparing protein identifiers found to be differentially expressed for mouse genotype with DIGE and iTRAQ data. For iTRAQ data, differential search was combined from the results of all three used methods: Mascot, X!Tandem/OpenMS and PPINGUIN. Details for differential search are described in Section 4.4.7. Only three protein identifier are found with both technologies: FABPL, NLTP and METK1.

6.1.3 *Pathways*

Instead of inspecting (large) lists of gene identifiers, it is often easier to interpret found biomarkers in their functional context. Cellular processes often affect sets of biomolecules at the same time. A rather small alteration of all biomolecules in a metabolic pathway may dramatically alter the flux through the pathway and may be more important than high alterations of a single biomolecule[150]. Many different tools have been developed to search for associated pathways (68 tool have been surveyed by Huang et al.[71]). Huang et al. et al.[71] divided the tools into three different classes:

(i) Singular enrichment analysis tools based on a list of genes and a set of reference genes. These tools are based on hypergeometric distribution and Fisher's exact test[49].

(ii) Gene set enrichment analysis based on a list of genes and their experimentally measured expression values. Statistics is mostly based on t-test or wilcoxon test and resampling/permutation tests.

(iii) Modular enrichment analysis which is similar to singular enrichment but with consideration of gene-gene and term-term relationships.

Since for DIGE data the number of identified proteins is limited, the first class of enrichment tools (Singular enrichment analysis) is favorable for the detection of associated pathways. For pathway enrichment we adopted gossip[16] and implemented the whole functionality in R. The big advantage of gossip is an exact determination of the expected Number of False Discoveries (NFD). Multiple testing corrected p-values of gossip are more reliable compared to Bonferroni or Benjamini-Hochberg correction.

For the detection of metabolic pathways, we used the 39 differentially regulated proteins identified with at least one technology as test set and all available UniProt proteins are reference set. We included three different pathway repositories: KEGG[78], Reactome[106] and WikiPathways[124]. The best 8 pathways with the most significant p-values for each pathway repository are shown in Table 22. All three repositories result in similar pathways: Metabolic pathways, especially pathways related to amino acids metabolism are found with highly significant p-values.

As seen in the previous section (6.1.2) the overlap of differentially expressed proteins found with DIGE (17 proteins) and iTRAQ (25 proteins) is small (overlap of 3 proteins). Searching for pathways from Reactome using the 17 DIGE proteins the most significant pathway is *Metabolism* with a p-value of $2.2 \cdot 10^{-05}$ (pFDR = 0.0037). Searching for Reactome pathways using the 25 iTRAQ proteins, the most significant pathway is also *Metabolism* with a p-value of $1.53 \cdot 10^{-11}$ (pFDR = $1.9 \cdot 10^{-09}$). Similar results are obtained for KEGG pathways. For both lists of identifiers, the most significant pathway is *Metabolic pathways*. Although the overlap at the level of protein ids is small, the overlap on the functional level is bigger since the most significant pathway is identical.

6.2 GENERALLY DETECTED PROTEINS

In 2008 Petrak et al.[123] published a review about repeatedly identified differentially expressed proteins. Therefore they analyzed 186 different 2D PAGE studies with various experimental

	Pathway	n test	n ref	pValues	pFDR	NFD
KEGG Pathways	Metabolic pathways	17	986	3.681e-12	1.86e-10	1.86e-10
	Arginine and proline metabolism	4	53	7.197e-06	0.0002298	0.00046
	Pentose phosphate pathway	3	25	3.11e-05	0.0006423	0.00193
	Cysteine and methionine metabolism	3	32	6.148e-05	0.0009477	0.00379
	Legionellosis	3	48	0.000191	0.00249	0.0124
	PPAR signaling pathway	3	64	0.0004289	0.004926	0.0296
	Ascorbate and aldarate metabolism	2	18	0.00088	0.008147	0.057
	Amoebiasis	3	85	0.000952	0.00746	0.0597
Reactome Pathways	Metabolism	17	814	1.747e-13	2.046e-11	2.05e-11
	Metabolism of amino acids and derivatives	5	145	1.809e-05	0.001849	0.0037
	Bile acid and bile salt metabolism	3	29	4.68e-05	0.002801	0.0084
	ChREBP activates metabolic gene expression	2	4	7.084e-05	0.003718	0.0149
	Metabolism of carbohydrates	4	99	7.501e-05	0.003045	0.0152
	Metabolism of lipids and lipoproteins	5	207	9.443e-05	0.00277	0.0166
	Fatty Acyl-CoA Biosynthesis	2	6	0.0001319	0.0042	0.0294
	Pentose phosphate pathway (hexose monophosphate shunt)	2	7	0.0001693	0.004298	0.0387
Wiki Pathways	Amino Acid metabolism	6	75	2.07e-08	7.865e-07	7.86e-07
	One Carbon Metabolism	3	27	3.843e-05	0.000815	0.00163
	One carbon metabolism and related pathways	3	41	0.0001227	0.001516	0.00455
	Pentose Phosphate Pathway	2	7	0.0001693	0.001838	0.00735
	Selenium metabolism/Selenoproteins	2	16	0.0007106	0.005766	0.0288
	Urea cycle and metabolism of amino groups	2	19	0.0009713	0.006366	0.0382
	Fatty Acid Biosynthesis	2	20	0.001067	0.006046	0.0423
	MAPK signaling pathway	3	136	0.00352	0.01976	0.158

Table 22: Best 8 pathways for each of the three used pathway repositories: KEGG, Reactome and WikiPathways. As test set we used the 39 differentially regulated proteins identified with at least one technology. As reference we used all available UniProt proteins. Significance (corrected for multiple testing) of the pathways is given by pFDR values. NFD gives the expected number of false discoveries.

objectives. They identified a list of proteins that are repeatedly differentially expressed regardless of tissue, species or experimental objective.

Independently one year later Wang et al.[169] published a paper with similar results. They identified a list of generally detected proteins in comparative proteomics involving both, 2-D gels and MALDI MS experiments. They used 66 proteomics studies including different species, tissues and experimental objectives. The list of proteins found differentially expressed in many studies from Wang et al. is in good agreement with the list of Petrak et al..

6.2.1 *iTRAQ proteins*

During the analysis of the iTRAQ data we have identified a list of 12 proteins differentially expressed for the combination of genotype and diet (see Section 4.5.9). From these 12 proteins two are part of the list of generally detected proteins:

- aldo-keto reductase
- glutathione S-transferase

6.2.2 *DIGE proteins*

In the DIGE analysis we identified 17 proteins differentially expressed for genotype. From these proteins 5 are part of the lists of Petrak et al. or Wang et al.:

- Heat shock cognate 71 kDa
- ATP synthase
- Annexin A5
- Serum albumin
- fatty acid binding protein 1, liver

Pathway results as seen previously in section 6.1.3 are not changed if these 7 generally detected proteins are excluded. The top results of each pathway repository remain the same.

6.2.3 *Consequences*

The ubiquitous differential identification of proteins implies that they are not related to a specific study condition like species, tissue and experimental objective. Consequently it is questionable if they are specifically related with the experimental factors of the Sys-Prot project. Both, Petrak et al. and Wang et al. conclude

that cellular stress response might be a universal reason for these proteins to be generally expressed differentially. Besides biological explanation, in the case of DIGE a protein specific dye effect might be a technical cause for proteins to show up differentially expressed in many studies (e.g. if experimental design does not include randomized sample design or dye swap). However, as investigated in Section 5.4.2, the dye had only minor effects on our data.

7 CONCLUSION

This thesis comprehensively described the data analysis strategies we developed to evaluate complementary proteomics data acquired within the Sys-Prot project. The primary objective of the project was the investigation of obesity induced Type-2 Diabetes Mellitus (T2DM) mouse model at proteomic level. To this end, the three complementary proteomic technologies Matrix-Assisted Laser Desorption/Ionization (MALDI), Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) and 2D Difference Gel Electrophoresis (DIGE) were used. All technologies were applied to the same mouse individuals and a similar experimental design is used for each experiment. The primary goal of this thesis was to develop, apply and assess specifically tailored approaches for data evaluation for each technology. Compared to standard strategies for data evaluation, the approaches developed in this work show more convincing results for the specific problems of our three data sets. Application of our methods facilitates the interpretation of the results and allows to draw adequate conclusions. Although the developed approaches are specifically tailored for our three data sets, they are still applicable to a broader range of data analysis problems. Therefore they might help to improve the results of other experiments as well.

7.1 MALDI

For evaluation of multi-factorial MALDI-TOF MS data we developed a method for biomarker identification and feature selection. Our approach combines ANOVA and clustering-based redundancy reduction. An appropriate pre-processing was developed and applied in order to guarantee that all requirements for statistical evaluation are fulfilled. Applying out method, we were able to identify peaks that are characteristic for the combination of genotype and diet as well as peaks that are significant for a single experimental factor. These results are significant even when applying rigid multiple testing corrections. We showed that ANOVA is an adequate approach for the identification of biologically interesting biomarkers from MS profiling data based on multi-dimensional experimental design. Furthermore, classification based on features selected with our approach performs similarly well to those generated with more complex global optimization methods.

The method can easily be applied to other MALDI MS datasets or in general to all kinds of classification problems. The combination of cluster based redundancy reduction and ANOVA is very promising and is certainly helpful for a variety of feature selection tasks.

7.2 ITRAQ

As for the evaluation of iTRAQ data we developed Peptide Profiling Guided Identification of Proteins (PPINGUIN). This method exploits correlation of quantitation profiles of spectra to address the problem that in contrast to our expectations relative concentrations of peptides derived from the same protein are often not correlated. It proceeds by first clustering MS spectra based on their quantitation profiles. Protein identification is performed afterwards for each cluster independently. The quality of our approach is assessed in terms of increased reproducibility of independent experiments and better accordance with prior knowledge. Besides leading to a more reliable protein quantitation, PPINGUIN can reveal new protein isoform candidates. PPINGUIN can detect potential novel splice variants and thus it may help to improve protein or even nucleotide databases. We see our work as a promising step towards quantitation guided identification. We expect that our approach will be still more valuable with an increasing number of parallel quantified samples (e.g. 8-plex iTRAQ). An increased multi-plexing capability leads to an increase in covered experimental states that leads to gain in specificity.

The proposed approach can also be very useful for other quantitative proteomics technologies like e.g. SILAC. In general, we recommend to use our method in case accurate quantitation is a major objective of research. Regarding the increasing importance of quantitative proteomics we think that this method will be useful in practical applications like model fitting or functional enrichment analysis.

