QuPE

An Integrated Bioinformatics Platform for Quantitative Proteomics

Zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
an der Technischen Fakultät der Universität Bielefeld
vorgelegte Dissertation

von

Stefan P. Albaum

April 16, 2012

Supervisors: apl. Prof. Dr.-Ing. Tim Wilhelm Nattkemper
Dr. Alexander Goesmann

Stefan P. Albaum
Dorotheenstr. 26
33615 Bielefeld

alu@cebitec.uni-bielefeld.de

Summary

Proteins are the workhorses of life acting as molecular machines, structural elements, and transporters. They receive and transmit signals in and between cells, help in the construction of cell structures, and decompose our diets. Due to their omnipresence and importance, it is no wonder that research activities in the fields of biology, medicine, and biotechnology focus on their analysis. Nowadays, developments in mass spectrometry provide researchers with a comprehensive inventory of methods to gain qualitative and quantitative knowledge about these integral components of life. Techniques such as liquid chromatography coupled to tandem mass spectrometry in combination with isotopic labeling, finally, paved the way for the analysis of complete proteomes in a high-throughput manner. With the number of mass spectra that are produced in such experiments running easily into the thousands, there is, undoubtedly, a strong demand for appropriate processing and analysis strategies.

Aim of this work was to tackle this computational challenge in mass spectrometry-based quantitative proteomics. In this work, therefore, the concept of a software application for quantitative proteomics experiments was devised and put into practice. This envisaged a platform, firstly, to manage all data and meta data related to these experiments, and secondly, to ease the development and integration of novel analysis methods. Based on the capabilities of the system a variety of new methods has been designed and implemented, starting from procedures for the assessment of protein identifications, to optimized but also novel algorithms for protein quantification, to the first-time derivation of a workflow for the multivariate statistical analysis of quantitative proteomics experiments.

The resulting platform named QuPE has been developed as a rich internet application to provide data management capabilities as well as analysis functionality for protein identification, quantification, and in particular statistical evaluation from any location in the world via a standard web browser. It is one of the most characteristic features of the system that

also locally dispersed users, once they have uploaded their data, can start and continue their analysis in a collaborative manner, whenever an internet connection is available. This advantage of QuPE, which is best expressed by the concept of 'Software as a Service' (SaaS, Mell and Grance 2010), has led to cooperations, *inter alia*, in the frame of the BMBF-funded QuantPro initiative with workgroups at the universities of Bochum and Greifswald [grant 0313812], with the Heart and Diabetes Center in Bad Oeynhausen, the University College Cork in Ireland, and the Palacký University in the Czech Republic.

A significant part of the work of this thesis was dedicated to the optimization and enhancement of algorithms for the calculation of (relative) abundance values from isotope-labeled protein samples. This started with the implementation of a rather simple single-spectrum based approach, which nevertheless achieves competitive results, and ended with a new method that now allows to compare the abundances of two differentially labeled peptides, i. e. a partially-labeled peptide and its fully-labeled or fully-unlabeled counterpart in a high-throughput manner. Overall, the newly developed algorithms allow to accurately and precisely determine relative abundance values of metabolically stable isotope-labeled data and furthermore represent a significant improvement in terms of quality in comparison to other existing approaches.

The next step after protein identification and quantification concerns the interpretation of the data. Therefore, methods of statistics and data mining are indispensable. The provision of user-friendly and conceivable statistical analysis methods is, however, only 'half the battle'—moreover, it is necessary to elucidate which statistical analysis strategy promises success for stable isotope-labeled proteomics data, and allows to draw accurate and valid conclusions from the data. The two central questions posed in a multitude of quantitative proteomics experiments are, firstly, which proteins are differentially regulated regarding the selected experimental conditions, and secondly, whether there are groups of proteins that show similar abundance ratios and thus might have a similar turnover. To answer these questions, a comprehensive evaluation was conducted within the scope of this work taking into account three real-world datasets from recently published experiments. This finally led to the derivation of a workflow for quantitative proteomics data analysis.

Different statistical analysis methods were evaluated regarding their suitability to identify up- or down-regulated proteins in multivariate experimental data. In the same manner, cluster algorithms were investigated and their outcomes compared to each other in order to determine the method that best fitted to this type of data. The evaluation assessed not only the cluster algorithms itself but also their validation to obtain the optimal number of clusters for a specific dataset. In this context, the inclusion of external information such as functional categories turned out to be a key element to gain meaningful clusterings, both from a biological and a computational point of view.

In summary, QuPE constitutes a comprehensive platform for the analysis of quantitative proteomics experiments, especially of metabolic stable isotope labeling approaches. Due to its extensible nature, the system can easily be extended to cope with future developments in this field of research, e. g. with regard to the emerging interest in posttranslational protein modifications or novel quantification methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Since its invention in the middle of the 20th century mass spectrometry has been closely linked with humankind's interest in biomolecules, in particular proteins—the main actors of life. But it was not until the end of the century that both fields of research finally found a way together. The invention of the so called soft ionization methods paved the way for a thorough understanding of the localization and function of proteins in a cell (Karas et al. 1987; Tanaka et al. 1988; Whitehouse et al. 1985). Today, mass spectrometry is probably the most important method to characterize individual proteins extracted from a biological sample, and thanks to high-throughput methods such as MudPIT (Wolters et al. 2001) it is nowadays possible to identify hundreds of proteins simultaneously within a few hours.

When Wilkins introduced the concept of the proteome as "the entire PROTein complement expressed by a genOME, or by a cell or tissue type" (Wilkins et al. 1996, p.20), he particularly highlighted that in contrast to the genome—the entirety of all genes in an organism—the proteome is highly dynamic showing changes under different conditions and even in the course of time. There is no better example to demonstrate this fact—that the state of an organism is reflected by the proteome—than the life of a butterfly, which starts as a rather inconspicuous caterpillar and metamorphoses into a (predominantly) beautiful insect. "While genes fundamentally shape physiology and pathophysiology, proteins are the final executive force of all cellular processes that finally drive physiology and behavior in health and pathology" (Frank et al. 2009, p.1), or in other words "there is more to paella than the recipe, more to Bach than

ink on paper, and more to a society than its code of laws" (Anderson and Anderson 1998, p.1854).

Four years ago, in one of the top-ranking scientific journals, it was admitted that "Mass spectrometry (MS)-based proteomics has become a formidable tool for the investigation of posttranslational modifications to proteins, protein interactions, and organelles", and furthermore questioned "Is it now ready to tackle comprehensive protein expression analysis?" (Cox and Mann 2007, p.395). Today, in 2012, the answer to this question is certainly yes, and it is in many ways the utilization of stable isotopes in mass spectrometry-based experiments that opened the gates to a new era of quantitative proteomics. "Proteomics has shifted from the analysis of small sets of proteins towards the comprehensive investigation of a much larger number of proteins expressed in a cell, tissue, or organism" (Gouw et al. 2010, p.11). Being one of the main building blocks of systems biology, current proteomics research aims to uncover the functional networks of genes and proteins at the level of the whole cell and to scrutinize the effects of changing environmental conditions on the concentration of proteins (Wolters et al. 2001; Ong et al. 2002; Zhu et al. 2002; MacCoss et al. 2003; Hufnagel and Rabus 2006; Bantscheff et al. 2007; Mueller et al. 2008; Mallick and Kuster 2010). A wide field of applications has evolved ranging from medical diagnosis and the investigation of pharmaceutical effects to the optimization of biotechnologically-relevant production processes, e. g. of amino acids in the bacterium *Corynebacterium glutamicum* (Kalinowski et al. 2003; Rehm 2006; Fränzel et al. 2010b; Poetsch et al. 2011).

1.2 Aims and objectives

A typical mass spectrometry-based proteomics experiment is characterized by three defining steps: firstly, protein identification, secondly, protein quantification, and thirdly and most importantly, the generation of statistically valid conclusions from the data. Methods such as the MudPIT approach produce enormous amounts of data and can easily encompass several thousands of individual mass spectra, which in turn account for hundreds of proteins. It seems obvious that a thorough analysis of these data masses demands computational assistance and requires automatic processing of the data (Matthiesen 2007b).

In the frame of this work a software application, named QuPE (d. v. 'Quantitative Proteomics data Explorer'), was designed and implemented that presents a comprehensive and extensible software solution to support researchers in the analysis of quantitative proteomics data and to gain thorough and in-depth analysis results (Albaum et al. 2009a). The development was made possible in the frame of the BMBF-funded QuantPro initiative [grant 0313812] and, in particular, through the integrative character of the Center for Biotechnology (CeBiTec), an academic institution at Bielefeld University.

Driven by a close cooperation with biologists located in Bielefeld but also at the universities of Bochum and Greifswald, the weaknesses of existing software solutions have quickly been identified, and manifested in the following requirements:

1. A strong need for adequate data management capabilities to organize and structure mass-spectrometry datasets but also associated meta data, such as descriptions of the experimental setup or lists of identified peptides.
2. In particular as a result of the rapidly advancing development in mass spectrometry, novel instruments place higher demands on algorithms for the calculation of abundance values from isotopically-labeled protein samples. It is, for example, a higher resolution of these instruments that needs to be reflected in the quantification procedure.
3. To the end, datasets resulting from such quantitative proteomics experiments are often very complex and consist of lists of measured abundance values for hundreds (or thousands) of proteins. As a manual exploration of such large datasets is practically impossible, there is a keen demand for computational approaches concerning statistical data analysis and data mining in order to support experimenters.

The devised QuPE system does not only provide data management capabilities but, moreover, serves as a platform that eases the development and integration of novel analysis methods starting from the assessment of protein identifications from mass spectra to multivariate statistical analysis and data mining. Using the engineered platform, a wide range of methods has been devised and evaluated which, in summary, contributed, on the one hand, to novel computational approaches for the analysis of quantitative proteomics data, and on the other hand, to a better understanding of regulation at the protein level (Albaum et al. 2011b; Trötschel* et al. 2012).

In view of the distributed locations of users, it was, furthermore, a central objective of this work to bring the developed tools and methods closer to the researcher. Therefore, QuPE was created as a rich internet application, which addresses the limitations in "the richness of the application interfaces, media and content" (Allaire 2002, p.1) of classical web applications. Based on Asynchronous JavaScript and XML (AJAX, Garrett 2005), the user interface behaves similar to the user interface of a standalone software application started on a personal computer. Requiring only a standard-compliant web browser, the application is independent from any operating system. Data stored in the system, such as mass spectra, or analysis results may be accessed on any computer connected to the internet. A local installation is self-evidently not necessary, either.

1.3 Structure of this work

The thesis at hand describes the ideas and their implementations towards a comprehensive and complete solution for the analysis of mass spectrometry-based quantitative proteomics data. In the second chapter of this work, the biological processes underlying the synthesis but also the degradation of proteins are briefly introduced. Starting at the level of gene transcription the focus is turned on the various influences that affect the total amount of proteins in a cell. There then follows a target-oriented overview of mass spectrometry

explaining structure and generation of the data including methods for protein identification as well as methods for protein quantification. Chapter four is devoted to the current state of the art in proteomics software tools and applications. This comprises, firstly, the provision of data management functionality to organize and structure experimental datasets, and secondly, current algorithms and methods for data analysis, in particular regarding the calculation of protein abundances from isotopically-labeled data. In the following chapter five, the requirements that necessitated the development of the rich internet application QuPE are formulated. The considerations go a step further in chapter six, in which methods for the analysis of quantitative proteomics data are introduced. Chapter seven describes in detail the implementation of the system QuPE. Apart from the apparent user interface, the various layers of the application are presented and portrayed. In this context, the extensibility of the system that facilitates the integration of novel methods for the processing of stored data shall particularly be highlighted. A significant proportion of the work aimed at the development and implementation of protein quantification methods. Therefore, in chapter eight, a comprehensive evaluation of the devised algorithms is elucidated. Based on the flexible and extensible application programming interface of QuPE, a workflow was contrived to analyze quantitative proteomics experiments. This workflow is presented in chapter nine. A final and critical reflection of this work can be found in chapter ten.

1.4 Related publications

Publications as first author:

C. Trötschel*, S. P. Albaum*, D. Wolff, S. Schröder, A. Goesmann, T. W. Nattkemper, M. Rögner, and A. Poetsch (2012). Protein turnover quantification in a multi-labeling approach–from data calculation to evaluation. *Molecular and Cellular Proteomics* 11.8. (*contributed equally), pp. 512–526

S. P. Albaum, H. Hahne, A. Otto, U. Haußmann, D. Becher, A. Poetsch, A. Goesmann, and T. W. Nattkemper (2011). A guide through the computational analysis of isotope-labeled mass spectrometry-based quantitative proteomics data: an application study. *Proteome Science* 9.1, p. 30

S. P. Albaum, H. Neuweger, B. Fränzel, S. Lange, D. Mertens, C. Trötschel, D. Wolters, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2009). Qupe–a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics* 25.23, pp. 3128–3134

Publications as co author:

J. Toepel, S. P. Albaum, S. Arvidsson, A. Goesmann, M. la Russa, K. Rogge, and O. Kruse (2011). Construction and evaluation of a whole genome microarray of Chlamydomonas reinhardtii. *BMC Genomics* 12, p. 579

H. Neuweger, M. Persicke, S. P. Albaum, T. Bekel, M. Dondrup, A. T. Hüser, J. Winnebald, J. Schneider, J. Kalinowski, and A. Goesmann (2009). Visualizing post genomics data-sets on customized pathway maps by ProMeTra—aeration-dependent gene expression and metabolism of Corynebacterium

glutamicum as an example. *BMC Systems Biology* 3, p. 82
M. Dondrup, S. P. Albaum, T. Griebel, K. Henckel, S. Jünemann, T. Kahlke, C. K. Kleindt, H. Küster, B. Linke, D. Mertens, V. Mittard-Runte, H. Neuweger, K. J. Runte, A. Tauch, F. Tille, A. Pühler, and A. Goesmann (2009). EMMA 2–a MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics* 10, p. 50
J. Blom, S. P. Albaum, D. Doppmeier, A. Pühler, F.-J. Vorhölter, M. Zakrzewski, and A. Goesmann (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10, p. 154
H. Neuweger, S. P. Albaum, M. Dondrup, M. Persicke, T. Watt, K. Niehaus, J. Stoye, and A. Goesmann (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics* 24.23, pp. 2726–2732
H. Neuweger, J. Baumbach, S. Albaum, T. Bekel, M. Dondrup, A. T. Hüser, J. Kalinowski, S. Oehm, A. Pühler, S. Rahmann, J. Weile, and A. Goesmann (2007). CoryneCenter – an online resource for the integrated analysis of corynebacterial genome and transcriptome data. *BMC Systems Biology* 1, p. 55
D. Bartels, S. Kespohl, S. Albaum, T. Drüke, A. Goesmann, J. Herold, O. Kaiser, A. Pühler, F. Pfeiffer, G. Raddatz, J. Stoye, F. Meyer, and S. C. Schuster (2005). BACCardI–a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. *Bioinformatics* 21.7, pp. 853–859

Supervised bachelor and master theses:

S. Schröder (2010). "Entwicklung, Implementierung und Optimierung von Verfahren zur quantitativen Analyse von stabilisotop markierten, massenspektrometrischen Protein-Daten". MA thesis. Bielefeld University
M. Westermeyer (2008). "Implementierung von Wizards für ein Laboratory Information Management System in einer quantitativen Proteomanalyse-Plattform". BA thesis. Bielefeld University
D. Mertens (2008). "Global quantitative proteomics by stable isotope labeling and tandem mass spectrometry". MA thesis. Bielefeld University
M. Koch (2008). "Interactive visualization and gene selection tool based on webservices". BA thesis. Bielefeld University

Conference contributions:

S. P. Albaum, B. Linke, S. Jaenicke, J. Blom, N. Kessler, S. Juenemann, and A. Goesmann (2011). "Tools for Genome and Post-Genome Data Analysis Developed by the Technology Platform Bioinformatics (Poster abstract)". In: *5th European Conference on Prokaryotic and Fungal Genomics*. Göttingen, Germany
C. Trötschel, C. Lange, S. Albaum, A. Goesmann, R. Krämer, and K. Marin (2011). "Oxidative Stress Response in Corynebacterium glutamicum – the Proteome in Focus (Poster abstract)". In: *5th European Conference on Prokaryotic and Fungal Genomics*. Göttingen, Germany
S. Albaum, H. Neuweger, S. Lange, D. Mertens, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2009). "ProSE – a Rich Internet Application to securely Store, Organise, and Analyse Quantitative

Proteomics Experiments (Poster abstract)". In: *17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference on Computational Biology*. Stockholm, Schweden

S. P. Albaum, H. Neuweger, S. Lange, D. Mertens, K. Runte, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2008). "ProSE – "Software as a Service" for Quantitative Proteomics (Poster abstract)". In: *Human Proteome Organisation 7th Annual World Congress*. Amsterdam, Netherlands

Chapter 2

From genomics to proteomics

The aim of this chapter is to provide a brief insight into the biological processes underlying the synthesis of proteins. Starting at the level of gene transcription the focus is turned onto the various influences that affect the total amount of proteins in a cell.

2.1 Protein synthesis – from genes to proteins

With the benefit of hindsight, it is rather remarkable that "for a long time, biologists thought that 'genes', the units of inheritance, were made up of protein" (McCarty 2003, Editor's note). Until the middle of the 20th century, the rumor that deoxyribonucleic acid (DNA) would be "too limited in its diversity to carry genetic information" (McCarty 2003, p.1) remained stubbornly. Interestingly, this was still widely believed after Avery and his colleagues at the Rockefeller Institute had already proven that DNA is the true carrier of genetic information (Avery et al. 1944). The turning point and thereby the birth of molecular genetics is marked by Watson and Crick and their identification of the three dimensional structure of DNA (Watson and Crick 1953). In half a century, many secrets of this sequence of chemical letters and the (almost) inscrutable relationships between the genes, their transcripts and the proteins as well as their products, the metabolites, have been revealed, yet many secrets are still undisclosed.

It was the regularity of the diffraction by the exposure to X-rays that led to the characteristic double stranded structure of the DNA with its building blocks—the nucleotides. Although it is nowadays known that DNA may also look like "a telephone cord after a kink" (Pearson 2003, p.310), the molecule, in its most common form, is figuratively shaped like a spiral

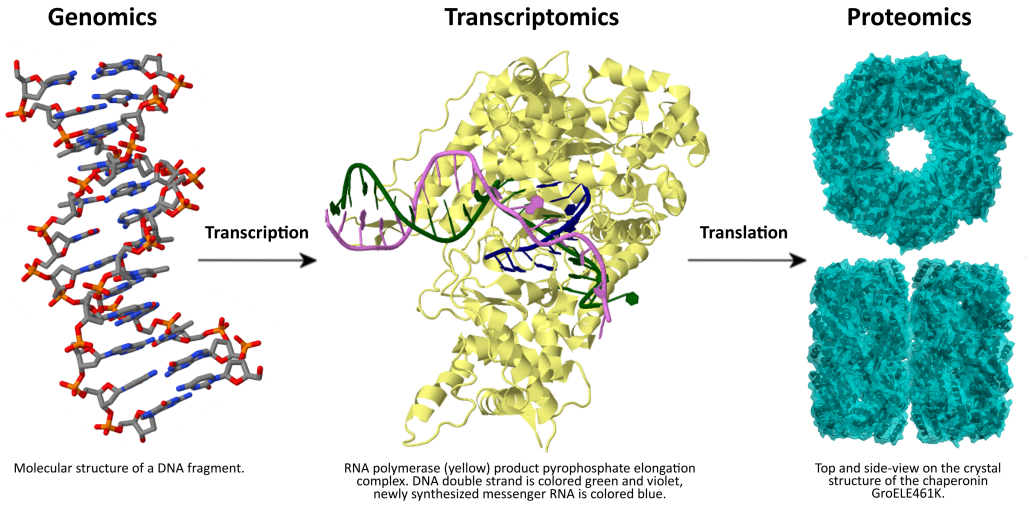


Figure 2.1 – The central dogma of genetic information transfer: through the process of transcription DNA specifies RNA, which in turn specifies the proteins of a cell (Alberts 2003). The molecules are taken from the Protein Data Bank and depicted using Jmol (Jmol Entwicklerteam 2010); entries from left to right: 1D28, Narayana et al. (1991); 1S77, Yin and Steitz (2004); 2EU1, Cabo-Bilbao et al. (2006)

staircase wherein the oppositely placed nucleotides adenine and thymine as well as cytosine and guanine each represent the stairs (Knippers 2001). During the process of transcription, which is, in general, initialized by characteristic promoter regions, the genes of the DNA are rewritten into a single-stranded ribonucleic acid (RNA). Utilizing the genetic code as its conversion table the linear order of the nucleotides on this macro molecule termed messenger RNA (mRNA) encodes for the linear order of amino acids in a resulting protein. This step of translation takes place at the ribosomes—the location of proteins synthesis (see Figure 2.1).

Mulder (1839) introduced the word "protein" in a German journal, derived from the Greek word "$\pi\rho\omega\tau\varepsilon\iota\varsigma$", "primarius", as he falsely suspected protein to be one, uniform substance, universal for both animals and plants. Proteins are chains of on average 100 up to 800 amino acids, which each consist of an amine group, a carboxylic acid group, and a variable side chain together bonded to a central C-atom. Currently, 23 amino acids are known that constitute the building blocks of proteins. Two important amino acids both belonging to a group of alkaline amino acids are arginine (R) and lysine (K). Their final amine groups are often ionized and thereby positively charged (at normal pH level). The primary structure of a protein is made up from its sequence of amino acids, where the carboxylic acid group of one amino acid covalently binds to the amine groups of its neighbor to form a so called peptide binding. Thus, the polypeptide chain is headed in a certain direction: from the left side with a free amine group (N-terminus) to the right side with a free carboxylic acid group (C-terminus). While the secondary structure refers to inner shapes formed by alpha helices and beta sheets, the correct folding of a protein, its tertiary structure, is essential for a protein's functionality. Protein domains are defined as the smallest units of a protein with an unambiguous and independently folded structure, each consisting of up to 150 amino

acids, and often responsible for a distinct reaction. Altogether, these domains characterize a protein (Knippers 2001).

2.2 Protein turnover: degradation and synthesis

Whereas the genome—the entirety of all genes of an organism or a cell—constitutes a static entity, the analogously defined transcriptome and proteome are highly dynamic. Depending on the developmental stage of an organism and on reaction to changing environmental conditions, a variety of regulatory mechanisms influence the rate at which genes are transcribed and later on translated into proteins. A lot of attention is paid to uncover these differences in the expression of genes and proteins, for example to optimize the industrial production of amino acids in different *Corynebacterium glutamicum* strains (Kalinowski et al. 2003; Fränzel et al. 2010b).

Studies utilizing the Microarray technology to analyze the transcriptome of a cell give a detailed picture of the rates at which the genes of an organism are expressed at the moment of measurement. One might assume that thereby also the current amounts of proteins in the cell might indirectly be determined. However, when the abundances of all protein in a sample are directly measured and compared to transcriptome data, the observed correlations are typically rather moderate. In a very early study on this topic, Anderson and Seilhamer (1997) found a correlation coefficient of only $r = 0.48$ between mRNA abundances, which were measured by expressed sequence tag (EST) counting, and corresponding protein abundances arising from two-dimensional electrophoresis. Jayapal et al. (2010) summarized the results of

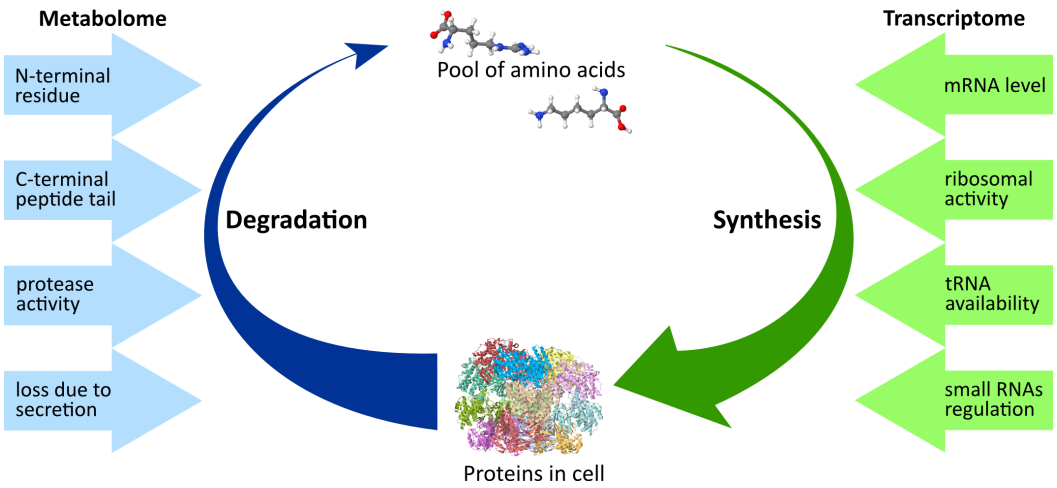


Figure 2.2 – This Figure displays a simplified model of protein turnover describing its two opposing components—synthesis and degradation. While the rate of protein synthesis is determined, mainly, by the amount of mRNA but also other factors such as the ribosomal activity in terms of the rates of initiation of translation and elongation, protein degradation is influenced, *inter alia*, by specific peptide modifications and protease activity.

multiple comparative studies and yielded Spearman rank correlations ranging from $r = 0.2$ to 0.7. Anderson and Anderson (1998, p.1855), therefore, conclude that the "necessity to measure protein levels is inescapable".

The observed differences between mRNA and protein levels are caused by the fact that two opposing processes are responsible for the quantity of a protein in a cell—synthesis as well as degradation (see Figure 2.2, Beynon 2005). While amino acids are, on the one hand, assembled into new proteins at the ribosomes, these biomolecules are, on the other hand, subject to intracellularly regulated degradation processes, e. g. via the ubiquitin pathway. Protein synthesis is, firstly, influenced by the concentration of mRNA in a cell but, obviously, there are other prevailing circumstances that have an impact on this process. This includes, *inter alia*, the availability of tRNA molecules as well as the overall ribosomal activity, which is not least limited by the highest possible rate of translation initiation and elongation. Moreover, gene expression can be modulated at the post-transcriptional level by small non-coding RNAs (Eddy 2001; Storz 2002).

The primary function of intracellular protein degradation is the elimination of old, irregular, damaged, or superfluous proteins. In several situations, this process is attenuated or amplified, e. g. in response to changing environmental conditions such as heat stress (Araki 1992). Features of a protein that influence its stability are the N-terminal's residue of a protein (Tobias et al. 1991) or its C-terminal peptide tail as it was, for example, found in *Escherichia coli* (Gottesman et al. 1998; Herman et al. 1998; Lies and Maurizi 2008). Degradation is conducted by proteolytic enzymes, which can further be separated in exo- and endoproteases depending on their starting point of destruction within a protein.

It can be summarized that, whenever quantities of proteins are measured directly, synthesis as well as degradation as the two components of protein turnover both affect the measurement and must, therefore, be taken into consideration.

Chapter 3

Mass spectrometry-based proteomics

At first sight, the range of available mass spectrometry-based methods and technologies that are used to investigate and scrutinize proteins seems to be immense. This chapter wants to shed light on the two main building blocks of mass spectrometry-based proteomics: the identification of proteins and their quantification in a relative or absolute manner. Therefore, selected methods, which are of particular importance for this field of research, are described in detail.

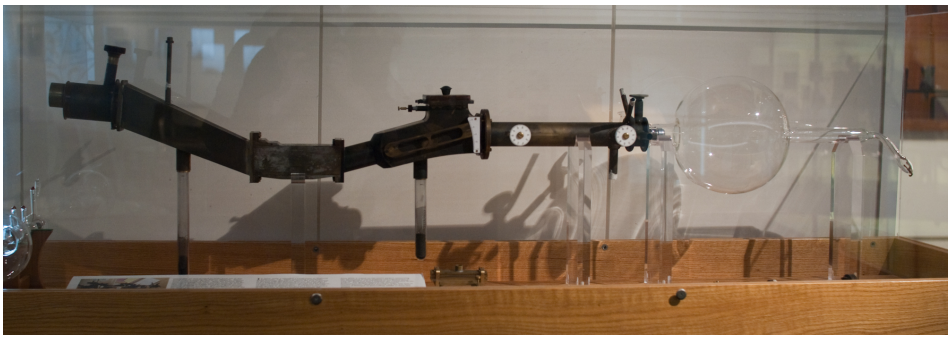


Figure 3.1 – Mass spectrometry is the key technology for the analysis of proteins. The idea originates from the two scientists J. J. Thomson and F. W. Aston. A replica of one of their first mass spectrometers is shown in this picture (© Jeff Dahl `http://commons.wikimedia.org`).

3.1 A historical view on mass spectrometry

Since the 1960s mass spectrometers have been used in laboratories to determine the weight of chemical compounds (Rehm 2006). The underlying idea goes back to the beginning of the 20th century. At that time, the physicist Joseph J. Thomson, better known for his discovery of the 'corpuscles', explored the effects of electromagnetic fields on cathode rays (see glossary for further details, Falconer 1987). Spurred by the findings, his research assistant Aston built the first functional mass spectrometer (Figure 3.1 shows a replica of one of the first instruments). The mass spectrograph, as it was initially called, allowed him to successfully measure the atomic weights of over 200 isotopes, for example of chlorine and bromine (Aston 1922), and finally earned him the Nobel prize.

In a commonly applied setup, gas chromatography is used to isolate molecules from a gaseous mixture of compounds, which are then subjected to electrons emitted from a heating filament. With a typical energy of 70 eV, a molecule M is ionized as described in the following reaction:

$$M + e^- \rightarrow M^+ + 2e^- \tag{3.1}$$

While perfectly suitable for small organic molecules, the application of electron ionization to large macromolecules such as proteins, peptides, or DNA proved to be elusive for a long time: such big molecules are not volatile, apart from that the high energy results in a break-up of the molecules into thousands of small pieces. From the perspective of proteomics, the breakthrough in mass spectrometry (MS) was achieved with the invention of the two ion sources matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI, see next paragraph). These so called soft ionization techniques both allow for a fast and accurate determination of protein and peptide masses.

3.2 Mass spectrometry for the identification of proteins

The first step of any mass spectrometry-based proteomics experiment—an exception is intact cell mass spectrometry (ICM, Feng et al. 2010)—is, doubtlessly, the isolation and extraction of proteins from a sample under investigation. Necessary experimental steps involve the purification of proteins and their solubilization in sample buffer. Here, it is important to take into account differing solution behaviors e. g. of cytoplasmic and membrane proteins, or the prohibition of protease activity to prevent unwanted digestion of proteins. Before extracted proteins can then be analyzed in a mass spectrometry instrument it is, in most cases, necessary to reduce the sample's complexity, and to separate the proteins in a sample from each other. Typical approaches for this purpose are 2D-electrophoresis or liquid chromatography.

The invention of soft ionization methods marked the turning point in mass spectrometry-based proteomics. Beginning with matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) a variety of different methods is nowadays available to determine molecular weights of proteins and peptides. Simply put, all mass spectrometers

consist of three parts: an ion source, one or more analyzers, and a detector. In general, different types of ion sources and analyzers can be combined. However, there are two setups that are most often found: a MALDI ion source is commonly used in conjunction with a time-of-flight (TOF) mass analyzer, while an ESI is usually connected to one or more ion trap or quadrupole analyzers.

Without loss of generality, and with the knowledge that this leaves a variety of experimental setups unattended, in the following two mass spectrometry-based workflows are exemplary described in detail. The two workflows belong to the most commonly applied approaches and are found in many proteome laboratories. First, there is presented the traditional approach which combines two-dimensional electrophoresis (2D-electrophoresis) with MALDI-TOF mass spectrometry. The second workflow aims at the usage of liquid chromatography (LC) in combination with ESI. Beforehand, a few fundamentals regarding mass spectrometry data processing are briefly introduced.

3.2.1 Fundamentals of mass spectrometry data processing

3.2.1.1 Peak detection – profile vs. centroid data

In general, the raw data that is directly recorded by a mass spectrometer is available in form of a continuous signal or spectrum (Hansen and Smedsgaard 2004). Preprocessing of raw mass spectra, often referred to as peak detection, involves several steps, starting from the application of a smoothing function such as a Savitzky–Golay smoothing filter (Savitzky and Golay 1964) to remove noise from the data, up to baseline correction and peak finding. The purpose of baseline correction is to adjust for a potential offset of recorded values over time. It resets drifting values and results in a flattened baseline of a mass spectrum. Peak finding refers to the conversion of the raw signal from the mass spectrometer into a list of peak values. In this context, the original signal that results from one ion is herein referred to as the 'profile' peak (see Figure 3.2A). After "finding the vertical line passing through the center of gravity of the peak" (Matthiesen 2007a, p. 40), the result is a discrete value termed 'centroid' peak (see Figure 3.2B). Several methods have been suggested for the purpose of peak finding. They base, for example, on the weighted average of each profile peak's masses or the first derivative of the function that describes the continuous signal of the mass spectrum and its zeros (Matthiesen 2007a). Mass spectrometry vendors often provide own preprocessing methods and allow the direct conversion of mass spectra during recording. A comprehensive comparison of peak detection methods has been conducted by Yang et al. (2009). The authors recommend an algorithm based on continuous wavelet-transformation (Du et al. 2006).

In summary, the result of a mass spectrometry analysis can formally be described as a list of (centroid) peaks, each consisting of an intensity i and a mass to charge (m/z) ratio. A mass spectrum that consists of p discrete peaks, in other words p ions separated by the mass analyzer, can therefore be defined with two vectors $\mathbf{m} = \{m_1 \dots m_p\}$ and $\mathbf{i} = \{i_1 \dots i_p\}$.

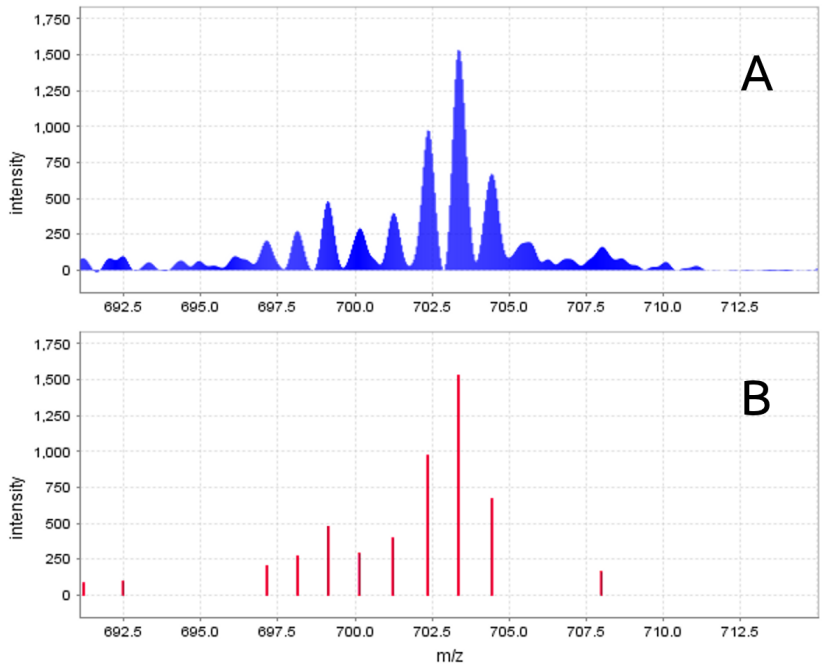


Figure 3.2 – Comparison between a raw mass spectrum recorded in profile mode (A) and the same spectrum after 'peak detection' (B) using the continuous wavelet-transformation proposed by Du et al. (2006).

3.2.1.2 Resolution and accuracy

The resolution of a mass spectrometer—measured in the unit Thomson (Th)—denotes the minimal difference in mass, which has to be present between two ions, so that the peaks of both ions will be clearly distinguishable in a recorded mass spectrum in profile mode. Illustrated in Figure 3.3 is a similar measure, the resolution power R, which is calculated as the ratio of a (profile) peak's mass m to the peak's width at half maximum Δm (full width at half maximum height, FWHM).

A second important measure that describes the characteristics of a mass spectrometer is its accuracy. It is usually denoted in the unit 'parts per million' (ppm). A MALDI system such as Bruker™'s (Bruker Daltonics) ultrafleXtreme is, for example, capable to determine the mass of an ion with an accuracy smaller than 1 ppm, hence, +/- 0.002 Da for a 2 kDa molecule. The resolution power of such a system is claimed to be higher than R = 40.000 (Schäfer 2009).

3.2.1.3 Purpose and function of mass spectrometry in proteomics

Resultant from any mass spectrometry analysis is a list of masses giving hint to the molecular weights of the analytes under investigation. Based on these findings, the conclusion needs to be drawn which molecules—proteins—may belong to these weights. Given an extracted and

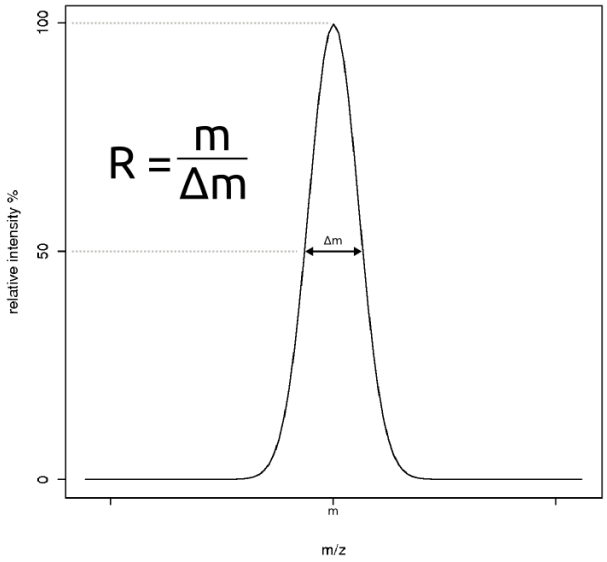


Figure 3.3 – The resolution power R of a mass spectrometer is calculated as the ratio of a (profile) peak's mass m to the peak's width at half maximum Δm (full width at half maximum height, FWHM).

purified protein sample, the basic principle of any experiment therefore is to separate the proteins contained in the sample and then to determine each protein's weight.

3.2.2 Two-dimensional electrophoresis in combination with matrix-assisted laser desorption/ionization and time-of-flight mass spectrometry

A classical approach to identify the proteins contained in a sample relies on two-dimensional electrophoresis (2D-electrophoresis) to separate a mixture of proteins combined with MALDI-TOF mass spectrometry for protein identification. Technologies such as DIGE allow, moreover, to compare two or more proteome samples and to gain a relative quantification of protein amounts. Figure 3.4 depicts the different steps in the workflow to gain the qualitative and quantitative information of proteins and peptides. Starting from a purified protein sample, the mixture is first separated by two-dimensional electrophoresis. The separated proteins are then cut out assuming that at each spot in the gel only one protein is located. If each picked-out protein is then subjected to a mass spectrometry analysis, this results in lists of the proteins' molecular weights. In many cases, the genome and thereby the amino acid sequences of all proteins of an organism are known. It is, hence, possible to calculate an expected molecular weight of each protein. A comparison of the expected and the observed weight might then yield a protein's identification. This is, however, not rarely ambiguous, and becomes almost impossible if proteins have multiple charges. To circumvent this problem,

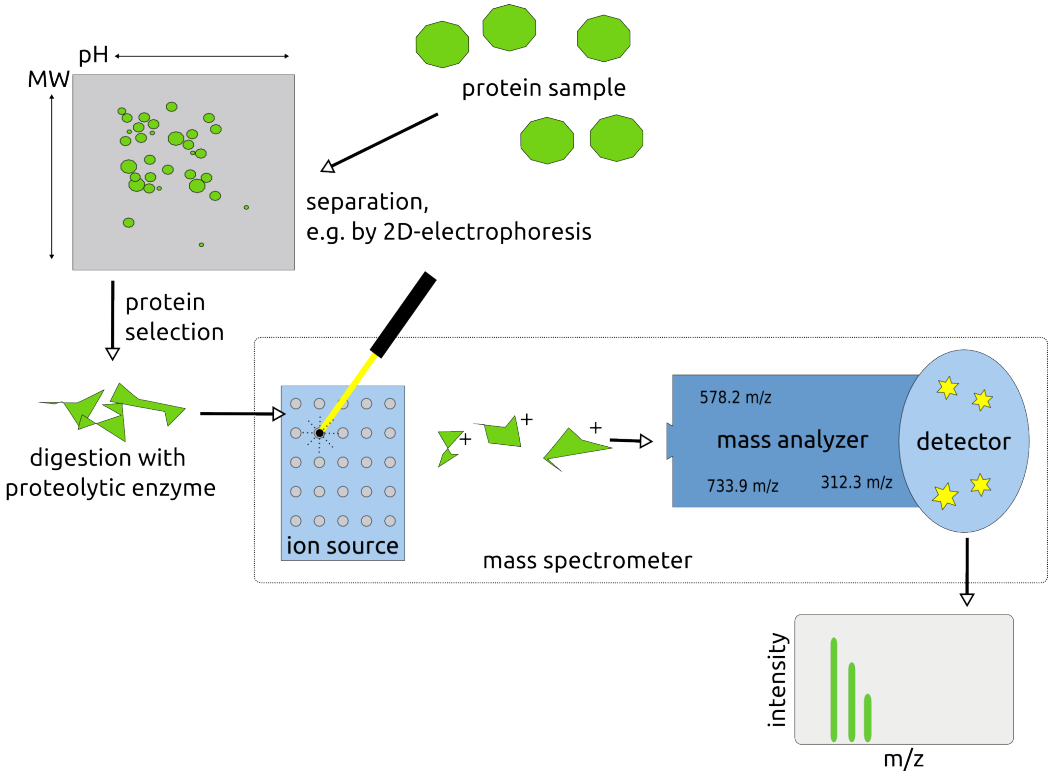


Figure 3.4 – This figure illustrates the typical workflow to identify the proteins contained in a sample utilizing MALDI-TOF mass spectrometry. Starting from a purified protein sample, the mixture is first separated by two-dimensional electrophoresis. Afterwards, the proteins are tryptically digested and co-crystallized with matrix compound on a plate. After ionization by MALDI, the molecular weights of the ionized peptide ions can then be determined in the TOF analyzer.

a method called peptide mass fingerprinting is utilized. Instead of complete proteins, the analyte is therefore enzymatically digested. The serine protease trypsin, for example, cleaves the amino acid structure of a protein at the C-terminal end after arginine or lysine (as long as no proline is following). As these enzymes produce a defined fragmentation of a protein, so to speak its fingerprint, they allow for a less ambiguous and improved protein identification. The cleaved peptides are co-crystallized with matrix compound on a plate. After ionization by MALDI, the molecular weights of the ionized peptide ions can then be determined in the TOF analyzer and thus give hint to the analyzed protein.

3.2.2.1 Protein separation – two-dimensional electrophoresis

The basic idea of two-dimensional electrophoresis (2D-electrophoresis) is to apply two different protein separation techniques on a mixture of proteins to spatially separate up to a complete proteome in two dimensions. The general approach goes back to the 1950s. Kaltschmidt and Wittmann (1970) used the method to purify individual proteins. The first

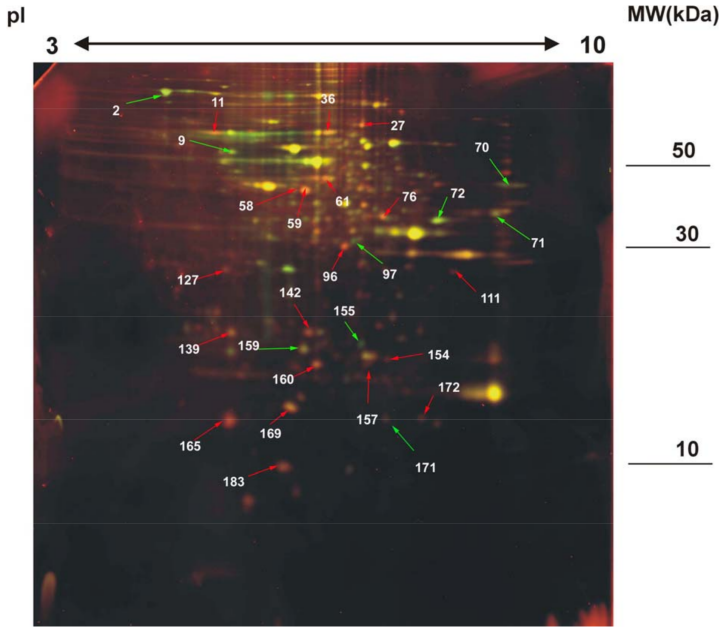


Figure 3.5 – This image of an DIGE experiment shows the comparison of the *Sorangium cellulosum* proteome in exponential (labeled with Cy3, red) and stationary (labeled with Cy5, green) growth phase. Positions in the gel where proteins have been identified are numbered. The experiment was performed by Alici (2007).

application of 2D-electrophoresis on complete proteome samples was performed by O'Farrell (1975) on *Escherichia coli* as well as by Klose (1975) on mice proteins.

While there are certainly more than two protein separation methods available, the first dimension of separation is typically isoelectric focussing (IEF). A mixture of zwitterionic compounds (ampholytes) with two or more differently charged functional groups serves as basis. Under the influence of an electric field the zwitterions tend to rearrange themselves in such a way that each individual compound shifts to a (spatial) position where it has a net charge of zero. Thereby, a pH-gradient is formed. If a protein is added to this gradient it also moves to that position where it reveals no electric charge, its so called isoelectric point. At this position—as it has a zero net charge—any applied electric field has no effect, the protein is focused in the gradient. An IEF-stripe typically corresponds to a certain pH range. The applied electric field ranges from 300 up to 5000V (Westermeier et al. 2008).

The second dimension of separation is usually SDS gel electrophoresis. This takes advantage of the property of most proteins to bind the soap Sodiumdodecylsulfat (SDS). SDS and protein form a complex with a constant ratio of about 1g protein to 1.4g SDS (given a typical 1% solution of SDS, cf. Rehm 2006). This results in proteins having an equal net charge and all proteins only differing in size. Moreover, SDS prevents interactions between proteins. In SDS-page, the IEF-gel from the first separation is directly placed in an SDS-gel which has a gradually increasing concentration e. g. of acrylamides—a substance often used as water-soluble thickener. Subjected to a second electric field the mixture of SDS-protein-

complexes is then separated by their molecular weight. In a typical workflow, protein spots in the 2D-gel are located and cut out. These "picked" spots are afterwards analyzed using mass spectrometry.

An improvement to this technology was made by Unlü et al. (1997) with the usage of fluorescence dyes allowing for the comparison of protein patterns between two samples in one gel. In differential gel electrophoresis (DIGE), two (or more) samples are each labeled with a fluorescence dye such as Cy3 and Cy5. After a scan process with different wavelengths in analogy to the utilized dyes differences in protein expression then become visible (cf. Figure 3.5).

Although 2D-electrophoresis has many advantages, *inter alia* its cost efficiency, the major problem of this method is its relatively high workload in terms of the number of necessary experimental steps, and hence many potential sources of error. Moreover, it is difficult—if not impossible using conventional methods—to investigate proteins with rather extreme properties in terms of size, hydrophobicity, acidity or alkalinity, which also includes the important group of membrane proteins (Rehm 2006).

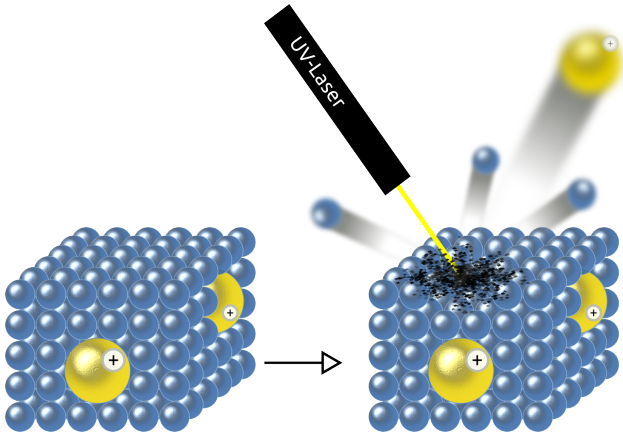


Figure 3.6 – MALDI: proteins or peptides are built into crystals of UV-absorbing molecules. During crystallization, protons are transfered to the sample ions. Irradiation with UV-laser explosively sets free matrix-compounds as well as charged sample ions.

3.2.2.2 Ionization – matrix-assisted laser desorption/ionization

Matrix-assisted laser desorption/ionization (MALDI) was first introduced by Michael Karas and Franz Hillenkamp (Karas et al. 1987) in the 1980s. At about the same time, a similar method to ionize large macromolecules called soft laser desorption (SLD) was introduced by Tanaka et al. (1988), and used to analyze, for the first time, an intact protein. The Japanese researcher won the Nobel prize for his invention, although it is, interestingly, the MALDI approach which is nowadays mainly employed.

The basic principle of MALDI is the incorporation of the macromolecules of interest into crystals built from acidic UV-absorbing molecules, called matrix. A solution consisting of crystallized molecules, a solvent, and the proteins under investigation is spotted on a plate. As the solvent vaporizes and the crystal growth, protons are transfered to the embedded molecules, which are thereby positively charged. The crystals are afterwards put into high vacuum and irradiated with an UV-Laser. During the laser pulse with a duration of 3-4 nanoseconds matrix compounds as well as charged protein ions are explosively set free as illustrated in Figure 3.6.

3.2.2.3 Analyzer – time-of-flight

In the most common setup, a MALDI ion source is directly coupled to a time-of-flight (TOF) analyzer. In this setup, the explosively released cloud of charged protein ions is accelerated by an electric field. Excited by the same field, each ion receives the same kinetic energy:

$$E = (m/z) \times v^2 \tag{3.2}$$

The velocity v of an ion is, thus, proportional to the ions mass to charge ratio: the lighter an ion or the higher its charge the stronger the acceleration.

$$v \propto \frac{1}{\sqrt{m/z}} \tag{3.3}$$

To ensure that all ionized proteins leave the ion source at the same time point and from the same location a technique called delayed extraction is employed. This focusing of the ion cloud is obtained by a delayed power up of the electric field that is responsible for the acceleration. A potential gradient compensates different starting energies and results in a simultaneous movement of all ions into the vacuum of the analyzer.

It is the general principle of the TOF analyzer to measure the time each accelerated ionized particle takes to cover a defined distance, until it finally hits a detector. As soon as the ions enter the apparatus, no further electric field has an effect on the ions. Thus, their velocity remains constant. All ions have to traverse the same distance and arrive at the detector only dependent on their m/z value. The detector is typically realized as a photomultiplier consisting of a number of glass capillaries with a diameter of about 25 μm. The inner surfaces of the capillaries are coated with electron absorbing materials. If a protein ion impinges the interior of a capillary, this triggers the release of electrons that can then be measured.

A substantial improvement of the TOF analyzer was the integration of a reflectron, a kind of mirror for ions. It may happen that two ions with the same mass to charge ratio leave the ion source with a slightly different kinetic energy. Aim of the reflectron, a static electric field, is to turn around all ions that dive into it. Due to the effect that ions travel into this field in a depth according to their energy, the reflectron causes ions with the same mass to charge ratio to arrive simultaneously at the detector.

A typical TOF has a length between 1 and 2 m, which results in a flight time of ions from 5 to 100 μs. A MALDI-TOF provides the possibility to measure very large molecules in a range up to 100.000 m/z.

3.2.2.4 Protein identification – peptide mass fingerprinting

Aim of peptide mass fingerprinting (PMF) is to identify a protein based on a defined fragmentation pattern. The method takes advantage of the fact that a proteolytic enzyme such as trypsin cleaves an amino acid sequence at specific positions. Applied on a protein, the resulting fragments characterize its source—they, so to say, represent the protein's fingerprint. To take an example, Figure 3.7 shows a mass spectrum as it has been recorded for a protein of the soil bacterium *Sorangium cellulosum*: based on the observed masses of the peptide fragments it can be concluded that the analyzed sample contained a phosphoglycerate kinase (sce7349). An algorithm to compare observed fragmentation patterns with the genome

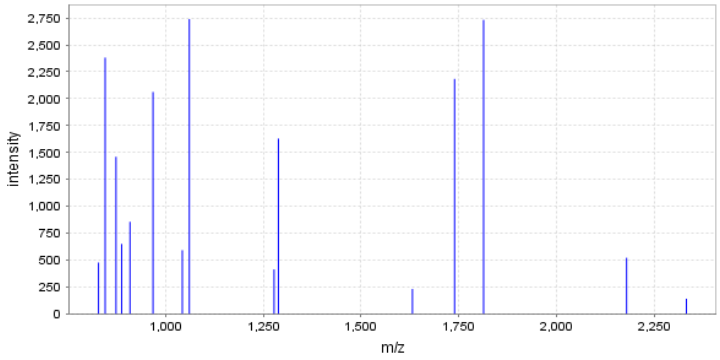


Figure 3.7 – The mass spectrum shows the m/z and intensity values recorded for a protein of the soil bacterium *Sorangium cellulosum*.

sequences of known organisms has, for example, been proposed by Perkins et al. (1999), and is flown in the commercial software product Mascot™ (see section 4.2.1).

Peptide mass fingerprinting has the serious disadvantage that it is very sensitive against missing or incorrect peaks in a mass spectrum, especially if the fragments of two or even more different proteins are mixed in the same spectrum. As 2D-electrophoresis in combination with MALDI-TOF mass spectrometry offers a good separation of individual proteins, it is best suited for this way to identify proteins, in contrast to other techniques such as liquid chromatography coupled to electrospray ionization. Nevertheless, tandem mass spectrometry as it is described in the next workflow delivers protruding advantages and considerably outperforms the PMF approach.

3.2.3 Liquid chromatography in combination with electrospray ionization

From today's point of view, it was mainly the introduction of electrospray ionization that, finally, enabled the high-throughput analysis of proteins. Due to the possibility to directly couple liquid chromatography (LC) to such an ion source, up to a complete proteome sample may automatically be analyzed. A typical workflow of such an LC-MS/MS experiment is illustrated in Figure 3.8. Apart from the extraction and preparation of the sample under investigation (including protein purification, tryptic digestion, column preparation etc.) all further steps starting from the separation by liquid chromatography to the generation of mass spectra can be performed in a fully-automatic manner.

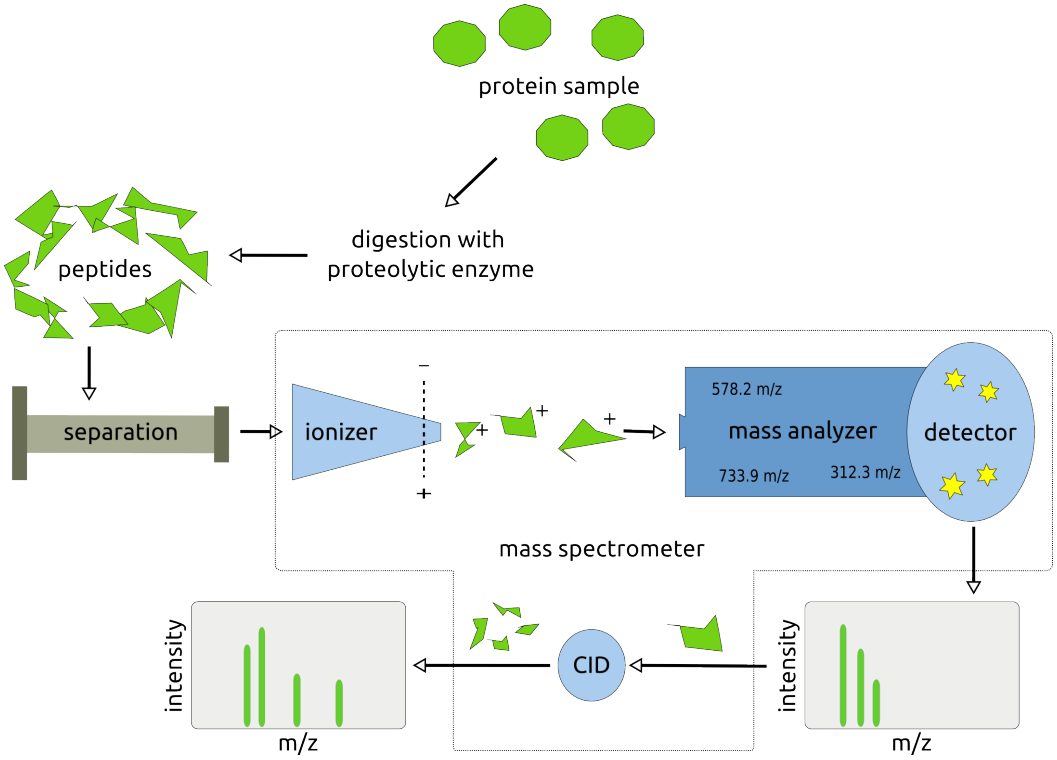


Figure 3.8 – A typical workflow of an LC-MS/MS experiment. An extracted and purified protein sample is typically subjected to enzymatical digestion e. g. using trypsin. Starting from the separation by liquid chromatography, to the generation of mass spectra all further analysis steps are then performed in a fully-automatic manner.

3.2.3.1 Protein separation – liquid chromatography

When a complete macromolecular complex from an organism with sequenced genome was analyzed for the first time, instead of a 2D-electrophoresis-based approach, liquid chromatography (LC) was utilized to purify the individual protein components (Neubauer et al. 1997).

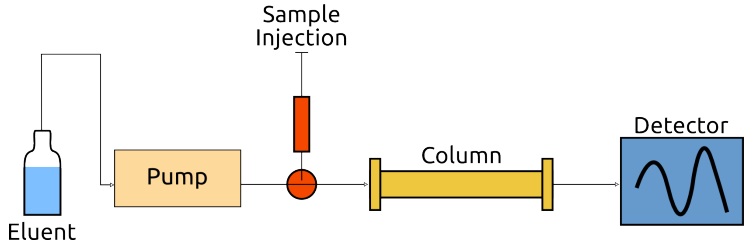


Figure 3.9 – Diagram showing the basic components of a high performance liquid chromatography system. In a typical LC-MS/MS experiment, instead of (or in addition to) a detector a mass spectrometer is directly coupled to the column.

In 1999, Link et al. proposed an approach based on two connected chromatography techniques, called 2D-chromatography, to directly analyze complex protein mixtures. In general, chromatography aims to separate proteins by their specific properties such as size, charge or hydrophobicity. The basic components of a high performance liquid chromatography (HPLC) system as depicted in Figure 3.9 are a pump, which is able to generate pressures of up to 400 bars, and a steel column with a diameter between a few millimeters down to nano-meter scale. The column is filled with a material such as a silica gel wherein each particle has a diameter of about 3-10 μm. This is called the stationary phase. In the commonly applied reverse phase (RP) HPLC, the particles are coated, leading to apolar and hydrophobic surfaces. After protein samples are injected with a so called injection loop, the biomolecules together with the eluent, an aqueous solution, pass through the column. In analogy to the material in the column, eluent and sample constitute the mobile phase. A typical experiment lasts up to one hour whereby the eluent's ratio e. g. of acetonitrile to water is constantly increased.

Apart from RP chromatography, ion exchange chromatography (IEX) is frequently utilized for the separation of proteins. The technique, invented already in the 1960s, allows to distinguish molecules that differ only by one charged amino acid (Westermeier et al. 2008).

The biggest advantage of liquid chromatography is its possibility to directly couple this protein separation method to a mass spectrometer (McCormack et al. 1997). This does not only avoid additional experimental steps but also eliminates a possible source of errors.

3.2.3.2 Advanced separation – multidimensional protein identification technology

Multidimensional protein identification technology (MudPIT) is an enhanced version of the liquid chromatography approach that finally gave rise to shotgun proteomics (Wolters et al. 2001). The separation technology is based on two different columns. The first utilizes a strong cation exchange material (SCX), similar to the material used in IEX. At second, a reversed phase (RP) material is used in the stationary phase. The approach is characterized in that the chromatography is conducted in cycles. In each cycle, a different salt concentration causes

specific peptides of the sample to pass the first column (SCX). A following eluent then allows peptides to elute from the RP directly into a coupled mass spectrometer. Compared to other approaches such as 2D-electrophoresis, MudPIT allows for a significant higher number of identified and quantified proteins (Netterwald 2007).

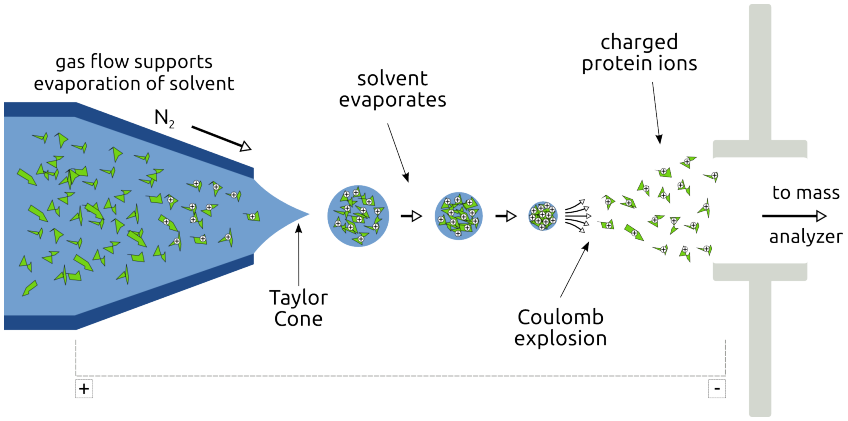


Figure 3.10 – ESI: a solution containing proteins is sprayed through a capillary. Due to an electric field between the capillary opening and a counter electrode, charged ions are moved to the surface forming the so called Taylor cone. Small drops separate from the cone, and supported e. g. by a nitrogen flow the solvent evaporates. As soon as the electrostatic repulsion within the droplets is bigger than their surface tension, the droplets explode (Coulomb explosion). Only charged, free gas-phase ions survive and are further pulled into the mass analyzer by the applied electric field.

3.2.3.3 Ionization – electrospray

When Koichi Tanaka was honored in 2002 with the Nobel prize for his invention of a soft ionization method, he had to share this award with a second scientist, John Fenn, who implemented a further method capable to ionize large biomolecules called electrospray ionization (ESI). Figure 3.10 illustrates the process. A solution consisting of protein ions and a solvent is sprayed through a capillary. This allows the direct coupling of a HPLC system, and thereby a continuous and automatic analysis of a complete proteome sample. An electric field with a typical voltage of 2-3 kV is installed between the capillary opening and a counter electrode at the entrance to the mass analyzer. Due to the applied electric potential, charged ions move to the surface as soon as the solution leaves the capillary. At this point, two contradicting forces interplay: the surface tension works against the electric potential, which pulls the solution to the counter electrode. This leads to a conical formation of the solution's surface, a so called Taylor cone. Small drops dissolve from the surface, and in the most widely accepted theory the ionization process can be explained as follows: Due to evaporation supported e. g. by a drying flow of nitrogen gas the volume of the drops decreases and included ions are crowded together. As soon as the Raleigh limit—given by the electrostatic repulsion within the droplets extending the surface tension—is reached the droplets break up in the so called Coulomb explosion. The process repeats until any

remaining solvent is evaporated. At the end, free gas-phase ions survive, and are further pulled into the mass analyzer (Whitehouse et al. 1985; Kebarle and Verkerk 2009). Here, the quadrupole as well as the ion trap and its variations are the analyzers that are most often found in proteomics laboratories.

3.2.3.4 Analyzer – quadrupole

Figure 3.11 illustrates the principle of a quadrupole mass analyzer. Four metal rods, arranged in parallel and in the same distance to each other give the analyzer its name (Paul and Steinwedel 1960). The rods located diagonally opposite are each operated with the same direct voltage (U) and an additional high-frequency alternating voltage (V). Due to this construction the analyzer has four poles characterized by the following two voltages at a time point t:

$$U_1(t) = -U + V \sin(\omega t) \text{ and } U_2(t) = U - V \sin(\omega t) \tag{3.4}$$

While the ratio between U and V has to remain constant, the value of U (and accordingly V) as well as those of the frequency ω are variable. During measurement, the polarity of these poles is constantly fluctuating. Thereby, a charged ion entering the mass analyzer is alternately pulled and pushed to the rods. The main concept is now based on the following principle: for each ion with a distinct m/z value there exists a particular value of U that allows the ion to fly through the quadrupole on an oscillating path, whereas the trajectories of other ions are instable. To record a full mass spectrum, e. g. in a range of 300 to 2000 m/z, the voltage level is constantly raised while a continuous flow of protein ions, commonly from a directly coupled HPLC and ESI ion source, moves into the analyzer.

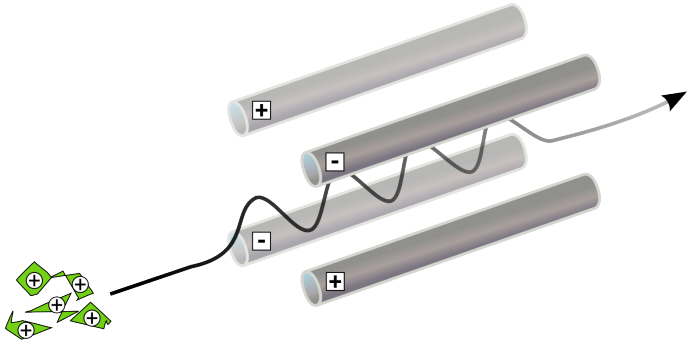


Figure 3.11 – This figure illustrates the principle of a quadrupole mass analyzer. Four metal rods are arranged in parallel, whereby the two rods located diagonally opposite are each operated with the same direct current (U) and an additional high-frequency alternating voltage (V). Given a charged ion with a distinct m/z there exists a corresponding value of U (and V) that allows the ion to pass the quadrupole. Due to fluctuating polarity the ions do not move on a distinct path but oscillate with a fixed amplitude. During a full mass scan the voltages are constantly raised allowing the successive measurement of all ions with a corresponding m/z value.

3.2.3.5 Analyzer – ion trap

The principle of the ion trap, also called quadrupole ion trap, is closely related to the quadrupole mass analyzer with its four metal rods. It is also called Paul trap named after its inventor, Wolfgang Paul (Paul and Steinwedel 1953), who was therefore awarded with the Nobel prize in 1989. As illustrated in Figure 3.12 the ion trap consists of a ring electrode and a pair of hyperboloid-shaped end cap electrodes. Similar to the quadrupole a direct voltage U superimposed with a radio frequency alternating voltage V is applied on the ring electrode. The resulting electric field forces incoming protein ions to traverse on a circular path within the inner region of the ring electrode—the ions are trapped. This is usually supported by Helium gas in the inside of the analyzer leading to a deceleration of the ions moving into the trap through a hole in the end caps. To measure the m/z values of the ions, the amplitude of V is constantly raised. At a particular value of V the trajectory of all ions having a specific m/z value gets unstable. The ions are, figuratively speaking, thrown out of the trap through a hole at the second end cap, and can then be measured at a detector (McLuckey et al. 1994).

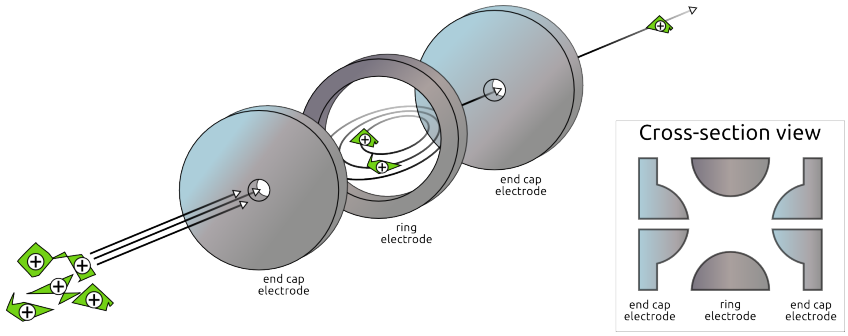


Figure 3.12 – Simplified illustration of the principle of an ion trap. In the cross section view the ring electrode as well as as the two hyperboloid-shaped end cap electrodes are shown. An electric field at the ring electrode forces incoming protein ions to traverse on a circular path—the ions are trapped. At a particular voltage the trajectory of all ions having a specific m/z value get unstable, and the ions are, figuratively speaking, thrown out of the trap.

There are two further developments of this technology that have exerted a great influence on mass spectrometry-based proteomics. The Fourier transform ion cyclotron resonance (FT-ICR) constitutes an improvement of the traditional ion trap in terms of accuracy with an impressive resolution of up to $R = 1,000,000$. While Lawrence and Livingston already introduced the basic idea of the method on a conference in 1931, the first applications of FT-ICR-MS were made possible not before the 1970s (Comisarow and Marshall 1974). The technique allows to distinguish ions that differ only by a few atoms, and is thereby ideally suited for the detection of post-translational modifications. During measurement, ions are trapped in a magnetic field and forced into an orbital (or cyclotron) movement. By the application of a radio frequency pulse parallel to the magnetic field, ions are subsequently forced into larger orbits, and finally pass a detector, which then leads to the determination of their masses. For this purpose, however, a strong magnetic field has to be applied that

is typically achieved by the utilization of superconducting magnets. For that reason, this kind of analyzer belongs to the most expensive types of mass spectrometers (Hamdan and Righetti 2005).

A second important development in this field was the introduction of the so called Orbitrap (Makarov 2000). The operating principle of this analyzer is similar to the FT-ICR, but instead of a magnetic field, ions are trapped in an electrostatic field, in which they orbit around an electrode. As the system does not demand the expensive cooling effort of a superconducting magnet, this type of mass analyzer is comparably less cost-intensive. Nevertheless, a comparably high mass accuracy and resolution power of up to $R = 150,000$ can be reached. The LTQ Orbitrap by Thermo Scientific is probably one of the most often found mass spectrometers in laboratories that conduct high-throughput proteome experiments.

3.2.3.6 Fragmentation – collision-induced dissociation

An ESI or MALDI ion source coupled to any kind of analyzer delivers the m/z values of the complete proteins or its peptides under investigation. The structure of the protein in terms of its amino acid composition and order is, however, not determined. This is achievable by a combination of different mass analyzers. A first analyzer is used to filter all ions having a certain m/z value. As the filtered ions pass a following chamber, usually a second analyzer, a collision gas is induced that leads to a defined fragmentation of the ions. This is referred to as collision-induced dissociation (CID). A third mass analyzer then enables the measurement of the molecular weights of the fragments. Typical combinations include two quadrupole analyzers, one operated in scanning mode and used for filtering and another one operated only with radio frequency for fragmentation. In this so called RF-only mode any ion is allowed to pass through while the collision gas causes the fragmentation. To scan the fragments' m/z values either a third quadrupole may be employed (Yost and Enke 1978) or a TOF analyzer (Chernushevich et al. 2001). The resulting mass spectrum is commonly referred to as MS/MS or MS^2 spectrum. In a typical setup, at first a full MS scan is performed from which, at second, one or more of the most abundant ions are chosen for fragmentation—the so called precursor ions. For this reason, the full scan is sometimes termed the parent spectrum.

3.2.3.7 Protein identification – MS/MS ion search

The fragmentation pattern resultant from collision-induced dissociation (CID) provides a fundamentally improved solution to identify the proteins contained in a sample in comparison to the peptide mass fingerprinting approach. A first information that is known about a protein is the exact weight of its peptide as its m/z value has been used for filtering in the first mass analyzer. In combination with the knowledge gained from the proteolytic digestion this gives a first hint to the protein's identification. A second information is the fragmentation pattern that has been induced by the collision gas. Figure 3.13 illustrates the possible 'pieces of this puzzle'. Similar to peptide mass fingerprinting the observed molecular weights can be compared to expected fragment sizes of known proteins of the same or a

related organism (Eng et al. 1994). This is generally referred to as MS/MS ion search (MIS). The most sophisticated approaches promise to determine the sequence of amino acids based solely on these fragment masses without taking any additional information such as sequence data into account (Rehm 2006).

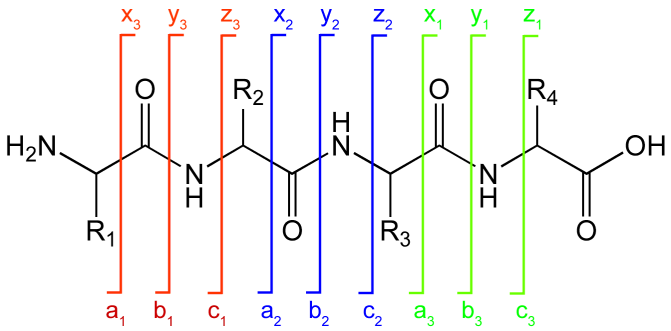


Figure 3.13 – Nomenclature for possible peptide fragments that result from desorption ionization methods such as collision-induced dissociation (©Roepstorff and Fohlman 1984, p.1). In general, the backbone of a peptide is cleaved at one of three distinct positions. Dependant on this position, each N-terminal fragment is denoted with the letter "a", "b", or "c", each C-terminal fragment with the letter "x", "y", or "z". In the depicted example, each potential fragment of a peptide consisting of four amino acids is shown.

3.3 Protein quantification

How wonderful it would be if the intensities recorded by a mass spectrometer would directly mirror the abundances of the proteins contained in a sample. Obviously, it may be thought that the more ions are in a sample the more ions may hit a detector. This is however and unfortunately not true. It is even worse: "Intensities can vary greatly across peptides from the same protein" (Karpievitch et al. 2009, p.2028) and gets worser as "the same sample can result in differences in the peak intensities of the peptides from run to run" (Zhu et al. 2010, p.1). The reasons therefore are manifold but are mostly caused by the fact that in both soft ionization technologies, MALDI as well as ESI, the energy transfered in the ionization process may vary between different peptides—particular ions have a higher ionization efficiency. Moreover, some ions are even suppressed in the analysis (Tang et al. 2004).

As a solution to this problem it may therefore be beneficial to add an internal standard, or more generally, a second sample as a reference to the measurement. Accordingly, a protein's quantity is, thus, not measured in an absolute manner but relative to this reference. In mass spectrometry, stable isotopes can be employed for this purpose. Other methods to gain relative abundance values of proteins are based on 2D-electrophoresis and the utilization of different dyes.

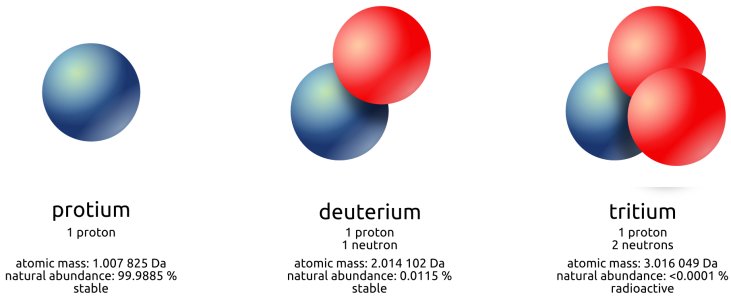


Figure 3.14 – The three isotopes of hydrogen and their atomic nuclei.

3.3.1 Stable isotope labeling

Isotopes are atoms of the same chemical element having the same number of protons but a differing number of neutrons. As illustrated in Figure 3.14 for the element hydrogen, this has an influence on the atomic mass of an isotope. While some isotopes are radioactive, the so called stable isotopes rarely show deviating chemical properties. Isotopes occur with varying natural abundances. The lightest isotope of the element nitrogen ^{14}N, for example, has a natural occurrence probability of about 99.63%, and only approximately 0.37% of all nitrogen isotopes have a weight of ≈15 Dalton (^{15}N). The impact of isotopes in mass spectrometry becomes apparent by the fact that the measurement of one ion does usually not only produce one distinct signal but instead a series of signals. As the difference in mass between most isotopes is approximately one Dalton in addition to the so called monoisotopic peak, which is constituted of the most abundant isotopes, several isotope peaks occur at regular intervals (see Figure 3.15 for an example). Due to this feature, it seems reasonable to utilize stable

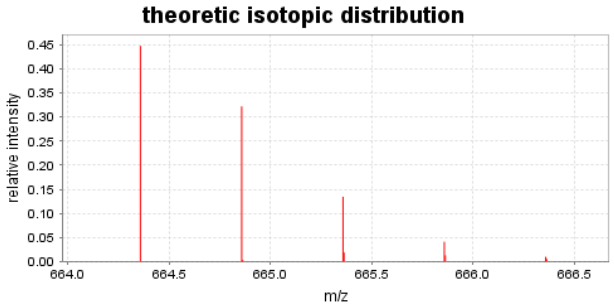


Figure 3.15 – Mass spectrum of a single peptide (FNYDSVMQVPK) showing the peaks resultant from different isotopes.

isotopes for the labeling of two or more biological samples. After mixture of these differently labeled samples their signals are then distinguishable in a mass spectrum. Among the most frequently used strategies to incorporate an isotopic label are metabolic labeling (Oda et al. 1999), SILAC (Ong et al. 2002), and the iTRAQ (Ross et al. 2004) approach. Their advantages

and disadvantages are mainly characterized by the way in which a mass tag is incorporated—either in vivo or in vitro—and their labeling efficiencies in terms of the number of proteins that may have the label built in. A special labeling approach that finally allows to draw conclusions on the absolute concentration of a protein in a cell is the so called AQUA method introduced by Gerber et al. (2003).

3.3.1.1 Metabolic labeling using stable isotopes

The metabolic incorporation of a label (Oda et al. 1999) is often termed the gold standard of labeling (Haegler et al. 2009). Gouw et al. (2010, p. 13) constitute that "clearly, the best place to introduce an internal standard is by metabolically incorporating the stable isotope into living organisms or cells, thereby producing the lowest variation before any sample processing occurs". Frank et al. (2009, p.1) state: "For an accurate and sensitive comparative proteome analysis metabolic labeling of one sample with a stable isotope is the preferred approach. This method results in an enrichment of the stable isotope in every protein in vivo, which can be compared with an unlabeled proteome by combining the two samples prior to MS analysis."