7.3 DIGE

For the analysis of DIGE data, a specialized pre-processing was developed that benefits from the common reference pool present on each gel. This preprocessing increases between-gel comparability. We showed that ANOVA is well suited to detect biomarker candidates in DIGE data. We were able to detect spots that are differentially expressed for single experimental factors and combination of genotype and diet. Some of the spots were identified and assigned to protein ids. Correlation between spots was used to cluster spots according their quantitation profile. Spots in the

same cluster are often in close proximity in the gel, which may be caused by labeling process or PTMs. The other way around this implies that spots in close proximity in the gel may represent the same protein or protein isoforms.

7.4 CHOICE OF TECHNOLOGY

The results presented in this work underline that the three technologies MALDI, iTRAQ and DIGE are complementary proteomic approaches. Each technology has certain advantages and disadvantages.

iTRAQ is certainly the most advanced technology. It allows for simultaneous quantitation and identification of a large number of proteins. But iTRAQ is often limited to a small number of samples due to high experiment efforts and costs (only 3 iTRAQ experiments with a total of 12 different mice where performed). The low number of samples also hampers the statistical assessment.

On the other hand, MALDI and DIGE are characterized by simplicity and allow for a large number of samples to be processed. More than 150 distinct biological samples were used to create more than 1100 MALDI spectra. But MALDI and DIGE are often restricted to qualitative results or the number of identified proteins is limited.

The choice of technology clearly depends on the experimental goal. If the experimental objective is to quantify the proteome (or at least a large number of proteins) iTRAQ is very well suited for this task. If the experimental objective is to find (only a few) biomarkers e.g. for a certain disease state using as much samples as possible to guarantee a reasonable statistical assessment than MALDI most suitable.

7.5 POSSIBLE IMPROVEMENTS

Although the developed methods are specifically tailored for the corresponding experimental technique, the basic ideas are applicable to many different problems. E.g. it would be very interesting to validate the performance of the feature selection technique developed for the MALDI data also with other classification tasks (e.g. classification of different cancer types based on microarray data). PPINGUIN may be also very interesting for the analysis of metabolite data because the basic situation is comparable. Many metabolites are very similar (e.g. simple derivatives from each other) eventually leading to similar expression patterns. PPINGUIN might also help to resolve ambiguities in the analysis of metabolite data.

Beside the application of the methods to other data sets or other problems, each techniques has several critical points.

7.5.1 *MALDI*

One of the most critical aspects of the MALDI data evaluation chapter is certainly the missing evidence in terms of protein identification. Only a very limited number of peaks has been identified in an follow-up experiment. Still, we have seen that tree peaks identified as hemoglobin are located in close proximity in the cluster dendrogram. This supports our hypothesis that peaks (peptides) derived from the same protein show similar intensity profiles and are located in close proximity in the dendrogram. However, to really validate this hypothesis, additional identification of the peaks in a cluster would be necessary.

Furthermore we have found many peaks with interesting biological patterns (e.g. peaks that are only present in SJL mice). But none of the most significant peaks has been assigned to protein identifiers. So we do not know which proteins are responsible for the observed biological pattern. At the end, for an adequate biological interpretation especially in terms of pathways or modeling the most significant peaks have to be identified. A direct comparison with the results of the other two techniques will require further protein identification.

7.5.2 *iTRAQ*

The central element of PPINGUIN workflow is the clustering of not yet identified spectra. We have applied k-means clustering since it is a well established clustering technique and sufficient for our needs. K-means allocates each point (in our case each spectra) to exactly one cluster. Alternatively, a fuzzy clustering allocates each point (spectra) with a certain probability to all clusters. Such a strategy would certainly improve our results because there might be some spectra located between two clusters. If these spectra are considered simultaneously in different clusters than protein quantitation might be improved.

Furthermore, for the results presented in this thesis we have used a rather restrictive FDR of 0.1%. For a more permissive FDR of 1% or 5% the number of identified proteins would be higher.

During the analysis of the DIGE data we have seen that a common reference pool is very helpful for the analysis especially considering normalization and ratio calculation. Such a common reference pool can also be used for iTRAQ experiments. A common pooled reference would facilitate data normalization strategy and ratio calculation and would increase experimental comparability.

As already discussed, PPINGUIN is certainly more effective with an increasing number of channels (e.g. 8-plex iTRAQ). Application of PPINGUIN to such a dataset would be very promising.

7.5.3 *DIGE*

The number of DIGE gels is rather small since only 10 different gels were analyzed. Especially compared to the > 1100 MALDI spectra this number is very small. With an increasing number of samples the significance of the statistical evaluation increases.

Similar to MALDI the number of spots with identified proteins is limited. At least the most significant spots for mouse genotypes were identified enabling the comparison between DIGE and iTRAQ. However, an increasing number of associated protein accessions also for experimental factor diet facilitates the biological interpretation and increases the comparability with the other two proteomic techniques.

7.6 FUTURE PERSPECTIVE

The three technologies MALDI, iTRAQ and DIGE are only parts of the complete spectrum of proteomic technologies. There are, of cause, many other promising proteomics technologies such as microarray based proteomics, label-free MS/MS approaches or metabolic labeling by SILAC. Especially the latter has proven to be capable of identifying and quantifying several thousands of proteins simultaneously.

However, towards fully understanding complex disease mechanisms such as Type-2 Diabetes Mellitus (T2DM), proteomics is surely essential but other 'omics' technologies (genomics, transcriptomics or metabolics) are required as well.The concurrent investigation from multiple points of view is indispensable to understand complex diseases. In the recent years especially genomics approaches based on next generation sequencing showed a great potential. But also the investigation of metabolites gained more attention. However, a major challenge for the future is the combination and integration of all the different aspects in order to see the whole picture. A starting point for this integration of multi-omics sources may be given by biochemical pathways since they already incorporate metabolites, proteins and genes. Furthermore, we have seen that the overlap between DIGE and iTRAQ is very limited at the level of protein IDs while at the level of pathways we obtained similar results.

A

APPENDIX

EXPERIMENTAL REPLICATES

Two main types of experimental replicates are typically performed: replicated measurements from the same biological sample (technical replications) and measurements of different biological samples with identical combination of experimental factors e.g. same treatment group (biological replications). Both types of replications are a distinct source of noise and as an effect, the observed experimental variance is composed of technical and biological variance. This holds true even if no replicates were performed. The type of replicate affects the outcome of statistical analysis[82] which is discussed in more detail in the following sections.

Impact of Technical Replicates

In this section we describe the effects of technical replicates on the outcome of statistical tests using standard t-test. Let g_i be a hypothetical gene that is differentially expressed in two different disease states (e.g. healthy vs. diseased). The group means of g_i are different: $\mu_{healthy} - \mu_{disease} = \Delta\mu \neq 0$. Knowing the group mean difference ($\Delta\mu$), variances and group size (n), the t-value can be calculated using the t-test formula (for equal variances and equal group sizes):

$$t = \frac{\Delta\mu}{\sqrt{\frac{\sigma^2_{healthy} + \sigma^2_{disease}}{2}} * \sqrt{\frac{2}{n}}}$$

The calculated t-value follows a t-distribution with $2n - 2$ degrees of freedom which can be used to define the corresponding p-value. For example g_i has the following measurements: $(2, 3, 2, 3, 2, 3)$ for healthy group and $(3, 4, 3, 4, 3, 4)$ for disease group. Then $\Delta\mu = 1$, $\sigma = \sqrt{0.5^2 \cdot \frac{6}{5}} = 0.55$, the t-value $= 3.16$ and the corresponding p-value $= 0.01$.

But how does the kind of replication affect the t-test? T-test assumes the sample measurements to be independent and normally distributed. Using technical replicates, the samples are correlated and therefore not independent anymore. Considering the case of $\sigma_{tech} < \sigma_{bio}$ and low number of biological repeats, the resulting distribution is a multi-modal distribution with peaks around the biological repeats. So using technical replicates the data is neither

independent nor distributed normally and hence, t-test must not be applied.

The effects of technical replicates on t-test p-values is demonstrated in a small simulation experiment. We simulate a differentially regulated gene ($\Delta\mu = 1$) with 10 biological ($\sigma_{bio} = 1$) and 10 technical replicates ($\sigma_{tech} = 0.1, 0.5, 1, 2, 5$) and perform a standard t-test. A box plot of p-values for 1000 replications is shown in Figure 63.

The expected t-value for biological samples can be calculated using the t-test formula from above:

$$t = \frac{1}{\sqrt{\frac{1^2+1^2}{2} \cdot \sqrt{\frac{2}{10}}}} = 2.23$$

With the degree of freedom = 18, the expected p-values = 0.038 (horizontal dashed line in Figure 63). Considering only biological replicates (first bar of Figure 63) the simulation experiment is consistent with the expected p-value. For low and medium technical variance the t-test p-value is underestimated strongly. This little simulation shows that the technical replicates indeed affect the statistical tests and substantiates the need for an adequate treatment of technical replicates.

Handling Technical Replicates

The standard approach of handling technical replicates is to calculate the mean value in order to reduce the technical noise. Unfortunately, this can lead to loss of information[147]. We will discuss two different strategies for handling of technical replicated without the loss of information. At first we develop a method to directly calculate the proportion of the variance that is derived from biological variability. Second, we present a standard approach based on mixed models.

Assessing biological Variance

Typically a biologist is interested in biological variance without the technical noise. For assessing biological variance we first investigate the mutual influence of biological and technical variances. Assuming a two-staged process where the biological intensity of the gene g_i is distributed as $N(\mu_{bio}, \sigma_{bio})$ followed by a technical measurement adding an error distributed as $N(0, \sigma_{tech})$, the measured (effective) intensity of g_i is distributed as $N(\mu_{bio}, \sqrt{\sigma^2_{bio} + \sigma^2_{tech}})$. The intensity values for g_i can be written as $m \times n$ matrix with m = number of technological replicates and n = number of biological samples with every row reflecting a replicated experimental run and a column reflecting a mouse individual (see figure 64).