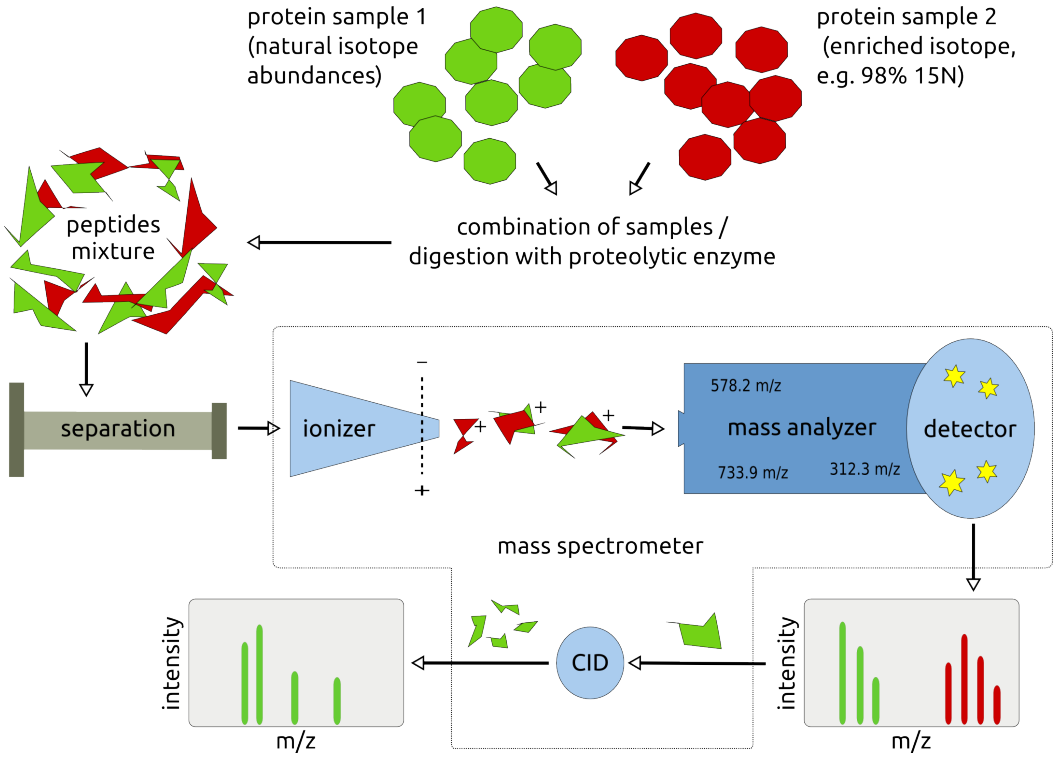


Figure 3.16 – This illustration gives a simplified overview of the typical workflow to gain relative abundance values of two stable isotope labeled proteins. While a first sample is grown on media containing isotopes in naturally abundances, a second sample has incorporated a stable isotope such as ^{15}N. After protein extraction the samples are mixed and analyzed using LC-MS/MS.

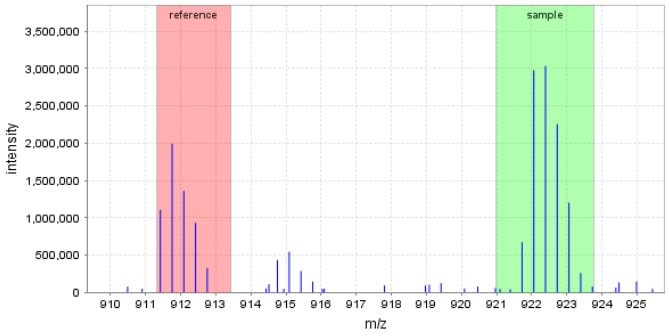


Figure 3.17 – The shown example illustrates a subsection of a mass spectrum recorded for two peptides of the protein lcl|BSU00690. While the first peptide (red) is expected to have the natural occurring distribution of isotopes, for the second peptide (green) the isotope ^{15}N is approximatively at 98 percent.

It is a crucial prerequisite for the applicability of metabolic labeling that there is a possibility to metabolically incorporate the label in an organism. This can be achieved by the addition of isotopically-enriched amino acids or salts as for example ^{15}N-labeled tryptophan or sulphate to the growth medium (Otto et al. 2010). Although fast-growing organisms such as bacteria are most suitable for this labeling strategy (Haußmann et al. 2009; Fränzel et al. 2010a) the number of labeled organisms is steadily increasing. Fed with ^{15}N-labeled bacteria or yeast the list of labeled organisms includes the nematode *Caenorhabditis elegans*, the fly *Drosophila melanogaster* as well as the mouse *Mus musculus* (Frank et al. 2009; Krijgsveld et al. 2003).

Figure 3.16 illustrates the typical workflow of an experiment that compares two samples utilizing stable isotope metabolic labeling. While a first sample is grown on media containing isotopes in naturally abundances, a second sample has incorporated a stable isotope such as ^{15}N. After protein extraction the samples are mixed and analyzed using LC-MS/MS. Protein identification is obtained by MS/MS ion search. A relative protein abundance value can then for example be calculated if the intensities observed for both variants of a protein are summed up and set in relation (see Figure 3.17 for an example mass spectrum).

3.3.1.2 Metabolic labeling using amino acids: SILAC

Ong et al. (2002) proposed a widely-applied labeling method that utilizes isotopically enriched amino acids. As the name—stable isotope labeling by amino acids in cell culture (SILAC)—suggests the label is metabolically incorporated in the cells during their cultivation. Common amino acids used for this purpose are $^{13}C_6$ arginine and $^{13}C_6$ lysine both introducing a mass shift of six Dalton. In contrast to metabolic labeling employing e. g. a ^{15}N-labeled salt, the successful comparison of two SILAC-labeled samples depends on the replacement of a specific amino acid by its heavy variant. Looking at this the other way around, this obviously allows to analyze only those peptides that comprise the labeled amino acid. Furthermore, the organism under investigation needs to be auxotroph for the targeted amino acid as otherwise

a complete incorporation could not be guaranteed (Gouw et al., 2010). An advantage of SILAC (in comparison to metabolic labeling using stable isotopes) is the fact that a labeled peptide is shifted by a definite m/z value while the general distribution of the isotopic peaks remains unchanged—a circumstance that facilitates the quantification. This is exemplarily illustrated in Figure 3.18.

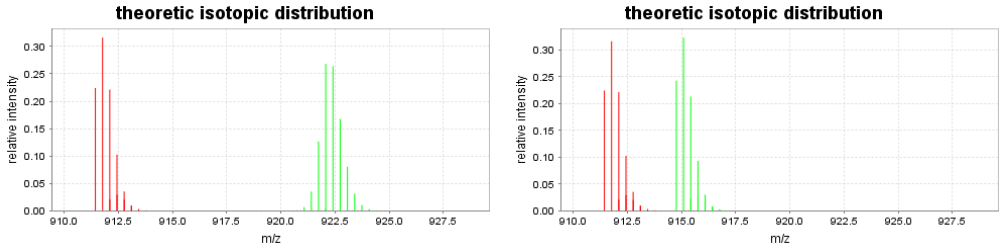


Figure 3.18 – This Figure illustrates the differences in mass but also in the form of the isotopic distribution for two different metabolic labeling approaches: while on the left side the heavier peptide (green) is labeled with ^{15}N with a common enrichment of 98%, on the right side $^{13}C_6$ $^{15}N_4$ arginine has been utilized for labeling.

3.3.1.3 Chemical tags: ICAT, ICPL, iTRAQ

With their proposal of a novel labeling approach based on a class of chemical reagents called isotope-coded affinity tags (ICAT) Gygi et al. (1999) belong to the pioneers in quantitative proteomics. The application of their reagent leads to a transformation of the side chains of all cysteinyl residues in a protein. After extraction, proteins are labeled with a light and a heavy form of the ICAT molecule containing either deuterium or 'normal' hydrogen. This differential labeling results in a mass shift of eight Dalton. The samples are combined and tryptically digested. A clear limitation of the labeling method is its specificity to peptides containing the amino acid cysteine. Prior to MS/MS analysis ICAT-labeled peptides are, therefore, isolated by a specific affinity chromatography to remove any bias from untagged peptides.

An improvement to ICAT that circumvents this limitation was developed by Schmidt et al. (2005). The approach termed isotope-coded protein label (ICPL) causes a derivatization reaction of the free amino group of proteins. It thereby allows to chemically label all proteins contained in an extracted protein sample.

A third, commonly applied chemical labeling strategy constitutes iTRAQ (Ross et al. 2004). While the original ICAT is practically limited to two labels, iTRAQ includes a set of up to eight isobaric reagents. These chemically modify peptides at the N-termini as well as at lysine side chains. In contrast to the aforementioned labeling methods quantification is based on the peptide's fragmentation after CID. Each reagent yields individual signatures that are distinguishable in the mass spectrum with mass shifts ranging from 113 to 121 Dalton.

Both advantage and disadvantage of ICAT, ICPL as well as iTRAQ is the fact that proteins have to be extracted before any labeling can be performed. While this allows the comparison

of organisms that are difficult to cultivate, it introduces at the same time an incalculable source of variation due to potentially different sample handling.

3.3.1.4 Absolute quantification: AQUA

While all aforementioned methods are only able to measure the relative abundances of proteins contained in a sample, Gerber et al. (2003) introduced a strategy termed AQUA to obtain an absolute quantification. Their idea is based on the utilization of artificially synthesized peptides with incorporated stable isotopes as an internal standard. Following protein harvesting, these synthesized peptides are added to the sample in a known volume. It is the crucial point of this analysis that a measurement can only be performed if both a labeled as well as an unlabeled variant of a specific peptide are available. The applicability of AQUA is therefore, in general, limited to a small number of proteins of interest.

3.3.2 Special application: analysis of protein turnover

Protein quantification using a metabolically incorporated label such as heavy stable nitrogen isotopes allows to determine protein abundance values by setting protein amounts, for example, under two different environmental conditions into relation. Since these measurements are relative, they, however, do not allow to formulate any statement about the causes of deviating protein amounts. An increased ratio, for instance, may originate from a faster synthesis rate of a protein at one condition but it may also be resultant from a reduced protein degradation. Pulse chase experiments allow to gain knowledge about the synthesis and degradation rates of a protein, which is vital for an in-depth interpretation of the protein turnover changes that occur during physiological adaptation processes or an emerging disease. In such experiments, a label is impulsively introduced in a living organism or cell, for example, in form of an essential nutrient such as ^{15}N-labeled glucose. It is then investigated whether and to what extent the label is incorporated in newly translated proteins. Already in the late 1940's, Sprinson and Rittenberg (1949) employed ^{15}N-labeled glycine as a diet to measure the utilization of nitrogen for protein synthesis. The analysis of protein degradation can be conducted in a similar manner by comparing the amounts of a protein before and after an induced pulse given that no newly synthesized protein influences the measurement, for example, if cell cultures are grown in a chemostat at steady-state. Several experiments have been devised and implemented to monitor protein synthesis to degradation rates, starting from a ^{15}N-algae diet for mice (Price et al. 2010), to $^{13}C_6$-Arginine for the investigation of human cells (Pratt et al. 2002; Doherty et al. 2009). A very interesting approach on this issue was carried out by Jayapal et al. (2010) on *Streptomyces coelicolor*: The transfer of $^{13}C_6$ $^{15}N_4$-arginine-labeled cells into unlabeled medium allowed the tracing of newly synthesized proteins. In addition, the chemical labeling method iTRAQ was employed to tag all SILAC-labeled proteins at four time points after the chase, which in turn made it possible to monitor protein degradation.

3.3.3 Two-dimensional electrophoresis

In the first instance, 2D-electrophoresis serves as a means to separate complex protein mixtures. Moreover, the technique also allows to make a statement about protein quantities—spot sizes give hints on the relative amount of each separated protein. While the traditional approach based on IEF and SDS-page is limited in terms of validity and reproducibility, e. g. due to inhomogeneity in different gels, a large proportion of the difficulties is avoided by the DIGE technique, which utilizes different dyes to compare two or more samples in the same gel (Lilley and Dupree 2006). However, the applicability of 2D-electrophoresis has several drawbacks concerning, for example, proteins with exceptional properties such as a high molecular weight. This is particularly problematic with regard to the identification of, in general, hydrophobic membrane proteins—doubtlessly a group of proteins that fulfills most important biological functions in a cell. Another limitation of the technology results from the typical complexity of proteomics samples. It is not unusual that a single spot in a gel does not contain only one individual enzymatically digested peptide species but instead a mixture of peptides, which moreover not necessarily have to belong to the same protein (Hamdan and Righetti 2005).

3.3.4 Label-free approaches

Apart from labeling approaches, which are often time-consuming, laborious, and comparatively cost expensive, a variety of strategies have been conceived (and not seldom discarded) to quantify proteins in a sample. A simple but nevertheless reasonable method, which is often practiced in LC-MS/MS experiments, is to count the number of times a peptide has been observed during a measurement. The idea of 'spectral counting' is based on the assumption that the more of a protein is in a cell, the more enzymatically digested peptides should be present in the investigated sample, which in turn should result in an increase in detected mass spectra for this protein. In fact, Liu et al. (2004) found correlations between the number of mass spectra and known protein abundances greater than $r^2 = 0.9997$, but only for a mixture of six protein markers. Hendrickson et al. (2006) performed a comparison of two datasets on *Methanococcus maripaludis* gained by spectral counting, on the one hand, and metabolic labeling using ^{15}N, on the other hand. In summary, a rather low correlation of $r = 0.58$ was observed between all investigated abundance values for all proteins. If only regulated proteins were taken into account, the correlation increased to $r = 0.89$, but still there remained large differences in the data. They draw the conclusion that spectral counting "performs poorly when counts are low [...], but performs quite well when counts and signal-to-noise are high", but also continue noting that "the low counts and (or) low signal-to-noise portion of the data is often of the greatest experimental interest" (Hendrickson et al. 2006, p.7).

In recent times, the idea of spectral counting has been improved and extended. A very similar method is not to count each spectrum individually but instead add up the number of all uniquely identified peptide sequences per protein. 'Absolute Protein Expression', abbreviated APEX (Lu et al. 2007), goes a step further and combines spectral counting with additional

information derived from the amino acid sequence of a peptide including the confidence in the peptide's detection (see section 4.2). In this way, peptides that are expected to occur only rarely in an experiment are weighted with a higher score. Another approach, the 'normalized spectral index (Griffin et al. 2010), considers both the spectral as well as the aforementioned peptide count in conjunction with intensity of the fragment ion (MS/MS) recorded by the mass spectrometer.

3.4 Shedding light on the importance of mass spectrometry for proteome research

This chapter aimed at highlighting the importance of mass spectrometry on proteomics not only for the identification of biomolecules but also for the assessment of their quantities, in particular, using heavy stable isotopes. Two different workflows have been exemplary introduced, and demonstrate the diversity of approaches to gain knowledge about the proteins contained in a cell or organism. As shown, both workflows have their advantages and disadvantages. MALDI-TOF mass spectrometry, typically used in combination with 2D-electrophoresis, allows to precisely characterize individual proteins. However, even though individual worksteps can be automated, e. g. using picking roboters, the experimental procedures are comparably tedious and time-consuming. Moreover, protein separation using 2D-electrophoresis often excludes interesting groups of proteins such as membrane proteins. In contrast, LC-MS/MS, in particular the MudPIT approach, offers to investigate complete proteomes in high-throughput experiments, which also facilitates the utilization of stable isotopes for labeling.

With the introduction to these workflows, the challenges are pointed out that have to be tackled by an integrated software solution to support experimenters in the conduction of these experiments. For example, a data model for mass spectra but also a user interface to view and process this type of data needs to take into account that a sample in a MALDI-TOF experiment, which has been extracted from a spot in a 2D-gel, results in only one mass spectrum, while an LC-MS/MS experiment often generates thousands of mass spectra from a single sample.

Chapter 4

State of the art in proteomics data analysis

A typical proteomics experiment may comprise more than one hundred liquid chromatography runs on a mass spectrometer, which in turn can consist of up to 40,000 individual mass spectra. The number of identified peptides (in case of MS/MS) may then easily reach hundreds or thousands, accounting, for their part, for several hundred proteins. It seems obvious that a thorough analysis of these amounts of data which constitutes a pile of relevant but also irrelevant information demands computational assistance. For this reason, the proteomics community offers a plenitude of software solutions aiming at two primary objectives: firstly, the provision of comprehensive data management capabilities to organize and structure the data and all associated meta data, and, secondly, the provision of analysis functionality to extract all relevant and important features, so to say, to pick the cherries out of the proteomics information cake.

4.1 Data standards in proteomics

An integral part of every proteomics experiment is the qualitative assessment of the proteins contained in a sample. In many cases, this is enhanced by the determination of protein abundance values to gain quantitative information about a proteome. The availability of a profound data basis represents a key element for the provision of subsequent analysis methods. Within the proteomics community, therefore, efforts are being made to create common data standards for the storage of experimental data and meta-data. This appears particularly important with regard to the manifold vendor-specific formats in which mass

spectrometry instruments produce their results. Often neglected, it is also the long-term archiving of experimental data that places particular demands on the storage formats. There is the risk of a 'semantic' data loss if a stored file is not readable any more due to the absence of a software application that could interpret its content. At this point, proprietary data formats have a considerable disadvantage in comparison to open document standards as implementation details are often inaccessible and typically bound to a specific company (Neuroth et al. 2009).

4.1.1 Human Proteome Organization: PSI, MIAPE and MIBBI

Inspired by the Human Genome Organization (HUGO), which supports collaborations between genome scientists in international projects, the Human Proteome Organization (HUPO) was founded in 2001 "to help increase the awareness of proteomics across society and biomedicine—in particular, the benefits that are offered by knowledge of the human proteome" (Huber 2003, p.75). The consortium unites several national and international research groups with an academic, governmental as well as industrial background. An ambitious project initiated by the organization is the "Human Proteome Project". Their declared objective is no less than the determination of the quantities and locations of all human proteins and their interactions (Pearson 2008).

The HUPO's Proteomics Standards Initiative (PSI) aims at the development of common data standards in proteomics (Orchard et al. 2003). Intended to standardize the information about conducted experiments and to facilitate the exchange of data, the project suggests a set of guidelines known as the minimum information about a proteomics experiment (MIAPE, Taylor et al. 2007). Individual workgroups focus on all aspects of proteomics experiments starting from data generation and analysis to the description of protein interactions and protein modifications (Taylor et al. 2006). As an example, the mass spectrometry group targets "the minimum information required to report the use of a mass spectrometer in a proteomics experiment, sufficient to support both the effective interpretation and assessment of the data and the potential recreation of the work that generated it" (Taylor et al. 2008a, p.1). The guidelines are following other efforts for example in the field of transcriptomics. Here, MIAME—the minimum information about a microarray experiment (Brazma et al. 2001)—beared fruit as software is developed in accordance to the guidelines and journals require MIAME-compliant transmission of data (Dondrup et al. 2009). MIAPE is registered with the "minimal information for biological and biomedical investigations"-project (MIBBI) which provides a general resource for "collaborative minimum information checklist development projects" (Taylor et al. 2008b, p.889). The project aims to harmonize the developments in different fields of research from genome sequencing to flow cytometry.

4.1.2 Institute for systems biology

Apart from the HUPO's initiatives, other groups are engaged in the development of data standards in proteomics. The Seattle Proteome Center (SPC) at the Institute for Systems Biology

aims to bring together knowledge in the field of proteomics. In 2005, the group proposed a set of guidelines similar to the ideas pursued with MIAPE stating which information should be included in the publication of experiments (Bradshaw 2005). The approach covers the reporting of the method which generated the mass spectrometry raw data, the algorithm used to identify peaks and its parameters as well as lists of identified and quantified proteins and their significance.

4.1.3 Data standards for mass spectra

The variety of mass spectrometer systems is vast, and their range of application in the field of proteomics mainly depends on the intended research objective of an experiment. Although most vendors aim to provide comprehensive software tools for the analysis of their data, in many cases specialized solutions have to be developed and implemented—a work that can often only be done by public research institutes or universities. In this connection, it is of course problematic if instruments produce their data in a multitude of data formats, which are moreover not seldom proprietary. For this reason, there is a strong need to unite the storage of mass spectra.

4.1.3.1 mzXML

Developed at the Institute of Systems Biology, mzXML (Pedrioli et al. 2004) represents an open and generic data format to store mass spectra in form of an extensible markup language (XML) document. The format supports MS data e. g. from MALDI-TOF analysis but also tandem mass spectrometry and even MS^n. A range of tools is provided to convert vendor-specific formats, parsers that are aware of an index-structure integrated into the format, and tools for visualization and validation. An interesting implementation detail concerns the storage of peak information. This would best be stored in a binary format, simply to preserve disk space. However, as binary data cannot be directly incorporated in XML documents, this was circumvented by base64-coding of the data.

4.1.3.2 mzData and mzML

In 2004, the HUPO's PSI proposed a data model for the storage of mass spectra data called mzData (Orchard et al. 2004). Similar to the aforementioned mzXML, the format utilizes XML to structure its content. In addition, tools are provided for data conversion and visualization.

It was not until 2008 before both initiatives to implement a common standardized data format for mass spectrometry data finally realized that two independently developed formats are obviously incomprehensible and counterproductive. Under the roof of the HUPO the two efforts were joined, and resulted in the new format mzML (Martens et al. 2010). This combines advantages of both formats, and furthermore integrates new aspects such as the possibility to assign different instrument configurations to individual mass spectra.

4.1.4 Ontologies and controlled vocabularies

Closely related to the effort to implement common data standards in proteomics is the necessity for a controlled vocabulary. Purpose of such an ontology is the unambiguous annotation of an experiment's datasets and associated meta data. The OBO Foundry aims to coordinate the development of ontologies that support data integration in biological and medical applications (Smith et al. 2007). In the context of genomics, the Gene Ontology provides a standardized way for the description of gene functions (Ashburner et al. 2000). Similarly, the individual workgroups of the PSI are publishing proteomics-specific controlled vocabularies.

4.2 Software for protein identification

Mass spectrometry allows to determine the masses of any analyte under investigation (see section 3.2). In a typical experimental workflow proteins are subjected to a digesting enzyme, which accordingly leads to the analysis not of a protein on the whole but instead of its peptide fragments. It further depends on the applied technology and workflow whether the observed peptide masses are directly utilized to identify a protein by using its so called 'peptide fingerprint' (see section 3.2.2), or whether a peptide undergoes a further, second, fragmentation as it is the case in MS/MS ion search (see section 3.2.3).

A variety of algorithms and software applications have been developed for the purpose of protein identification. The most ambitious and powerful approach allows to determine a peptide's amino acid sequence based solely on the spectral information of its MS/MS fragmentation pattern. Granted that each possible peptide fragment (see Figure 3.13) gives a distinct peak in the mass spectrum, each mass can be matched to a certain combination of amino acids and it is thus possible to derive the complete sequence of amino acids of this peptide. However, as soon as peaks are missing e. g. due to non-ionized fragments or incorrect peaks caused by foreign substances such as the solvent occur, the direct derivation gets difficult if not impossible. Moreover, the large number of possible amino acid combinations renders the determination of a peptide's complete amino acid sequence based on its mass spectrum expensive and is, thereby, in many cases impracticable.

The classical 'identification strategy' in mass spectrometry utilizes a library of prerecorded mass spectra for a list of known compounds (Martinsen and Song 1985). Identification then relies on the comparison of an observed mass spectrum with this library. Perfectly fitted for the field of metabolomics the approach renders useless in the face of the enormous number of different proteins and their possible enzymatic digestions.

A practicable way to gain knowledge of the proteins contained in a sample links the spectral information with the information about proteins yielded by genome sequencing projects. In peptide mass fingerprinting, the expected fragment masses resulting from an enzymatic digestion are compared to the observed masses in a mass spectrum (James et al. 1993). If MS/MS fragmentation patterns are available this, additionally, can be taken into account.

Mann and Wilm (1994) claim that in each MS/MS spectrum there exists at least a series of peaks allowing to determine parts of the peptide's amino acid structure. They proposed a software tool named PeptideSearch for protein identification based on these 'peptide sequence tags' in combination with additional information such as the molecular weights of other peptide fragments. Other approaches do not aim to determine an amino acid structure directly from the mass spectrum, but instead compare a recorded mass spectrum with theoretically expected patterns derived from sequence databases. The list of search engines that follows this approach includes OMSAA (Geer et al. 2004), ProbID (Zhang et al. 2002), X!Tandem (Craig and Beavis 2004), and the two most prominent members Mascot™ (Perkins et al. 1999) and Sequest™ (Eng et al. 1994; Yates et al. 1995).

4.2.1 Mascot™

Pappin et al. (1993) were the first to extend peptide mass fingerprinting with a scoring algorithm that takes the non-uniform distribution of peptide fragment sizes into account. As depicted in Figure 4.1, which shows the molecular weight distribution of all possible tryptic peptides of *Xanthomonas campestris pv. campestris*, smaller-sized peptides are, generally, more frequent than heavier fragments. Moreover, the relative frequencies depend on the overall protein size: for *Xanthomonas campestris pv. campestris* it seems to apply that the higher the molecular weight of a protein the less frequent are smaller peptides.

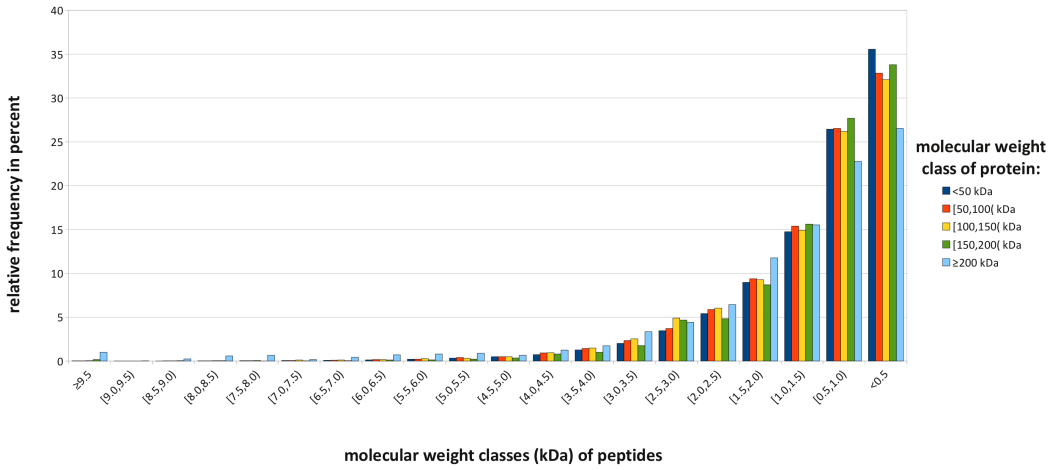


Figure 4.1 – Histogram showing the distribution of all possible tryptic peptides of *Xanthomonas campestris pv. campestris* (as of January 2010). Molecular weights for five different classes of molecular weights. Smaller-sized peptides are for example slightly more frequent for proteins having a maximal molecular weight below 50 kDa compared to heavier proteins, in particular above 200 kDa.

Given a list of observed mass values $\mathbf{m} = \{m_i\}$ for each m_i, $i \in \{1 \dots p\}$, a corresponding peptide mass m'_i is searched in an *in silico*-digested sequence database such that $m'_i - \varepsilon \leq m_i \leq m'_i + \varepsilon$. Here, ε defines an error window, typically termed 'peptide tolerance'. A score is then assigned to each 'hit' based on the frequency with which peptides with the

same mass m_i are found in other proteins with a similar molecular weight. The scores of all matching peptide hits aggregate to a total score for each possible matching protein. In the original publication, the final score was achieved by multiplication of the individual distribution frequency scores. In addition, the score was normalized to "an 'average' protein molecular weight of 50kD to reduce the influence of random score accumulation in large proteins" (Pappin et al. 1993, p.331). The original software tool was called MOWSE, and later on commercialized as Mascot™ by the company Matrix Science Ltd. In subsequent years, the 'search engine' was extended to support MS/MS ion search (Perkins et al. 1999). Unique to Mascot is the assignment of a probability to each protein score that reflects in how far a match may have occurred randomly taking into account, *inter alia*, the database size.

4.2.2 Sequest™

The algorithm proposed by Eng et al. (1994) and Yates et al. (1995) is based on the idea that CID fragmentation patterns generated by tandem mass spectrometry are reproducible and, moreover, predictable. In an initial step, the experimentally determined mass of a precursor ion is used to screen a sequence database for potential peptides with the same or at least a similar molecular weight. The algorithm then "converts the character-based representation of amino acid sequences in a protein database to a fragmentation pattern" (Eng et al. 1994, p.977). Therefore, the fragmentation pattern of each matching peptide is predicted, which yields e. g. a list of type-b and type-y ions for a typical ion trap or quadrupole analyzer (see Figure 3.13). The m/z values of each fragment that might occur are calculated by summing up the weights of the corresponding amino acids: given a peptide consists of J amino acids a_j, $j \in \{1 \dots J\}$, the i-th type-b ion corresponds to the sum of the mass values of its amino acids $b_i = \sum_{p=1}^{i} a_p + 1$. Analogously, it applies for type-y ions that $y_i = MW - \sum_{q=i}^{J} a_q$, where MW denotes the peptide's overall weight.

All candidate peptides that matched the mass of the precursor ion are ranked according to their total number of actually observed fragment ions. Thereby, a bonus is given for consecutive matching fragments, in terms of the peptide's amino acid sequence, in order to further reduce the search space. To finally identify a peptide (with a certain probability), a cross-correlation analysis is utilized that directly compares the experimentally determined MS/MS spectrum to the top-ranking candidates. For this purpose, artificial mass spectra are constructed based on the lists of fragment ions that have been predicted for each potential peptide. Both the recorded mass spectrum $\mathbf{d}$ and each artificially constructed mass spectrum $\mathbf{c}$ constitute discrete signals, which can furthermore be converted to comprise each N individual peaks d_n, c_n with $n \in 1 \dots N$ (e. g. by rounding on nominal masses). Given that $\tau \in \mathbb{R}$ represents an arbitrary offset between both signals, a cross correlation can then be calculated as follows:

$$X_{\mathbf{d},\mathbf{c}} = \sum_{n=0}^{N} d_n \, c_{n+\tau} \tag{4.1}$$

In the end, the best matching hit is characterized by the highest achieved cross-correlation value, and—best of best—with a sufficient difference between this score value and that of a

potentially second ranked hit. An improved version of the algorithm is nowadays used in the commercially distributed software Sequest™.

4.2.3 Evaluation of search results

Several problems may arise in protein identification using sequence databases. Difficulties start if the genome of an organism is not fully known. In such cases, a close relative may be used or a comprehensive database that includes a diversity of sequenced organisms such as the protein databases of the Universal Protein Resource (UniProt, The UniProt Consortium 2008) or the National Center for Biotechnology Information (NCBI, Wheeler et al. 2008). Difficulties continue, if search results are ambiguous: in peptide mass fingerprinting observed masses may match more than one peptide fragment, and in the worst case, the peaks of a mass spectrum account not for a single protein but a mixture of proteins. Hufnagel and Rabus (2006) argue that at least five and up to 50 peptides are necessary for an assured identification.

In MS/MS ion search another profound issue has to be addressed: even if a peptide has been identified based on its tandem mass spectrum with a high degree of certainty it may be questionable which protein it belongs to. In eukaryotes, this is beset with particular difficulties as different splice variants of one gene may result in proteins that share a common set of peptides.

Decoy databases: determining the false discovery rate of protein identification

A common strategy to evaluate the quality of peptide identifications is based on the utilization of so called decoy databases (Moore et al. 2002; Peng et al. 2003; Elias and Gygi 2007). The approach can be best illustrated with a metaphorical example: a search for the word 'BOOT'—as a synonym for a certain peptide—in a German dictionary—analogously to a sequence database—would yield a highly confident hit for a 'small sea vessel'. If, however, the same word is also searched in an English dictionary—analogous to the genome of another organism—the confidence dwindles as a second meaning 'footwear' becomes evident. In protein identification, the second 'dictionary' is usually a copy of the original sequence database where each protein's amino acid sequence has been either randomized or simply reversed. Using a modified and, thereby, equally-sized version of the same database ensures that also the probability of a false positive hit in both databases is, in theory, equal. Mass spectra are then searched in both the original and this (target-)decoy database. The resulting hits give hint on the reliability of the identifications: If a search revealed FP number of hits in the decoy database and TP hits in the original database a false discovery rate (FDR) can be estimated as follows:

$$\mathrm{FDR} = \frac{FP}{FP + TP} \tag{4.2}$$

The practical application of the FDR manifests itself in the implementation of filter criteria for protein or peptide hits. Hits are, for example, ranked by their score or correlation value,

and only those hits with a score above a threshold are accepted that result in an acceptable FDR, e. g. of $p \leq 0.05$ or 0.01.

4.3 Quantitative analysis of isotopically labeled data

The incorporation of isotopic labels in protein samples allows for a relative and under certain circumstances even absolute quantification of protein amounts (see section 3.3.1). While, in the end, the utilized label decides about the concrete implementation of a quantification method, all approaches to calculate relative abundance values between a labeled and an unlabeled sample follow a common procedure as depicted in Figure 4.2. In general, the information gained from a preceding protein identification is used as basis to determine a ratio, in the following also denoted as *M* value, for each identified peptide. Beginning from the mass spectrum that gave rise to the peptide identification, the crucial task is to demarcate the peaks that belong to the specific peptide and thereby determine its abundance. In case of the chemical labeling method iTRAQ the MS/MS scan can be utilized for quantification, whereas in most other approaches the full MS scan has to be used. One reason for this is the fact that in many cases only for one of the two peptides—either the labeled or the unlabeled variant—a corresponding MS/MS scan has been (automatically) recorded. Therefore, the complete information for quantification is only available in the parent spectrum (see 3.2.3.6). The known amino acid composition of the peptide can be translated into a molecular composition (4.2A), and under consideration of the utilized label the theoretically expected isotopic distribution resultant from a mass spectrometry analysis of this peptide can be estimated (4.2B). This gives hint to the exact m/z ranges in which the peaks of the labeled as well as the unlabeled variant have to be expected (4.2C) in the corresponding mass spectrum. In this context, it is important to consider a peptide's charge state as it leads to a division of the observed mass by a factor equal to the value of the charge.

If protein separation has been performed by liquid chromatography, the gained temporal information provides a possibility to significantly improve a following quantification—the elution of a peptide can be taken into account. This is demonstrated in Figure 4.3 for a number of subsequently recorded mass spectra: the fact that a peptide elutes in a time frame of a few seconds allows the extraction of ion chromatograms (EIC or XIC) for each labeled and unlabeled peptide which can then be used for quantification. In the following a selected choice of software tools is introduced. In accordance with the objectives of this work, the mentioned algorithms will mainly concentrate on metabolic labeling approaches. A detailed discussion and comparison of different applications can be found for example in Nesvizhskii et al. (2007), and Mueller et al. (2008).

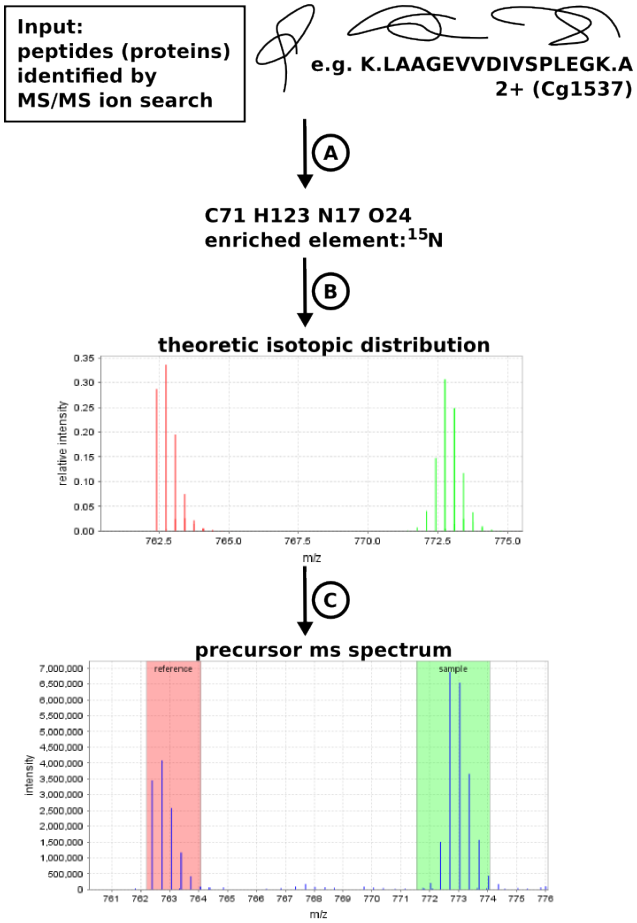


Figure 4.2 – General procedure to calculate relative abundance ratios of two peptides: one unlabeled, and one fully labeled (in the example with heavy stable nitrogen). A) In general, the information gained from protein identification, namely the amino acid sequence, charge state, modifications and the associated protein accession number, are used to calculate B) each peptide's molecular composition and its theoretical isotopic distribution. In this context, it is of no relevance if both peptide variants have been identified in a preceding database search. In general, the type of labeling is known *a priori* and the missing information can be deduced from the available molecular composition of a peptide. C) Based on the m/z ranges determined by the isotopic distribution, the associated peptide intensities can be extracted from the recorded mass spectrum.

4.3.1 ASAPRatio

ASAPRatio (Li et al. 2003) has been developed for the quantification of ICAT and SILAC labeled proteins from LC-MS/MS data. The process to gain quantitative values not only for a specific peptide but complete proteins consists of several steps.

At first, based on present MS/MS peptide identifications, ion chromatograms are extracted using the first three peaks of each unlabeled peptide and its labeled partner. The two chromatograms are smoothed using a Savitzky-Golay filter (Savitzky and Golay 1964) to eliminate

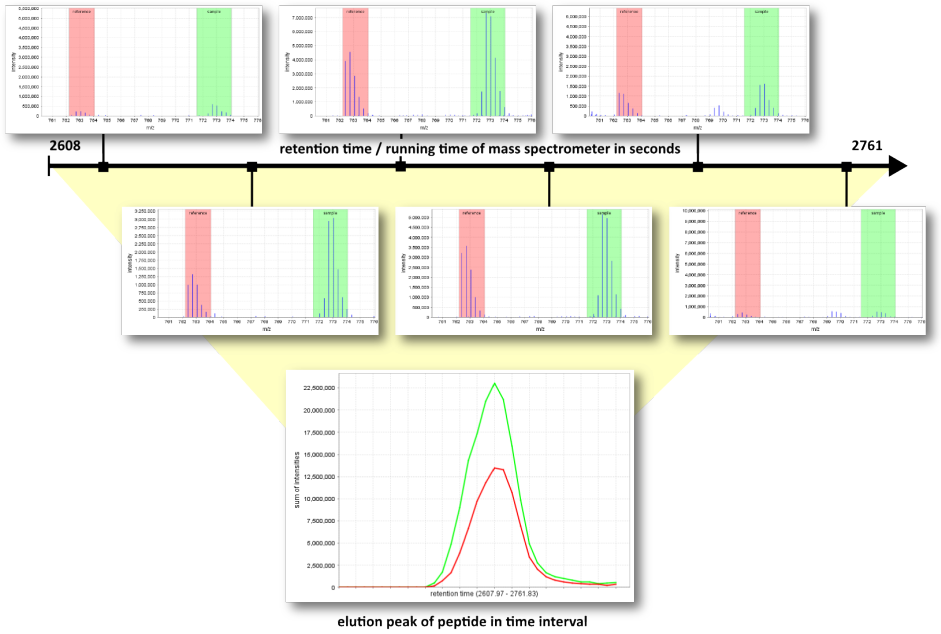


Figure 4.3 – If peptides have been separated using chromatography, quantification can significantly be improved if a peptide's elution is factored into the calculation. The fact that a peptide elutes not only at a defined time point but, typically, in a time frame of a few seconds ion chromatograms can be extracted (EIC) for each labeled and unlabeled peptide.

unwanted differences in the measurements. After both elution peaks for the labeled as well as the unlabeled peptide are detected from the smoothed chromatograms, the relative abundance of a peptide is calculated by setting the areas under the elution peaks in relation. Before a measurement is accepted, peak intensities in the neighborhood of each detected elution peak are used as an estimate of present background noise. It is then investigated whether the ratio between the peak apexes and this background exceeds a given threshold. ASAPRatio, in addition, allows to take into account any elution peak shift between the labeled and the unlabeled peptide. This may, for example, result from the labeling e. g. with hydrogen isotopes. Furthermore, different charge states of the same peptide are considered in the quantification. Particularly in LC-MS/MS experiments, it may happen that a peptide is identified only in one specific charge state, although the same peptide may also be found differently charged.

At second, peptide measurements are combined to form an overall protein abundance ratio. As a special feature of ASAPRatio, for each peptide abundance ratio an error measure "is estimated by the signal difference of the raw and the smoothed chromatogram" (Li et al. 2003, p. 6651). Using this error, and in addition an outliers test, a weighted protein abundance ratio is averaged. Results are presented in a CGI generated web page.

4.3.2 RelEx

Belonging to the first implemented software tools for the quantification of proteins, RelEx (MacCoss et al. 2003) allows the calculation of relative abundance ratios from isotopically—predominantly metabolically—labeled samples. The program is written in the programming languages C and Visual Basic and runs only on the Microsoft™ Windows platform. RelEx demands existing protein identifications as input. Initially, a tool named EXTRACT-CHRO is used to construct extracted ion chromatograms (EICs) based on these peptide identifications and the corresponding raw files from the mass spectrometer for each labeled and unlabeled peptide pair. A user-defined window, which defaults to 100 MS scans around any identifying MS/MS spectrum, is used as starting point for peak detection.

RelEx differs from other approaches in that the calculation of peptide abundance ratios is not based on the calculation of the area under the elution peak's curve but on least squares regression: the values of both EICs (in the range of the most intense elution peak) are set in relation and the slope of the regression line between these values gives hint to the peptide's relative abundance. In the end, RelEx allows the aggregation of all peptide abundance ratios to finally output a list of protein abundance ratios. Similar to ASAPRatio (Li et al. 2003), chromatographic shifts may be taken into account, and outliers can be removed using a Dixon's Q-test.

4.3.3 ProRata

Pan et al. (2006) designed a Microsoft™ Windows based software tool named ProRata for the quantitative analysis of metabolically labeled proteomics samples. The approach closely follows the idea of RelEx but replaces linear regression analysis with a principal component-based approach for the calculation of relative abundance ratios. It is proposed that the first principal component refers to the 'true' ratio between the signals of the labeled and the unlabeled peptide while the second principal component explains any present noise in the data. Based on the eigenvalues of both components as measures of the variance of the 'true' ratio and the noise, a so called profile signal to noise value is defined that can be used to evaluate the quality of any calculated ratio. As an additional feature, ProRata introduces an algorithm termed "parallel paired covariance algorithm" to detect the elution peaks of both the EIC of the labeled and the unlabeled peptide. In contrast to other software tools, elution peaks are not searched separately in the original EICs, but in a derived chromatogram that consists of the covariances between both signals.

4.3.4 Census

Census (Park et al. 2008) is the successor of RelEx (MacCoss et al. 2003) and has been developed in the same laboratory. Although the program has been written in Java parts of the application, e. g. a tool for the creation of EICs from mass spectrometry data, depend on Microsoft™ Windows. Census supports a variety of quantification approaches for LC-MS/MS

data starting from iTRAQ to metabolic labeling using stable isotopes, SILAC and even label-free quantification. In comparison to RelEx, the tool has, particularly, been improved with regard to high-resolution mass spectrometry data.

4.3.5 QN

QN (Andreev et al. 2006) is another software tool that has been developed for the quantification of proteins containing metabolically incorporated heavy stable nitrogen isotopes. The Microsoft™ Windows based software, which has been written in Visual Basic and MATLAB, puts its focus on high-resolution mass spectrometry data from LC-MS/MS experiments, particularly using an LTQ-FT MS instrument. Based on existing protein identification results, EICs are created for both the labeled and the unlabeled peptide. In contrast to most other approaches, only the monoisotopic peak is included, while any other isotopic peaks are used to validate the correctness of the peptide identification. QN features the calculation of a reliability score for each resultant peptide abundance ratio based, *inter alia*, on the height of the included peak intensities and the validity (score or probability) of the original peptide identification.

4.3.6 QuantiSpec

The tool QuantiSpec has been developed for the relative quantification of MALDI-TOF data (Haegler et al. 2009). In contrast to all other aforementioned approaches it allows to set in relation unlabeled and fully-labeled but also partially-labeled samples. Currently, the software tool supports only heavy stable nitrogen isotopes. QuantiSpec is implemented in Perl and provides a GTK2-based graphical user interface. It is intended to run on Microsoft™ Windows.

4.3.7 MaxQuant

MaxQuant is denoted "an integrated suite of algorithms" (Cox and Mann 2008, p.1367) developed for the quantification of SILAC-labeled proteins. It follows a fundamentally different approach in that protein quantification precedes protein identification. The three-dimensional feature space spanned by an LC-MS/MS dataset, which is characterized by each spectrum's m/z to intensity values and, in addition, the temporal dimension given by the elution time, is searched for characteristic peak patterns that indicate pairs of labeled and unlabeled peptides. This is, in particular, possible since the mass shift that is introduced by a labeled amino acid such as $^{13}C_6$ arginine (see section 3.3.1.2) but also the form of the isotopic distribution (see Figure3.18) are to a great extent predictable and consistent. For each identified pattern that denotes a potential peptide the change in abundance is calculated using linear regression analysis. Only at this point, protein identification is started based on the—so far unconsidered—MS/MS spectra of the dataset. MaxQuant utilizes the Mascot™ search engine for this purpose. Not necessarily being a disadvantage, one has to denote in this

connection that certainly not for all quantified peptides an appropriate MS/MS spectrum is available, which could be used for a reliable determination of each peptide's sequence. Conclusively, there may remain quantified but not identified peptides in the final result. The software is implemented using the Microsoft™ .NET framework and designed to be run on Microsoft™ Windows.

4.4 Data storage and management solutions

Starting from the preparation, extraction, and measurement of protein samples in the wet lab, a typical quantitative proteomics experiment involves several—often laborious and time-consuming—experimental steps. In general, the data generation is followed by a database search to identify the proteins contained in a sample, and—if applicable—the calculation of relative (or absolute) abundance values. Information collected during an experiment includes mass spectra and lists of reported proteins but also the documentation of a chosen experimental setup.

Already during the conduction of an experiment, it is in general advantageous that all data and meta-data is brought together and stored in a common place. This facilitates not only the retrieval of information but also its validation, e. g. of protein identifications by setting search results from different databases in comparison. Furthermore, in many fields of application, special attention must be paid to the issue of long-term archiving and access to all experiment-relevant information.

A number of software applications have been developed that provide a standardized way of data storage and data management. Typically these applications are referred to as laboratory information management systems. Within these systems data and meta-data are stored either using specific flat file formats or in a database. As experiments are frequently conducted in cooperative work within larger teams, information has to be shared between a number of participants. Therefore, user management and data access control are vital components of these systems. Examples of applications for this purpose are CPAS (Rauch et al. 2006), MASPECTRAS (Hartler et al. 2007), Proteios (ProSE) (Gärdén et al. 2005; Levander et al. 2009), and the command-line based Trans-Proteomics pipeline (TPP) (Keller et al. 2002; Nesvizhskii et al. 2003).

In addition to these laboratory information management systems and with particular regard to the steadily increasing amounts of data, it is desirable that information about conducted experiments is shared within the proteomics community. The exchange of information does not only enable the global-scale comparison of experiments but it can also avoid the unnecessary duplication of experimental data. This task is performed e. g. by PRIDE (Vizcaíno et al. 2009)—a data repository for the proteomics community.

4.4.1 Laboratory information management systems

The term laboratory information management system, abbreviated LIMS, refers to software applications that provide data storage and retrieval capabilities for all data and meta-data related to an experiment. This typically covers experimental procedures as conducted in the wet lab as well as parameters and settings of any software application used for data analysis. Strictly speaking, a LIMS provides only capabilities for data management. From this point of view, the selected choice of applications that is presented in the following goes beyond the scope of a LIMS expanding the meaning of this term to also include analysis functionality for proteomics data. A comprehensive comparison of different LIMSs has been made available for example by Stephan et al. (2010).

4.4.1.1 ProDB

Andreas Wilke designed ProDB as a system that "integrates the analysis and storage of mass spectra with a detailed description of the experimental setup" (Wilke et al. 2003, p.155). The software, which has been developed at the CeBiTec, aims at the evaluation of 2D-electrophoresis data, mainly, in combination with MALDI-TOF mass spectrometry. It does not support modern high-throughput methods such as MudPIT. Great emphasis was placed on a comprehensive LIMS system to describe the experimental procedures that have been employed to obtain and analyze the samples of an experiment. ProDB has been implemented in Perl and provides both a GTK- as well as a CGI-based graphical user interface (GUI). It is such available as a stand-alone desktop and a web application. Data storage is based on a relational database system. ProDB is, in a way, the predecessor of QuPE, which incorporates aspects of the extensive LIMS system.

4.4.1.2 CPAS

The 'Computational Proteomics Analysis System' (CPAS) constitutes "an open-source, web-based analysis platform that organizes and annotates general biological experiments and provides capabilities for managing and analyzing LC-MS/MS proteomics data" (Rauch et al. 2006, p.112). As a special feature, CPAS does not only allow the import of Mascot™ as well as Sequest™ results but also includes its own database search engine X!Tandem (Craig and Beavis 2004).

CPAS has been written in Java and demands an Apache Tomcat web server (The Apache Software Foundation 2011b) for running. The user interface has been designed using Struts (The Apache Software Foundation 2011a). User authentication can refer to an existing authentication provider with the help of the Lightweight Directory Access Protocol (LDAP). The authors claim that the web application is adaptable to any type of database management system—at least Microsoft™ SQL Server (Microsoft 2011a) and PostgreSQL (PostgreSQL-Team 2011) are currently supported. Distributed under the Apache 2.0 license the system is open for further community-extensions.

Although CPAS features comprehensive possibilities to describe a proteomics experiment, and even provides an integrated pipeline to conduct database searches for protein identification it does not offer a module for protein quantification. CPAS has recently been renamed to 'LabKey Server platform' (Nelson et al. 2011). The new version of the platform features, *inter alia*, advanced data visualization capabilities using the R statistical programming language.

4.4.1.3 MASPECTRAS

Hartler et al. (2007) developed the 'MAss SPECTRometry Analysis System' (MASPECTRAS) for the management and analysis of LC-MS/MS data. The web application allows the import of search results from various search engines such as Sequest™, Mascot™, X!Tandem (Craig and Beavis 2004) or OMSAA (Geer et al. 2004). The analysis pipeline of MASPECTRAS includes extended functionality to verify protein identifications based on a probability score, which is computed for all imported database search results. The approach follows an idea that has been suggested by Keller et al. (2002). As an additional and unique feature, the authors have implemented a cluster algorithm to evaluate the assignment of peptide identifications. This addresses the aforementioned severe problem in proteomics (see section 4.2.3): given a search engine has identified a specific peptide based on an MS/MS spectrum, it is often questionable which protein this peptide belongs to. This applies, in particular, to eukaryotes in which different isoforms of the same protein may exist that all share a common peptide. Protein quantification can be conducted using an integrated implementation of the algorithm used in ASAPRatio (Li et al. 2003).

MASPECTRAS is implemented in Java and utilizes a three-tier architecture model. It is compliant to the Java Platform Enterprise Edition (Java EE) specification. Similar to CPAS, the user interface has been designed using Struts (The Apache Software Foundation 2011a). A MySQL (Oracle 2011b), PostgreSQL (PostgreSQL-Team 2011) as well as an Oracle™ Database (Oracle 2011c) can be utilized as database backend. The underlying data scheme is based on the PEDRo data model (Garwood et al. 2004), which also influenced the developments of the HUPO's PSI (4.1.1). The web application allows the outsourcing of computational extensive tasks on a compute cluster. This is mediated via an own implementation of a web service-based interface. The graphics library JFreeChart is utilized to display for example imported mass spectra (JFree.org 2011).

Although MASPECTRAS provides very comprehensive capabilities for data management and analysis in terms of protein identification and quantification, the web application lacks advanced statistical data analysis features.

4.4.1.4 ProSE/Proteios

In 2005, Gärdén et al. introduced the software application 'Proteios' as a web-based application for proteomics data management. Four years later, the platform was rewritten and republished as the 'Proteios Software Environment' (ProSE) (Levander et al. 2009) augmented

by comprehensive data analysis functionality. ProSE features the integration of search results from different search engines including OMSAA (Geer et al. 2004) and X!Tandem (Craig and Beavis 2004). It provides user interfaces to define search parameters and allows the initiation of database searches directly from a web browser. In addition, Mascot™ results can be imported. Using the application, search results from different engines can be combined and verified based on the utilization of decoy databases and the calculation of false discovery rates (see 4.2.3).

ProSE is written in Java and utilizes Hibernate (JBoss Inc. 2011) to map Java objects on a relational MySQL database (Oracle 2011b). The web application is designed to run on a Tomcat web server (The Apache Software Foundation 2011b). Apart from this web server, a file transfer protocol (FTP) server is integrated that facilitates data import and export. Similar to MASPECTRAS, computationally intensive tasks are placed in a queue and automatically processed to reserve sufficient compute resources for the user interface. ProSE provides an application programming interface (API) that eases the extension of the platform in form of plug-ins.

4.4.1.5 Trans-Proteomics pipeline

The 'Trans-Proteomics Pipeline' (TPP) consists of a number of individual tools (Keller et al. 2002; Nesvizhskii et al. 2003) including the aforementioned software application ASAPRatio (Li et al. 2003), which allows to determine relative abundance ratios from isotopically labeled proteins. A software named XPRESS (Han et al. 2001) is integrated for the quantification of chemically, e. g. ICAT, labeled proteins. The user interface of the TPP is provided through a Perl/CGI-based web application running on an Apache web server. Installation is supported under Microsoft™ Windows as well as the Linux operation system. The pipeline supports the import of protein identifications from the three search engines X!Tandem (Craig and Beavis 2004), Mascot™, and Sequest™. As part of the pipeline, the two applications PeptideProphet and ProteinProphet allow the assessment of search results on the peptide as well as on the protein level. This can be based on the utilization of decoy databases and thereby estimated false discovery rates, but in addition Keller et al. (2002) developed and implemented a statistical model that takes into account various search related scores and parameters such as the number of cleavage sites to evaluate the accuracy of peptide identifications.

4.4.2 Data repositories

Data repositories for proteomics data and meta-data aim at the long-term archiving of experiments. In general, all information stored in these databases is freely accessible (in some cases a membership is required) and comprehensive data retrieval functionality are provided. The most important repository is PRIDE, which is hosted by the European Bioinformatics Institute (EBI).

4.4.2.1 PRIDE – proteomics identifications database

The 'Proteomics Identifications Database' (Vizcaíno et al. 2009), abbreviated PRIDE, has been designed as an online-available repository to fulfill two main purposes: researchers can submit their experimental data, which is oftentimes desirable in connection with a publication of the work, and thereby allow others to retrieve this data and better comprehend the work or compare it to own results. However, experiments can also be stored in a private context being made available only for privileged partners, e. g. for data exchange. All uploaded data is stored in a special purpose data management system termed BioMart (Smedley et al. 2009), which facilitates complex data retrieval operations. The repository supports all major open source data formats including mzData and mzXML (see section 4.1.3) for raw mass spectrometry data as well as for the representation of protein identifications. Currently (as of March 2012), PRIDE comprises more than 21,000 experiments consisting of over 8.1 million identified proteins and 280 million mass spectra.

4.4.2.2 PeptideAtlas

PeptideAtlas is the name of a project that has been designed and implemented by the Seattle Proteome Center (SPC) at the Institute for Systems Biology. It focuses at the collection and provision of peptide sequence data, which has been yielded in LC-MS/MS experiments, and its mapping onto genome sequences. The project's motivation originates from the idea that the information about genes and their products "can be enhanced through the collection of different types of experimental data and its integration and validation in a genomic context" (Desiere et al. 2005, p.R9). To ensure a common basis for data comparison, all contributed mass spectra datasets are processed by the TPP (see section 4.4.1.5) upon submission. As of March 2012, the repository contains more than 750 samples, of which 91 are from human plasma alone making up more than 3 million identified peptides/mass spectra (Farrah et al. 2011).

4.5 Identification, quantification, ... and next?

Apparently, the initial analysis steps of any mass spectrometry-based quantitative proteomics experiment concern, firstly, the identification, and secondly, the quantification of the proteins that are contained in a number of investigated samples. For this purpose, there exist a variety of software applications that support not only the aforementioned two tasks but also allow to store, query, and combine datasets and additional information. A selected choice of these tools and algorithms was introduced in this chapter. At this point of analysis, the preliminary result is, simply spoken, a list of identified proteins together with their abundance ratios. A typical experimental setup, however, includes in general more than one condition, and thus the resulting values actually need to be combined to form e. g. a data matrix (Kumar and Mann 2009). This can then be further processed using methods of statistical analysis and

data mining. Arising questions are, for example, whether proteins are up- or down-regulated regarding the investigated conditions, whether clusters of proteins show similar expression profiles, or whether observed differences in the proteomes of different individuals allow the prediction of diseases or defects.

Software such as spreadsheet programs or statistical programming languages, albeit generally usable for this purpose, demand a high level of background knowledge and training, or do not adapt to the complexity of proteomics data. Their most important drawback is, however, that they do not allow to directly link analysis results to the originating raw data. If data and associated meta-data are found connected all at the same place, it would for example be possible to thoroughly investigate a differentially regulated protein taking into consideration every individual peptide abundance ratio including potential quality measures of the calculation or even the underlying mass spectra.

4.5.1 Spreadsheet-alike analysis of proteomics data: DAnTE, StatQuant, GProX

Different applications have been introduced that provide a range of statistical methods for proteomics data. Two stand-alone software applications which feature comprehensive statistical analysis and visualization methods and which are based on the Microsoft™ .NET framework (Microsoft 2011b) are DAnTE (Polpitiya et al. 2008) and GProX (Rigbolt et al. 2011). StatQuant (Breukelen et al. 2009) fulfills a similar purpose. It is implemented in the Java programming language. The workflow of all three applications starts with the import of one or more data matrices in form of a Microsoft™Excel sheet or as a tab- (or character-) delimited text file. DAnTE has the unique feature that it is not only limited to proteomics datasets but also supports other types of PolyOmics data. GProX claims to support particularly complex experimental setups. None of the applications does, however, allow to integrate any originating data such as mass spectra or search results from a database search. Moreover, they do not provide data management functions, e. g. to group and archive all data that belongs to an experiment.

4.5.2 Integration of functional annotation data: PIPE

Tools such as the 'Protein Information and Property Explorer' (PIPE, Ramos et al. 2008) do not aim at the statistical evaluation of quantitative proteomics data but focus at the functional analysis of identified proteins. The principal task of PIPE is the collection and integration of additional information from different databases such as Uniprot (The UniProt Consortium 2008) or the Gene Ontology (Ashburner et al. 2000). Therefore, lists of protein accession numbers can be uploaded and securely stored with access restriction. PIPE is implemented using the Google Web Toolkit (Google 2011) and makes use of the R-programming language (R Development Core Team 2011).

4.6 An inventory of the current state of proteomics software tools and applications

This chapter took an inventory of the current state of the art in proteome data analysis, and showed that there is a broad range of software applications available to support experimenters in the conduction of their experiments, especially concerning the identification and quantification of proteins. There are, however, several significant drawbacks: first of all, there exists no software application that integrates all data and meta-data of an experiment in one place and, moreover, provides the functionality for the complete workflow of a quantitative proteomics experiments up to the statistical analysis of complex experimental setups. Besides that, algorithms for the quantification of proteins are still worth of improvement. Targeting the specific application of pulse chase experiments there has up to now no integrated software solution been introduced that allows to calculate synthesis and degradation rates from isotopically labeled LC-MS/MS data in a high-throughput manner.

Chapter 5

Requirements: computational support for quantitative proteome experiments

As has already been noted, current mass spectrometers such as Thermo Scientific's LTQ Orbitrap Velos™ are able to record up to 40,000 mass spectra in a single run, thereby producing data files with several hundred megabytes in size. Since a typical experiment consists of dozens of these files, there is, undoubtedly, a strong need for computational assistance in handling, organizing, and particularly, analyzing these amounts of data. Over the past decade, a variety of software tools and applications has been introduced by the scientific community to cope with this quantitative data (see chapter 4). On closer examination of the current state of the art in proteomics software, it becomes clear that there is, however, no application available that provides an integrated solution covering all aspects from data acquisition to data evaluation. Moreover, several issues concerning an in-depth and comprehensive analysis of quantitative proteomics experiments remain unaddressed. This applies, for example, to the quality and accuracy of methods for the calculation of relative protein abundance values from isotopically-labeled protein samples including specific applications such as pulse chase experiments. Furthermore, methods for the statistical analysis of this multivariate type of biological data have not been evaluated yet, not to mention the design and validation of a comprehensive workflow to draw reliable conclusions from stable isotope labeled samples.

Following key requirements can be formulated that need to be addressed by a software solution to fully support experimenters in the conduction of mass spectrometry-based quantitative proteomics experiments:

1. A basic requirement is the provision of data management capabilities in terms of data storage and archiving. This includes, firstly, the retrieval not only of raw datasets but also of filtered and prepared information derived from the data, and secondly, an addressing visual representation of data and analysis results: to reveal the underlying meaning of data an appropriate visualization is indispensable. "Modern data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information" (Tufte 2007). An integral requirement of such a software is to ensure the security and the integrity of any entrusted data. Therefore, user management, authentication, and access control have to be implemented.
2. The analysis and processing of quantitative proteomics datasets demands the provision of appropriate analysis functions and tools. In order to facilitate the integration of existing methods but also the development and evaluation of novel algorithms and tools for the processing of mass spectrometry data and meta-data, it is necessary that developers can make use of a standardized programming interface. In this connection, particularly with regard to the enormous amounts of data, it is furthermore advantageous to ease the use of distributed compute resources.

In a wider sense, such a software application can be regarded as an information system about mass spectrometry data (Parker et al. 1994). As such, it allows researchers to organize and annotate the data that belongs to an experiment and helps to find an answer to the overarching question: what can be learned from the data?

5.1 Use case analysis

The basic data type of any proteomics experiment is the mass spectrum. Yet, their concrete format depends on the utilized instrument, i. e. the ion source such as MALDI or ESI and the characteristics of the mass analyzer, as well as the instrument's structure, e. g. in form of a tandem mass spectrometer (see 3.2). In general, instrument vendors rely on their own, proprietary data format. As discussed in 4.1, data exchange and community-driven support require data to be independent of any protected format—it should be easily readable and editable. Therefore, the community has developed various open source data formats such as mzXML and mzData (see section 4.1.3). An information system for mass spectrometry data must be capable of understanding these data formats, and must allow to browse and display its contents.

Obviously, a proteomics experiment is more than the sum of its parts—the individual mass spectra. Based on the mass spectra a multitude of information is to be yielded not only about the identified proteins in a sample but also their relative or absolute quantities. These types of data have to be stored in the system. Moreover, it is necessary to provide interaction and connectivity, i. e. an identified protein should be linked to its mass spectrum and vice versa. The system, furthermore, needs to provide extensive capabilities to group and integrate

all data relevant to a particular experiment. This comprises mass spectra data but also a description of the experimental setup as well as all analysis results.

The typical workflow of a quantitative proteomics experiment involves several steps from data acquisition to analysis. Although the sequence of the individual steps naturally differentiates between different experiments, there is still a high degree of congruence. Referring to the analysis of LC-MS/MS data (see section 3.2.3), but without loss of generality, e. g. regarding MALDI-TOF experiments 3.2.2, a typical workflow is depicted in Figure 5.1.

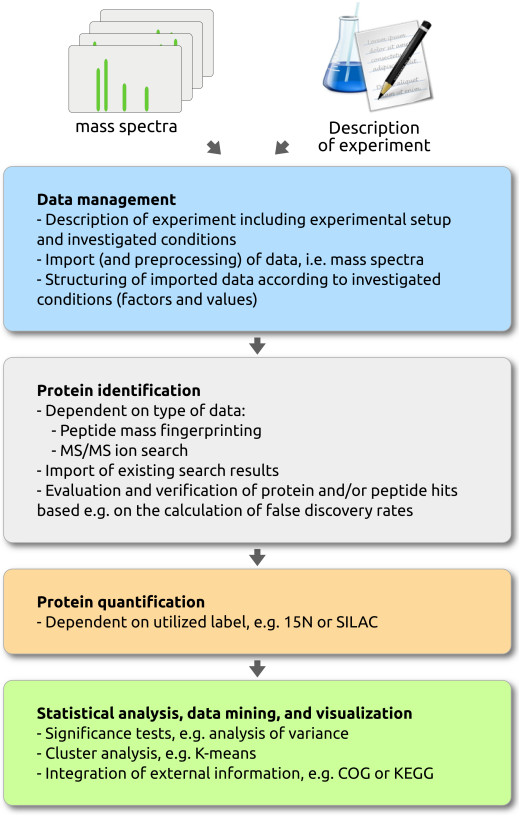


Figure 5.1 – The diagram depicts a typical workflow to quantitatively analyze isotopically labeled data from mass spectrometry-based experiments: starting with the import of data and the description of the experimental setup to protein identification, protein quantification and further statistical analysis, data mining and visualization.

5.1.1 Data organization and structuring

To structure all data relevant for an experiment—beginning from raw mass spectra to lists of identified peptides—appropriate data representations need to be devised and made available. Projects might be used to group related experiments. An important aspect is data security: it is necessary to ensure that only authorized persons are allowed to read datasets. This involves authentication of a user at login using a password and, in addition to that, an authorization

before certain methods are allowed to be processed, such as delete or edit operations. Access control list directives (ACLs) provide an opportunity to implement fine-granular privileges on individual objects allowing, for example, an experimenter to read-access an experiment and all associated datasets but denying the same user any modification of the experiment's content.

Comprehensive descriptions of an experimental setup allow for future retrieval of the worksteps carried out by an experimenter in the laboratory and the way individual samples have been treated. Within the scope of a practical training (Gau 2008) a set of five worksteps has been identified providing a sufficient description of a typical proteomics experiment: "Cultivation", "Protein extraction", "Protein filtration", "Digestion", and "Mass spectrometry". From this point of view, a workstep would, for example, describe the cultivation of an organism including parameters such as the optical density at time of harvesting and the utilized growth medium.

Mass spectrometry data is generated in various formats and instrument vendors often distribute their own, mostly proprietary, data format as does the company Bruker (Bruker Daltonics, Billerica, MA) by using a binary format for single-stage mass spectrometry data (luckily relevant peaks are written out in an XML-based format). Nevertheless, a range of freely available open source formats has emerged, namely mzXML (Pedrioli et al. 2004), mzData (Orchard et al. 2004), and mzML (Martens et al. 2010). As most vendor specific formats can be converted, an information system for proteomics has to target these data formats. Appropriate visualizations of imported mass spectra allow to validate that measurements were successful but also to compare individual scans. Furthermore, it is often necessary—as a first step in data analysis before protein identification and quantification can be applied—to preprocess mass spectra using peak detection methods such as MassSpecWavelet (Du et al. 2006) or Bruker's SNAP™ algorithm (Köster and Holle 1999).

5.1.2 Protein identification

Dependent on the type of imported data, e. g. tandem mass spectra, peptide mass fingerprinting (PMF) or MS/MS ion search (MIS) are the methods of choice to identify the proteins contained in a sample. The usual way is to employ a search engine such as Mascot™ or Sequest™ for this purpose, which base on the comparison of the recorded mass spectra with theoretical fragmentation patterns derived from sequence databases (see section 4.2). In an ideal situation, several sequence databases, such as an organism-specific and a general-purpose sequence database like UniProtKB/Swiss-Prot (The UniProt Consortium 2008), can be searched at once—at best using different sets of parameters, e. g. prohibiting one or more missed cleavage sites of an utilized proteolytic enzyme.

Afterwards, search results have to be combined and it is necessary to assess the hits reported from one or more search engines, found in different databases using different sets of parameters. Extensive research focused on the development of methods to ensure that further analysis can rest on a solid ground of valid protein identifications. A commonly accepted

strategy, which has been put forth by Peng et al. (2003) as well as Elias and Gygi (2007), bases on the utilization of decoy databases to estimate and control the false discovery rate (FDR) for database search results (Reidegeld et al. 2008).

5.1.3 Protein quantification

The combination of mass spectrometry and isotopic labeling techniques represented a key milestone on the way towards a comprehensive comparison of protein abundances under different environmental conditions, disease states, or physiological adaptation processes (Mallick and Kuster 2010; Gouw et al. 2010; Hufnagel and Rabus 2006; Bantscheff et al. 2007; Mueller et al. 2008; Zhu et al. 2002; Ong et al. 2002; MacCoss et al. 2003; Wolters et al. 2001). Various software programs have been introduced to quantify protein amounts (see section 4.3). A software application for the analysis of quantitative proteomics data needs to integrate the results of these tools or provide own implementations of quantification algorithms.

A specific problem, still seeking a solution, is the analysis of metabolically labeled samples, in which a label has not been fully incorporated. There is, hitherto, no publicly available and easy-to-use algorithmic approach available that allows to perform an automated, high-throughput protein quantification of this type of data. On top of that, as it will be shown in chapter 8, the accuracy and performance of currently available quantification methods is still worthy of improvement.

5.1.4 Statistical analysis, data mining, and visualization

The next step after protein identification and quantification is, undoubtedly, the most important (and probably exciting) one: while the data basis is—in a manner of speaking—prepared and ready, the focus now turns towards a thorough interpretation of the data. Therefore, methods of statistics and data mining are indispensable. As already stated in section 4.5, there is currently no software application available that supports the complete workflow of a quantitative proteomics experiment and allows to integrate raw data and analysis results in the same place. Moreover, it is, of course, not only important to make statistical analysis methods available in a user-friendly and conceivable way but also to provide guidance regarding their application in the context of quantitative proteomics experiments (cf. chapter 6). In view of the broad range of methods that could be applied, there is a strong need to assess which approaches are pointing to success (cf. chapter 9).

A reasonable interpretation of analysis results in a biological context requires information about the functions of individual proteins and the relationships between different proteins in a cell. Whatever reason proteins are of interest, e. g. because they are found significantly differentially regulated by a statistical test, the information about a protein's function should be as detailed as possible in order to understand the results of an experiment such as the impact of a stress stimulus or the influence of changing environmental conditions. It seems, for example, logical that proteins which fulfill a similar function are also similarly regulated.

Hence, if a group of proteins reveals a similar pattern of abundance in an experiment, this might indicate that these proteins also play a comparable role in the metabolism of an organism. While the outcomes of database searches, typically, only provide the accession numbers of identified proteins, it is necessary to query and integrate this data from different resources such as UniProt (The UniProt Consortium 2008). This particularly applies to functional annotations of proteins, e. g. in form of affiliations to specific clusters of orthologous groups of genes (Tatusov et al. 2003) or in form of a structured and ontology-controlled description of gene products as found in the Gene Ontology (Ashburner et al. 2000). To take another example, the KEGG (Kanehisa and Goto 2000) database allows to gain a detailed picture of protein regulation in the context of known metabolic pathways such as the glycolysis—the conversion of glucose to pyruvate. At the Center for Biotechnology, the GenDB annotation system (Meyer et al. 2003) constitutes a valuable resource for gene annotation data allowing to map protein and gene information via the BRIDGE layer (Goesmann et al. 2003), which mediates data access across different applications.

Chapter 6

Methods for the statistical analysis of quantitative proteomics data

The scientific questions that are posed in the field of proteomics and which are to be answered with the help of stable-isotope labeling methods are without any doubt manifold. A closer investigation of the typical experimental setups reveals, however, that there are in particular two questions that are most frequently asked by experimenters: firstly, "which proteins are differentially regulated regarding the selected experimental conditions", and secondly, "are there groups of proteins that are characterized by similar abundance ratios, indicating a common regulation?" (Albaum et al. 2011b, p.1). The aim of this chapter is to introduce a set of methods that can be used to adequately answer these two questions, especially with regard to the particular nature of quantitative proteomics data.

6.1 Detection of differentially regulated proteins

Statistical measures such as a mean value or a standard deviation aim to provide an estimate of the true conditions in a population under investigation, and thereby help to draw reliable conclusions from observed protein abundance measurements, among others. Statistics can be used in a descriptive way to summarize—in a numerical or a graphical manner—the essential characteristics of a series of measurements. The derived parameters are supposed to explain the properties of the frequency distribution of all observations and, applied to samples taken for example under varying conditions, to expose similarities and differences

in the data. Statistics goes a step further when inferences about the overall population are deduced from the chosen sample. Techniques such as regression or hypothesis testing base on the best fit of a model and its parameters to the data, where the best model is meant to explain most of the differences and relations of all measurements—"the model that produces the least unexplained variation (the minimal residual deviance)" (Crawley 2007, p.4).

Looking at the different types of data that are involved in a quantitative proteomics experiment, one typically finds calculated ratio measurements on the one side, and—without loss of generality—categorical variables on the other. In this connection, a categorical variable is typically termed a factor, which is furthermore separated in one or more individual levels. This can for example be the factor "strain" with two levels "wildtype" and "mutant" or a time series experiment with different levels corresponding to various time points a sample has been taken at (strictly speaking, time is of course interval-scaled).

It is important to consider that a sample represents only a subset of the overall population, and moreover, that the repeated measurement e. g. of a protein's abundance may result in a slight but noticeable different value. Obviously, this kind of variance is unavoidable when living organisms are under examination, and the only way to cope with this biological variance is to take several (biological) replicate measurements. However, also the process of measuring data may itself be error-prone, thus, the measurement contains defects. In this case, one typically speaks of technical variance. So called technical replicates, that is the repeated measurement of the same biological replicate, can be employed to balance out this source of variation (Rocke 2004; Levin 2011).

A typical experiment includes several replicate measurements for each protein. Furthermore, a protein's abundance is usually not measured directly, but instead observations are made at the peptide level. It depends on the type of analysis whether it is advantageous to combine peptide measurements to a protein ratio or whether peptide measurements are interpreted—in broader terms—as a replicate measurement for a protein.

Certainly, the detection of differently regulated proteins can be based on averaged protein abundance ratios for each investigated condition, but it is more than beneficial to scrutinize the variances that occur between measurements. "Given a number of measured abundance ratios for a protein, a small variation between these values could mean that the strict enforcement of the protein's quantity is of key importance, e. g. for the development of an organism. Contrary, a rather high variation could indicate a weak influence of regulatory elements and lead to the assumption that the exact dosage e. g. of an enzyme regarding a metabolic pathway may not be important. If, for a protein, repeated measurements are obtained under different conditions, i. e. can be separated into two or more groups, it can be questioned whether variations are larger between two groups than within the same group" (Albaum et al. 2011b, p.2).

A statistical test aims to investigate the significance of deviations between samples, or more generally, whether a hypothesis, the null hypothesis H_0, concerning the population under investigation can be verified or falsified—in which case the alternative hypothesis H_1 would be valid. An essential aspect of this approach is the definition of significance, which determines

if a result is likely to occur just by chance under the provision that the defined hypothesis H_0 is true. Therefore, the highest acceptable significance level α has to be set in advance to the application of a statistical test. Typical values for this purpose are $\alpha \leq 0.05$ or 0.01.

6.1.1 Up- or down- regulation of an individual protein

In the simplest case, an experiment consists of a direct comparison between two samples, in which one is labeled and one unlabeled. Then, the focus of interest is on the similarity (or dissimilarity) between both of them or to be more precise between the abundances of each individual protein present in both samples. Since relative quantification results in one ratio value (cf. section 3.3.1), the H_0 hypothesis can simply be defined as the deviation of this logarithmically-scaled relative abundance value M from the theoretical mean value $\mu = 0$ —synonymous to 'no differential regulation'. Given that $\{x_l, l = 1, \ldots, L\}$ denotes a series of abundance measurements for a protein (M-values), a t-statistic, and thereby a measure for significance, can be calculated as follows:

$$t = \frac{|\,\overline{x} - \mu\,|}{s} \cdot \sqrt{L} \tag{6.1}$$

where

$$\overline{x} = \frac{1}{L}\sum_{l=1}^{L} x_l \quad \text{(arithmetic mean)}$$

$$s = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(x_l - \overline{x})^2} \quad \text{(standard deviation)}$$

Depending on the degrees of freedom ($df = L - 1$) and taking into account the corresponding t-distribution, a probability p can be derived stating whether the measured value is significantly differing from the theoretical mean $\mu = 0$ (cf. Koehler et al. 2002).