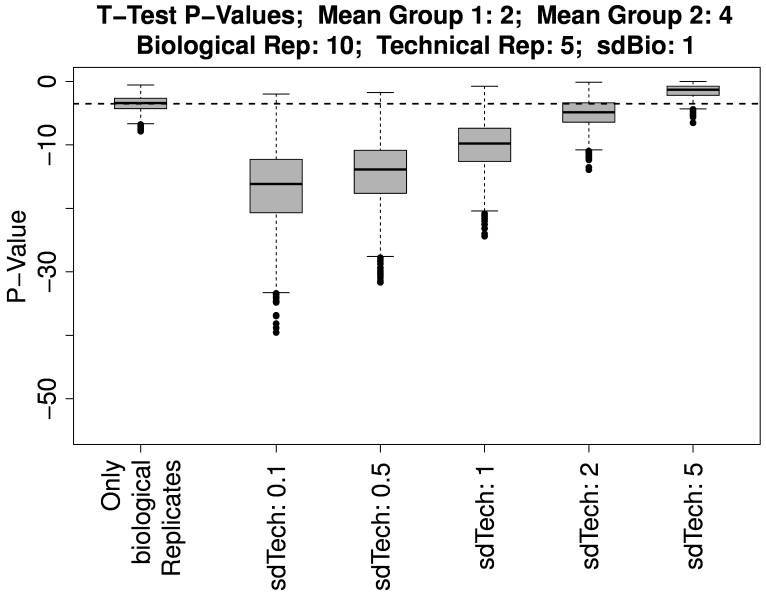


Figure 63: Box plot of simulations of t-test p-values for a differentially regulated gene ($\Delta\mu = 1$) with 10 biological ($\sigma_{bio} = 1$) and 10 technical replicates ($\sigma_{tech} = 0.1, 0.5, 1, 2, 5$). The horizontal line shows the expected p-value (= 0.038) for biological samples. First box refers to biological samples without technical replicates. The p-value is disturbed by technical replicates and depends on the ratio of biological and technical variance. Especially for low technical variance the p-values are estimated too optimistically.

The technical variance can be estimated from the data by calculating the mean of column-wise standard deviation:

$$sd_{tech}(g_i)^2 = \text{mean}\left(\text{var}\left(g_{i,k}\right)\right)$$

$$sd_{tech}(g_i)^2 = \frac{1}{n(m-1)} \sum_{k=1}^{n} \sum_{j=1}^{m} \left(g_{i,j,k} - \mu_{g_{i,k}}\right)^2$$

where j,k refer to a row or a column in the intensity data matrix for g_i (figure 64). To estimate the biological variance we estimate the overall variance and subtract the technological variance:

$$sd_{bio}(g_i)^2 = \left(\text{sd}\left(\text{mean}\left(g_{i,k}\right)\right)\right)^2 - \frac{1}{m} * sd_{tech}(g_i)^2$$

$$sd_{bio}(g_i)^2 = \left(\frac{1}{n-1} \sum_{k=1}^{n} \left(\mu_{g_{i,k}} - \mu_{g_i}\right)^2\right) - sd_{tech}(g_i)^2$$

This allows for assessing the biological variation. But if the technological variance is much bigger than the biological, the estimation of the biological variance becomes critical. The difference

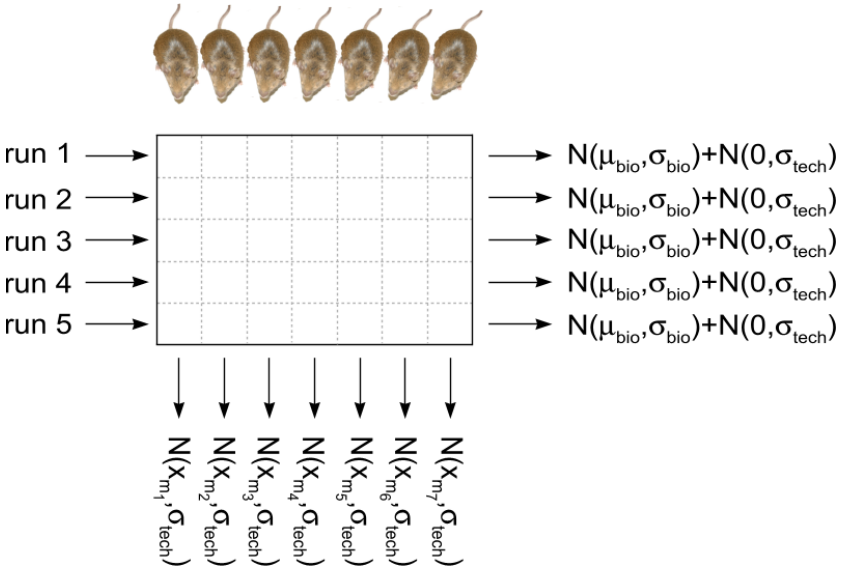


Figure 64: Measurements for a single gene displayed in a matrix of biological (columns) and technical replicates (rows). We assume that σ_{tech} is similar for each experiment and independent of the mouse individual. Technical repeats for a given mouse, e.g. m_1, are distributed $\sim N(x_{m_1}, \sigma_{tech})$, whereas x_{m_1} is the real intensity of the gene in mouse m_1. Biological repeats are composed of biological variance and technical variance: $\sim N(\mu_{bio}, \sigma_{bio}) + N(0, \sigma_{tech})$.

of overall and technical variance is close to 0 and the estimated biological variance might even be estimated as negative. On the contrary, a very high technical variance makes the effective variance to be dominated by the technological error and therefore the biological effect is more or less invisible.

Mixed-effect Model

A more sophisticated way to handle technical replicates without loss of information are mixed-effects models, incorporating fixed-effects parameters applied to the entire population and random effects applied to particular experimental units or subunits (e.g. technical replicates). Mixed effect models have already been applied for a variety of data analysis tasks including proteomics[108, 36]. The `lme4` package[10, 11] for the open-source language and environment for statistical computing `R`[128] offers fast and reliable algorithms for parameter estimation and model evaluation. For theoretical background of mixed effect models and exemplary application of `lme4` package see Baayen et al..

Comparing the different Methods

We will compare the three methods presented above:

1. averaging technical replicates (mean)

2. mixed models (R package `lme4`)
3. directly assessing biological variance as presented in Section A.

To compare the three methods we continued the simulation experiment (c.f. section A) of a differentially expressed biomolecule, evaluated using the three methods. Simulations were performed with 1000 replications with the following properties: 10 biological replicates, 5 technical replicates, mean difference of $\Delta x = 2$, biological variance of sdBio $= 1$ and three different technical variances: sdTech $= \{0.5, 1, 2\}$.

Using mean difference, biological variance and the number of biological samples, the expected p-value (for the biological effect) can be calculated. Figure 65 shows a box plot of the calculated p-values of the three methods for three different technical variances (low, medium and high technical variance). The horizontal dashed line represents the expected p-value of the biological effect. For low technical variance all three methods are similar and the mean p-value for every method is close to the expected value. For higher technical variances the resulting p-values of the three methods rather are different. Averaging technical replicates and mixed models result in more conservative p-values while the direct estimation of biological variance seems to result in a good average p-values close to the expected p-value. However, variance of the p-values is strongly increased when trying to estimate biological variances which leads to an increased number of too optimistic p-values. Furthermore, this method is vulnerable to higher technical variances as the estimated biological variance might become negative, which makes the statistical test impossible. This effect occurred 3 times (0.3%) for medium technical variance and 54 times (5.4%) for high technical variance.

This simulation experiment suggests that the standard approach of handling technical replicates by averaging is rather similar to more sophisticated approaches like mixed models. Although the p-values estimation is too conservative especially for high technical variance, averaging technical replicates is preferable to a direct estimation of biological variance since the latter leads to a high variance of p-values and thus increases the number of false positive hits. Due to these reasons and because of the simplicity of the approach, technical replicates are averaged prior to statistical testing in any statistical evaluation presented in this thesis.

IMPLEMENTATION OF ACO

The feature selection approach using ACO by Ressom et al.[133] was implemented in MATLAB. Following this ACO approach

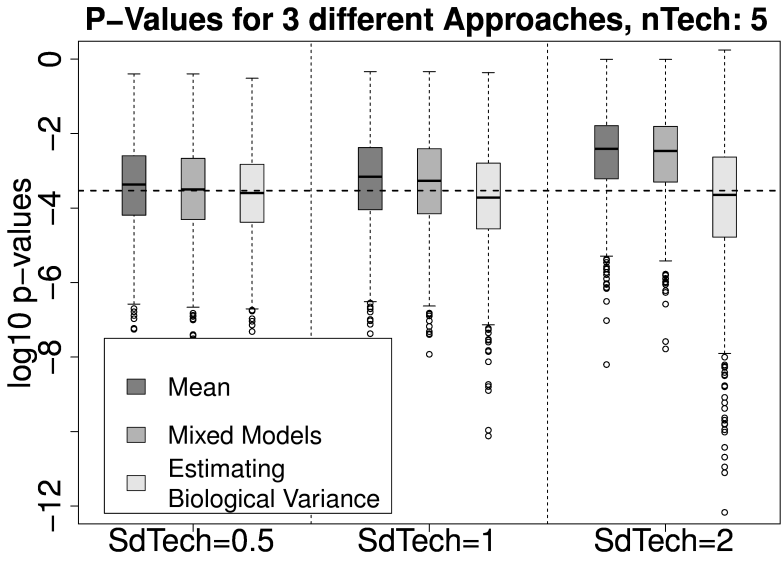


Figure 65: Comparison of the three different methods to handle technical replicates: 1. averaging technical replicates (mean), 2. mixed models (R package `lme4`), 3. directly assessing biological variance as presented in Section A. Box plot of p-values for 1000 simulation experiments with 10 biological replicates, 5 technical replicates, mean difference of $\Delta x = 2$, sdBio $= 1$ and three different technical variances: 0.5, 1 and 2. Horizontal dotted line represents the correct p-value for biological effect without considering technical variance. The third method of directly estimating biological variance results in a good average p-values while the other two methods show more conservative p-values especially for high technical variance. But the third methods leads to an increased range of p-values and an increased number of too optimistic p-values.

we implemented ACO-based feature selection in an in-house R-package. Basically our implementation is similar to Ressom et al. but beside SVM classification can be performed using RF or Bayes classifier as well as logistic regression. The header of the R function and input parameter description extracted from package documentation is given below:

```
performFeatureSelection (data,classes,
    cv=1:length(classes), classifier = "randomForest",
    decay=0.5, nFeatures=1:min(5,ncol(data)), nAnts=100,
    nIter=50,cpus=NULL)
```

`data` Feature matrix.