6.1.2 Comparison of multiple-condition proteome data

The biologist and statistician Fisher (1918) coined the term "variance" as the square of the standard deviation, and described analytical methods to measure the impact of various sources of variance on a dependent variable, hence, with regard to a quantitative proteomics experiments, on the calculated (relative) abundance values of a protein. Considering the case of having protein abundance measurements that can be separated in G groups (levels) regarding a specific type of treatment (factor), and furthermore, letting μ_g denote the mean vector of all abundance values associated to a specific level $g \in [1, \ldots, G]$, a hypothesis H_0 can be formulated as follows:

$$\mu_1 = \mu_2 = \ldots = \mu_G = \mu \tag{6.2}$$

On the opposite, the alternative hypothesis H_1 is given as the divergence of at least two parameters:

$$\mu_i \neq \mu_{i'} \text{ for } i, i' \in [1, \ldots, G], i \neq i' \tag{6.3}$$

Aim of the analysis of variance (ANOVA) is to test whether these group means significantly differ or, in case H_0 is valid, equal the expected population's mean value μ. Formally, a variance analysis could also be replaced by pairwise t-tests (Bortz 2005), which might, however, lead to an inflation of the type I error (cf. 6.1.3). Characteristic trait of the method—hence the name—is not to directly consider the mean vector of each level but instead to investigate in how far the overall variability of the data, in other words, the total variance $\mathbf{QS_{total}}$, can be explained by the variability caused by the factor, namely the treatment variance $\mathbf{QS_{treatment}}$. The remaining variance, termed $\mathbf{QS_{error}}$, must consequently originate from other sources, such as errors in measurements:

$$\mathbf{QS_{total}} = \mathbf{QS_{treatment}} + \mathbf{QS_{error}} \tag{6.4}$$

Using the estimations of the population's variance $\hat{\sigma}^2_{treatment}$ as the squared sum of all differences of each groups' mean value from the overall mean value, and $\hat{\sigma}^2_{error}$, given by the squared sum of all variances observed within each level, in analogy to the t-statistic, a so called F-statistic can be calculated and transformed into a probability value p depending on a specific F-distribution:

$$F = \frac{\hat{\sigma}^2_{treatment}}{\hat{\sigma}^2_{error}} \tag{6.5}$$

Whereas the decomposition of $\mathbf{QS_{total}}$ in $\mathbf{QS_{treatment}}$ and $\mathbf{QS_{error}}$ is not subject to any conditions, a valid and meaningful interpretation of the F-statistics demands the following three prerequisites (cf. Ellison et al. 2009; Crawley 2007):

i) Gaussian-distributed error components: within a group, deviations from the group's mean vector should follow a normal distribution. To investigate whether this precondition is fulfilled, a Shapiro-Wilks test (Shapiro and Wilk 1965) may be considered.

ii) Homogeneous error variances: the samples should be taken from equally distributed populations. Therefore, variances within different samples are not allowed to differ significantly. The Fligner-Killeen test (Fligner and Killeen 1976) provides a measure to examine this condition.

iii) Independent error components: Certainly, each measurement is subject to confounding variables. The influence of these error components has to be independent for each measurement. This should be the case if biological replicates are considered, it might be problematic in case of technical replicates, however.

"Infringements of these premises, in particular of ii), might result in the false assessment of proteins as significantly differentially regulated. Although the ANOVA has more power in terms of discovering significant differences, in cases of violated assumptions a non-parametric method such as the Kruskal–Wallis one-way analysis of variance has to be applied" (Albaum et al. 2011b, p.2).

6.1.3 Error in hypothesis testing

Considering one single statistical test, an error of as much as the *a priori* defined significance level α is allowed to falsely reject the null hypothesis. Albeit small for a single test, this error increases dramatically if multiple tests have been performed—and this is certainly always the case when hundreds of proteins are being investigated in a single experiment. Given m hypotheses H_0^1 to H_0^m of which m_0 hypotheses are true and accordingly $1 - m_0$ false, m significance tests can, consequently, be performed to verify or falsify each of these null hypotheses. In the outcome, a number of hypotheses, hereinafter defined S, may then be found valid with regard to a significance threshold, or, in terms of proteomics S proteins are presumably not significantly differentially regulated. It has to be assumed that TN number of hypotheses are correctly accepted as well as TP hypotheses correctly rejected. In addition to that, however, few hypotheses might be wrongly declared invalid (FP) or valid (FN), respectively, and these two types of errors must be taken into account in the calculation of p-values:

	H_0 accepted	H_0 rejected	total
H_0 true	TN	FP (type I error)	m_0
H_0 false	FN (type II error)	TP	$m - m_0$
	$m - S$	S	m

The 'family-wise error rate' (FWER) defines the probability that at least one type I error might occur:

$$FWER = Pr\left(FP > 0\right) \tag{6.6}$$

A well-known, albeit conservative, instrument to control the FWER is based on Bonferroni's inequality (Hochberg 1988). Let p_1 to p_m be the probabilities that correspond to the hypotheses H_0^1 to H_0^m, then a hypothesis H_0^i ($i \in [1, \ldots, m]$) has to be rejected, if the following condition holds: $p_i \leq \frac{\alpha}{m}$. Bonferroni's inequality "ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α" (Hochberg 1988, p.800):

$$Pr\left(\bigcup_{i=1}^{m} p_i \leq \frac{\alpha}{m}\right) \leq \alpha \quad (0 \leq \alpha \leq 1) \tag{6.7}$$

An alternative and less conservative method to control the FWER has been formulated by Holm (1979). Given that all p-values p_1 to p_m are increasingly sorted, so that $p_1 \leq p_2 \leq \ldots \leq p_m$, a hypothesis H_i is rejected, if the following is valid for $j = 1, \ldots, i$:

$$p_j \leq \frac{\alpha}{m - j + 1} \tag{6.8}$$

Another instrument to control the error in hypothesis testing is given by the 'false discovery rate' (FDR), which is defined as the expected number of false positives (FP) regarding the overall number of rejected hypotheses (S):

$$FDR = E\left(Q\right) \tag{6.9}$$

where

$$Q = \begin{cases} \frac{FP}{S}, if\ S > 0 \\ 0\ otherwise \end{cases} \tag{6.10}$$

While, undoubtedly, the exact determination of the expectation value E of the variable Q is rather impossible, Benjamini and Hochberg (1995) postulated a limitation of the FDR to a level q, $q\frac{m_0}{m} \leq q$, which has to be defined *a priori* in analogy to the significance level α. Given m hypotheses H_0^1 to H_0^m, with corresponding p-values p_1 to p_m, increasingly sorted ($p_1 \leq p_2 \leq ... \leq p_m$), the procedure of Benjamini and Hochberg defines an index i_{max} as follows ($i \in [1, \ldots, m]$):

$$i_{max} = \max\{i : p_i \leq \frac{i}{m} \cdot q\} \tag{6.11}$$

All hypotheses H_0^j having an index in the range of $j = 1, \ldots, i_{max}$ have to be rejected.

6.2 Identification of co-regulated proteins

"One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification[1]" (Everitt et al. 2001, p.2). As soon as a factor with two or more levels is analyzed in an experiment, the question commonly arises whether a group of proteins shows similar abundance ratios related to these factor levels and thus might have a similar protein turnover. It seems reasonable to suppose that these proteins fulfill a similar function or play a comparable role in the metabolism of a cell or organism. Aiming to aggregate a number of proteins, each characterized by a series of measurements, i. e. relative (or even absolute) abundance values, in groups or clusters, at first, a solution needs to be found that determines a measure of similarity (or dissimilarity) between each two proteins. Based on the similarity values computed for all proteins, the aggregation procedure can then be performed in such a way that all proteins in a cluster are as homogeneous as possible, whereas between all members of each two clusters there is a considerable heterogeneity (Bacher 1996). A trivial solution to this optimization problem would be the successive assorting, evaluation, and re-sorting of proteins to clusters until an optimal grouping fulfilling these criteria has been found. This approach, however, is being bought with an enormous computing effort, since N elements result in $B_N = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}$ possible combination. Given for example only $N = 50$ elements, $B_N = 23.9 \cdot 10^{21}$ different clusterings would need to be taken into consideration.

Cluster analysis methods provide a heuristic approximation of the optimal assorting of a dataset. They belong to the group of unsupervised learning methods, which are characterized by being independent from any external information. The calculation is solely performed on inherent features of the data—clusters are not known *a priori* but discovered during the clustering process. Clustering techniques are traditionally divided into three distinct classes: firstly, hierarchical; secondly, partitioning or vector quantization; and thirdly, probabilistic or density-based methods (Cormack 1971; Everitt et al. 2001). Applied to proteomics data,

[1] Everitt here uses the term classification in the sense of clustering.

hierarchical approaches group proteins into clusters, which are then, iteratively, grouped into larger clusters. Thereby, a hierarchical tree structure is formed. Partitioning approaches, in contrast, follow a given optimization strategy to assign each protein to one of an *a priori* specified number of groups. Density-based approaches differ from the other two strategies in the way that each protein is not necessarily belonging to a single cluster but instead is assigned a probability that specifies its membership to a group.

6.2.1 Measures of similarity between two proteins

As mentioned above, a first prerequisite for the application of a cluster analysis on proteomics data is the specification of a similarity or distance measure between each two proteins. In principle, a plethora of similarity measures is available, their applicability, however, depends on the scale of the data as well as the relation between two objects that an experimenter is interested in. In the context of proteome experiments, one may, for example, think of an experiment, in which the actual difference in the abundances of two proteins over time is negligible but instead a positive or negative correlation between the two protein's series of measurement is of considerable importance. In such a case, correlation-based distances can be taken into consideration.

In general, measures of similarity or distance fulfill the properties of a metric. Given three proteins $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, a measure of distance $d : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ therefore has to satisfy the following conditions:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &\geq 0 \\ d(\mathbf{x}, \mathbf{y}) = 0 &\Leftrightarrow \mathbf{x} = \mathbf{y} \\ d(\mathbf{x}, \mathbf{y}) &= d(\mathbf{y}, \mathbf{x}) \\ d(\mathbf{x}, \mathbf{z}) &\leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \end{aligned}$$

A well-known and commonly used metric is the Minkowski distance, which is well-suitable for the typically interval-scaled protein abundance values (cf. Bortz 2005). It is defined as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt[r]{\sum_{i=1}^{n} \mid x_i - y_i \mid^r} \tag{6.12}$$

The Minkowski distance can be considered as a generalized form of the Manhattan metric—here the value of r is equal to 1. For $r = 2$ the distance is better known as the Euclidean distance, which has the particular characteristic of representing the physical distance between two points in space.

As already mentioned above, in some cases correlation coefficients can be a favorable measure of similarity such as Pearson's (centered) correlation coefficient (cf. Hastie et al. 2001):

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}}) (y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{\mathbf{y}})^2}} \tag{6.13}$$

with

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(arithmetic mean of vector } \mathbf{x}\text{)} \tag{6.14}$$

Assuming a linear interrelation between the series of measurements of two proteins the correlation coefficient measures their degree of correlation. Resulting values ($[-1 \ldots 1]$) can then be transformed in a distance value d:

$$d(\mathbf{x}, \mathbf{y}) = 1 - cor(\mathbf{x}, \mathbf{y})^2 \tag{6.15}$$

With a slight modification—the subtraction of each protein's mean abundance value is omitted—Pearson's uncentered correlation coefficient provides another possibility to measure similarities between two proteins:

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{6.16}$$

6.2.2 Formal definition of cluster analysis

Based on pairwise-computed similarity measures for a set of proteins, cluster analysis can formally be described as the partitioning of these proteins in K clusters $\{C_k, k = 1, \ldots, K\}$. Given N proteins with abundance values, which are described by a matrix $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\}$ with x_{ij} denoting the j-th measurement of the i-th protein, the clustering can be defined by a matrix:

$$\mathbf{W}(\mathbf{X}) = [w_{ki}]_{K \times N} \tag{6.17}$$

In hierarchical and partitioning cluster analysis, the association of a protein to a cluster is unique. It applies, therefore, that $w_{ki} \in \{0, 1\}$ according to:

$$w_{ki} = \begin{cases} 1, \text{if protein } \mathbf{x}_i \in \text{cluster } C_k \\ 0, \text{ otherwise} \end{cases} \tag{6.18}$$

In addition, the following restriction has to be imposed to ensure this uniqueness:

$$\sum_{k=1}^{K} w_{ki} = 1 \quad \text{for } i = 1, \ldots, N \tag{6.19}$$

From these two conditions, it can, consequently, be deduced an equation expressing the number of proteins that belong to each cluster C_k:

$$|C_k| = \sum_{i=1}^{N} w_{ki} \quad , k \in \{1, \ldots, K\} \tag{6.20}$$

For probabilistic approaches, a protein may, in principle, belong to more than one cluster with a certain probability such that $w_{ki} \in [0 \ldots 1]$. In order to enable the comparison and evaluation of cluster results, it is, however, necessary to assign each protein i to one specific cluster, i. e. typically $\max_k w_{ki}$.

6.2.3 Hierarchical cluster analysis

In hierarchical cluster analysis groups of objects are successively merged together to form greater clusters. An important aspect of this process is that a grouping is regarded permanent once it has been made. It cannot be reversed (Bacher 1996). Strictly speaking, two contrary approaches have to be distinguished since apart from the mostly used agglomerative cluster analysis algorithms there exists also the group of divisive algorithms. While the first-mentioned approach initially begins with each object forming a singleton cluster, all of which are then successively merged, the last mentioned approach starts with one cluster containing all objects, which is then successively divided into two parts (Gordon 1987).

All (agglomerative) hierarchical cluster analysis methods share a common algorithm. Applied to the context of proteomics and given a set $\{\mathbf{x}_i, i = 1, \ldots, N\}$ of proteins as well as d_{C_p,C_q} as an arbitrary measure of similarity or distance between two clusters C_p and C_q (initially each object forms a singleton cluster), for $p, q \in [1, \ldots, N]$ and $d \in \mathbb{R}$ the hierarchical clustering procedure can be outlined as follows:

(1) for $i := 1$ to N do
(2) define cluster $C_i := \mathbf{x}_i$ od
(3) define $k := N$
(4) while $k \neq 1$ do
(5) find $\{p, q\} := \min_{p,q}\{ d_{C_p,C_q} \mid p, q \in [1, \ldots, N] \}$
(6) define $C_{p,q} :=$ merge(C_p, C_q)
(7) $k := k - 1$ od

The formation of clusters generates a monotonic hierarchical structure between the individual proteins. Dependent on the engaged clustering method, the structure fulfills the properties of an ultrametric space, where in addition to the triangle inequality the so called strong triangle or ultrametric inequality applies (Milligan 1979):

$$d(\mathbf{x}, \mathbf{y}) \leq \max\{d(\mathbf{x}, \mathbf{z}), d(\mathbf{z}, \mathbf{y})\} \tag{6.21}$$

This is valid for Single- and Complete-linkage (see below), in which distances between successively merged clusters are monotonically increasing, but may not be satisfied for example in case of the Centroid approach. This has to be considered, therefore, if results of a cluster analysis are displayed in form of a tree as a so called dendrogram.

Whereas a distance measure, as mentioned above, is utilized to determine the similarity between two individual proteins during the process of clustering, an additional measure is required that defines the distance between two clusters consisting of more than one object. Therefore, a number of different approaches has been proposed, each having its advantages and disadvantages concerning their utilization in the frame of proteomics data.

In the following a selected choice of cluster algorithms is presented. At this, $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\}$ will denote a number of proteins, and C_p and C_q, $p, q \in [1, \ldots, K]$, two arbitrary clusters. Furthermore, for each two elements $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ a distance function $d : \mathbf{X} \times \mathbf{X} \mapsto \mathbb{R}$ shall be predefined, e. g. based on a coefficient of correlation or a Minkowski metric.

6.2.3.1 Single- and Complete-linkage

The general idea behind Single- as well as Complete-linkage is derived from the application of computers to taxonomy, e. g. to cluster different strains according to similar and dissimilar features (Sneath and Sokal 1973). Both approaches, basically, represent opposite models in the sense that either the two objects of two clusters to be fused which are nearest to each other (Single-linkage) or which are furthest apart from each other (Complete-linkage), are chosen for the determination of an inter-cluster distance (Bacher 1996). "Transfered to the context of proteomics this can be seen as the conflict between the two ideas to, on the one hand, combine as many proteins as possible if they reveal only a slight similarity and to form compact clusters that contain only those proteins that are utmost similar, on the other hand" (Albaum et al. 2011b, p.8).

Formally, the inter-cluster distance in Complete-linkage is computed as follows:

$$d_{ic}(C_p, C_q) = \max_{\mathbf{x},\mathbf{y}} \{ d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_p \wedge \mathbf{y} \in C_q \} \tag{6.22}$$

Using this approach a high degree of homogeneity within each cluster is achieved. In contrast to that, Single-linkage defines a measure of distance between two clusters with the equation:

$$d_{ic}(C_p, C_q) = \min_{\mathbf{x},\mathbf{y}} \{ d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_p \wedge \mathbf{y} \in C_q \} \tag{6.23}$$

The approach, however, has one decisive disadvantage, thus it is conceivable, that two, obviously inhomogeneous, clusters may be merged solely due to the spacial neighborhood of two of their representatives, which then "leads to the notorious chaining effect [...]" (Kaufman and Rousseeuw 1990, p.226).

6.2.3.2 Average-linkage

Hierarchical cluster analysis algorithms that rely on averaging can be understood as a variation of Complete- and Single-linkage. Figuratively speaking, they constitute an intermediate between both approaches, in which the distance between two clusters is calculated as the average distance between all pairs of objects. Most frequently, "Weighted-Average"-linkage (Sokal and Michener 1958; McQuitty 1966) is being applied, which employs the following formula:

$$d_{ic}(C_p, C_q) = \frac{1}{|C_p|\,|C_q|} \sum_{\mathbf{x} \in C_p, \mathbf{y} \in C_q} d(\mathbf{x}, \mathbf{y}) \tag{6.24}$$

While Kaufman and Rousseeuw (1990) refer to this approach as "Grouped Average", another commonly used term is "Unweighted Pair Group Average Method" (UPGMA, cf. Everitt et al. 2001). In contrast to other averaging methods, it is ensured that during the fusion of clusters, inter-cluster distances are monotonically increasing. This is, for example, not the case if "Within-Average"-linkage is being utilized (Bacher 1996).

6.2.3.3 Centroid-linkage

Closely related to the previous method, inter-cluster distances in Centroid-linkage are defined as the (squared) Euclidean distances between each cluster's so called centroid, which basically represents the mean element—also termed prototype—of all elements of a cluster:

$$d_{ic}(C_p, C_q) = \| \overline{c_p} - \overline{c_q} \|_2 \tag{6.25}$$

where

$$\overline{c}_k = \frac{1}{|C_k|} \sum_{i=1}^{|C_k|} \mathbf{x}_i \mid \mathbf{x}_i \in C_k \text{ (cluster centroid)} \tag{6.26}$$

Whether the method can be applied depends on the data's type of scale. It is necessary that all measurements are interval-scaled as otherwise a (meaningful) mean value computation would not be possible. This is certainly the case for quantitative proteomics data. It also has to be noted that Centroid-linkage can only be employed in combination with Euclidean distances (cf. Bacher 1996).

6.2.3.4 Ward-linkage

Ward (1963) proposed another approach, which can only be applied on interval-scaled data, similar to Centroid-linkage. Distances between two clusters are calculated using the Euclidean metric, but furthermore the algorithm penalizes an increase in the error sum of squares (ESS):

$$d_{ic}(C_p, C_q) = \frac{2\,|C_p|\,|C_q|}{|C_p| + |C_q|} \left(\overline{c_p} - \overline{c_q}\right)^2 \tag{6.27}$$

6.2.4 Partitioning cluster analysis

Partitioning cluster algorithms iteratively re-sort objects into a specified number of groups, thereby attempting to minimize the error—given by a certain numerical criterion—within each group while differences between individual clusters shall be as large as possible. K-means is a typical and popular representative of this type of cluster algorithms. It is important to note that these methods all demand a preceding estimate of the 'correct' number of clusters the data shall be partitioned in.

6.2.4.1 K-means

Strictly speaking, the term K-means refers to several closely related algorithms. Common element of all of these approaches is the iterative construction of so called cluster centers as kind of anchor points for the formation of a partitioning of a set of objects. The basic idea is to minimize the error sum of squares in all clusters in a way that at the same time the differences between all clusters are maximal—an approach similar to Ward-linkage (cf.

6.2.3.4) in hierarchical cluster analysis. While such an optimization is hardly feasible for large datasets, various algorithmic approaches have been proposed striving to find at least a suboptimal solution, e. g. by Forgy (1965) and Lloyd (1982). He already introduced the method in 1957 at a conference. Following the procedure of MacQueen (1965) initially K random objects are selected as cluster centers. Each object is then assigned to one of these cluster centers on condition that the (Euclidean) distance between both is minimal. Once the assignment has been completed, for each of the generated K clusters the new center is computed. The process is iteratively repeated until it converges in a way that, after an iteration, no object would be assigned to any other cluster center.

Applied to proteomics data, the K-means algorithm can formally be described as follows: Given $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\} \in \mathbb{R}^m$ as a set of N proteins, $d(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \mathbb{R}$ define the (Euclidean) distance between two proteins $\mathbf{x}_i$ and $\mathbf{x}_j$, $ij \in [1, \ldots, N]$, and, furthermore, let $K \in \mathbb{N}$ indicate the number of clusters, the following procedure aims to find a partitioning $C_1, \ldots, C_K$:

(1) //K random proteins $\mathbf{c_1}$ to $\mathbf{c_K}$ are selected as initial cluster centers:
(2) for i := 1 to K do
(3) $\quad \mathbf{c_i}$:= randomly chosen vector $\in \mathbb{R}^m$ od
(4) //the procedure is repeated as long as at least one protein
(5) //can be re-assigned to any other cluster center:
(6) while changes in any cluster C_i occure do
(7) $\quad$ //(re-)assign each protein to a cluster on condition that the distance is minimal:
(8) $\quad$ foreach k := 1 to Kdo
(9) $\quad\quad C_k = \{\ \mathbf{x}_i \mid d(\mathbf{c_k}, \mathbf{x}_i) \leq d(\mathbf{c_h}, \mathbf{x}_i),\ \forall \mathbf{x_i} \in \mathbf{X} \wedge h = 1, ..., K, k \neq h\ \}$
(10) $\quad$ od
(11) $\quad$ //re-calculation of cluster centers:
(12) $\quad$ foreach k := 1 to Kdo
(13) $\quad\quad \mathbf{c_k} := \frac{1}{|C_k|} \sum_{i=1}^{|C_k|} \mathbf{x}_i \in C_k$
(14) $\quad$ od
(15) od

While the algorithm is comparatively efficient—its complexity depends on the number of iterations I and can be specified with $O(KNI)$—the method has one clear disadvantage: MacQueen (1965, S.282) had already noted, that "in general, the k-means procedure will not converge to an optimal partition, although there are special cases where it will". For that reason, Hartigan and Wong (1979) proposed another K-means approach that "goes further, and ensures that there is no single switch of an observation from one group to another group that will decrease the objective" (Hastie et al. 2001, p.462). However, it may, consequently, be advisable to repeat the algorithmic procedure using different sets of initial cluster centers. The solution offering the minimal within as well as maximal between cluster error sum of square should then be chosen as a solution to the clustering problem.

6.2.4.2 Neural-Gas

The cluster algorithm Neural-Gas (Martinetz et al. 1993) was inspired by the K-means approach—the authors claim it an extension—but also conceals elements of Kohonen's self-organizing map (Kohonen 1990), whose idea goes back to the beginning of the sixties. Hubel and Wiesel (1962) found neurons in a cat's visual cortex responding to complex patterns of light. They were able to show that (visual) stimuli occurring in nearby locations are also processed in nearby areas of the cortex. Self-organizing maps attempt to model these characteristics of the nervous systems, thereby aiming to "very closely resemble the topographically organized maps found in the cortices of the more developed animal brains" (Kohonen 1990, p.1464). In this sense, the Neural-Gas network consists of a number K of predefined neurons—synonymous with the cluster centers in K-means—represented by weight vectors. In an iterative process, termed the learning procedure, all objects—the input data—are, figuratively speaking, projected onto this network, where in each iteration the neurons of the network adapt to the presented data.

In the following, $\mathbf{x} = (x_1, x_2, \ldots, x_P)^T$ denotes a series of measurements for one protein. Given the *a priori* specified number of clusters that shall be found is K, for each $j = 1, \ldots, K$ a cluster center or weight vector $\mathbf{w}_j = (w_{j1}, w_{j2}, \ldots, w_{jP})^T$ has to be determined, e. g. by sampling from the input space.

During the I-step learning procedure ($I \in \mathbb{N}$) each protein is iteratively selected and subject to the following procedure: Given $\mathbf{x}$, all cluster centers or weight vectors are increasingly sorted according to their Euclidean distance to this vector of protein abundance values:

$$d_{\mathbf{x},\mathbf{w}_j} = \| \mathbf{x} - \mathbf{w}_j \|_2 \quad \text{for } j \in [1, \ldots, K] \tag{6.28}$$

so that the following condition holds:

$$d_{\mathbf{x},\mathbf{w}_1} < d_{\mathbf{x},\mathbf{w}_2} < \ldots < d_{\mathbf{x},\mathbf{w}_K} \tag{6.29}$$

Let $\mathbf{w}_j$ denote the weight vector at position j in this sequence, dependent on the input and the current iteration $i \in [1, \ldots, I]$ of the learning procedure, a kind of "neighborhood ranking"

$$h_{\mathbf{w}_j,\mathbf{x}}(i) = \exp\left(\frac{\mathbf{w}_j}{\sigma(i)}\right) \tag{6.30}$$

can be applied on all weight vectors with the result that in the next iteration $i + 1$ each vector is defined as follows:

$$\mathbf{w}_j(i+1) = \mathbf{w}_j(i) + \eta(i) h_{\mathbf{w}_j,\mathbf{x}}\left(\mathbf{x} - \mathbf{w}_j(i)\right) \tag{6.31}$$

where η and σ denote, typically, exponentially or linearly decaying functions that determine either the range of adaption regarding neighboring weight vectors or specify a learning-rate. As the cluster analysis aims to find a partitioning of the data, after the learning procedure, each object is assigned to its nearest cluster, which is represented by its weight vector or cluster center.

6.2.5 Cluster validation

Cluster analysis has the potential to reveal hidden structures in the data, which—in the sense of quantitative proteomics—might be groups of proteins having a similar turnover. In contrast to supervised learning methods, where success can, in general, directly be measured based on the error between a prediction and a suspected outcome, it "is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly." (Hastie et al. 2001, p.439). In the run-up to the analysis, in general, no information regarding a true partitioning is available. Moreover, the results produced by different algorithms are not rarely dissimilar: the hierarchical structures for example obtained by Single- and Complete-linkage are seldom characterized by a strong congruence (see section 9.3 for a detailed evaluation). A fundamental part of the clustering process is therefore an evaluation of the algorithms' results (Halkidi et al. 2002).

Cluster algorithms such as K-means partition a set of objects in a specified number of groups, which is therefore required as input parameter. However, before execution, a determination of this number is, strictly speaking, not possible. An elegant solution to this problem is to iteratively compute clusterings of the data in different sizes. The quality of each resultant cluster structure may then be evaluated using some kind of numerical criterion—in case of K-means this might for example be the partitioning offering the minimal within as well as maximal between cluster error sum of square. In addition to this approach, a variety of techniques have been developed for this purpose. Milligan and Cooper (1985) compared the performance of more than 30 of these so called cluster indexes on simulated datasets, and found the procedure of Calinski and Harabasz (1974) giving the best results. In a more recent study, another approach, the index I, has been suggested and recommended as "more consistent and reliable in indicating the correct number of clusters" (Maulik and Bandyopadhyay 2002, p.1654).

In the context of hierarchical cluster analysis the investigator may also be interested in a certain partitioning of the input data. This can directly be achieved if the algorithm (cf. 6.2.3) is stopped as soon as the desired number of clusters has been merged—figuratively speaking, one could also imagine that the resulting tree structure is cut at the desired level. An evaluation of different cluster solutions—to find the 'correct' number of clusters—might for example be conducted by plotting the increase (or decrease, respectively) in distance gained from the merge of each two clusters against the current iteration of the algorithm. Of course, cluster indexes might also be considered.

In general, it is recommended to apply several different algorithms on the input data and to compare their outcomes to each other. An evaluation includes both a comparison of the different algorithms regarding the significance of their results as well as the determination of the 'correct' number of groups.

In the following, a number of cluster indexes are described. This includes the aforementioned measures from Calinski and Harabasz, Maulik and Bandyopadhyay, a commonly used index from Davies and Bouldin (1979), as well as another 'classical' approach introduced by Krzanowski and Lai (1988). In addition to these, an index, called Figure of Merit (FOM), is listed, that has been delineated by Yeung et al. (2001) particularly for the analysis of gene expression data. The special feature about this method is the idea to integrate a kind of bootstrapping approach (cf. Hastie et al. 2001) and, thereby, to estimate the predictive power of a cluster algorithm.

Terms and definitions: $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\}$ denotes a set of objects, as well as $K \in \mathbb{N}$ a number of clusters which the set is partitioned in and which is described by the matrix $\mathbf{W}(\mathbf{X}) = [w_{ki}]_{K \times N}$. As already mentioned before,

$$|C_k| = \sum_{i=1}^{N} w_{ki} \tag{6.32}$$

equals the number of objects assigned to a cluster C_k. While $\overline{C_k}$ defines the mean vector (or cluster center) of the k-th cluster (cf. 6.2.3.3),

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \tag{6.33}$$

is the overall mean vector of all objects.

6.2.5.1 Calinski-Harabasz

Calinski and Harabasz (1974)'s cluster index is calculated using the following equation:

$$CH(K) = \frac{[trace\ \mathbf{B}_K \ / \ K-1]}{[trace\ \mathbf{W}_K \ / \ N-K]} \quad \text{for} \quad K \in \mathbb{N} \tag{6.34}$$

where **B** denotes the error sum of squares between different clusters (inter-cluster)

$$trace\ \mathbf{B}_K = \sum_{k=1}^{K} \frac{1}{|C_k|} \parallel \overline{C}_k - \overline{\mathbf{x}} \parallel_2 \tag{6.35}$$

and **W** the squared differences of all objects in a cluster from their respective cluster center (intra-cluster)

$$trace\ \mathbf{W}_K = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{ki} \parallel \mathbf{x}_i - \overline{C_k} \parallel_2 \tag{6.36}$$

Calculated for each possible cluster solution the maximal achieved index value indicates the best clustering of the data. It is an important characteristic of the index that $trace\ \mathbf{W}_K$ will start at a comparably large value while $trace\ \mathbf{B}_K$ should behave in the opposite direction: With an increasing number of clusters K, approaching the optimal clustering solution of K^* groups, the value of $trace\ \mathbf{W}_K$ should significantly decrease due to an increasing compactness

of each cluster. As soon as the optimal solution is exceeded an increase in compactness and thereby a decrease in value might still occur; this decrease, however, should be notably smaller. On the other hand, $trace\ \mathbf{B}_K$ is expected to get higher as the number of clusters K increases but will reveal a kind of softening in its rise if K gets larger than K^*.

6.2.5.2 Index-*I*

Maulik and Bandyopadhyay (2002) proposed a cluster index that is, in principal, composed of three individual elements:

$$I(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^p \quad \text{for} \quad p, K \in \mathbb{N} \tag{6.37}$$

While the first factor simply normalizes each index value by the overall number of clusters K, the second term sets the overall error sum of squares of the complete datasets in relation to the intra-cluster error of a given clustering:

$$E_K = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{ki} \parallel \mathbf{x}_i - \bar{\mathbf{x}}_k \parallel \quad \text{for} \quad K \in \mathbb{N} \tag{6.38}$$

A third factor takes into account the maximally observed difference between two of the K clusters:

$$D_K = \max_{p,q=1,\ldots,K \wedge p \neq q} \parallel \bar{\mathbf{x}}_p - \bar{\mathbf{x}}_q \parallel \quad \text{for} \quad K \in \mathbb{N} \tag{6.39}$$

The index computation includes a variable parameter $p \in \mathbb{N}$ that is "used to control the contrast between the different cluster configurations" (Maulik and Bandyopadhyay 2002, p.1651). The authors recommend a value of $p = 2$.

6.2.5.3 Davies-Bouldin

Instead of simply proposing a cluster index, Davies and Bouldin (1979) formulated a general framework for the evaluation of the outcomes of cluster algorithms. In analogy to Halkidi et al. (2002) an instance of their index $DB(K)$ may be defined as follows:

$$DB(K) = \frac{1}{K} \sum_{k=1}^{K} R_k \quad \text{for} \quad K \in \mathbb{N} \tag{6.40}$$

where

$$R_k = \max_{j=1,\ldots,K, j \neq k} \left(\frac{S_k + S_j}{d_{kj}}\right) \quad \text{for} \quad k \in [1, \ldots, K] \tag{6.41}$$

and

$$S_k = \frac{1}{\sum_{i=1}^{N} w_{ki}} \sum_{i=1}^{N} w_{ki} \parallel \mathbf{x}_i - \bar{\mathbf{x}}_k \parallel \quad \text{for} \quad k \in [1, \ldots, K] \tag{6.42}$$

as well as

$$d_{kj} = \| \bar{\mathbf{x}}_k - \bar{\mathbf{x}}_j \| \tag{6.43}$$

For each cluster C_k an utmost similar cluster—regarding their intra-cluster error sum of square—is searched, leading to R_k. The index then defines the average over these values. In contrast to the aforementioned cluster indexes, here, the minimal observed index indicates the best cluster solution.

6.2.5.4 Krzanowski-Lai

Krzanowski and Lai (1988) developed a cluster index that, similar to the index of Calinski and Harabasz (1974), is based on the squared differences of all objects in a cluster from their respective cluster center—$trace\ \mathbf{W}$. The authors define $\mathrm{DIFF}(K)$ as the difference between a clustering of the data in K and a clustering in $K-1$ clusters. Let J be the number of variables that has been measured on each $\mathbf{x}_i \in \mathbf{X}$ and $trace\ \mathbf{W}_K$ the sum of squares function that corresponds to the clustering in K clusters, their measure $\mathrm{DIFF}(K)$ is than defined as follows:

$$\mathrm{DIFF}(K) = (K-1)^{\frac{2}{J}} \cdot trace\ \mathbf{W}_{K-1} - K^{\frac{2}{J}} \cdot trace\ \mathbf{W}_K \tag{6.44}$$

In this formula, a normalizing factor $\frac{2}{J}$ is included, which is derived from the observation that, given independently uniformly distributed measurements on each variable $j \in [1, \ldots, J]$, the optimal clustering of the data will reduce the sum of squares exactly by this factor (Krzanowski and Lai 1988, p.25).

The authors claim that if there exists an optimal clustering solution in K^* groups, the value of $\mathrm{DIFF}(K^*)$ should be comparably large and positive (see index of Calinski and Harabasz for further explanation). In contrast, all values of $\mathrm{DIFF}(K)$ for $K > K^*$ will have rather small values (maybe even negative ones), whereas values for $K < K^*$ will be rather large and positive. Bringing these observations together the index $\mathrm{KL}(K)$ is defined as follows:

$$\mathrm{KL}(K) = \left| \frac{\mathrm{DIFF}(K)}{\mathrm{DIFF}(K+1)} \right| \tag{6.45}$$

The optimal cluster solution is then indicated by the highest value of $\mathrm{KL}(K)$.

6.2.5.5 Figure of Merit

Coming from a gene expression background, the Figure of Merit (Yeung et al. 2001) is based on the assumption that the validity of a cluster is certainly increasing in value if in a second experiment the same genes would group together and reveal a similar pattern of expression. Following a bootstrapping or jackknife approach, this situation may be simulated by successively applying a cluster algorithm on a set of proteins whereby in each iteration one experimental condition—in terms of a feature of each object/ a column of the data matrix—is left out. If a cluster algorithm would have assigned an object to a cluster just by chance, it

seems logical that the emission of a condition will produce different results. Otherwise, it is likely that two cluster results reveal a similar structure if the dependence on the left-out feature is small.

Let in the following $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\}$ denote a set of N objects, each having the dimension $P \in \mathbb{N}$, so that x_{ij} is the j-th feature of $\mathbf{x}_i$, $j \in 1, \ldots, P$; furthermore, let there be a number of clusters $K \in \mathbb{N}$ whereby $\mathbf{W}(\mathbf{X}) = [w_{ki}]_{K \times N}$ describes the clustering of the data. Assuming that a clustering has been performed with a data matrix where the j-th feature has been omitted, the Figure of Merit is defined as follows:

$$\mathrm{FOM}(j, K) = \sqrt{\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} w_{ki} \left(x_{ij} - \overline{C_{kj}} \right)^2} \tag{6.46}$$

where

$$\overline{C_{kj}} = \frac{1}{N} \sum_{i=1}^{N} w_{ki} x_{ij} \tag{6.47}$$

denotes the arithmetic mean in feature j of all objects of cluster k.

To avoid a bias towards the overall number of clusters, the so called "adjusted Figure of Merit" takes this amount K into account:

$$\text{adjusted } \mathrm{FOM}(j, K) \cdot \frac{1}{\sqrt{\frac{N-K}{N}}} \tag{6.48}$$

If the calculation is iterated over all P features of the objects, the "aggregate Figure of Merit" can be computed:

$$\text{aggregate } \mathrm{FOM}(K) = \sum_{j=1}^{P} \mathrm{FOM}(j, K) \tag{6.49}$$

The authors state that in the outcome "A small figure of merit indicates a clustering algorithm having high predictive power. We compare clustering algorithms with the same number of clusters, and over a range of number of clusters" (Yeung et al. 2001, p.310).

6.2.6 A measure to determine the congruence between clustering results

The Rand measure (Hubert and Arabie 1985) gives an indication of the congruence between two clusterings, for example, representing the outcomes of two cluster algorithms. Likewise, the index may also be used to compare a cluster solution to the true partitioning of a dataset, if this is known. Assuming that the application of a cluster algorithm on a dataset $\mathbf{X}$ with N proteins, produced a list of clusters $\mathbf{C} = \{C_1 \ldots C_K\}$ and, additionally, a second clustering of the same dataset $\mathbf{X}$ is given by the clusters $\mathbf{P} = \{P_1 \ldots P_J\}$, for each pair of proteins from the dataset $\mathbf{X}$, namely $(\mathbf{x}_v, \mathbf{x}_u)$, one of the following conditions must be met:

- **SS**: in both clusterings $\mathbf{C}$ and $\mathbf{P}$, the two proteins both belong to the same cluster

- **SD**: in **C** both proteins are in the same cluster, while they are separated in **P**
- **DS**: similar to **SD**, except that in **C** both belong to different clusters, while they group together in **P**
- **DD**: in both clusterings both proteins are in separate clusters

The Rand measure is derived from these conditions: For all possible pairwise combinations of proteins from the dataset **X** it is counted which of the three conditions holds, the measure is then given by

$$R = \frac{|\mathbf{SS}| + |\mathbf{DD}|}{|\mathbf{SS}| + |\mathbf{SD}| + |\mathbf{DS}| + |\mathbf{DD}|} = \frac{|\mathbf{SS}| + |\mathbf{DD}|}{N\frac{N-1}{2}} \tag{6.50}$$

The domain of definition of the resulting value is restricted between $[0, \ldots, 1]$ and it is valid that the higher the value the greater the similarity.

A derivative of the measure takes into consideration that two different proteins may be grouped into a cluster not as a result from any similarities between these two proteins but rather at random. The adjusted Rand index (cf. Hubert and Arabie 1985) therefore corrects the original Rand index by the expectation value $E(R)$ that the clustering occurred just by chance:

$$R_{adjusted} = \frac{R - E(R)}{1 - E(R)} \tag{6.51}$$

In analogy to the aforementioned unadjusted Rand index the domain of definition is restricted between $[0, \ldots, 1]$, while a high value indicates a high similarity.

6.3 Data analysis: more questions than answers

In this chapter, a set of methods was presented, which allow to answer two of the most frequently asked questions arising in quantitative proteomics experiments: firstly, "which proteins are differentially regulated regarding the selected experimental conditions", and secondly, "are there groups of proteins that are characterized by similar abundance ratios, indicating a common regulation?" (Albaum et al. 2011b, p.1). The analysis of variance allows to detect differentially regulated proteins regarding a number of experimental conditions. This method, however, demands certain prerequisites, whose fulfillment with regard to quantitative proteomics data cannot be taken for granted. Similarly, the identification of groups of proteins, which show similar abundance ratios, can be based on a variety of cluster analysis methods, which produce—in the worst case scenario—completely different results.

The aim of this work is to provide a software platform that allows to apply the introduced set of methods on the data of a quantitative proteomics experiment. This requires both the persistent storage and the appropriate representation of analysis results, e. g. in form of plots and tables. It also demands the provision of a framework to perform computationally intensive tasks. However, the aim of this work is not only to allow the application of these methods but also to investigate their applicability on this particular type of data. The final

objective is to provide straight answers to the central two questions. For this purpose, an evaluation study has been carried out taking into account three real-world datasets (Haußmann et al. 2009; Hahne et al. 2010; Otto et al. 2010). The results have been published in Proteome Science in 2011. Based on these results, a workflow for the comprehensive analysis of quantitative proteomics data is presented in chapter 9.

Chapter 7

Implementation of the QuPE system

This chapter describes in detail aspects of the implementation of the rich internet application QuPE as well as related methods and algorithms. The application consists of several modules that target independent services, ranging from the presentation of data in a web browser-based graphical user interface to the execution of computationally intensive tasks on a compute cluster. On the whole, the modules complement each other to form a versatile and extensible system for the storage and analysis of quantitative proteomics data and for the development and in-depth evaluation of data processing and analysis methods. Apart from the overall system design, this chapter addresses, in particular, the implementation of algorithms for the calculation of relative abundance values of metabolically stable isotope labeled protein samples.

7.1 System design

QuPE is based on Spring (Johnson 2003; SpringSource, a division of VMware 2011), which provides a framework for the development of applications compliant with the Java Platform, Enterprise Edition (Java EE, Oracle 2011a) specification. The traditional implementations of a server-side architecture model are Enterprise JavaBeans. They are, however, regularly suspected[1] of placing too high restrictions on the design of components for data management and processing logic as well as of demanding too complex configurations (keyword:

[1] Many disadvantages have been addressed in the newest EJB definitions 3.0 and 3.1.

'deployment descriptors', Höller 2005). The Spring framework, in contrast, offers a—in this sense—lightweight alternative to manage a number of loosely-coupled objects or rather a number of plain old Java objects (POJOs), as simple objects are called today in the Java world. The most characterizing and interesting aspect of this framework is the utilization of the so called 'Inversion of control' software pattern in the form of a technique termed 'dependency injection'. This allows for a centralized configuration and administration of the more than 450 Java-classes, which constitute the entire QuPE system (over 850 including auto-generated code). In this regard, the relationships between different objects are defined by means of XML-based configuration files.

To take an example of this concept: an instance of the class *ExcelExporter* provides functionality to export database search results in form of an Microsoft™ Excel sheet. According to the multilayer architecture model of QuPE, its business logic such as the functionality to retrieve protein identifications and other search information is implemented in classes belonging to the logic layer (*ObservationBusiness* and *ProteinBusiness* in this example). In the context of the Spring framework, the *ExcelExporter* instantiation and wiring is then performed solely based on the following code fragment:

```
<bean id="excelExporter" class="de.cebitec.qupe.export.ExcelExporter">
    <property name="observationBusiness"><ref bean="observationBusiness" /></property>
    <property name="proteinBusiness"><ref bean="proteinBusiness" /></property>
</bean>
```

All implemented classes follow the technical recommendations described by the JavaBean conventions (Hamilton 1997). In the case of the *ExcelExporter* this manifests itself in a public constructor as well as public get- and set-methods for all private attributes. In keeping with the definition of a JavaBean to be reusable, each class has, in general, no dependencies on the Spring framework, which allows to deploy a class and its instantiations in different and unrelated contexts, such as JUnit tests.

7.2 System architecture

The design of QuPE resorts to the multi-tier architecture model. As depicted in Figure 7.1, the system consists of three layers responsible for data access and retrieval, application logic and data processing, and the presentation of data.

7.2.1 Data access layer

The data model of an application targeting the analysis of quantitative proteomics data has to cover a variety of different data types starting from the representation of mass spectra, to protein and peptide identifications in terms of matches to sequence databases, to analysis results such as lists of statistically significant proteins or plots of expression values. Whereas the design of parts of the data model could follow recommendations made, in particular, by

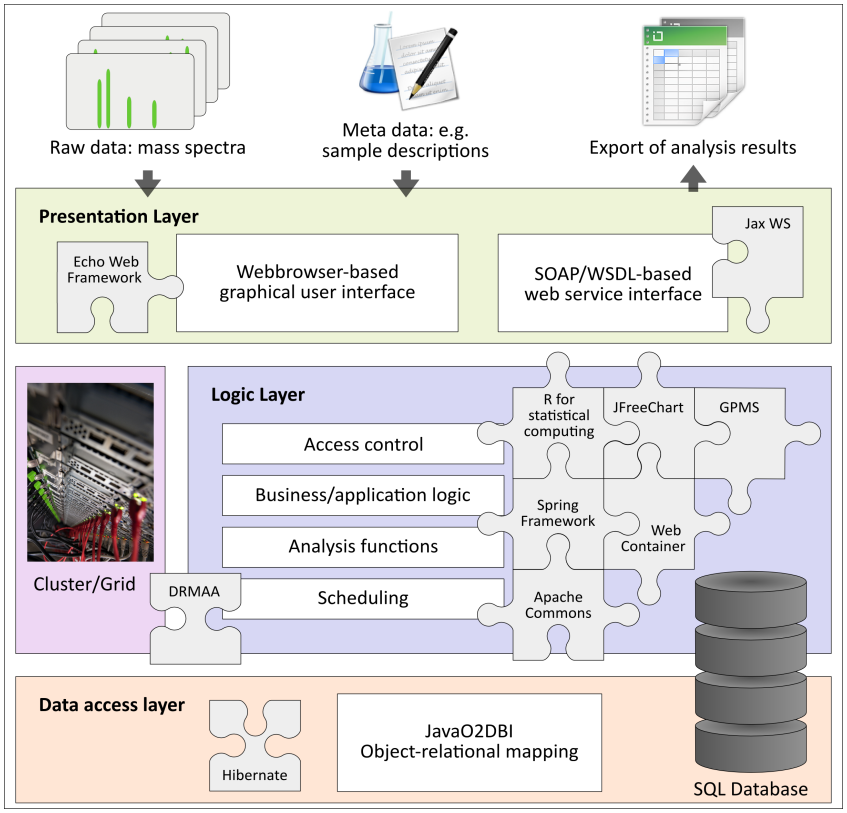


Figure 7.1 – This diagram depicts the three tier architecture model of the QuPE system. The data access layer provides an object-relational mapping utilizing Hibernate. The implementation of the application or business logic is located in the second layer, including the framework for the execution of computationally intensive tasks. The presentation layer is separated in two distinct components: a graphical user interface that allows the interaction with the system through a standard web browser, and a SOAP/WSDL-based web service, which can be utilized by other applications for data exchange.

the Proteomics Standards Initiative (PSI) at the HUPO (Orchard et al. 2003), there has not yet been proposed a standardized way to store higher-level analysis results.

The overall development of the data model pursued the model driven architecture approach (MDA, Object Management Group 2008) using the model designer O2DBI (Linke 2002). The implementation itself bases on Hibernate (JBoss Inc. 2011), a wide-spread object-relational mapping library. Therefore, a tool named JavaO2DBI developed by Kai Runte (Benölken 2007) has been extended to cooperate with the Spring framework. The XML-based modeling produced by the O2DBI designer is translated into the appropriate Java classes and Hibernate mappings. Objects responsible for data access, the so called DAOs, directly extend the class *org.springframework.orm.hibernate3.support.HibernateDaoSupport*, which provides support for database transaction handling, data retrieval, and database-related error handling. The model driven architecture approach yields a substantial benefit as it allows to cope easily with future requirements for the analysis of proteomics data. Thus, it facilitates not only the

addition of further attributes to existing classes but also the integration of new classes into the data model. In the following, selected aspects of the data model are presented in detail. For a complete description, the reader may refer to the QuPE API.

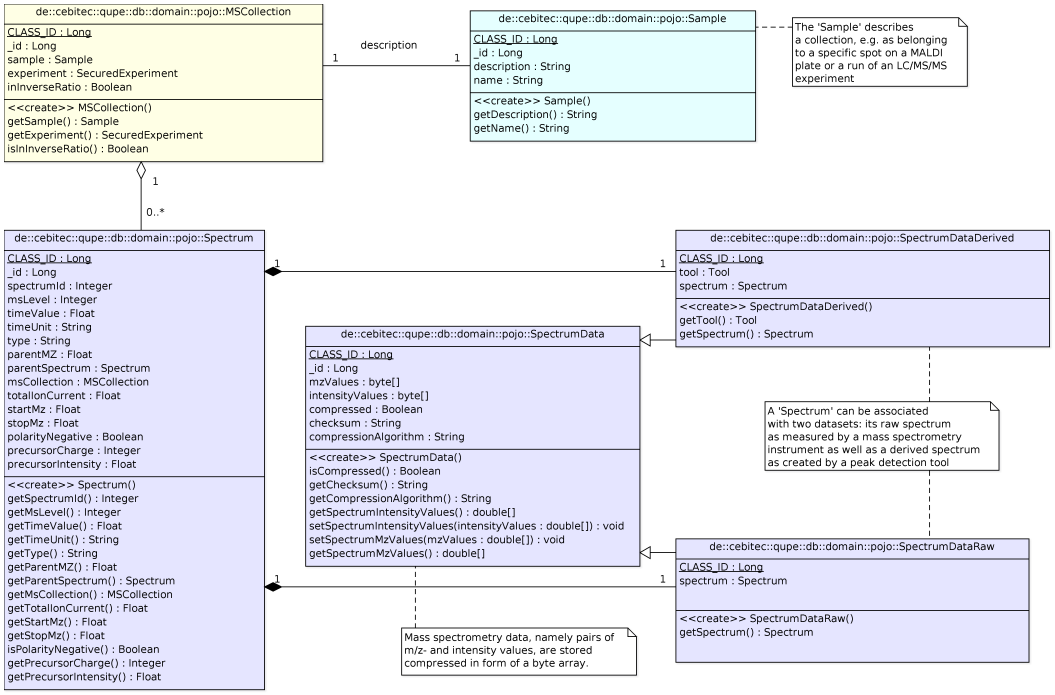


Figure 7.2 – This class diagram explains the data model used for the storage of mass spectra. An instance of *MSCollection* is a container for one or more mass spectra of type *Spectrum*. While the data itself is stored in an instance of *SpectrumData*, each spectrum may have associated both a raw dataset as measured by a mass spectrometer and a derived dataset as created, for example, by a peak detection tool. Please note, that some attributes and methods—in particular setter-methods and obvious methods such as 'equals' and 'hashCode'—are omitted in the visualization.

7.2.1.1 Object model for mass spectra

The classes designed to represent a collection of mass spectra (see Figure 7.2) in the data model closely follow the open source format mzData (Orchard et al. 2004) developed by the PSI and include attributes for the total ion current, the time of recording of a spectrum, as well as its numeric identifier. Aiming at an efficient storage of the raw datasets, namely pairs of m/z and intensity values in form of two arrays, data is stored as a compressed byte array (based on ZIP as implemented in the package *java.util.zip*). Each spectrum may have associated two different datasets, the raw mass spectrum as recorded by a mass spectrometry instrument, on the one hand, and a therefrom derived mass spectrum as result, for example, from a peak detection algorithm, on the other hand. In combination with a technique called 'lazy loading' (child objects are retrieved from database only on demand) the modeled parent-child relationship ensures an efficient retrieval of the data. An additional class *Sample* has

been introduced to describe the origin of a collection of mass spectra, such as a particular spot on a MALDI target plate or a specific run of an LC-MS/MS experiment. In summary, the implemented data model allows to efficiently store and retrieve mass spectra produced by different types of mass spectrometry instruments. The class *MSCollection* is able to represent both a single MALDI-TOF spectrum with only a few child spectra up to huge LC-runs comprising thousands of individual mass spectra.

7.2.1.2 Object model for protein and peptide identifications

Similar to the representation of the basic data type *Spectrum*, classes for the storage of peptide and protein identifications model themselves on PSI recommendations, which are nowadays described in the 'mzIdentML' data exchange standard (Proteomics Informatics Standards Group 2011). To base reported hits from a sequence database search using for example Mascot™ (see section 4.2.1) on common ground, all designed classes extend a common superclass named *Observation* (see Figure 7.3). Specializations are then found either in the class *ObservationProteinHit*, which refers to database search results from peptide mass fingerprinting (PMF), or *ObservationPeptideHit* to describe MS/MS ion search (MIS) results. Owing to their increasing importance, particular attention was paid to the representation of protein modifications in the data model. These may origin from a post-translational modification (PTM) of the protein or may have been introduced due to a specific chemical or physical treatment. For this purpose, an observation can have associated a variable number of objects of type *Modification*. Each modification, in turn, has to be described by an instance of the class *ModificationType* to define occurring changes in the molecular composition of the protein. In addition, references to further descriptions of the type of modification may be assigned as found in Unimod (Creasy and Cottrell 2004) or the RESID database (Garavelli 2003), for example.

A special feature of QuPE's data model is the fact that an observation can be labeled with an annotation, and thereby marked as a valid protein or peptide identification. At this, the annotation may, in addition, be given a level of certainty. This can either be done manually by a user or automatically based on certain criteria (see section 7.4.2 for further details). All evidential observations that 'proof' the identity of a specific protein are aggregated by an object of class *Protein*. Based on the final list of proteins determined for an experiment, information from external resources such as UniProt (The UniProt Consortium 2008) or KEGG (Kanehisa and Goto 2000) is collected and integrated into the 'pool of knowledge' about the samples under investigation.

7.2.1.3 Object model for analysis results

In contrast to the aforementioned data models, yet no binding recommendations have been made regarding the representation of higher-level analysis results obtained, for instance, by a statistical test or a hierarchical cluster analysis. Only recently, an initiative has been started to create a common data exchange standard for protein amounts quantified in a cell

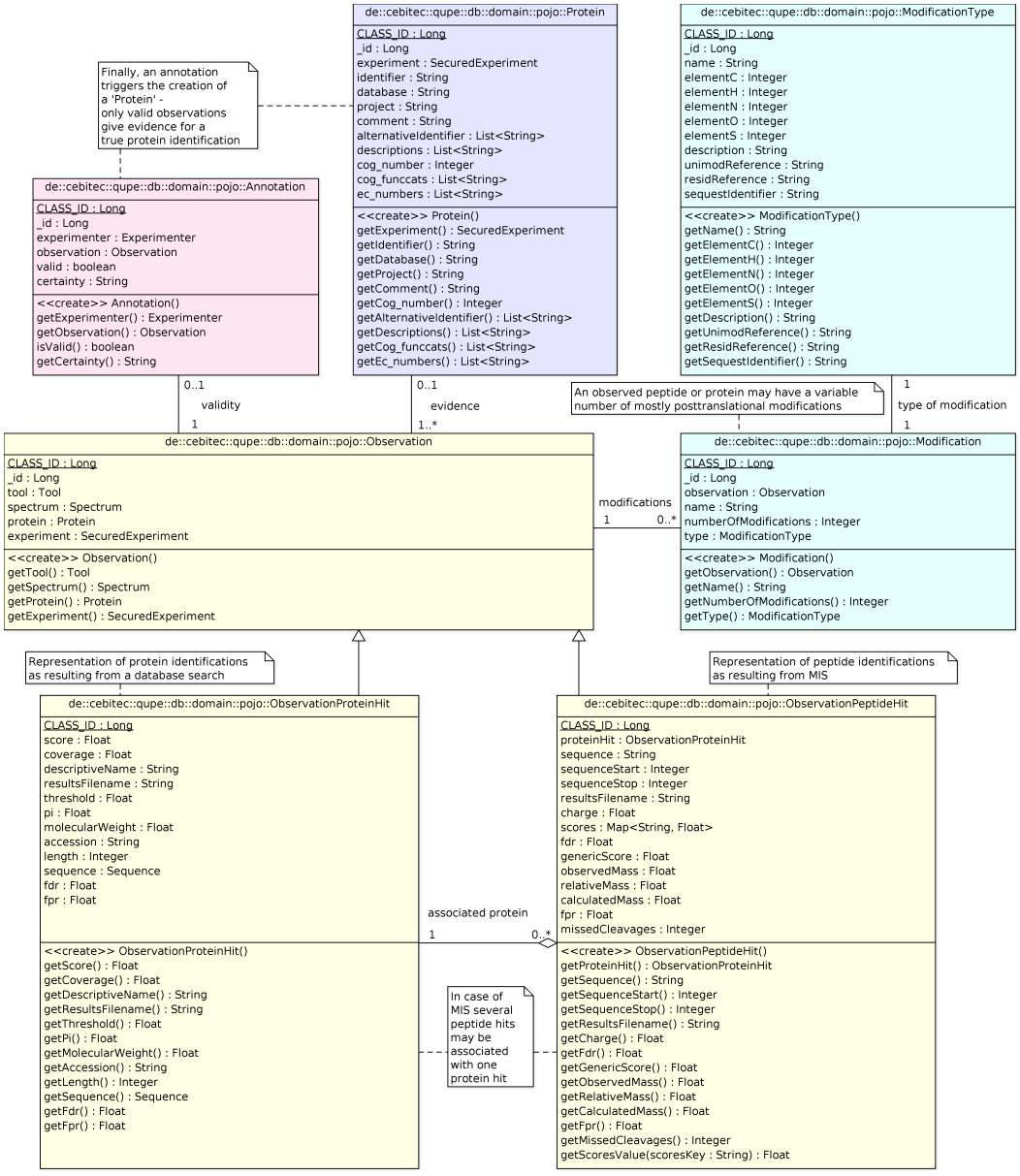


Figure 7.3 – This class diagram explains the data model implemented for protein and peptide identifications. Extending the common superclass *Observation*, the two classes *ObservationProteinHit* and *ObservationPeptideHit* refer to database search results from peptide mass fingerprinting (PMF) or MS/MS ion search (MIS), respectively. An observation can have associated a variable number of protein modifications, which have to be described by an instance of class *ModificationType*. A special feature of QuPE's data model is the fact that an observation can be labeled with an annotation, and thereby marked as valid. An object of type *Protein* groups all information related to a protein identification. Due to reasons of space some attributes and methods are omitted in the diagram. This includes setter-methods and obvious methods such as 'equals' and 'hashCode' but also 'convenience methods' e. g. to add or remove an element to or from, respectively, a collection or to check whether a list is empty.

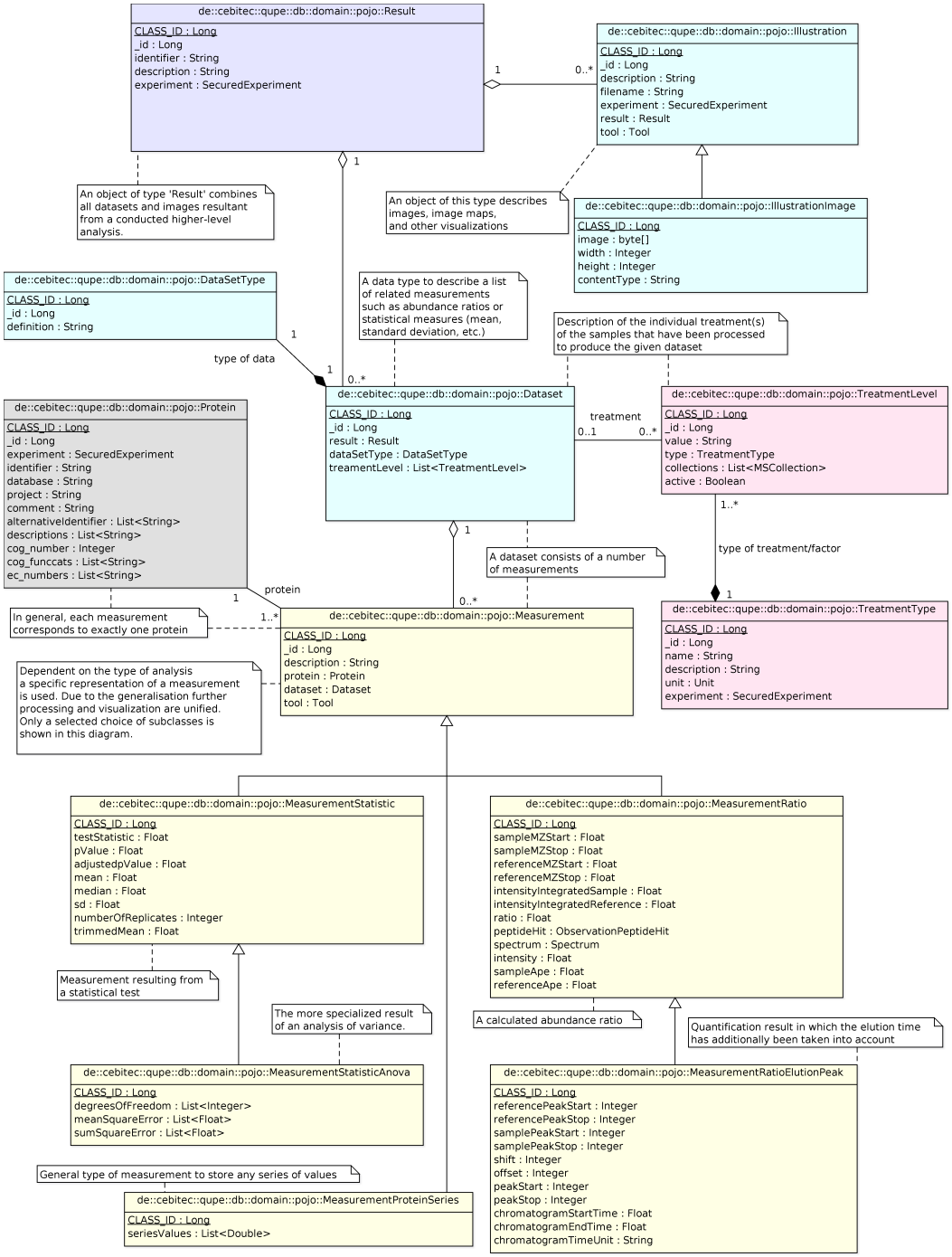


Figure 7.4 – This class diagram explains the data model used to store analysis results such as calculated abundance ratios, statistical measures including a mean value and its standard deviation, but also images such as box-and-whisker plots or heatmaps. Please note that due to reasons of space some attributes and, in this case, all methods have been omitted.

or organism. The development is predominantly driven by the PSI and will presumably be named 'mzQuantML'; however, no version has been released until now.

Along with the inventory of methods for the analysis of quantitative proteomics data (see chapter 6), the requirements have been identified for an object model to store the data arising from these analyses. First of all, this aimed at an adequate representation of calculated protein abundance ratios, which also included protein quantification utilizing the elution time of a peptide. At second, statistical measures needed to be taken into account in the object model, such as the mean value of all abundance ratios found for a protein under a specific condition together with the standard deviation, or analogous, the median. Thirdly, an analysis may produce any kind of plot or, more generally, visualization as a result. This can, for example, be box-and-whisker plots, or heatmaps as originating from a hierarchical cluster analysis.

In the design of the object model (see Figure 7.4), all results of a performed analysis are combined in one *Result* object, which in turn may consist of datasets of measurements and/or any number of illustrations. While a general type of measurement termed *MeasurementProteinSeries* has been devised to store any series of values for a protein, more specific types of measurements are included in the data model for the most common types of analyses. This involves, *inter alia*, statistical measures and test results including, for example, the ANOVA or a Kruskal-Wallis rank sum test. A dataset groups any number of measurements. In order to distinguish the results of an analysis regarding different conditions or sample treatments, a dataset can be further described using a list of objects of type *TreatmentLevel*. At this point, a type of treatment refers to the factor or condition which has changed during the experiment. An example would be 'heat' or 'time'. A level of treatment then defines the specific value or characteristic of this type, e. g. one hour or 30°C.

7.2.1.4 Object model to structure experiments and related data

To structure all data and meta-data relevant to a specific experiment, the data model of QuPE provides a corresponding representation (see Figure 7.5). Several experiments may in turn belong to a project[2] to group all information gathered, for example, by an individual experimenter or a particular working team. Each user is represented in the object model by an instance of the class *Experimenter*. Upon creation a specific object is firstly only accessible by its owner, yet differentiated access rights can be assigned to other experimenters allowing them to read, modify or delete an experiment or even a complete project. The use of a common superclass (*Secured*) for all kinds of 'secured' objects, allows to provide generic implementations of the necessary functions for this level of application security, which is based on access control list (ACL) directives. Given an object of type *Secured*, there may exist a variable number of objects of type *ACL* that specify whether particular privileges are granted or denied.

[2] To avoid name clashes and misunderstandings with instances of the class *Project* within the central CeBiTec project management system (GPMS, see section 7.4.1 for further details), a project in QuPE is internally referred to as *Subproject*.

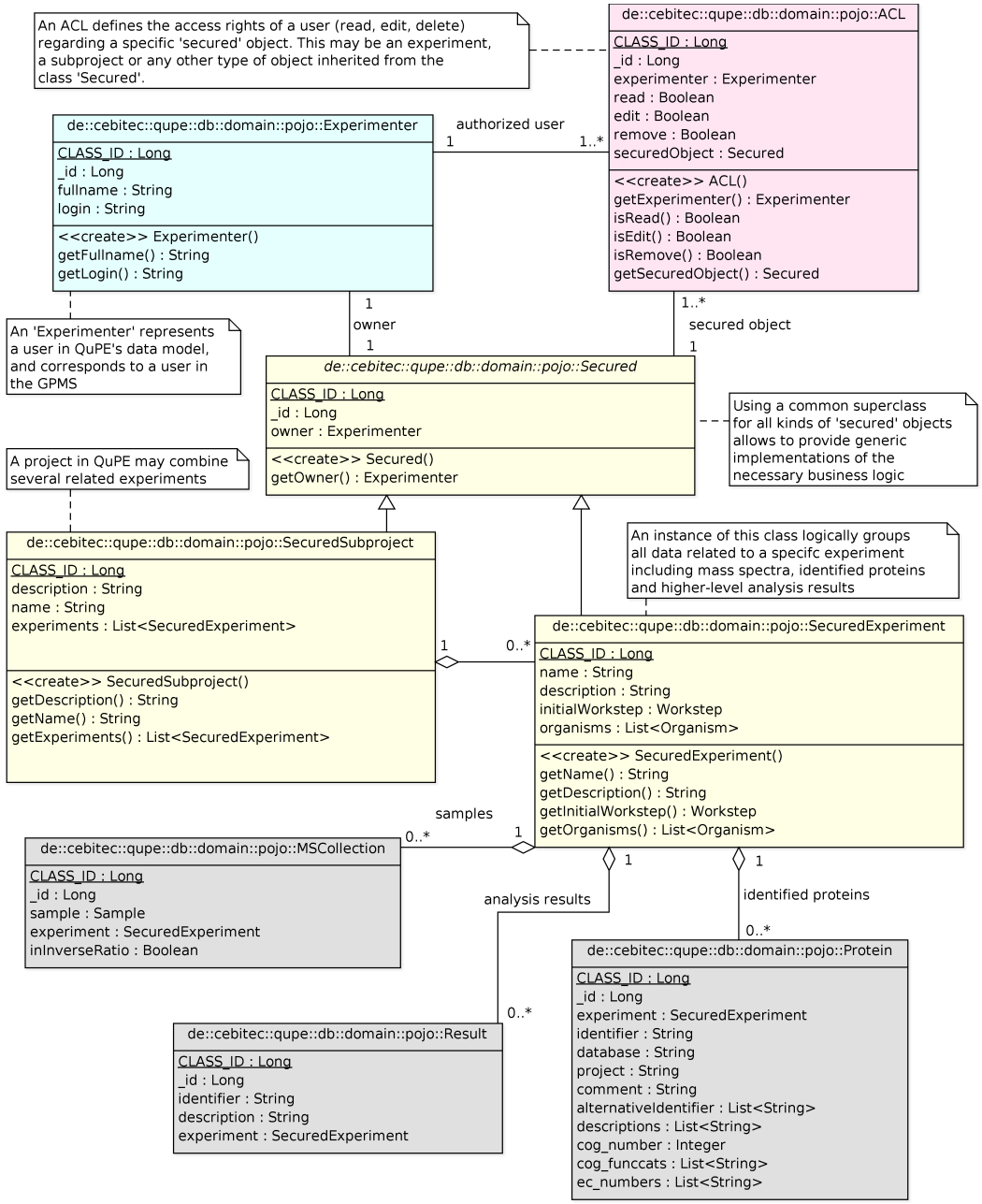


Figure 7.5 – The classes designed to group all data relevant to a specific experiment are described in this diagram. This may include mass spectra as well as lists of identified proteins, and higher-level analysis results. In QuPE, several related experiments may belong to one project (internally referred to as 'Subproject'). To allow the assignment of fine-granular privileges to individual objects, as for example an experiment, access control list (ACL) directives have been set up.

7.2.2 Logic layer

At the heart of the QuPE system are the classes and methods that provide the overall business logic, *inter alia*, to process and validate interactive requests, to distribute workload between a web server and a compute cluster, to initiate and prepare storage and retrieval of data, or to perform calculation tasks.

Classes in the Java package *de.cebitec.qupe.business* mediate the connection between the data access and the presentation layer. They act as an interface to process user actions received by the graphical user interface but also those of the provided web service, and accordingly initialize data transfer from and to the database. This includes the composition of information e. g. in form of newly instantiated objects, the validation of received data, and the verification of user permissions regarding a user's right to perform a requested operation.

7.2.2.1 Job and tools framework

A main pillar of this middle layer of QuPE is the framework for the execution of tasks, e. g. to import data, to perform calculations, or to conduct a database search for protein identification (package *de.cebitec.qupe.task*). The objective of this framework is, firstly, to provide programmers with a well-defined programming interface (API) that eases the integration of new functions and that offers frequently used methods for the retrieval, processing, and storage of data. In addition, the API facilitates the integration of routines written in the programming language R (R Development Core Team 2011; Chair for computeroriented statistics and data analysis 2008), and thereby allows developers to resort to a wealth of established data analysis methods. Secondly, the framework enables generic and unified views for both the configuration and initiation of a task, as well as the monitoring of a task during its execution. Similar, the access to resulting data objects and their presentation in a graphical user interface are unified, and do not need to be addressed by a developer.

Central parts of this framework are the two classes *Tool* and *Job*, which represent a specific unit of work, on the one hand, and a complete task consisting of these work units processed in a defined order, on the other hand (see Figure 7.6). In combination, both are used to collect and describe all aspects of a computation. Obviously, this includes the input in form of individual mass spectra, complete samples, or datasets that resulted from a previous calculation. In the end, each job may be associated with its resulting output in form of an object of type *Result*—if appropriate—consisting, for example, of datasets of calculated abundance ratios or a number of plots (see section 7.2.1.3).

If a tool demands additional configuration, e. g. to let a user choose a specific method for a calculation, a list of parameters of the type *ToolParameter* may be defined for each tool, in practical terms, pairs of parameter names and associated values. Currently, *String*, *Float*, *Integer*, and *Boolean* values are supported for this purpose. To ensure a valid configuration, a programmer may set, in case of numeric values, upper and lower limits as well as an allowed

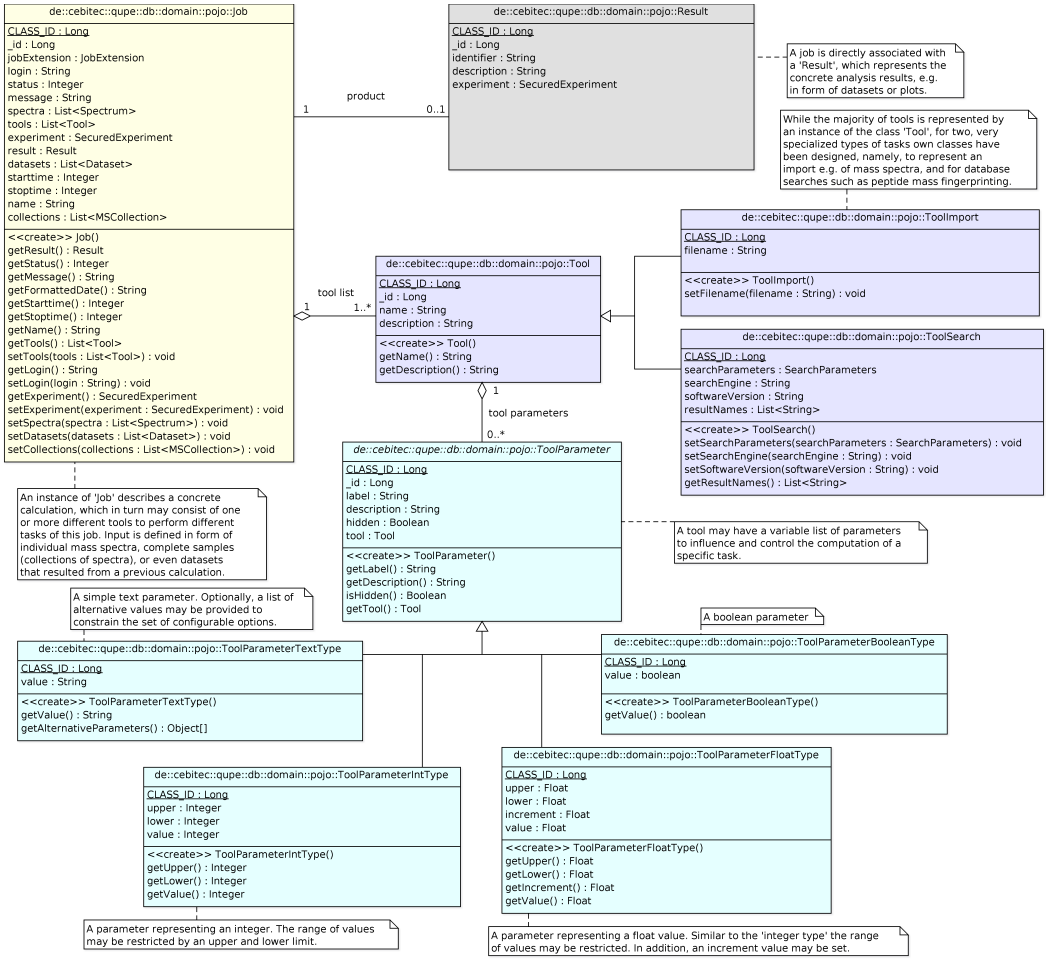


Figure 7.6 – This diagram describes all classes representing any kind of computational tasks performed either on data in the QuPE system or to import data into the system. An instance of the class *Job* consists of one or more individual tools (*Tool*), each designed to portray a specific part of a calculation. The input of a job can be a list of individual mass spectra, complete samples, or even datasets that resulted from a previous calculation. Each tool can have a variable list of parameters to control its behavior during the execution.

increment value. In case of textual parameters, optionally, a list of terms may be provided to restrict the value of a parameter to these alternatives.

While the majority of work units is represented by an instance of *Tool*, for two, very specialized types of units of work, own classes have been designed. This is, firstly, the import of data into the system, in particular, of mass spectra, and secondly, the identification of proteins using a database search engine such as Mascot™.

The necessary application logic to perform a specific unit of work is strictly segregated from its representation in the data model and has to be implemented in a class of type *ToolTask*. In analogy to the two classes *Job* and *Tool* such a *ToolTask* or any combination of these assemble to a *JobTask*. An important characteristic of each *JobTask* is the implementation of the Java

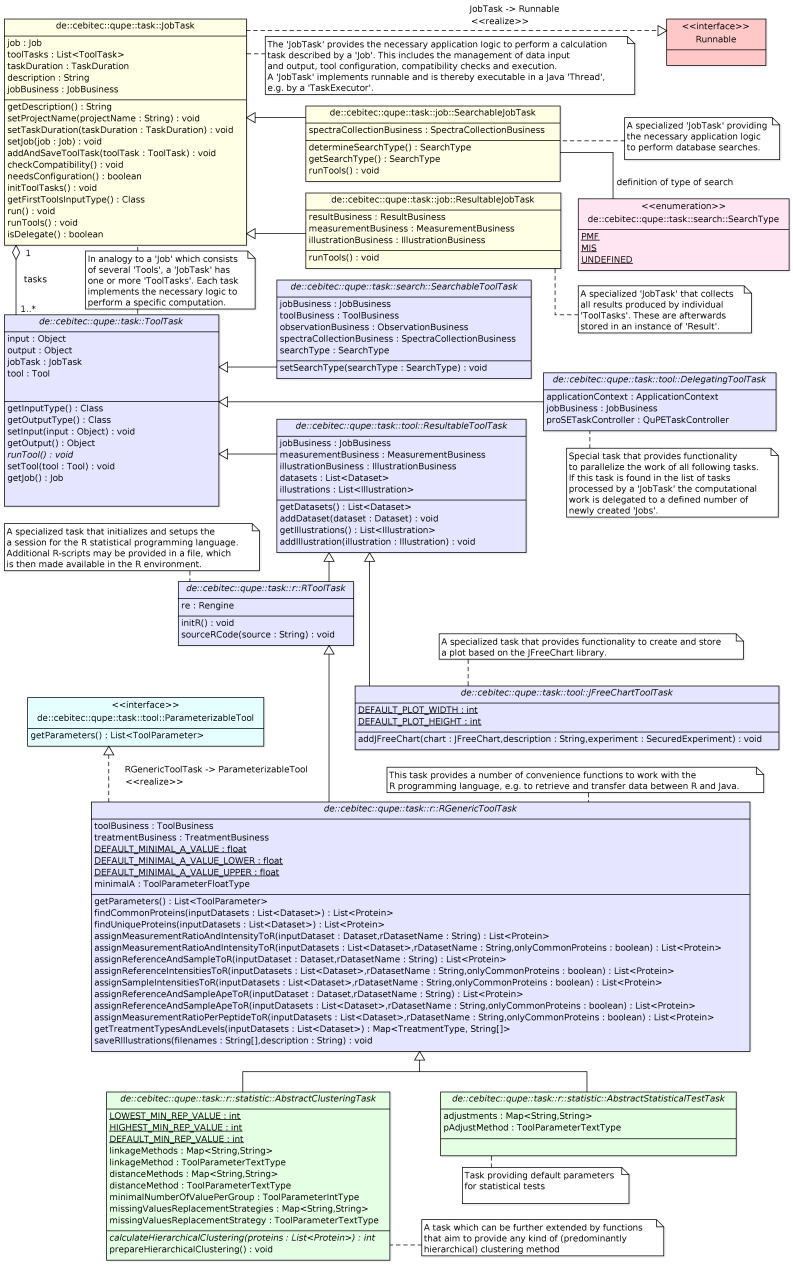


Figure 7.7 – Together with the data model shown in Figure 7.6, the herein described classes complement the framework for the execution of tasks, e. g. to import data, to perform calculations, or to conduct database searches. The class *JobTask* and its specializations, which all implement the Java interface *Runnable*, provide the application logic to perform these tasks. Concrete implementations for all aspects of these computations are found in tools each represented by a *ToolTask*. This is, for example, the calculation of isotopic distributions for a list of peptides, which are then used as basis for a second tool to determine abundance ratios.

interface *Runnable*. Each task is, hence, executable in its own Java *Thread*, and the system can thereby take care of the execution of a *JobTask* in a separate process.