`classes` A vector with class identifiers.

`cv` This influences the cross-validation. The cross validation will be performed according to cv vector. Default: leave one out

`classifier` The used classifier. Possible values are: randomForest, svm, bayes, plr, LogitBoost. Default: randomForest

`dacay` The percentage of the pheromone that remains after decay after each iteration.

`nFeatures` Vector with number of features for classification. Every ant decides how much features it takes by taking one value of this vector. A single number forces to a fixed feature Number.

`nAnts` Number of ants

`nIter` Number of iterations

`cpus` Enables multi cpu usage via snowfall package

ALGORITHM FOR PEAK MATCHING

Algorithm 5 Algorithm for peak matching of MALDI data. The peaks are aligned to a reference profile (mean spectrum). The alignment is performed by an index shift.

```
1: function getMeanSpectrum(SpectraMat_{i=1..n,l=1..m})
2:     RefSpec ← SpectraMat_{1,}
3:     for i = 2 to n do                                   ▷ every spectrum
4:         for l = 1 to m do                               ▷ length of a spectrum
5:             RefSpec_l ← RefSpec_l + SpectraMat_{i,l}
6:         end for
7:     end for
8:     for l = 1 to m do
9:         RefSpec_l ← RefSpec_l/m
10:    end for
11:    return RefSpec
12: end function
13:
14: function getAlignedSpecta(SpectraMat_{i=1..n,l=1..m})
15:    ▷ calculation reference peaks and distances to reference peaks
16:    RefSpec ← getMeanSpectrum(SpectraMat)
17:    RefPeakList_{1..nrp} ← GaussCorPeakPicking(RefSpec)   ▷ 43 peaks
18:    Distances_{i=1..n,j=1..nrp} ← NA                      ▷ Store Distances
19:    for i = 1 to n do                                     ▷ every spectrum
20:        for j = 1 to nrp do                               ▷ every reference peak
21:            peak ← GaussCorPeakPicking(SpectraMat_{i,j−d..j+d})
22:            ▷ d reflects a small environment around the peak (d = 40)
23:            if peak ≠ NA then
24:                Distances_{i,j} ← peak − RefPeakList_j
25:            end if
26:        end for
27:    end for
28:
29:    ▷ calculating average of distances for every spectra
30:    Displacements_{i=1..n} ← 0          ▷ Displace indices for each spectrum
31:    maxDis ← 0                                            ▷ maximal Distance
32:    for i = 1 to n do                                     ▷ every spectrum
33:        for j = 1 to nrp do                               ▷ every reference peak
34:            Displacements_i ← Displacements_i + Distances_{i,j}
35:        end for
36:    end for
37:    for i = 1 to n do                                     ▷ every spectrum
38:        Displacements_i ← Displacements_i/npr
39:        if Displacements_i > maxDis then
40:            maxDis ← Displacements_i
41:        end if
42:    end for
43:
44:    ▷ index shift according to displacements
45:    AlignedMat_{i=1..n,l=1..(m+2·maxDis)} ← NA
46:    for i = 1 to n do                                     ▷ every spectrum
47:        for l = 1 to m do                                 ▷ length of a spectrum
48:            AlignedMat_{i,l+maxDis+Displacements_i} ← SpectraMat_{i,l}
49:        end for
50:    end for
51:    return RefSpec
52: end function
```

CURRICULUM VITAE

For data privacy reasons, this online version does not contain the CV.

BIBLIOGRAPHY

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, Mar 2003. [DOI:10.1038/nature01511] [PubMed:12634793]. (Cited on pages 8 and 27.)

[2] E. Ahrne, M. Muller, and F. Lisacek. Unrestricted identification of modified proteins using MS/MS. *Proteomics*, 10:671–686, Feb 2010. [DOI:10.1002/pmic.200900502] [PubMed:20029840]. (Cited on page 83.)

[3] A. Alban, S. O. David, L. Bjorkesten, C. Andersson, E. Sloge, S. Lewis, and I. Currie. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, 3:36–44, Jan 2003. [DOI:10.1002/pmic.200390006] [PubMed:12548632]. (Cited on pages 11 and 115.)

[4] T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. Tollenaar, P. Maass, and H. Thiele. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25: 643–649, Mar 2009. [PubMed Central:PMC2647828] [DOI:10.1093/bioinformatics/btn662] [PubMed:19244390]. (Cited on page 8.)

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, pages 245–248, 2009. (Cited on page 77.)

[6] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, L. Bower, P. Browne, W. M. Chan, E. Dimmer, R. Eberhardt, A. Fedotov, R. Foulger, J. Garavelli, R. Huntley, J. Jacobsen, M. Kleen, K. Laiho, R. Leinonen, D. Legge, Q. Lin, W. Liu, J. Luo, S. Orchard, S. Patient, D. Poggioli, M. Pruess, M. Corbett, G. di Martino, M. Donnelly, P. van Rensburg, A. Bairoch, L. Bougueleret, I. Xenarios, S. Altairac, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, L. Bollondi, E. Boutet, S. B. Quintaje, L. Breuza, A. Bridge, E. deCastro, L. Ciapina, D. Coral, E. Coudert, I. Cusin, G. Delbard, M. Doche, D. Dornevil, P. D. Roggli, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, S. Gehant, N. Farriol-Mathis, S. Ferro, E. Gasteiger,

A. Gateau, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hulo, J. James, S. Jimenez, F. Jungo, T. Kappler, G. Keller, C. Lachaize, L. Lane-Guermonprez, P. Langendijk-Genevaux, V. Lara, P. Lemercier, D. Lieberherr, T. de Oliveira Lima, V. Mangold, X. Martin, P. Masson, M. Moinat, A. Morgat, A. Mottaz, S. Paesano, I. Pedruzzi, S. Pilbout, V. Pillet, S. Poux, M. Pozzato, N. Redaschi, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, E. Stanley, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, L. Yip, L. Zuletta, C. Wu, C. Arighi, L. Arminski, W. Barker, C. Chen, Y. Chen, Z. Z. Hu, H. Huang, R. Mazumder, P. McGarvey, D. A. Natale, J. Nchoutmboube, N. Petrova, N. Subramanian, B. E. Suzek, U. Ugochukwu, S. Vasudevan, C. R. Vinayaka, L. S. Yeh, and J. Zhang. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38:D142–148, Jan 2010. [PubMed Central:PMC2808944] [DOI:10.1093/nar/gkp846] [PubMed:19843607]. (Cited on page 2.)

[7] C. Arima, K. Hakamada, M. Okamoto, and T. Hanai. Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering. *J. Biosci. Bioeng.*, 105:273–281, Mar 2008. [DOI:10.1263/jbb.105.273] [PubMed:18397779]. (Cited on page 77.)

[8] R.H. Baayen, D.J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390 – 412, 2008. ISSN 0749-596X. doi: DOI:10.1016/j.jml.2007.12.005. Special Issue: Emerging Data Analysis. (Cited on page 148.)

[9] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, 37:D885–890, Jan 2009. [PubMed Central:PMC2686538] [DOI:10.1093/nar/gkn764] [PubMed:18940857]. (Cited on page 105.)

[10] Douglas Bates. Fitting linear mixed models in R. *R News*, 5(1):27–30, May 2005. URL `http://CRAN.R-project.org/doc/Rnews/`. (Cited on page 148.)

[11] Douglas Bates and Deepayan Sarkar. *lme4: Linear mixed-effects models using S4 classes*, 2007. (Cited on page 148.)

[12] C. Bauer, R. Cramer, and J. Schuchhardt. Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol. Biol.*, 696:341–352, 2011.

[13] C. Bauer, F. Kleinjung, C. J. Smith, M. W. Towers, A. Tiss, A. Chadt, T. Dreja, D. Beule, H. Al-Hasani, K. Reinert, J. Schuchhardt, and R. Cramer. Biomarker discovery and redundancy reduction towards classification using a multifactorial MALDI-TOF MS T2DM mouse model dataset. *BMC Bioinformatics*, 12:140, 2011.

[14] C. Bauer, F. Kleinjung, D. Ruthishauser, C. Panse, A. Chadt, T. Dreja, H. Al-Hasani, K. Reinert, R. Schlapbach, and J. Schuchhardt. PPINGUIN: Peptide Profiling Guided Identification of Proteins Improves Quantitation of iTRAQ Ratios. *BMC Bioinformatics*, 13:34, Feb 2012.

[15] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL `http://dx.doi.org/10.2307/2346101`. (Cited on page 78.)

[16] N. Bluthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, and D. Beule. Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform*, 16:106–115, 2005. [PubMed:16362912]. (Cited on page 134.)

[17] A. M. Boehm, S. Putz, D. Altenhofer, A. Sickmann, and M. Falk. Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinformatics*, 8:214, 2007. [PubMed Central:PMC1940031] [DOI:10.1186/1471-2105-8-214] [PubMed:17584939]. (Cited on pages 11 and 68.)

[18] Leo Breiman. Random Forests. *Machine Learning*, 45(1): 5–32–32, October 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL `http://dx.doi.org/10.1023/A:1010933404324`. (Cited on page 58.)

[19] I. A. Brewis and P. Brennan. Proteomics technologies for the global identification and quantification of proteins. *Adv Protein Chem Struct Biol*, 80:1–44, 2010. [DOI:10.1016/B978-0-12-381264-3.00001-1] [PubMed:21109216]. (Cited on page 1.)

[20] A. K. Callesen, W. Vach, P. E. Jurgensen, S. Cold, O. Mogensen, T. A. Kruse, O. N. Jensen, and J. S. Madsen. Reproducibility of mass spectrometry based protein profiles for diagnosis of breast cancer across clinical studies: a systematic review. *J. Proteome Res.*, 7:1395–1402, Apr 2008. (Cited on page 66.)

[21] S. M. Carlson, A. Najmi, and H. J. Cohen. Biomarker clustering to address correlations in proteomic data. *Proteomics*,

7:1037–1046, Apr 2007. [DOI:10.1002/pmic.200600514] [PubMed:17390293]. (Cited on page 41.)

[22] J. Casado-Vela, M. J. Martinez-Esteso, E. Rodriguez, E. Borras, F. Elortza, and R. Bru-Martinez. iTRAQ-based quantitative analysis of protein mixtures with large fold change and dynamic range. *Proteomics*, 10:343–347, Jan 2010. (Cited on page 67.)

[23] A. Chadt, K. Leicht, A. Deshmukh, L. Q. Jiang, S. Scherneck, U. Bernhardt, T. Dreja, H. Vogel, K. Schmolz, R. Kluge, J. R. Zierath, C. Hultschig, R. C. Hoeben, A. Schurmann, H. G. Joost, and H. Al-Hasani. Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. *Nat. Genet.*, 40:1354–1359, Nov 2008. [DOI:10.1038/ng.244] [PubMed:18931681]. (Cited on page 16.)

[24] H. L. Chan, S. Gharbi, P. R. Gaffney, R. Cramer, M. D. Waterfield, and J. F. Timms. Proteomic analysis of redox- and ErbB2-dependent changes in mammary luminal epithelial cells using cysteine- and lysine-labelling two-dimensional difference gel electrophoresis. *Proteomics*, 5:2908–2926, Jul 2005. [DOI:10.1002/pmic.200401300] [PubMed:15954156]. (Cited on pages 11 and 116.)

[25] L. H. Choe, K. Aggarwal, Z. Franck, and K. H. Lee. A comparison of the consistency of proteome quantitation using two-dimensional electrophoresis and shotgun isobaric tagging in Escherichia coli cells. *Electrophoresis*, 26:2437–2449, Jun 2005. [DOI:10.1002/elps.200410336] [PubMed:15924362]. (Cited on pages 68 and 72.)

[26] T. A. Clark, A. C. Schweitzer, T. X. Chen, M. K. Staples, G. Lu, H. Wang, A. Williams, and J. E. Blume. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, 8:R64, 2007. [PubMed Central:PMC1896007] [DOI:10.1186/gb-2007-8-4-r64] [PubMed:17456239]. (Cited on pages 1 and 106.)

[27] W.S. Cleveland, E. Grosse, and W.M. Shyu. *Local regression models*, pages pp. 309–376. Wadsworth, 1992. (Cited on page 29.)

[28] Alberto Colorni, Marco Dorigo, and Vittorio Maniezzo. Distributed optimization by ant colonies. In *European Conference on Artificial Life*, pages 134–142, 1991. (Cited on page 21.)

[29] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M. C. Hung, and H. 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*, 5:4107–4117, Nov 2005. [DOI:10.1002/pmic.200401261] [PubMed:16254928]. (Cited on pages 29 and 30.)

[30] J. Cox and M. Mann. Is proteomics the new genomics? *Cell*, 130:395–398, Aug 2007. [DOI:10.1016/j.cell.2007.07.032] [PubMed:17693247]. (Cited on pages 2 and 7.)

[31] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26:1367–1372, Dec 2008. [DOI:10.1038/nbt.1511] [PubMed:19029910]. (Cited on page 3.)

[32] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466–1467, Jun 2004. (Cited on pages 70 and 73.)

[33] R. Cramer, J. Gobom, and E. Nordhoff. High-throughput proteomics using matrix-assisted laser desorption/ ionization mass spectrometry. *Expert Rev Proteomics*, 2:407–420, Jun 2005. [DOI:10.1586/14789450.2.3.407] [PubMed:16000086]. (Cited on page 8.)

[34] M. J. Crawley. *Statistics. An Introduction using R*. Wiley, New York, NY, 2005. (Cited on page 17.)

[35] D. M. Creasy and J. S. Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4:1534–1536, Jun 2004. (Cited on pages 78 and 85.)

[36] D. S. Daly, K. K. Anderson, E. A. Panisko, S. O. Purvine, R. Fang, M. E. Monroe, and S. E. Baker. Mixed-effects statistical model for comparative LC-MS proteomics studies. *J. Proteome Res.*, 7:1209–1217, Mar 2008. [DOI:10.1021/pr070441i] [PubMed:18251496]. (Cited on page 148.)

[37] M. E. de Noo, B. J. Mertens, A. Ozalp, M. R. Bladergroen, M. P. van der Werff, C. J. van de Velde, A. M. Deelder, and R. A. Tollenaar. Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur. J. Cancer*, 42:1068–1076, May 2006. [DOI:10.1016/j.ejca.2005.12.023] [PubMed:16603345]. (Cited on page 8.)

[38] M. Dorigo, V. Maniezzo, and A. Colorni. Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern B Cybern*, 26:29–41, 1996.

[DOI:10.1109/3477.484436] [PubMed:18263004]. (Cited on page 21.)

[39] T. Dreja, Z. Jovanovic, A. Rasche, R. Kluge, R. Herwig, Y. C. Tung, H. G. Joost, G. S. Yeo, and H. Al-Hasani. Diet-induced gene expression of isolated pancreatic islets from a polygenic mouse model of the metabolic syndrome. *Diabetologia*, 53:309–320, Feb 2010. [PubMed Central:PMC2797618] [DOI:10.1007/s00125-009-1576-4] [PubMed:19902174]. (Cited on pages 17, 51, and 62.)

[40] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22: 2059–2065, Sep 2006. [DOI:10.1093/bioinformatics/btl355] [PubMed:16820428]. (Cited on pages 29 and 32.)

[41] Sandrine Dudoit, Juliet P. Shaffer, and Jennifer C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1):71–103, 2003. doi: 10.2307/3182872. URL `http://dx.doi.org/10.2307/3182872`. (Cited on pages 78 and 118.)

[42] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, Dec 1998. [PubMed Central:PMC24541] [PubMed:9843981]. (Cited on page 126.)

[43] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4:207–214, Mar 2007. (Cited on pages 73 and 113.)

[44] J. Eriksson and D. Fenyo. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J. Proteome Res.*, 3:32–36, 2004. [PubMed:14998160]. (Cited on page 71.)

[45] Y. Fan, T. B. Murphy, and R. W. Watson. digeR: a graphical user interface R package for analyzing 2D-DIGE data. *Bioinformatics*, 25:3033–3034, Nov 2009. [DOI:10.1093/bioinformatics/btp514] [PubMed:19706743]. (Cited on page 128.)

[46] D. Fenyo and R. C. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75: 768–774, Feb 2003. [PubMed:12622365]. (Cited on page 70.)

[47] J. Fievet, C. Dillmann, G. Lagniel, M. Davanture, L. Negroni, J. Labarre, and D. de Vienne. Assessing factors for reliable quantitative proteomics based on two-dimensional gel electrophoresis. *Proteomics*, 4:1939–1949, Jul 2004. [DOI:10.1002/pmic.200300731] [PubMed:15221754]. (Cited on page 66.)

[48] Finnigan. Maldi-tof mass analysis, 2009. [WebSite: www.biotech.iastate.edu/facilities/protein/maldi.html]. (Cited on page 9.)

[49] R. A. Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):pp. 87–94, 1922. ISSN 09528385. URL `http://www.jstor.org/stable/2340521`. (Cited on page 134.)

[50] P. Flicek, B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y. A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S. P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, J. Smith, and S. M. Searle. Ensembl's 10th year. *Nucleic Acids Res.*, 38:D557–562, Jan 2010. [PubMed Central:PMC2808936] [DOI:10.1093/nar/gkp972] [PubMed:19906699]. (Cited on page 1.)

[51] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, Feb 2007. [DOI:10.1126/science.1136800] [PubMed:17218491]. (Cited on page 111.)

[52] D. B. Friedman, S. Hill, J. W. Keller, N. B. Merchant, S. E. Levy, R. J. Coffey, and R. M. Caprioli. Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics*, 4:793–811, Mar 2004. [DOI:10.1002/pmic.200300635] [PubMed:14997500]. (Cited on page 116.)

[53] T. Frohlich, D. Helmstetter, M. Zobawa, A. C. Crecelius, T. Arzberger, H. A. Kretzschmar, and G. J. Arnold. Analysis of the HUPO Brain Proteome reference samples using 2-D DIGE and 2-D LC-MS/MS. *Proteomics*, 6:4950–4966, Sep 2006. [DOI:10.1002/pmic.200600079] [PubMed:16927427]. (Cited on page 116.)

[54] A. Gamez-Pozo, I. Sanchez-Navarro, M. Nistal, E. Calvo, R. Madero, E. Diaz, E. Camafeita, J. de Castro, J. A. Lopez, M. Gonzalez-Baron, E. Espinosa, and J. A. Fresno Vara. MALDI profiling of human lung cancer subtypes. *PLoS ONE*, 4:e7731, 2009. [PubMed Central:PMC2767501] [DOI:10.1371/journal.pone.0007731] [PubMed:19890392]. (Cited on page 8.)

[55] C. S. Gan, P. K. Chong, T. K. Pham, and P. C. Wright. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J. Proteome Res.*, 6:821–827, Feb 2007. (Cited on pages 66, 69, and 112.)

[56] G. Ge and G. W. Wong. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9:275, 2008. [PubMed Central:PMC2440392] [DOI:10.1186/1471-2105-9-275] [PubMed:18547427]. (Cited on page 33.)

[57] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3:958–964, 2004. [DOI:10.1021/pr0499491] [PubMed:15473683]. (Cited on page 70.)

[58] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80, 2004. [PubMed Central:PMC545600] [DOI:10.1186/gb-2004-5-10-r80] [PubMed:15461798]. (Cited on page 30.)

[59] M. Gentzel, T. Kocher, S. Ponnusamy, and M. Wilm. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*, 3:1597–1610, Aug 2003. [DOI:10.1002/pmic.200300486] [PubMed:12923784]. (Cited on page 71.)

[60] S. Gharbi, P. Gaffney, A. Yang, M. J. Zvelebil, R. Cramer, M. D. Waterfield, and J. F. Timms. Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Mol. Cell Proteomics*, 1:91–98, Feb 2002. [PubMed:12096126]. (Cited on page 126.)

[61] K. Gruden, M. Hren, A. Herman, A. Blejec, T. Albrecht, J. Selbig, C. Bauer, J. Schuchardt, M. Or-Guil, K. Zupancic,

U. Svajger, B. Stabuc, A. Ihan, A. N. Kopitar, M. Ravnikar, M. Knezevic, P. Rozman, and M. Jeras. A 'crossomics' study analysing variability of different components in peripheral blood of healthy caucasoid individuals. *PLoS ONE*, 7(1): e28761, 2012.

[62] M Guilhaus. Principles and Instrumentation in Time-of-flight Mass Spectrometry. *JOURNAL OF MASS SPECTROMETRY*, 30:1519–1532, 1995. (Cited on page 8.)

[63] I.M. Guyon, S.R. Gunn, M. Nikravesh, and L. Zadeh. *Feature Extraction, Foundations and Applications*. Springer, 1 edition, 2006. (Cited on page 21.)