Conception and design of a specific task rely on the capabilities of the Spring framework and its unified, centralized configuration. To take an example, the following XML-code defines a method to quantify protein amounts in isotopically-labeled samples. The complete task consists of four different tools, each fulfilling a specific purpose from the calculation of isotopic distributions, to the extraction of ion chromatograms, to finally, the calculation of protein abundance ratios.

```
<bean id="elutionPeakQuantification" parent="resultableJobTask" scope="prototype">
    <property name="description">
        <value>
            Quantification utilizing peptide elution (Linear regression/RelEx approach)
        </value>
    </property>
    <property name="taskDuration"><value>LONG</value></property>
    <property name="toolTasks">
        <list>
            <ref bean="isotopicDistributionTask" />
            <ref bean="elutionPeakTask" />
            <ref bean="elutionPeakChromatogramTask" />
            <ref bean="elutionPeakQuantificationTask" />
        </list>
    </property>
</bean>
```

Typically for many calculation tasks, as can also be seen in this example, the job inherits from the provided class *ResultableJobTask*. Thereby, the creation of an appropriate *Result* object to store the datasets written out by this computation is taken care of. In this way a developer is only required to implement the four *ToolTasks*, which are then, of course, reusable in other contexts. Before and during execution of a *JobTask*, the provided implementation verifies that the different tools and their inputs and outputs are compatible to each other and manages the transfer of data objects. In a job configuration, an expected duration of the overall computation may be specified and expressed by the enumeration *TaskDuration*, which currently supports three stages: short, medium, and long.

At this point, a further characteristic of the Spring framework comes into play: given that the implementation of an object relies on Java interfaces, the concrete realization of an object is exchangeable solely based on a modified XML-configuration. This is made use of for the delegation of computationally-intensive and long-running tasks: while the production version of the QuPE server may refer to an Oracle™ Grid Engine (Oracle 2011d) to benefit from the advantages of a distributed computing solution, during development, task execution is performed locally using a *ThreadPool*:

```
public class QuPETaskController implements Serializable {

    /* ... */

    /*
     * Task execution is categorized dependent on its expected
     * duration and consumption of resources in three categories.
     * The concrete implementation of an executor is
     * interchangeable, but has to base on the interface
     * org.springframework.core.task.TaskExecutor
     */
    private TaskExecutor shortTaskExecutor;
    private TaskExecutor mediumTaskExecutor;
    private TaskExecutor longTaskExecutor;

    /**
```

```
     * Setter for property longTaskExecutor
     * @param longTaskExecutor
     */
    public void setLongTaskExecutor(TaskExecutor longTaskExecutor) {
        this.longTaskExecutor = longTaskExecutor;
    }

    /* ... */
}
```

```
<bean id="proSETaskController" class="de.cebitec.qupe.task.controller.QuPETaskController">
    <property name="shortTaskExecutor"><ref bean="shortTaskExecutor"/></property>
    <property name="mediumTaskExecutor"><ref bean="mediumTaskExecutor"/></property>
    <property name="longTaskExecutor"><ref bean="longTaskExecutor"/></property>
    <property name="jobBusiness"><ref bean="jobBusiness" /></property>
</bean>
...
<-- Locally a ThreadPool is used for the execution of computationally intensive tasks -->
<bean id="longTaskExecutor" class="org.springframework.scheduling.concurrent.ThreadPoolTaskExecutor">
    <property name="corePoolSize" value="${task.short.executor.core.pool.size}" />
    <property name="maxPoolSize" value="${task.short.executor.max.pool.size}" />
</bean>
```

```
<-- In production mode this is instead replaced by a
      binding to the Oracle(TM) Grid Engine via DRMAA -->
<bean id="longTaskExecutor" class="de.cebitec.qupe.task.drmaa.DRMAAExecutor" />
```

7.2.3 Presentation layer

The main interface that allows users to interact with the QuPE system is implemented using the Echo2 web framework (NextApp, Inc. 2011). An example screenshot is depicted in Figure 7.8. In addition, a web service interface is provided based on SOAP and the web service description language (WSDL, Gudgin et al. 2011), which can be utilized by other applications to retrieve and exchange analysis results as, for example, complete datasets of calculated abundance ratios. The capabilities of this interface have successfully been demonstrated by ProMeTra (Neuweger et al. 2009), a web application that allows to combine PolyOmics datasets from different sources and to project expression values on 'self-made' metabolic pathway maps.

7.2.3.1 Graphical user interface

It has been a matter of particular concern to take into account the distributed location of users. Aiming to enable the sharing of information and data not only between different departments such as a laboratory and an office but also between different universities or institutions, the QuPE system was developed as a rich internet application (Allaire 2002). The Echo web framework, which is based on Asynchronous JavaScript and XML (AJAX, Garrett 2005), allowed to develop a graphical user interface that, on the one hand, behaves similar to the user interface of a standalone software application started on a personal computer, but on the other hand, is accessible using a standard web browser whenever and wherever an internet connection is available. QuPE is installable on any Java EE-compliant web server. Using QuPE is independent from any web browser[3] or operating system. In the sense of

[3] Except for some minor display problems with the newest version (v9.0) of the Microsoft™ Internet Explorer

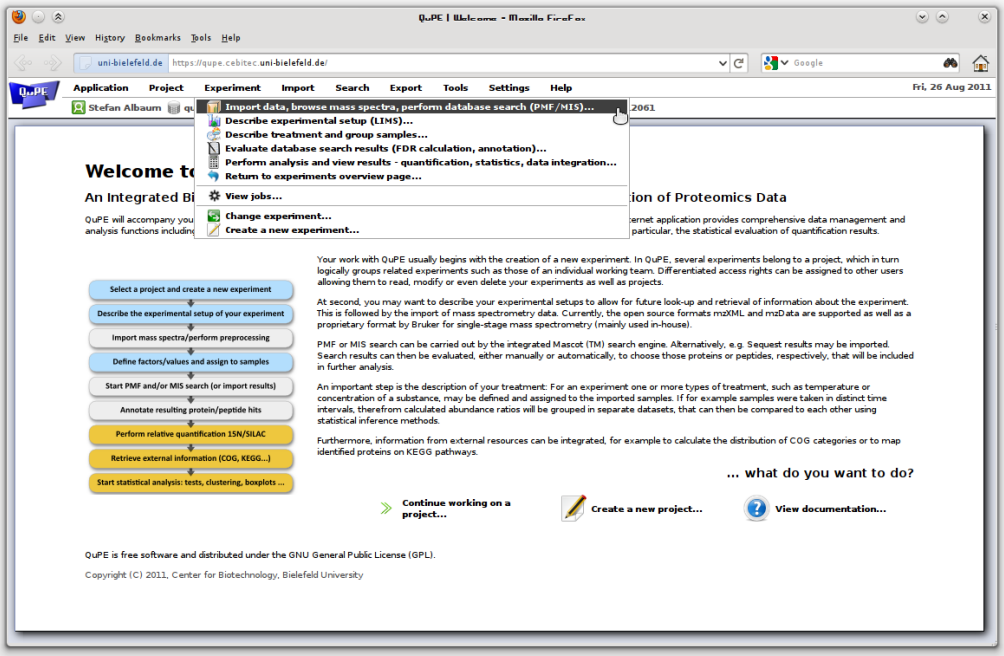


Figure 7.8 – This Figure shows a screenshot of QuPE's graphical user interface running in a web browser. On the top of the page, the main menu provides access to all functionalities of the system. Information about the logged-in user, as well as the currently selected experiment and data is displayed below in a status bar. In this example, the main part of the page is used to display a welcome message and some introductory text to QuPE.

the 'software as a service (SaaS)' concept software maintenance and further development are centralized (Mell and Grance 2010).

7.2.3.2 Design and control of the graphical user interface using a model-view-controller pattern

In contrast to many other frameworks for the development of web-based applications, Echo2 purely relies on the implementation of server-side components in the Java programming language. Communication between different components of the user interface is handled by an event-based programming paradigm akin to that of the Java Swing API. This allowed to implement a model-view-controller pattern to control the graphical user interface of QuPE. The use of this pattern facilitates the extension of the system by new input masks or visualizations of analysis results. In the context of the web application these are termed pages. As only the classes responsible for the view are directly referring to elements of the Echo2 web framework, a further advantage of this pattern is that it would—in the distant future—be conceivable to complement the presentation layer of QuPE with a graphical user interface implemented in Java Swing.

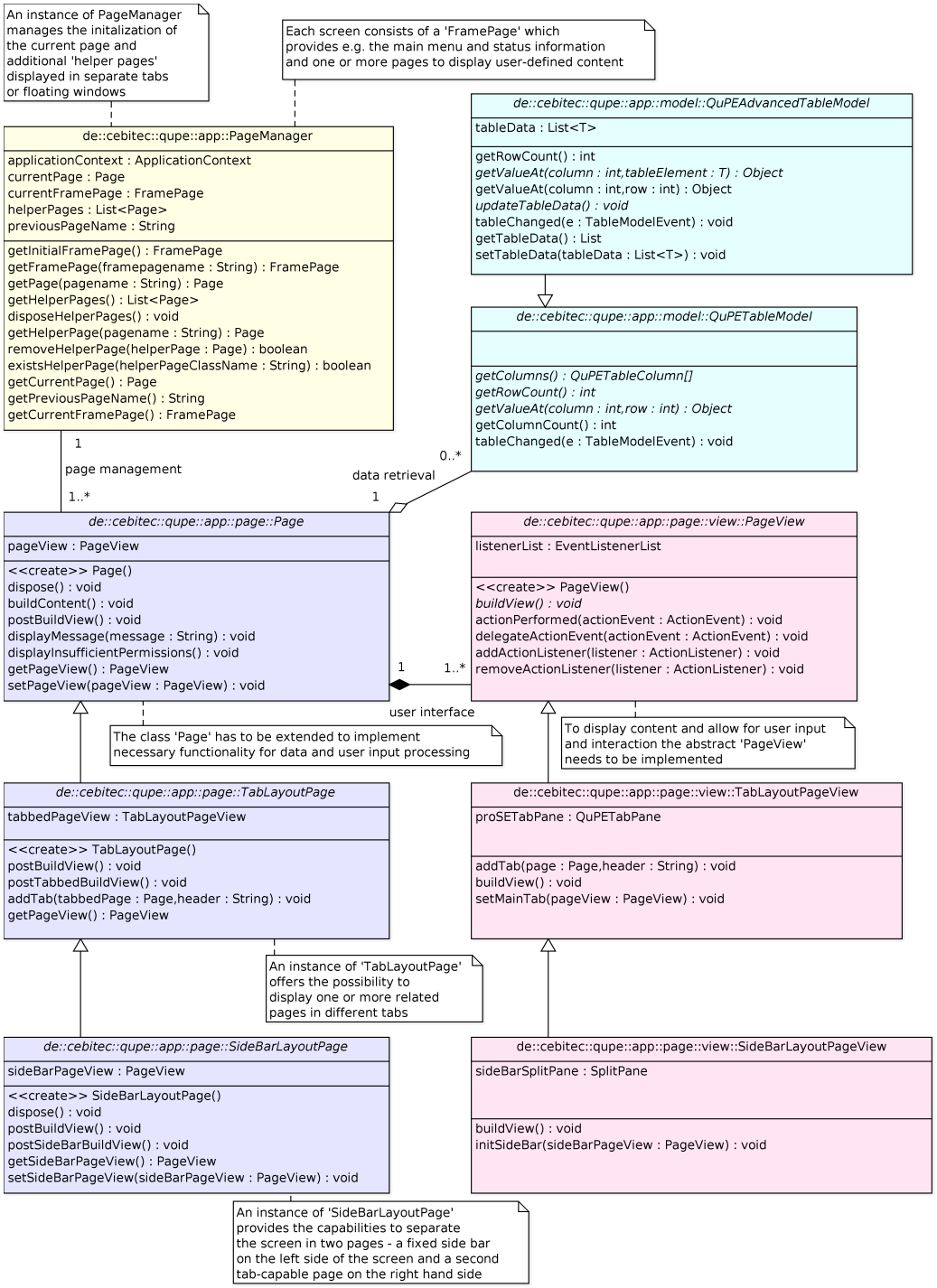


Figure 7.9 – This class diagram shows the relationship between the three classes *Page*, *PageView*, and *TableModel* and their specializations, which make up the model-view-controller pattern to retrieve data, process and display content, and allow for user interaction within the web browser-based graphical user interface of QuPE. An instance of the class *PageManager* manages the initialization of a page using the Spring framework.

Figure 7.9 gives an overview of the classes that have been designed to put this pattern into practice. A special type of page, the so called frame page, is responsible for the overall layout and structure. It provides a screen area to display the main content represented by an instance of the class *Page*. Currently, two different *FramePage* implementations are available in the context of QuPE. The first is used to draw a login screen providing a container to query a user's credentials and to select a GPMS project, respectively, database. A second frame page generates the basic structure of each subsequent screen including the main menu and additional status information.

Each individual page has to be based on the class *Page*, which acts as the controller to process user requests and at least one subclass derived from *PageView*, which is responsible for the layout and setup of the structural elements of the user interface. All events occurring at specific components, for example if a user performs a click on a button, are delegated from the *PageView* instance to the page controller of type *Page*. In addition, one or more model-implementations may be provided that act as mediator between the presentation layer and the application layer to retrieve data objects from the database. At this point, the design diverges from the original implementation of the model-view-controller pattern, as—in the name of simplicity—the model's realizations are direct specializations of the classes *TableModel* or *ListModel*. The reason for this variation is simply the fact that most of the data is displayed in form of lists and tables.

Apart from the simple display of a single page as it is the case in the screenshot of Figure 7.8, a developer may refer to presentations that are built up of several related pages using tabs and/ or using a sidebar.

All pages have to be registered within the Spring framework. During server startup, an instance of the Spring framework's class *ApplicationContext* is build up and configured according to the given configuration. If a user requests a new page, e. g. by selecting an entry in the main menu, setup and instantiation of a page are mediated by a singleton instance of the class *PageManager*, which has full access to the *ApplicationContext*. In the first instance, for each page and potential 'helper pages' displayed in a side bar or a tab the *buildView()* methods are invoked to install and setup each corresponding *PageView*'s content. At next, control is handled over to the controller, namely the instance of *Page* using the method *postBuildView()*. This can be used to finalize the display of content, in particular, in consideration of an actual selection, e. g. a currently chosen mass spectrum or protein hit.

7.3 Algorithms for the analysis of quantitative proteomics data

Even though a wide range of software tools for the quantification of isotopically-labeled proteins has been introduced in recent years, the accuracy and performance of currently available algorithms is still worthy of improvement, and, as already mentioned in chapter 5, regarding specific experimental setups, in particular concerning the quantification of

proteins in which a metabolic label has only been partly-incorporated, no algorithmic approaches have been conceived yet. Utilizing the application programming interface (API) of QuPE, therefore, a number of quantification algorithms have been developed, whose implementations are now explained in detail. A comprehensive evaluation of the herein introduced methods can be found in chapter 8. All implemented algorithms are accessible via the web interface of QuPE and have proven their applicability in several quantitative proteomics experiments (e. g. Grasse et al. 2011; Fränzel 2010; Albaum et al. 2011b; Haußmann and Poetsch in-press).

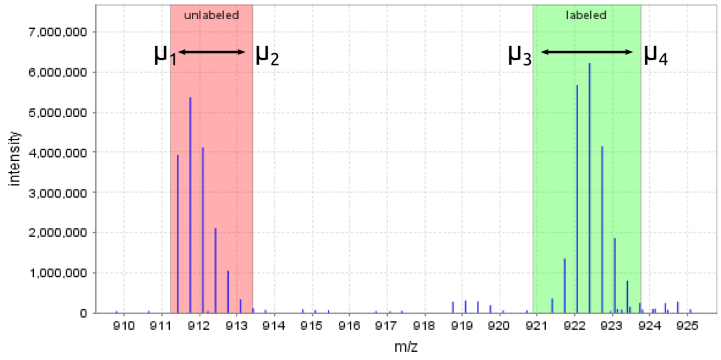
7.3.1 Sum quantification approach – simple but powerful

In the first instance, a simple and straightforward approach was taken to quantify metabolically-labeled samples in which one protein, or rather one peptide, is found fully-labeled and one completely unlabeled. In contrast to techniques such as iTRAQ (Ross et al. 2004), where the quantitative information can be extracted from a single isolated peptide and hence from the MS/MS scan that accounted for the peptide's identification, in metabolic labeling the full MS scan has to be used for quantification (cf. also section 4.3). Required input of the algorithm is a list of peptides that have been identified in a database search, e. g. using Mascot™. Based on each peptide's sequence, its actual charge state and any observed protein modification, the expected theoretical isotopic distributions are calculated for both the unlabeled peptide and its counterpart characterized by a specific label such as heavy stable nitrogen isotopes or a tagged amino acid and an estimated incorporation rate of this label. The information gained in this way about the m/z-values of each of the two peptides directly leads to their intensities in a mass spectrum, and subsequently, the intensities' ratio to a measurement of relative abundance. In principle, quantification could be based only on the monoisotopic or the most abundant peak of both isotopic distributions. However, since label incorporation rates typically do not reach 100 percent, the isotopic patterns reveal heavily varying forms. Therefore, the complete isotopic distribution[4] is used for quantification.

At the cost of efficiency but ensuring a high accuracy, the calculation of isotopic distributions is grounded on a polynomial algorithm derived from an open source program named 'Isotopic Pattern Calculator' (Nolting 2008, see Appendix A.1) that closely follows the method introduced by Yergey et al. (1983). The algorithm allows to determine the mass to charge positions and relative intensities of each isotope of a peptide with a user-definable but in all cases sufficient exactness. Atomic weights and isotope probabilities were taken from the database of atomic weights and isotopic compositions hosted at the National Institute of Standards and Technology (Coursey et al. 2005). The procedure of the sum quantification approach is implemented as follows: formally, a mass spectrum **S** that consists of p discrete peaks with each an m/z-value m and intensity i can be described by two vectors $\mathbf{m} = \{m_1 \dots m_p\}$ and $\mathbf{i} = \{i_1 \dots i_p\}$. Given the theoretical isotopic distribution calculations predict the peaks of the unlabeled peptide to be localized between the two m/z values μ_1 and μ_2, and for the labeled peptide between μ_3 and μ_4, a relative abundance value for a peptide

[4] Strictly spoken, all intensities above a reasonable threshold as for example one percent.

can be calculated and transformed into a logarithmic ratio, similar to the so called M-value known from microarray experiments (Dudoit et al. 2000):



$$M = \log_2 \frac{\sum i_k}{\sum i_j}, \forall k : \mu_1 - \varepsilon \leq m_k \leq \mu_2 + \varepsilon, \forall j : \mu_3 - \varepsilon \leq m_j \leq \mu_4 + \varepsilon \tag{7.1}$$

To ensure that the monoisotopic peak of the unlabeled peptide, in particular, is not missed, a user-defined tolerance value ε can be taken into account to extend the investigated ranges.

The ratio naturally ignores the overall intensities of both peptides. Although peak intensities are an unreliable predictor for the absolute amounts of proteins in a cell, they nevertheless provide a measure of the quality of a peptide identification as very low intensities may be affected by background noise. Therefore, a value, in the following termed A-value, is hereby proposed to assess a measurement:

$$A = \log_2 \sum i_k \bullet \sum i_j \tag{7.2}$$

Implementation of the sum quantification algorithm

For the concrete implementation of the sum quantification approach specific *ToolTasks* have been devised (see Figure 7.10). It was decided to separate the processing of samples labeled with the SILAC approach and those that utilize metabolic labeling with heavy stable isotopes such as ^{15}N. Accordingly, two different tools are provided for the calculation of theoretical isotopic distributions, which can then be combined with a second task to calculate peptide abundance ratios. For data exchange a mapping of peptide sequences on two isotopic distributions—one for the labeled and one for the unlabeled variant of a peptide—is utilized. As the quantification of MALDI-TOF data demands special processing of mass spectra and different default settings of parameters, the implementation is provided in an own *ToolTask*. Execution and configuration of the final calculation task is done as described in section 7.2.2.1.

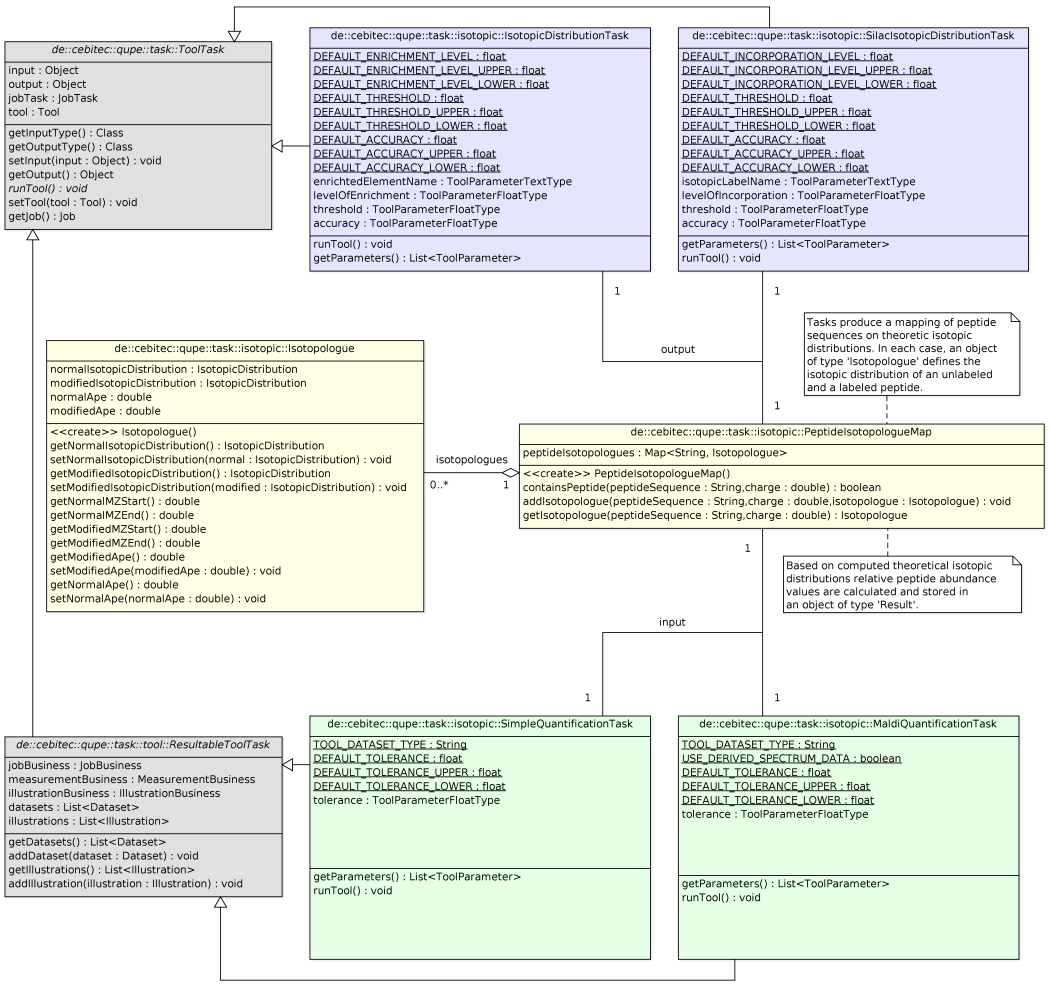


Figure 7.10 – The class diagram in this Figure displays details of the implementation of the described sum quantification algorithm. Dependent on the type of label, either SILAC or a heavy stable isotope such as ^{15}N, different implementations to calculate theoretical isotopic distributions are utilized. The output in the form of a mapping of peptide sequences on two distributions for the labeled as well as the unlabeled variant of a peptide are then used as input for the calculation of relative abundance values.

7.3.2 Utilizing the time

In case of an LC-MS/MS experiment, the calculation of relative abundance ratios can be significantly improved by taking the temporal information gained from a peptide's elution into account. This has successfully been shown by tools such as RelEx and ProRata (cf. section 4.3.2 ff.). In contrast to the sum quantification approach it is not only necessary to get to know the m/z positions of both the unlabeled and the labeled peptide but also the start and end point of their elution from the chromatographic column. The algorithm requires—in agreement with the previous approach—a list of identified peptides including their sequence, charge state and any present protein modification as its input. In addition to the concrete

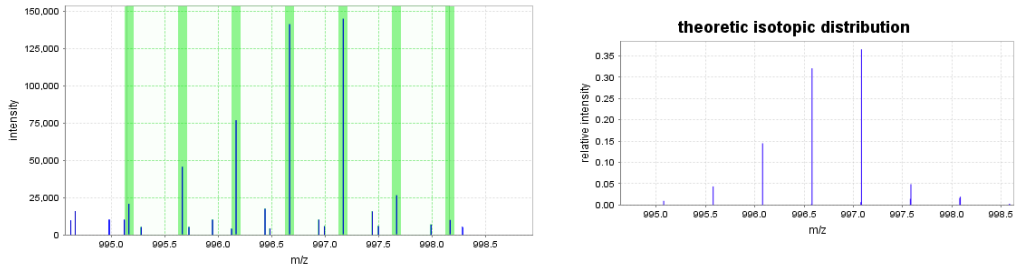


Figure 7.11 – A common problem in mass spectrometry is noise as illustrated in this example. To cope with this kind of error, the theoretical isotopic distribution (right side of the picture) can be used to extract the intensities only from a small-framed window around the exact positions of each peak (left spectrum, indicated by the highlighted bars).

mass spectrum that accounts for the identification of a peptide, the retention time is a further requisite information.

For the construction of extracted ion chromatograms (EICs), at first the theoretical isotopic distributions are calculated as previously described and used to determine the m/z ranges, in which the peptide pair can be found in a mass spectrum. Common problems in mass spectrometry are noise or interferences that occur, for example, if the peaks of two peptides with similar mass overlap in a spectrum (Hoopmann et al. 2007). Putting forward a proposal to reduce the impact of these errors, instead of the whole isotopic envelope that is spanned by the theoretical isotopic distributions exact peak positions can be utilized. The approach is illustrated in Figure 7.11. Given a spectrum $\mathbf{T}$ with q peaks, which is described by a vector of m/z-values $\mu = \{\mu_1 \dots \mu_q\}$, represents the theoretical isotopic distribution of a peptide, given further ε as a small-sized value, and furthermore, $\mathbf{S}$ denotes a recorded mass spectrum that has p peaks, $\mathbf{m} = \{m_1 \dots m_p\}$ and $\mathbf{i} = \{i_1 \dots i_p\}$, only those intensities of $\mathbf{S}$ are considered for the quantification that are defined by a new intensity vector $\tilde{\mathbf{i}}$:

$$\tilde{\mathbf{i}} = \{\tilde{i}_1 \dots \tilde{i}_q\} \text{ where each } \tilde{i}_j = \sum i_k, \forall k : \mu_j - \varepsilon \leq m_k \leq \mu_j + \varepsilon, j \in \{1 \dots q\} \qquad (7.3)$$

An important parameter in this connection is the utilized accuracy of the isotopic distribution calculation. A good setting of this value depends, on the one hand, on the accuracy of the instrument used to acquire the mass spectra, but is, on the other hand, also limited by an increasing computational effort due to the comparably high complexity of the algorithm.

Since it can be taken for granted that a peptide does not only elute at a definite time point but possibly within a distinct time interval, the next step is to extract the intensities not only from the spectrum that is responsible for the peptide's identification, but also from all scans recorded in the seconds before and after the detection. In this respect, it has to be taken into consideration that the same peptide with the same charge state has sometimes been identified not only once but two or more times, for instance in its labeled and its unlabeled variant. At best, these peptides have eluted in the same time interval, at worst there is a difference of several minutes between their retention times. In such a case, firstly those two peptides with the utmost different time interval are searched for. All remaining peptides (if there are

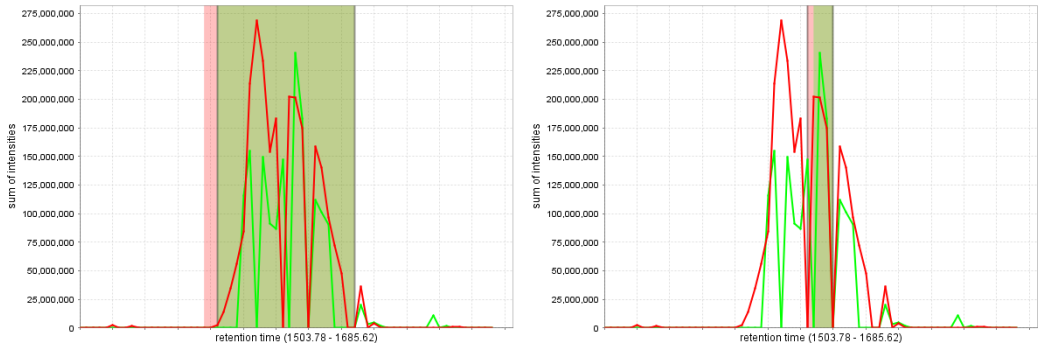


Figure 7.12 – The Figure illustrates the assets and drawbacks of different peak detection methods. While a Wavelet-based peak detection is able to cope with—admittedly extreme—signal instabilities as shown in the left picture, the right picture reveals the disadvantage of the top-down approach in this case as the peak detection is erroneous 'stuck' in a local minimum.

any) are then grouped according to their distance in time to these two peptides. Secondly, a majority rule is applied, and that peptide or group of peptides which has the lowest number of members is removed. If both groups are equally-sized, it is always the group with the highest retention time that is discarded since these have a higher probability to be inaccurate—not seldom peptides that somehow 'got stuck' in the chromatographic column elute in the last seconds of a liquid chromatography run. If the time interval spanned by the retention times of all peptides still exceeds a given threshold, the procedure is repeated.

After all intensities have been extracted and thereby the two EICs for the unlabeled and the fully labeled peptide were constructed, it is necessary to find the concrete first and last time point a peptide eluted at in the EICs, in other words, the borders of the peptides' elution peaks. A simple and straightforward approach searches for the peak's apex and follows the flanks on both sides until these ridge lines either reach the baseline or fall below a given threshold. Even though appropriate, this method has a significant weakness if deviations between the signals of the same peptide in two successive spectra occur, for instance, due to spray instabilities of the ESI ion source (Parvin et al. 2005). This can be countered by the application of a smoothing filter as has been proposed by Savitzky and Golay (1964). However, Yang et al. (2009) found an algorithm based on continuous wavelet transform having the best performance for the purpose of peak detection in chromatographic data. The algorithm of Du et al. (2006) utilizes a Mexican Hat wavelet, as it approximately describes the form of a peak. By this means, irregularities can be compensated for, as shown by the example in Figure 7.12.

Dependent on a retention time point τ this wavelet function $\psi(\tau)$ is defined by the following equation:

$$\psi(\tau) = \frac{2}{\sqrt{3\pi^{1/4}}}(1-\tau^2)e^{-\tau^2/2} \tag{7.4}$$

To apply the wavelet on an EICs, this has to be interpreted as a function of retention time, $c(\tau)$. Thus, a continuous wavelet transformation with the parameter a, which denotes a

scaling factor to shrink or stretch the width of the wavelet, in analogy to the width of a peak, can be conducted as follows:

$$CWT(\tau) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} c(t) \cdot \psi\left(\frac{t-\tau}{a}\right) dt\ t, a \in \mathbb{R} \tag{7.5}$$

To gain an optimal fit of the wavelet on the spectral data, convolution operations are conducted for a range of scales, each resulting in a list of wavelet coefficients. Leveraging the fact that only the most abundant peak at a time point close to the retention time of an identified peptide is of interest, the maximal observed coefficient can directly be used as indicator for the apex of this peak. Subsequently, the borders of the elution peak can be deduced from the corresponding scaling factor and the roots of the folded function.

On the one hand, peak detection can be conducted separately on the two EICs of the labeled as well as the unlabeled peptide. In this case, either the maximal peak width or the overlap provide both good estimates of the true peptide's elution start and end point. In this manner, non-overlapping peaks may also indicate an error in measurement or an incorrect peptide identification, and may be chosen to be omitted. On the other hand, both EICs may be combined before the peak detection, e. g. using the maximal intensity value for each time point τ. Peak detection is then performed on this merged EIC. A similar approach has been used in the tool ProRata (see section 4.3.3). Both methods have their advantages and disadvantages. Whereas the first allows to implement an additional verification, in particular if only overlapping peaks are taken into account, the second approach may give better results in case one peptide is only found with a very low abundance—the probability to correctly detect a barely existing elution peak of this peptide is likely to be increased.

Finally, the ion current ratio is estimated using the resulting elution peak borders. Various methods have been proposed for this purpose. For instance, the areas under the two curves described by the elution peaks may be set in relation to each other, or alternatively, MacCoss et al. (2003) and others (Li et al. 2003; Pan et al. 2006) proposed a linear regression approach for this purpose. Within the frame of this work, a Master and a Bachelor thesis have been conducted to evaluate different methods. Mertens (2008) successfully implemented a quantification algorithm based on the last mentioned linear regression approach, Schröder (2010) successfully devised a trapezoid-based procedure to calculate relative abundance values.

So far the best results have been achieved using the linear regression approach, especially due to its robustness against outliers as also illustrated in Figure 7.13. In QuPE's implementation a modified version of this method is employed, as instead of vertical offsets, which are commonly used in least squares fitting, perpendicular offsets are utilized to allow for uncertainties of the data points along both axes. Formally, the sections of the two chromatograms $c_1(\tau)$ and $c_2(\tau)$, attributable to the labeled and the unlabeled peptide, are plotted against each other. Aim is to fit the function $c_1(\tau) = a + b\, c_2(\tau)$ to the data (given, without loss of generality, $\|c_1\| = \|c_2\|$). In the outcome, the slope of the regression line a gives an estimate of the ratio of the abundances, as determined by minimization of the following equation:

$$r \equiv \sum_{\tau=1}^{\|c_1\|} \frac{(c_1(\tau) - (a + bc_2(\tau)))^2}{1 + b^2} \tag{7.6}$$

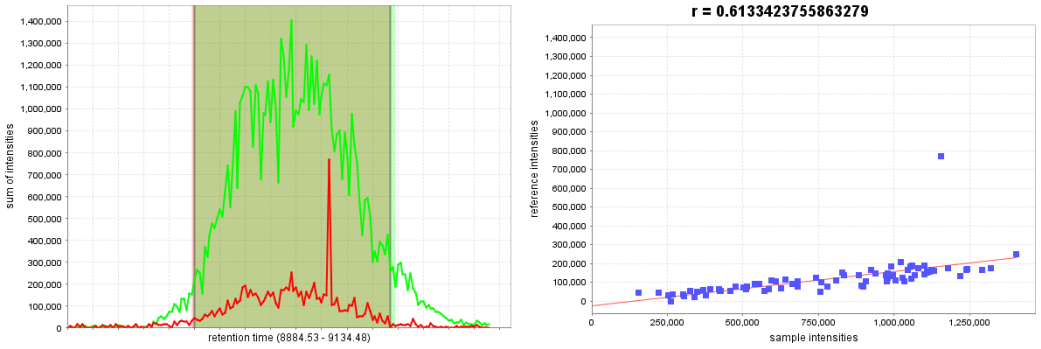


Figure 7.13 – This Figure clearly illustrates the advantages of the linear regression approach. It can be assumed that the outlier in the EIC that is drawn in red is an error in the measurement, e. g. due to spray instabilities. The picture on the right side shows the results of the linear regression analysis, and demonstrates that the outlier has only a minor (if any) influence on the calculated abundance ratio.

Calculated abundance ratios that exceed a given threshold of *r* are then stored in the QuPE database. In addition, for each ratio a signal-to-noise (S/N) value is computed and used for filtering that sets the overall peak intensity in relation to the mean signal intensity in a range before and after the detected peak borders.

Implementation of the elution peak quantification algorithm

The classes implemented for the elution peak quantification algorithm are described in Figure 7.14. Based on the theoretical isotopic distributions (see section 7.3.1 for further details), exported in form of a *PeptideIsotopologueMap*, at first, the tool implemented in the *ElutionPeakTask* groups all peptides having the same charge state according to their elution time. Co-eluting peptides are combined in one object of type *ElutionPeak*. A mapping of peptide sequences on these *ElutionPeaks* is subsequently used to extract the spectral information from the database. Therefore, both the m/z positions gained from the theoretical isotopic distribution calculation as well as the temporal information of each peptide's elution are taken into account. In the last step, the borders of the elution peak are determined using one of the available peak detection algorithms, and finally, relative abundance ratios are calculated and stored in database as an object of type *Result*.