[64] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, 17:994–999, Oct 1999. [DOI:10.1038/13690] [PubMed:10504701]. (Cited on pages 8 and 66.)

[65] J. Hakkinen, G. Vincic, O. Mansson, K. Warell, and F. Levander. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.*, 8:3037–3043, Jun 2009. [DOI:10.1021/pr900189c] [PubMed:19354269]. (Cited on page 117.)

[66] David J. Hand and Robert J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, November 2001. ISSN 0885-6125. doi: 10.1023/A:1010920819831. URL `http://dx.doi.org/10.1023/A:1010920819831`. (Cited on page 34.)

[67] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979. ISSN 00359254. [JSTOR:2346830]. (Cited on page 74.)

[68] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2 edition, 2009. (Cited on pages 41 and 74.)

[69] Z. He, H. Zhao, and W. Yu. Score regularization for peptide identification. *BMC Bioinformatics*, 12 Suppl 1:S2, 2011. [PubMed Central:PMC3044274] [DOI:10.1186/1471-2105-12-S1-S2] [PubMed:21342549]. (Cited on pages 68, 93, and 111.)

[70] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty. Optimal number of features as a function of sample size

for various classification rules. *Bioinformatics*, 21:1509–1515, Apr 2005. (Cited on page 58.)

[71] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37:1–13, Jan 2009. [PubMed Central:PMC2615629] [DOI:10.1093/nar/gkn923] [PubMed:19033363]. (Cited on page 133.)

[72] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104, 2002. (Cited on page 77.)

[73] R. A. amd Bhattacharyya G. K. Johnson. *Statistics: Principles and Methods*. John Wiley & Sons, 6 edition, 2009. (Cited on page 17.)

[74] K. Jung, A. Gannoun, B. Sitek, O. Apostolov, A. Schramm, H. Meyer, K. StÃ¼hler, , and Urfer W. Statistical Evaluation of Methods for the Analysis of Dynamic Protein Expression Data From a Tumor Study. *RevStat-Statistical Journal*, 4:67–80, 2006. (Cited on page 117.)

[75] H. S. Jurgens, A. Schurmann, R. Kluge, S. Ortmann, S. Klaus, H. G. Joost, and M. H. Tschop. Hyperphagia, lower body temperature, and reduced running wheel activity precede development of morbid obesity in New Zealand obese mice. *Physiol. Genomics*, 25:234–241, Apr 2006. [DOI:10.1152/physiolgenomics.00252.2005] [PubMed:16614459]. (Cited on page 16.)

[76] H. S. Jurgens, S. Neschen, S. Ortmann, S. Scherneck, K. Schmolz, G. Schuler, S. Schmidt, M. Bluher, S. Klaus, D. Perez-Tilve, M. H. Tschop, A. Schurmann, and H. G. Joost. Development of diabetes in obese, insulin-resistant mice: essential role of dietary carbohydrate in beta cell destruction. *Diabetologia*, 50:1481–1489, Jul 2007. [DOI:10.1007/s00125-007-0662-8] [PubMed:17437079]. (Cited on pages 17, 51, and 62.)

[77] L. Kall, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, 7: 29–34, Jan 2008. (Cited on page 73.)

[78] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and

M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34: D354–357, Jan 2006. [PubMed Central:PMC1347464] [DOI:10.1093/nar/gkj102] [PubMed:16381885]. (Cited on page 134.)

[79] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes*, 78:53 – 68, 1987. ISSN 0168-1176. doi: DOI:10.1016/0168-1176(87)87041-6. URL `http://www.sciencedirect.com/science/article/B6TG6-44KR0VJ-5/2/d9879520d4b899ed53c57ad1d3dd6c99`. (Cited on page 8.)

[80] N. A. Karp and K. S. Lilley. Design and analysis issues in quantitative proteomics studies. *Proteomics*, 7 Suppl 1:42–50, Sep 2007. [DOI:10.1002/pmic.200700683] [PubMed:17893850]. (Cited on page 116.)

[81] N. A. Karp, D. P. Kreil, and K. S. Lilley. Determining a significant change in protein expression with DeCyder during a pair-wise comparison using two-dimensional difference gel electrophoresis. *Proteomics*, 4:1421–1432, May 2004. [DOI:10.1002/pmic.200300681] [PubMed:15188411]. (Cited on page 117.)

[82] N. A. Karp, M. Spencer, H. Lindsay, K. O'Dell, and K. S. Lilley. Impact of replicate types on proteomic expression analysis. *J. Proteome Res.*, 4:1867–1871, 2005. [DOI:10.1021/pr050084g] [PubMed:16212444]. (Cited on pages 39 and 145.)

[83] N. A. Karp, W. Huber, P. G. Sadowski, P. D. Charles, S. V. Hester, and K. S. Lilley. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics*, Apr 2010. (Cited on pages 72, 75, 82, and 112.)

[84] M. Katayama, H. Nakano, A. Ishiuchi, W. Wu, R. Oshima, J. Sakurai, H. Nishikawa, S. Yamaguchi, and T. Otsubo. Protein pattern difference in the colon cancer cell lines examined by two-dimensional differential in-gel electrophoresis and mass spectrometry. *Surg. Today*, 36:1085–1093, 2006. [DOI:10.1007/s00595-006-3301-y] [PubMed:17123137]. (Cited on page 116.)

[85] J. Klose. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, 26:231–243, 1975. [PubMed:1093965]. (Cited on page 10.)

[86] M. Krogh, Y. Liu, S. Waldemarson, B. Valastro, and P. James. Analysis of DIGE data using a linear mixed model allowing for protein-specific dye effects. *Proteomics*, 7:4235–4244, Dec 2007. [DOI:10.1002/pmic.200700339] [PubMed:17979174]. (Cited on pages 117, 121, and 128.)

[87] K. Kultima, B. Scholz, H. Alm, K. Skold, M. Svensson, A. R. Crossman, E. Bezard, P. E. Andren, and I. Lonnstedt. Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: a proteomic study of L-DOPA induced dyskinesia in an animal model of Parkinson's disease using DIGE. *BMC Bioinformatics*, 7:475, 2006. [PubMed Central:PMC1635739] [DOI:10.1186/1471-2105-7-475] [PubMed:17067368]. (Cited on pages 11 and 117.)

[88] D. Kwon, M. Vannucci, J. J. Song, J. Jeong, and R. M. Pfeiffer. A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise. *Proteomics*, 8: 3019–3029, Aug 2008. [PubMed Central:PMC2855839] [DOI:10.1002/pmic.200701010] [PubMed:18615428]. (Cited on page 29.)

[89] C. M. Lacerda, L. Xin, I. Rogers, and K. F. Reardon. Analysis of iTRAQ data using Mascot and Peaks quantification algorithms. *Brief Funct Genomic Proteomic*, 7:119–126, Mar 2008. (Cited on page 67.)

[90] T. Laderas, C. Bystrom, D. McMillen, G. Fan, and S. McWeeney. TandTRAQ: an open-source tool for integrated protein identification and quantitation. *Bioinformatics*, 23:3394–3396, Dec 2007. (Cited on page 72.)

[91] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl,

K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, Feb 2001. [DOI:10.1038/35057062] [PubMed:11237011]. (Cited on page 1.)

[92] E. Lange, C. Gropl, K. Reinert, O. Kohlbacher, and A. Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac Symp Biocomput*, pages 243–254, 2006. [PubMed:17094243]. (Cited on pages 29 and 32.)

[93] K. H. Lee. Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol.*, 19:217–222, Jun 2001. [PubMed:11356283]. (Cited on page 66.)

[94] W. Li, L. Ji, J. Goya, G. Tan, and V. H. Wysocki. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. *J. Proteome Res.*, 10:1593–1602, Apr 2011. [DOI:10.1021/pr100959y] [PubMed:21204564]. (Cited on page 111.)

[95] X. J. Li, H. Zhang, J. A. Ranish, and R. Aebersold. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.*, 75:6648–6657, Dec 2003. [DOI:10.1021/ac034633i] [PubMed:14640741]. (Cited on pages 68 and 72.)

[96] K. Lilley and D. Friedman. Difference gel electrophoresis dige. *Drug Discovery Today: Technologies*, 3(3):347 – 353, 2006. ISSN 1740-6749. doi: DOI:10.1016/j.ddtec.2006.09.013. URL `http://www.sciencedirect.com/science/article/B75D6-4M51FHN-3/2/03d7389f8ff2e152742117964d92c68c`. (Cited on page 118.)

[97] J. Liu, W. Yu, B. Wu, and H. Zhao. Bayesian mass spectra peak alignment from mass charge ratios. *Cancer Inform*, 6:217–241, 2008. [PubMed Central:PMC2623297] [PubMed:19259411]. (Cited on pages 33, 36, and 38.)

[98] Q. Liu, A. H. Sung, M. Qiao, Z. Chen, J. Y. Yang, M. Q. Yang, X. Huang, and Y. Deng. Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics*, 10 Suppl 1:S3, 2009. [PubMed Central:PMC2709264] [DOI:10.1186/1471-2164-10-S1-S3] [PubMed:19594880]. (Cited on page 21.)

[99] X. Liu, Q. Feng, Y. Chen, J. Zuo, N. Gupta, Y. Chang, and F. Fang. Proteomics-based identification of differentially-expressed proteins including galectin-1 in the blood plasma of type 2 diabetic patients. *J. Proteome Res.*, 8:1255–1262, Mar 2009. [DOI:10.1021/pr800850a] [PubMed:19125585]. (Cited on pages 14 and 33.)

[100] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17:2337–2342, 2003. (Cited on page 67.)

[101] M. Mann. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.*, 7:952–958, Dec 2006.

[DOI:10.1038/nrm2067] [PubMed:17139335]. (Cited on pages 8 and 66.)

[102] M. Mann and O. N. Jensen. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21:255–261, Mar 2003. [DOI:10.1038/nbt0303-255] [PubMed:12610572]. (Cited on page 2.)

[103] D. Mantini, F. Petrucci, D. Pieragostino, P. Del Boccio, M. Di Nicola, C. Di Ilio, G. Federici, P. Sacchetta, S. Comani, and A. Urbani. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics*, 8:101, 2007. [PubMed Central:PMC1847688] [DOI:10.1186/1471-2105-8-101] [PubMed:17386085]. (Cited on page 30.)

[104] M. Martinez-Gomariz, M. L. Hernaez, D. Gutierrez, P. Ximenez-Embun, and G. Prestamo. Proteomic analysis by two-dimensional differential gel electrophoresis (2D DIGE) of a high-pressure effect in Bacillus cereus. *J. Agric. Food Chem.*, 57:3543–3549, May 2009. [DOI:10.1021/jf803272a] [PubMed:19338277]. (Cited on page 118.)

[105] A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6:386–398, May 2005. [DOI:10.1038/nrm1645] [PubMed:15956978]. (Cited on page 106.)

[106] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37:D619–622, Jan 2009. [PubMed Central:PMC2686536] [DOI:10.1093/nar/gkn863] [PubMed:18981052]. (Cited on page 134.)

[107] J. N. McGuire, J. Overgaard, and F. Pociot. Mass spectrometry is only one piece of the puzzle in clinical proteomics. *Brief Funct Genomic Proteomic*, 7:74–83, Jan 2008. [DOI:10.1093/bfgp/eln005] [PubMed:18308835]. (Cited on page 8.)

[108] C. Mercier, C. Truntzer, D. Pecqueur, J. P. Gimeno, G. Belz, and P. Roy. Mixed-model of ANOVA for measurement reproducibility in proteomics. *J Proteomics*, 72:974–981, Aug 2009. [DOI:10.1016/j.jprot.2009.05.005] [PubMed:19481188]. (Cited on pages 66 and 148.)

[109] F. Meyer. *Cytologie quantitative et morphologie mathï¿½matiques*. PhD thesis, Ecole des Mines de Paris, 1979. (Cited on pages 29 and 30.)

[110] J. S. Minden, S. R. Dowd, H. E. Meyer, and K. Stuhler. Difference gel electrophoresis. *Electrophoresis*, 30 Suppl 1:S156–161, Jun 2009. [DOI:10.1002/elps.200900098] [PubMed:19517495]. (Cited on pages 12, 116, and 128.)

[111] D. Mossman. Three-way ROCs. *Med Decis Making*, 19: 78–89, 1999. [PubMed:9917023]. (Cited on page 34.)

[112] J. L. Norris, D. S. Cornett, J. A. Mobley, M. Andersson, E. H. Seeley, P. Chaurand, and R. M. Caprioli. Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis. *Int J Mass Spectrom*, 260:212–221, Feb 2007. [PubMed Central:PMC1885223] [DOI:10.1016/j.ijms.2006.10.005] [PubMed:17541451]. (Cited on page 27.)

[113] A. L. Oberg and O. Vitek. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.*, 8:2144–2156, May 2009. [DOI:10.1021/pr8010099] [PubMed:19222236]. (Cited on page 115.)

[114] P. H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, 250:4007–4021, May 1975. [PubMed Central:PMC2874754] [PubMed:236308]. (Cited on page 10.)

[115] W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell Proteomics*, 4: 1487–1502, Oct 2005. [DOI:10.1074/mcp.M500084-MCP200] [PubMed:15979981]. (Cited on page 8.)

[116] J. V. Olsen and M. Mann. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.*, 101:13417–13422, Sep 2004. [PubMed Central:PMC518757] [DOI:10.1073/pnas.0405549101] [PubMed:15347803]. (Cited on page 71.)

[117] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127:635–648, Nov 2006. [DOI:10.1016/j.cell.2006.09.026] [PubMed:17081983]. (Cited on page 2.)

[118] S. E. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.*, 1:252–262, Oct 2005. [DOI:10.1038/nchembio736] [PubMed:16408053]. (Cited on page 7.)

[119] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, 1:376–386, May 2002. [PubMed:12118079]. (Cited on pages 8 and 66.)

[120] J. R. Ortlepp, R. Kluge, K. Giesen, L. Plum, P. Radke, P. Hanrath, and H. G. Joost. A metabolic syndrome of hypertension, hyperinsulinaemia and hypercholesterolaemia in the New Zealand obese mouse. *Eur. J. Clin. Invest.*, 30:195–202, Mar 2000. [PubMed:10691995]. (Cited on page 16.)

[121] S. Y. Ow, M. Salim, J. Noirel, C. Evans, I. Rehman, and P. C. Wright. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.*, 8:5347–5355, Nov 2009. (Cited on pages 66 and 75.)

[122] Magnus Palmblad, Ali Tiss, and Rainer Cramer. Mass spectrometry in clinical proteomics - from the present to the future. *PROTEOMICS - CLINICAL APPLICATIONS*, 3(1):6–17, 2009. ISSN 1862-8354. doi: 10.1002/prca.200800090. URL `http://dx.doi.org/10.1002/prca.200800090`. (Cited on page 8.)

[123] J. Petrak, R. Ivanek, O. Toman, R. Cmejla, J. Cmejlova, D. Vyoral, J. Zivny, and C. D. Vulpe. Déjà vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. *Proteomics*, 8:1744–1749, May 2008. (Cited on pages 134 and 136.)

[124] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: pathway editing for the people. *PLoS Biol.*, 6:e184, Jul 2008. [PubMed Central:PMC2475545] [DOI:10.1371/journal.pbio.0060184] [PubMed:18651794]. (Cited on page 134.)

[125] S. J. Pitteri and S. M. Hanash. Proteomic approaches for cancer biomarker discovery in plasma. *Expert Rev Proteomics*, 4:589–590, Oct 2007. [DOI:10.1586/14789450.4.5.589] [PubMed:17941811]. (Cited on page 25.)

[126] P. N. Pratapa, E. F. Patz, and A. J. Hartemink. Finding diagnostic biomarkers in proteomic spectra. *Pac Symp Biocomput*, pages 279–290, 2006. [PubMed:17094246]. (Cited on page 27.)

[127] J. Quackenbush. Microarray data normalization and transformation. *Nat. Genet.*, 32 Suppl:496–501, Dec 2002. [DOI:10.1038/ng1032] [PubMed:12454644]. (Cited on pages 77 and 117.)

[128] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0. (Cited on pages 30, 35, 75, and 148.)

[129] T. W. Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, 62:589–597, Jun 2006. [DOI:10.1111/j.1541-0420.2005.00504.x] [PubMed:16918924]. (Cited on pages 32 and 37.)

[130] P. V. Rao, A. P. Reddy, X. Lu, S. Dasari, A. Krishnaprasad, E. Biggs, C. T. Roberts, and S. R. Nagalla. Proteomic identification of salivary biomarkers of type-2 diabetes. *J. Proteome Res.*, 8:239–245, Jan 2009. [DOI:10.1021/pr8003776] [PubMed:19118452]. (Cited on page 14.)

[131] A. Rasche, H. Al-Hasani, and R. Herwig. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics*, 9:310, 2008. [PubMed Central:PMC2515154] [DOI:10.1186/1471-2164-9-310] [PubMed:18590522]. (Cited on pages 14, 97, and 98.)

[132] K. Reinert and O. Kohlbacher. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol. Biol.*, 604:201–211, 2010. [DOI:10.1007/978-1-60761-444-9_14] [PubMed:20013373]. (Cited on pages 30, 45, 72, and 73.)

[133] H. W. Ressom, R. S. Varghese, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, 23:619–626, Mar 2007. [DOI:10.1093/bioinformatics/btl678] [PubMed:17237065]. (Cited on pages 22, 33, 58, 149, and 150.)

[134] Robert Gentleman and Vince Carey and Wolfgang Huber and Rafael Irizarry and Sandrine Dudoit, editor. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer Verlag, 2005. (Cited on pages 30 and 32.)

[135] M. D. Robinson, D. P. De Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Liki. A dynamic programming approach for the alignment of

signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, 8:419, 2007. [PubMed Central:PMC2194738] [DOI:10.1186/1471-2105-8-419] [PubMed:17963529]. (Cited on page 33.)

[136] Joseph L. Rodgers and Alan W. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988. doi: 10.2307/2685263. URL `http://dx.doi.org/10.2307/2685263`. (Cited on pages 36, 41, and 119.)

[137] F. Rodriguez, E. Maire, P. Courjault-Rade, and J. Darrozes. The Black Top Hat function applied to a DEM: A tool to estimate recent incision in a mountainous watershed (Estibre Watershed, Central Pyrenees). *Geophysical research letters*, 29:9.1–9.4, 2002. (Cited on page 30.)

[138] E. Rodriguez-Suarez, E. Gubb, I. F. Alzueta, J. M. Falcon-Perez, A. Amorim, F. Elortza, and R. Matthiesen. Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage. *Proteomics*, 10:1545–1556, Apr 2010. (Cited on page 72.)

[139] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, 3: 1154–1169, Dec 2004. [DOI:10.1074/mcp.M400129-MCP200] [PubMed:15385600]. (Cited on pages 8, 9, 11, and 66.)

[140] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23: 2507–2517, Oct 2007. [DOI:10.1093/bioinformatics/btm344] [PubMed:17720704]. (Cited on page 21.)

[141] Abraham Savitzky and Marcel J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. (Cited on page 29.)

[142] M. Seike, T. Kondo, K. Fujii, T. Okano, T. Yamada, Y. Matsuno, A. Gemma, S. Kudoh, and S. Hirohashi. Proteomic signatures for histological types of lung cancer. *Proteomics*, 5:2939–2948, Jul 2005. [DOI:10.1002/pmic.200401166] [PubMed:15996008]. (Cited on page 118.)

[143] I. P. Shadforth, T. P. Dunkley, K. S. Lilley, and C. Bessant. i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics*, 6:145, 2005. (Cited on page 72.)

[144] J. Shaw, R. Rowlinson, J. Nickson, T. Stone, A. Sweet, K. Williams, and R. Tonge. Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. *Proteomics*, 3(7):1181–1195, Jul 2003. [DOI:10.1002/pmic.200300439] [PubMed:12872219]. (Cited on page 11.)

[145] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract.*, 87:4–14, Jan 2010. [DOI:10.1016/j.diabres.2009.10.007] [PubMed:19896746]. (Cited on page 13.)

[146] B. Sitek, O. Apostolov, K. Stuhler, K. Pfeiffer, H. E. Meyer, A. Eggert, and A. Schramm. Identification of dynamic proteome changes upon ligand activation of Trk-receptors using two-dimensional fluorescence difference gel electrophoresis and mass spectrometry. *Mol. Cell Proteomics*, 4: 291–299, Mar 2005. [DOI:10.1074/mcp.M400188-MCP200] [PubMed:15654083]. (Cited on page 116.)

[147] G. K. Smyth, J. Michaud, and H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21: 2067–2075, May 2005. [DOI:10.1093/bioinformatics/bti270] [PubMed:15657102]. (Cited on page 146.)

[148] D. Steinley. K-means clustering: a half-century synthesis. *Br J Math Stat Psychol*, 59:1–34, May 2006. [DOI:10.1348/000711005X48266] [PubMed:16709277]. (Cited on pages 74 and 77.)

[149] M. Sturm, A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9:163, 2008. (Cited on pages 30, 45, 72, and 73.)

[150] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102: 15545–15550, Oct 2005. [PubMed Central:PMC1239896] [DOI:10.1073/pnas.0506580102] [PubMed:16199517]. (Cited on page 133.)

[151] SysProt. Sysprot - system-wide analysis and modelling of protein modification, 2010. [Homepage:www.sysprot.eu]. (Cited on pages 13 and 14.)

[152] Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T. Matsuo. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2(8):151–153, 1988. doi: 10.1002/rcm.1290020802. URL `http://dx.doi.org/10.1002/rcm.1290020802`. (Cited on page 8.)

[153] M. J. Tenga and I. M. Lazar. Impact of peptide modifications on the isobaric tags for relative and absolute quantitation method accuracy. *Anal. Chem.*, 83:701–707, Feb 2011. [DOI:10.1021/ac100775s] [PubMed:21210697]. (Cited on page 85.)

[154] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. T. Le. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20:3034–3044, Nov 2004. [DOI:10.1093/bioinformatics/bth357] [PubMed:15226172]. (Cited on page 32.)

[155] Tibshirani, Robert and Walther, Guenther and Hastie, Trevor. Estimating the Number of Clusters in a Dataset via the Gap Statistic. *Journal of the Royal Statistical Society, Series B*, 63:411–423, 2000. (Cited on page 77.)

[156] N. Tiffin, E. Adie, F. Turner, H. G. Brunner, M. A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M. A. Andrade-Navarro, A. Adeyemo, M. E. Patti, C. A. Semple, and W. Hide. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.*, 34:3067–3081, 2006. [PubMed Central:PMC1475747] [DOI:10.1093/nar/gkl381] [PubMed:16757574]. (Cited on page 14.)

[157] J. F. Timms and R. Cramer. Difference gel electrophoresis. *Proteomics*, 8:4886–4897, Dec 2008. [DOI:10.1002/pmic.200800298] [PubMed:19003860]. (Cited on pages 12, 115, 126, and 129.)

[158] J. F. Timms, R. Cramer, S. Camuzeaux, A. Tiss, C. Smith, B. Burford, I. Nouretdinov, D. Devetyarov, A. Gentry-Maharaj, J. Ford, Z. Luo, A. Gammerman, U. Menon, and I. Jacobs. Peptides generated ex vivo from serum proteins by tumor-specific exopeptidases are not useful biomarkers in ovarian cancer. *Clin. Chem.*, 56:262–271, Feb 2010. [DOI:10.1373/clinchem.2009.133363] [PubMed:20093557]. (Cited on page 8.)

[159] A. Tiss, C. Smith, U. Menon, I. Jacobs, J. F. Timms, and R. Cramer. A well-characterised peak identification list of MALDI MS profile peaks for human blood serum. *Proteomics*, Jul 2010. [DOI:10.1002/pmic.201000100] [PubMed:20707003]. (Cited on page 28.)

[160] Z. Tombol, P. M. Szabo, V. Molnar, Z. Wiener, G. Tolgyesi, J. Horanyi, P. Riesz, P. Reismann, A. Patocs, I. Liko, R. C. Gaillard, A. Falus, K. Racz, and P. Igaz. Integrative molecular bioinformatics study of human adrenocortical tumors: microRNA, tissue-specific target prediction, and pathway analysis. *Endocr. Relat. Cancer*, 16:895–906, Sep 2009. [DOI:10.1677/ERC-09-0096] [PubMed:19546168]. (Cited on pages 105 and 109.)

[161] R. Tonge, J. Shaw, B. Middleton, R. Rowlinson, S. Rayner, J. Young, F. Pognan, E. Hawkins, I. Currie, and M. Davison. Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics*, 1:377–396, Mar 2001. [DOI:3.0.CO;2-6] [PubMed:11680884]. (Cited on page 11.)

[162] M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422:193–197, Mar 2003. [DOI:10.1038/nature01510] [PubMed:12634792]. (Cited on page 2.)

[163] M. Unlu, M. E. Morgan, and J. S. Minden. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, 18:2071–2077, Oct 1997. [DOI:10.1002/elps.1150181133] [PubMed:9420172]. (Cited on pages 10, 115, and 116.)

[164] G. Van den Bergh, S. Clerens, L. Cnops, F. Vandesande, and L. Arckens. Fluorescent two-dimensional difference gel electrophoresis and mass spectrometry identify age-related protein expression differences for the primary visual cortex of kitten and adult cat. *J. Neurochem.*, 85:193–205, Apr 2003. [PubMed:12641741]. (Cited on pages 11 and 116.)

[165] M. P. van der Werff, B. Mertens, M. E. de Noo, M. R. Bladergroen, H. C. Dalebout, R. A. Tollenaar, and A. M. Deelder. Case-control breast cancer study of MALDI-TOF proteomic mass spectrometry data on serum samples. *Stat Appl Genet Mol Biol*, 7:Article2, 2008. [DOI:10.2202/1544-6115.1352] [PubMed:18241195]. (Cited on page 8.)

[166] F. M. Veronese and G. Pasut. PEGylation, successful approach to drug delivery. *Drug Discov. Today*, 10:1451–1458, Nov 2005. [DOI:10.1016/S1359-6446(05)03575-0] [PubMed:16243265]. (Cited on page 38.)

[167] J. Voortman, T. V. Pham, J. C. Knol, G. Giaccone, and C. R. Jimenez. Prediction of outcome of non-small cell lung cancer patients treated with chemotherapy and bortezomib by time-course MALDI-TOF-MS serum peptide profiling. *Proteome Sci*, 7:34, 2009. [PubMed Central:PMC2746186] [DOI:10.1186/1477-5956-7-34] [PubMed:19728888]. (Cited on page 8.)

[168] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, Nov 2008. [PubMed Central:PMC2593745] [DOI:10.1038/nature07509] [PubMed:18978772]. (Cited on pages 1 and 106.)

[169] P. Wang, F. G. Bouwman, and E. C. Mariman. Generally detected proteins in comparative proteomics–a matter of cellular stress response? *Proteomics*, 9:2955–2966, Jun 2009. (Cited on page 136.)

[170] R. Webster, E. Didier, P. Harris, N. Siegel, J. Stadler, L. Tilbury, and D. Smith. PEGylated proteins: evaluation of their safety in the absence of definitive metabolism studies. *Drug Metab. Dispos.*, 35:9–16, Jan 2007. [DOI:10.1124/dmd.106.012419] [PubMed:17020954]. (Cited on page 38.)

[171] D. B. West, C. N. Boozer, D. L. Moody, and R. L. Atkinson. Dietary obesity in nine inbred mouse strains. *Am. J. Physiol.*, 262:R1025–1032, Jun 1992. [PubMed:1621856]. (Cited on page 16.)

[172] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N.Y.)*, 14:61–65, Jan 1996. [PubMed:9636313]. (Cited on page 1.)

[173] E. S. Witze, W. M. Old, K. A. Resing, and N. G. Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods*, 4:798–806, Oct 2007. [DOI:10.1038/nmeth1100] [PubMed:17901869]. (Cited on page 83.)

[174] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19:1636–1643, Sep 2003. [PubMed:12967959]. (Cited on page 33.)

[175] W. W. Wu, G. Wang, S. J. Baek, and R. F. Shen. Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D gel- or LC-MALDI TOF/TOF. *J. Proteome Res.*, 5:651–658, Mar 2006. [DOI:10.1021/pr0504050] [PubMed:16512681]. (Cited on page 132.)

[176] X L Xie and G Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991. (Cited on page 77.)

[177] Eric P. Xing, Michael I. Jordan, and Richard M. Karp. Feature Selection for High-Dimensional Genomic Microarray Data. In *Proc. 18th International Conf. on Machine Learning*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9512`. (Cited on page 21.)

[178] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics*, 10:4, 2009. [PubMed Central:PMC2631518] [DOI:10.1186/1471-2105-10-4] [PubMed:19126200]. (Cited on page 27.)

[179] J. R. Yates, J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, 67:1426–1436, Apr 1995. [PubMed:7741214]. (Cited on page 70.)

[180] C. Y. Yu, Y. H. Tsui, Y. H. Yian, T. Y. Sung, and W. L. Hsu. The Multi-Q web server for multiplexed protein quantitation. *Nucleic Acids Res.*, 35:W707–712, Jul 2007. (Cited on page 72.)

[181] Weichuan Yu, Baolin Wu, Tao Huang, Xiaoye Li, Kenneth Williams, and Hongyu Zhao. *Statistical Methods In Proteomics*, pages 623–638. Springer Verlag, 2006. Proteomics, PhysioSim. (Cited on page 27.)

[182] P. Zhang, X. Zhang, J. Brown, D. Vistisen, R. Sicree, J. Shaw, and G. Nichols. Global healthcare expenditure on diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract.*, 87: 293–301, Mar 2010. [DOI:10.1016/j.diabres.2010.01.026] [PubMed:20171754]. (Cited on page 13.)

[183] X. Zhang, X. Lu, Q. Shi, X. Q. Xu, H. C. Leung, L. N. Harris, J. D. Iglehart, A. Miron, J. S. Liu, and W. H. Wong. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC*

Bioinformatics, 7:197, 2006. [PubMed Central:PMC1456993] [DOI:10.1186/1471-2105-7-197] [PubMed:16606446]. (Cited on page 33.)

[184] D. Zosso, M. Podvinec, M. Muller, R. Aebersold, M. C. Peitsch, and T. Schwede. Tandem mass spectrometry protein identification on a PC grid. *Stud Health Technol Inform*, 126:3–12, 2007. [PubMed:17476042]. (Cited on page 71.)

DECLARATION

I declare that this thesis is my own work and has not been submitted anywhere for another degree or diploma. It contains no material previously published or written by any other person except where references are made in the text of the thesis.

Berlin, August 2012

Chris Bauer