7.3.3 Pulse chase quantification

Pulse chase experiments using metabolically incorporated stable isotopes provide a way to gain knowledge about the two components of protein turnover by determining synthesis as well as degradation rates of a protein (see section 3.3.2 for further details). This required the development of an algorithm that allows to calculate the ratio between the abundances of two differentially labeled peptides—first, a fully-labeled or unlabeled peptide, and second, a partially-labeled peptide that is synthesized either before or after a pulse chase. The approach

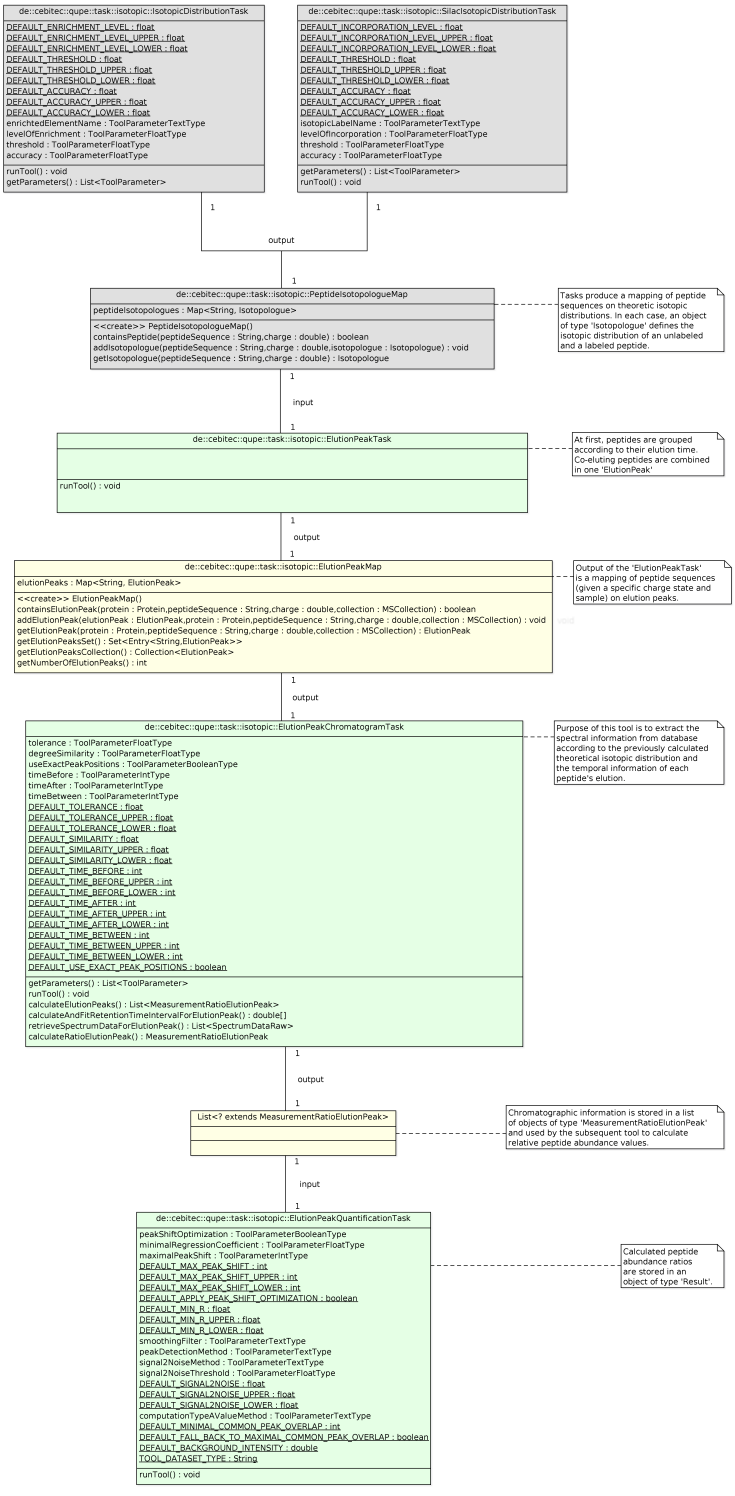


Figure 7.14 – This diagram shows the classes implemented for the elution peak quantification algorithm.

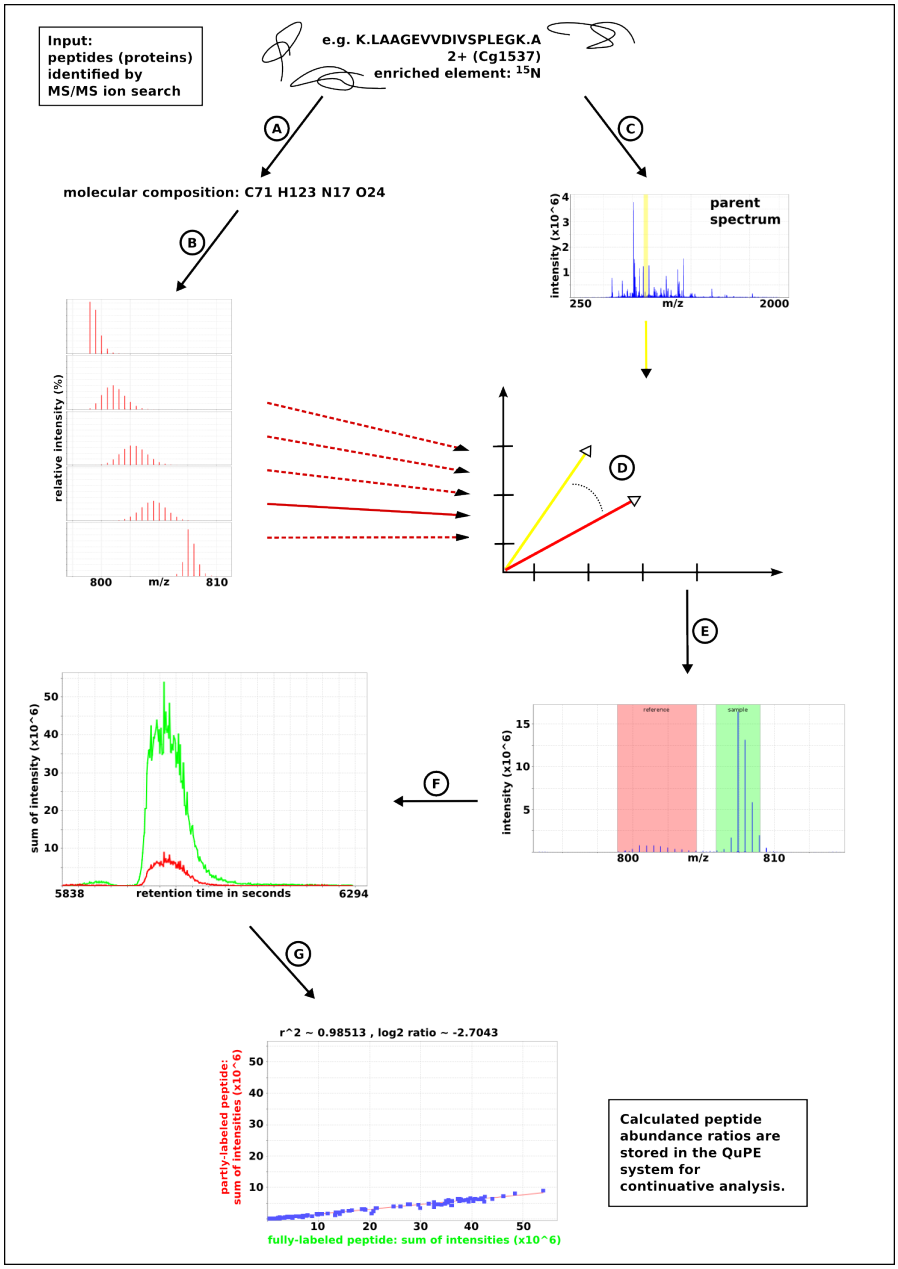


Figure 7.15 – Workflow of the algorithm allowing to quantify peptides with variable incorporation rates e. g. of nitrogen. A) The required input is a list of peptides identified by an MS/MS ion search along with information about their amino acid sequence, charge state, modifications and the associated protein accession number. B) Based on each peptides' molecular composition, theoretical isotopic distributions are computed for varying rates of ^{15}N or ^{13}C incorporation. C) The parent spectrum is retrieved and D) both, the theoretical as well as the "real" isotopic distributions are compared using the dot product. E) The best match of a theoretical distribution (e. g. at 30% enrichment) determines the m/z ranges for the extraction of ion chromatograms (EIC) for each the partly- and the fully-labeled peptide (F). Both EICs are compared by perpendicular linear regression (G), where the slope of the regression line leads to the peptide abundance ratio (e. g. -2.704).

described in the following has also been published in a journal article, in which also a comprehensive experimental setup for the analysis of protein turnover is explained in detail (Trötschel* et al. 2012).

The basic idea of the algorithm—analogous to the aforementioned elution peak quantification approach—is to extract ion chromatograms (EICs) for each of the two isotopologous peptides, and then to set these two EICs into relation. An immanent feature of a pulse chase approach that is based on metabolic labeling, e. g. using ^{15}N, is the fact that in the same spectrum a peptide is always found in a fully-unlabeled or—dependent on the applied approach—fully-labeled variant and, in addition, in a variant which is only partly-labeled at an unknown rate of enrichment. The crucial task is to determine this incorporation level, that then leads to the m/z positions of both peptides and subsequently allows the construction of EICs.

The complete workflow of the algorithm is depicted in Figure 7.15, starting as before with a list of identified peptides including their amino acid sequence, charge state and protein modifications (see Figure 7.15A). For each peptide, firstly, the expected theoretical isotopic distribution is calculated using the naturally occurring atomic weights and isotope probabilities. Secondly, further isotopic distributions are computed for a selected set of variable rates of enrichment of the isotope that has been used as label. Dependent on the applied pulse chase approach, this starts, for example, with a low incorporation of ^{15}N and ends with a distribution where almost all ^{14}N isotopes are replaced by their heavy counterpart (see Figure 7.15B). The next and crucial step is to determine the similarity between this set of theoretical isotopic distribution and the peptide's associated mass spectrum (see Figure 7.15D). From different investigated approaches, e. g. correlation-based measures, the comparatively 'simple' scalar product showed the best performance for this purpose. This also corresponds to the findings of Stein and Scott (1994), who investigated a closely related topic, namely, mass spectral library search algorithms. Given a spectrum $\mathbf{S}$ consists of p discrete peaks, each described by its m/z-value m and intensity i, formally $\mathbf{m} = \{m_1 \dots m_p\}$ and $\mathbf{i} = \{i_1 \dots i_p\}$ and, analogously, q peaks belong to a calculated theoretical distribution $\mathbf{T}$, ranging from the lowest m/z-value μ_1 to the highest m/z-value μ_q with intensities ι, a similarity is computed as

$$d_{\mathbf{S}\times\mathbf{T}} = \arccos\left(\frac{\mathbf{S}\times\mathbf{T}}{\sqrt{\mathbf{S}\times\mathbf{S}}\cdot\sqrt{\mathbf{T}\times\mathbf{T}}}\right) = \arccos\left(\frac{\tilde{\mathbf{i}}\times\iota}{\sqrt{\tilde{\mathbf{i}}\times\tilde{\mathbf{i}}}\sqrt{\iota\times\iota}}\right) \tag{7.7}$$

wherein $\tilde{\mathbf{i}}$ is derived from $\mathbf{i}$ as described above to, firstly, hold the necessary condition $\|\tilde{\mathbf{i}}\| = \|\iota\|$, and, secondly, to reduce noise, e. g. from overlapping peptides.The similarity $\mathbf{d}_{\mathbf{S}\times\mathbf{T}}$ is calculated for any theoretical isotopic distribution, with the highest value providing a clear indication of the correct rate of enrichment. In addition, the similarity is also used to verify whether a peptide identification is correct: If no theoretical isotopic distribution matches the given mass spectrum, i.e. all calculated similarities fall below a threshold, the peptide is omitted from further calculation. In this regard, it has to be added that in some cases the utilization of exact peak positions is not appropriate, e. g. if the resolution of a mass spectrometer is too low or a preprocessing algorithm was already applied to remove any noise from the spectra. In such a case, the value of ε may be chosen in such a way that in the

end all peaks of **S** are taken into account, thus, in principle, all values are combined to hold the condition $\|\tilde{\mathbf{i}}\| = \|\iota\|$.

In addition to the incorporation rate, the applied spectral matching also leads to the m/z-values and intensities of each of the two peptides, and the intensities' ratio to a measurement of relative abundance (see Figure 7.15E) following the same procedure as described for the elution peak quantification approach.

Implementation of the pulse chase quantification algorithm

The implementation of the pulse chase quantification algorithm is based to a large extent on the classes implemented for the elution peak quantification approach. As described in Figure 7.16 a tool has been added for the calculation of theoretical isotopic distributions given a range of incorporation rates. Therefore, a starting and an ending incorporation rate have to be given as well as an increment value. For the particular case that in addition to one isotope such as heavy stable nitrogen an additional isotope, for example heavy stable carbon, has been utilized to measure both protein synthesis and degradation rates, the special tool *DuplexIsotopicDistributionTask* is provided. Furthermore, the *ElutionPeakChromatogramTask* was extended to perform the task of determining the best fitting isotopic distribution.

7.4 Summary of features of the QuPE system

In the following, a short summary of the features is given that have been implemented in accordance to the requirements listed in section 5.1 and described in Figure 5.1 and which are provided by the QuPE system.

7.4.1 Data management: projects and experiments

In the first instance, QuPE provides the necessary functionality to organize and keep track of all data and meta-data relevant to a particular quantitative proteomics experiment. This includes the raw mass spectra that belong to an experiment as well as descriptions of the experimental setup and all further analysis results. Any access to the data is, firstly, secured by the CeBiTec's generalized project management system (GPMS), which has already proven its worth and functionality in hundreds of international PolyOmics projects (e. g. Neuweger et al. 2008; Dondrup et al. 2009). As explained in 7.2.1.4, a second level of application-based security utilizes access control list directives (ACLs) on selected database objects (see Figure 7.17A). The communication between a client's web browser and the QuPE server, furthermore, takes place using HTTP over Secure Sockets Layer (SSL).

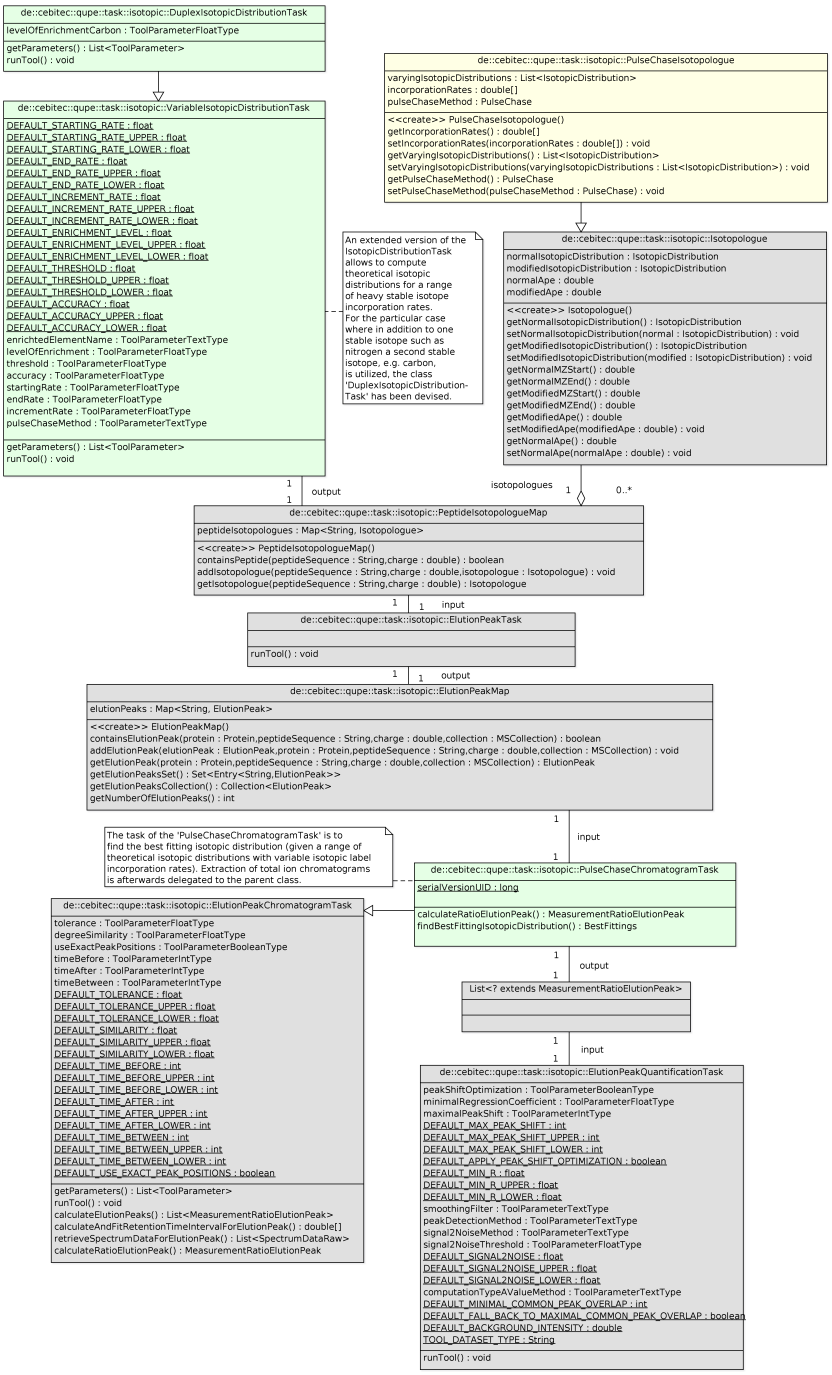


Figure 7.16 – The implementation of the pulse chase quantification algorithm closely follows the implementation of the elution peak quantification algorithm. Only a few classes have been added namely for the calculation of theoretical isotopic distributions for a range of variable incorporation rates and a tool to determine the best fitting incorporation rate of a stable isotope.

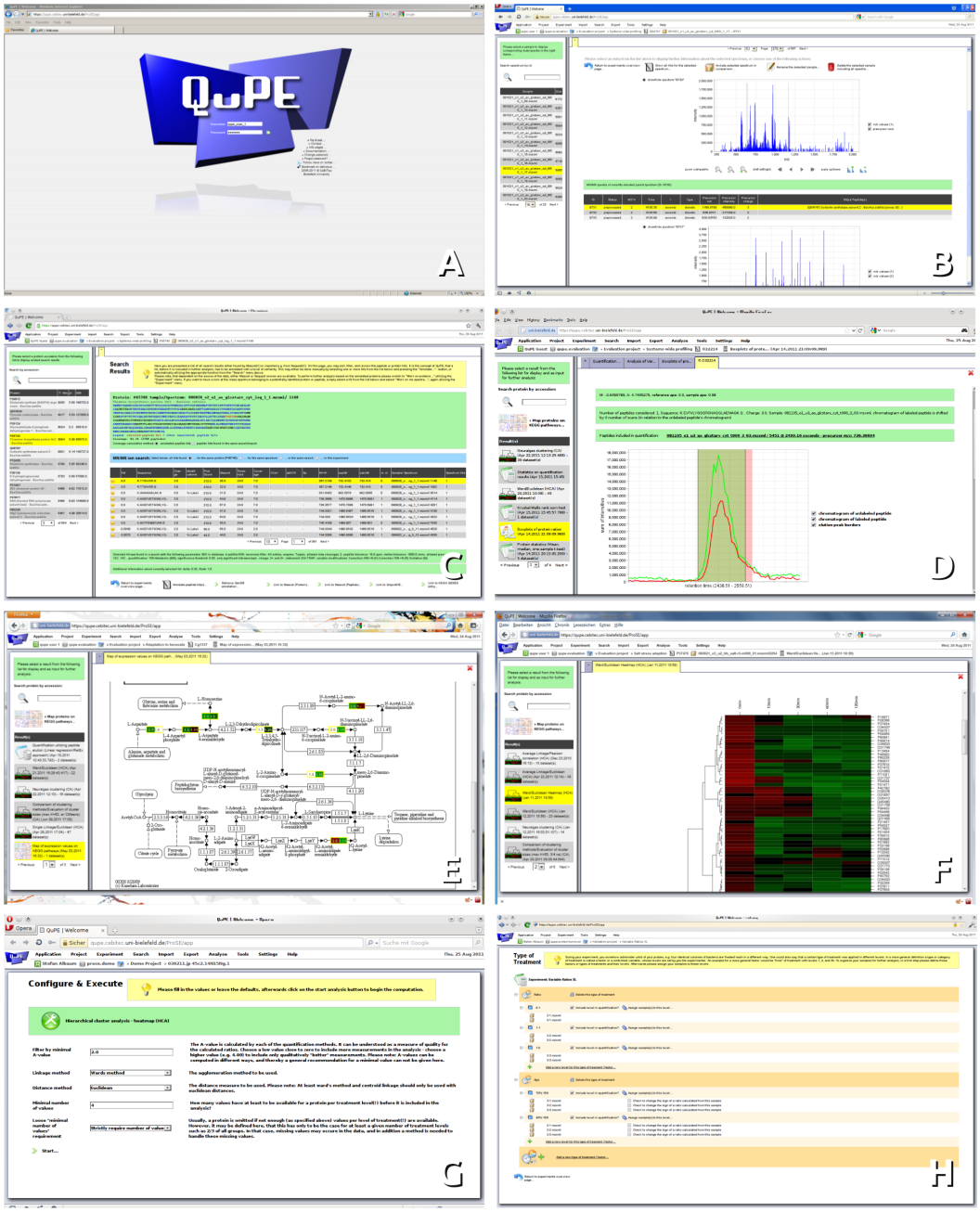


Figure 7.17 – A selection of screenshots showing QuPE's graphical user interface: A) login screen, B) view to browse and import mass spectra, C) visualization of database search results for protein identification, D) details about a protein quantification result, E) projection of abundance ratios on a metabolic pathway, F) view of analysis results, here, a heatmap as resulting from a hierarchical cluster analysis, G) configuration and start page of a calculation task, H) view to describe the treatment and grouping of samples.

Import and preprocessing of mass spectrometry raw data

QuPE supports the import of mass spectra data in the open source formats mzXML (Pedrioli et al. 2004), mzData (Orchard et al. 2004), the text-based Mascot™ generic format, and in addition a proprietary format of the company Bruker Daltonics™, which is used by particular MALDI-TOF instruments. Thereby, most of the mass spectrometers available on the market are compatible with QuPE as vendors, in general, provide appropriate data export and conversion tools in at least one of the aforementioned open source formats. This should, for example, include the majority of mass spectrometers from the companies Bruker Daltonics™, Thermo Scientific™and Agilent Technologies™.

Comprehensive capabilities are available to browse and search imported mass spectra (see Figure 7.17B). An interactive visualization allows to zoom into a spectrum, e. g. to closely investigate specific peaks or to overlay and compare two or more different mass spectra to each other. Before a database search to identify proteins can take place, it is often necessary to preprocess the imported raw data. Therefore, a number of methods are provided, which utilize QuPE's job and tools framework. For the task of peak detection (cf. section 3.2.1.1), the R-package MassSpecWavelet (Du et al. 2006) has been integrated. In addition, tools have been implemented to inspect and approve imported mass spectra according to the observance of certain criteria such as a minimal total ion current or number of peaks. A further use case concerns the filtering of individual peaks, e. g. if these have an intensity below a certain threshold and, hence, are potentially noise peaks. The application of these filtering methods oftentimes allows to greatly reduce the search space for peptide mass fingerprinting and MS/MS ion search.

Structuring of samples according to the experimental model

A meaningful analysis demands that samples, which have been measured under the same conditions, are grouped together and accordingly compared to each other. QuPE provides a mapping tool to describe this model of an experiment (see Figure 7.17H). After the experimental conditions have been defined, which are for instance different temperature levels such as 30 and 40°C, or concentrations of a substance, each imported sample can interactively be assigned to these conditions. A subsequent analysis is then based on the, in this way, described experimental model. At any time, it is possible to modify the created model and to temporarily exclude or include certain conditions, e. g. to investigate only a part of all samples according to a specific treatment. To support a user in finding an appropriate terminology for each condition the ontology lookup service of the EBI may be queried from the web interface (Côté et al. 2006; Martens et al. 2005).

7.4.2 Protein identification: peptide mass fingerprinting and MS/MS ion search

QuPE allows the import of search results, e. g. in form of a DTASelect-filter file, and has an integrated Mascot™ search engine to perform peptide mass fingerprinting and MS/MS ion search. As the local Mascot™ server installation is not directly accessible outside of the CeBiTec network, the integration is based on a self-written SOAP-based webservice as well as a kind of HTTP-proxy for communication with the server and the initial configuration of a search. Next, HTTP-Post/Get are used internally for the initiation and conduction of a search. Searches of the same set of mass spectra may be batch processed, for example, by means of the definition of ranges for peptide tolerance values or by querying several databases at once.

Automatic evaluation of database search results

To ensure that further analyses rest on a solid ground of verified peptide or protein identifications, it is necessary to assess the reported hits produced by database search tools. In QuPE, this can be based upon the calculation of false discovery rates (FDR) as suggested by Reidegeld et al. (2008). The precondition for this is that a concatenated decoy database (Peng et al. 2003; Elias and Gygi 2007) has been employed. In the first instance all peptide or protein hits that were either imported or reported by the integrated Mascot™ search engine are stored in a database. Based on user-defined parameters such as the exclusion of specific charge states, a certain FDR-threshold, or, alternatively, a minimal score value, reported hits are filtered to gain the set of proteins and peptides that will be included in further analyses.

7.4.3 Protein quantification

QuPE provides several quantification algorithms, which have been implemented and evaluated within the scope of this work. The list of quantitative methods which are addressed by these algorithms includes (see section 3.3.1 for further details regarding the different labeling strategies):

- The 'sum quantification' approach for ^{15}N and ^{13}C metabolically and SILAC-labeled data (see section 7.3.1). A special version of the algorithm has been designed for the quantification of MALDI-TOF mass spectra.
- The 'elution peak quantification' approach for ^{15}N and ^{13}C metabolically and SILAC-labeled data (see section 7.3.2).
- The 'pulse chase quantification' approach for variable ^{15}N or ^{13}C metabolically-labeled data (see section 7.3.3).
- The 'dual chase quantification' approach (a variant of the pulse chase approach) for both ^{15}N and ^{13}C metabolic labeling as used in Trötschel* et al. (2012).

- A spectral count quantification approach

In addition, QuPE facilitates the import of results as produced by two of the most commonly used quantification tools ProRata (Pan et al. 2006) as well as Census (Park et al. 2008). Figure 7.17D depicts an exemplary screenshot of the graphical user interface of QuPE showing an extracted ion chromatogram as calculated by the elution peak quantification approach.

7.4.4 Statistical analysis, data mining, and visualization

QuPE supports the complete range of analysis functions as introduced in chapter 6. This includes several methods to detect differentially regulated proteins and to verify these findings:

- Measures of descriptive statistics: mean, standard deviation, median (the implementation makes use of the R-package 'stats'; R Development Core Team 2011)
- One-sample t-test (R Development Core Team 2011)
- Analysis of variance (R Development Core Team 2011)
- Kruskal-Wallis rank sum test (R Development Core Team 2011)
- Shapiro-Wilk test of normality (R Development Core Team 2011)
- Fligner-Killeen test of homogeneity of variance (R Development Core Team 2011)
- Tools to visualize calculated abundance values and statistical measures including M/A (Ratio vs. intensity) plots, Box-and-Whisker plots, histograms, and scatter plots

Targeting the identification of co-regulated proteins a number of tools have been implemented to perform, *inter alia*, the task of cluster analysis (see Figures 7.17F and G):

- Hierarchical cluster algorithms (HCA) using Ward-, Complete-, Average-, Single-, Median-, or Centroid-linkage and either Euclidean or Correlation-based distance measures (implementation makes use of the R-packages 'stats' and 'amap'; Lucas and Jasson 2006; R Development Core Team 2011)
- Partitioning cluster algorithms including K-means and Neuralgas clustering (based on the R-packages 'stats' and 'cclust'; Dimitriadou 2009; R Development Core Team 2011)
- Fuzzy C-means clustering as a probabilistic method (Dimitriadou et al. 2011)
- Principal component analysis (R Development Core Team 2011)
- Cluster validity measures including the indexes Calinski-Harabasz, Krzanowski-Lai, Index-I, and Figure of Merit (Broberg 2012; Hennig 2010; Walesiak and Dudek 2012)
- Tools to visualize cluster results, e. g. in form of a heatmap or as a cluster profile plot

To extend the knowledge about identified proteins, information from external resources including Uniprot (The UniProt Consortium 2008), KEGG (Kanehisa and Goto 2000), and the NCBI entrez database (Schuler et al. 1996) can be integrated. This comprises COG or KOG (Tatusov et al. 2003) classes and numbers, EC numbers and pathway information, or GO terms (Ashburner et al. 2000). If protein identifiers have been derived from the GenDB annotation system (Meyer et al. 2003), a mapping onto regions via BRIDGE (Goesmann et al. 2003) is also available. This information can then be used, for example, to calculate the distribution of COG categories. Another function, which is integrated in QuPE, allows to map identified proteins and their calculated abundance ratios on KEGG pathways (see Figure 7.17E).

Chapter 8

Performance and accuracy of protein quantification

One of the central objectives of this work was the improvement of the accuracy and precision, in terms of reproducibility, of protein quantification methods. As described in the previous chapter in 7.3, different approaches have been undertaken—from the rather simple 'sum quantification' to an approach that utilizes the elution time of peptides and that is able to cope with variable isotope incorporation rates. To demonstrate the applicability and validity of the developed algorithms a comprehensive evaluation was conducted based on benchmark datasets made available by workgroups of the Ruhr-University Bochum and the University of Greifswald.

8.1 Protein mixtures – fully labeled vs. unlabeled

The University of Greifswald, Institute of Microbiology, provided five datasets containing mixtures of fully-labeled to unlabeled proteins in distinct ratios, each consisting of 14 individual runs. Therefore, *Bacillus subtilis* was grown on normal medium as well as media enriched to an extent of 98% with heavy stable isotopes of nitrogen. After protein extraction and digestion using the enzyme trypsin, samples were mixed in ratios of 1:1, 1:2, 2:1, 1:10, and 10:1 and analyzed by MudPIT (Wolters et al. 2001) coupled to a Thermo™ LTQ mass spectrometer. For further processing the resulting mass spectrometry data files were transformed from a proprietary vendor-specific format into the open source format 'mzXML' (see 4.1.3) using

the tool 'ReAdW' (Keller et al. 2002; Nesvizhskii et al. 2003). These data files were then imported into the QuPE system. Afterwards, MS/MS ion search was performed using the Mascot™ search engine in a decoy database specific for *Bacillus subtilis*. Search parameters were set to two allowed missed cleavage sides, a peptide tolerance of 10 ppm and an MS/MS tolerance of 1000 mmu. Only those hits were kept that were reported significant ($\alpha \leq 0.05$) by Mascot as well as below an equally set false discovery rate. Oxidation of methionine was considered as a variable modification. In addition to naturally-occurring nitrogen abundances, the database search was configured to also take into account a replacement of all ^{14}N isotopes by the heavy stable form ^{15}N. Following the recommendations of Zhang et al. (2009), a one Dalton shift of the fully-labeled peptide due to incomplete incorporation of the heavy isotope was set up for arginine as well as lysine as an additional variable modification.

8.1.1 Reference measurements

In order to gain a reference standard for comparison the tool ProRata (Pan et al. 2006) was utilized to calculate relative abundance ratios for the five protein mixtures provided by the University of Greifswald. In all cases the default parameter settings of the software were used, i. e. for the extraction of ion chromatograms a time interval of two minutes before and after the scan, which yielded the peptide identification, an m/z-error of 0.5, and an isotopic envelope cutoff of 0.1; a peak shift was not allowed (see Appendix for further details). The results are listed in Table 8.1. It becomes obvious that, although tendencies in the data are correctly estimated, the expected target values ($< M >$) are more or less clearly missed. Furthermore, calculated mean abundance values ($\bar{M}$) are characterized by a surprisingly high standard deviation, e. g. of $\sigma = 2.12$ in case of the uniform ^{14}N to ^{15}N mixture. In this connection, it has however to be noted that this does not necessarily indicate an error in the calculation as a bias might already have been introduced during sample preparation and measurement.

Secondly, the tool Census (Park et al. 2008) was applied on all benchmark datasets. In analogy to ProRata, the quantification method was configured with reference to the default values of the software. The calculated ratios were, afterwards, exported using the graphical user interface of Census (see Appendix for further details regarding the parameters), and then imported into QuPE. The achieved results, presented in Table 8.2, are however not satisfactory. In particular, the two mixtures having ratios of 2:1 and 10:1 could not be quantified accurately, and in relation to each other the calculated mean abundance values—the median abundance values are slightly more convincing—do not reflect the true ratios of the samples. A reason for these discrepancies might be an unfavorable configuration of the software tool. Due to comparably long running times[1], it was refrained from any subsequent usage of the software.

[1]Running time for Census: up to 10 days for each dataset, granted on the proviso that this was computed on Microsoft™ Windows XP operated on decent hardware (4x Quad-Core AMD Opteron™ 8356, 65GB RAM) but in Oracle™ VirtualBox (one virtual CPU, 2GB RAM assigned); for comparison: the tool ProRata has been used on the same system with processing times in the range of hours.

Table 8.1 – This table summarizes the quantification results achieved with the tool ProRata (Pan et al. 2006) on five benchmark datasets provided by the University of Greifswald. Due to different growth media, extracted protein samples of *Bacillus subtilis* were either fully labeled (98% ^{15}N) or completely unlabeled, i. e. with natural occurring nitrogen isotope abundances. Each two samples were mixed in distinct ratios of 1:1, 1:2, 2:1, 1:10 and 10:1 and analyzed using LC-MS/MS. $< M >$ denotes the expected mean value based on the given ratio. The column entitled $\bar{M}$ shows the mean value of all calculated peptide abundance ratios together with their standard deviation σ. The median is given in column $\tilde{M}$, while the sixth column contains the 95%-confidence interval. Finally, the last two columns denote the overall number of calculated peptide abundance ratios (#peptides) and the number of proteins these peptides account for (#proteins).

		ProRata					
^{14}N/^{15}N Ratio	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
1:1	0	**-0.31**	2.12	-0.42	-4.00;5.07	3320	352
1:2	-1	**-2.37**	1.89	-2.31	-5.98;2.17	5348	513
2:1	1	**1.63**	1.64	1.44	-1.48;4.92	4916	425
10:1	3.32	**3.78**	1.74	3.66	0.82;7.37	6648	529
1:10	-3.32	**-4.07**	2.32	-4.26	-8.06;2.22	7443	613

Table 8.2 – This table summarizes the quantification results achieved with the tool Census (Park et al. 2008) on five benchmark datasets provided by the University of Greifswald. See Table 8.1 for a description of the column headers.

		Census					
^{14}N/^{15}N Ratio	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
1:1	0	**-0.68**	1.36	-0.68	-4.32;1.89	2176	389
1:2	-1	**-1.54**	1.55	-1.56	-4.64;2.17	2975	496
2:1	1	**-0.51**	1.58	-0.15	-5.06;1.42	3616	409
10:1	3.32	**0.69**	2.48	1.73	-5.06;3.79	1791	345
1:10	-3.32	**-2.64**	2.01	-3.06	-5.64;2.16	4916	630

Table 8.3 – Spectral counting is a comparably simple but nevertheless powerful approach to gain abundance values of proteins. This table summarizes the Mascot™ search results for all five benchmark datasets provided by the University of Greifswald. In the columns titled "#unlabeled" and "#labeled" the overall number of peptides are listed that were found with or without enrichment of the heavy stable isotope of nitrogen. Given the simplicity of the approach, the calculated ratios (M) reflect the true ratios of the sample mixtures in a remarkably accurate way.

		Spectral counting approach				
^{14}N/^{15}N Ratio	$< M >$	#proteins	#peptides	#unlabeled	#labeled	M
1:1	0	655	20100	9699	10401	-0.10
1:2	-1	668	16986	5162	11824	-1.20
2:1	1	610	20319	13318	7001	0.93
10:1	3.322	642	17277	15916	1361	3.55
1:10	-3.322	765	16045	813	15232	-4.23

Table 8.4 – This table summarizes the quantification results achieved on all five benchmark datasets provided by the University of Greifswald with the simple and straightforward sum quantification approach. The accuracy of the isotopic distribution calculation and the tolerance value ε were both set to 0.1 m/z. See Table 8.1 for a detailed description of the column headers.

		Sum quantification approach					
^{14}N/^{15}N Ratio	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
1:1	0	**-0.3**	0.67	-0.31	-1.56;0.94	20088	655
1:2	-1	**-1.16**	0.83	-1.24	-2.57;0.79	16938	668
2:1	1	**0.44**	0.7	0.54	-1.17;1.51	20309	610
10:1	3.32	**1.94**	1.28	2.23	-1.22;3.77	17174	642
1:10	-3.32	**-2.89**	1.78	-3.21	-5.59;1.48	15548	765

To further investigate the five benchmark datasets, a variation of the spectral counting approach was applied. As resulting from the Mascot™ database search, the number of identified peptides was counted that were found unlabeled, on the one hand, and that were found fully labeled with the heavy stable nitrogen isotope ^{15}N, on the other hand. Following the idea of spectral counting (see section 3.3.4), the quotient of these counts then leads to an estimation of the protein abundances in the sample (M, logarithmized to base 2). Table 8.3 shows the results of this comparison, revealing a remarkably high agreement between the expected ratios and the calculated values, in particular, compared to the results of ProRata. However, as already mentioned before, broken down to the protein level spectral counting is known to perform rather poorly as soon as individual counts are low (Hendrickson et al. 2006). Nevertheless, these values confirm that the benchmark datasets are applicable to evaluate the accuracy of implemented quantification algorithms, and moreover, 'set the bar high' for any other approach.

8.1.2 Accuracy of the sum quantification

At first, the simple and straightforward sum quantification approach as described in section 7.3.1 was applied on the five benchmark datasets. Therefore, the algorithm was configured as follows: the accuracy for the isotopic distribution calculation was set to 0.1 m/z, which was also used for the tolerance value ε. As this algorithm does not implement any additional filter regarding, for example, the agreement between a theoretical isotopic distribution and the observed distribution of peaks in the recorded mass spectra, the number of quantified proteins and peptides is comparably high, and conforms in great measure with the number of identified proteins and peptides. At the same time, however, it should be noted that thereby an increased measurement error must be expected, in particular, due to deficient or noisy input data, which is not excluded from the calculation.

The quantification results are listed in Table 8.4, and show that in all cases tendencies in regulation are clearly differentiable and correctly estimated, although the expected mean values ($< M >$) are not exactly matched. In comparison to ProRata, which apparently overestimated

Table 8.5 – An evaluation was performed to investigate the impact of different parameters on the quantification results achievable with the elution peak quantification algorithm. Therefore, the 1:1 sample was analyzed in detail. The following settings were used in accordance with the characteristics of this experiment: accuracy of the isotopic distribution calculation: 0.1 m/z, tolerance value ε = 0.01 Da, investigated retention time interval for each peptide: 60 seconds before and after the identifying mass spectrum, CWT-based peak detection. The results of varying settings of the isotopic similarity $d_{S\times T}$ are shown in this table. For a detailed description of the column headers please refer to Table 8.1.

				Elution peak quantification - isotopic similarity					
$d_{S\times T}$	r	S/N	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
0.95	0.6	3.0	0	**-0.23**	0.57	-0.24	-1.33;0.80	3138	491
0.9	0.6	3.0	0	**-0.24**	0.57	-0.23	-1.34;0.80	3144	492
0.8	0.6	3.0	0	**-0.23**	0.57	-0.24	-1.34;0.80	3145	492

the 'true' ratio of the data, the sum quantification approach seems to underestimate all protein abundances. This can be clearly seen in view of the 2:1 data ($< M >= 1$): in this case, ProRata displayed a mean abundance ratio of 1.63 in contrast to a value of 0.44, which was achieved with this algorithm. The 'simple' approach, however, shows one indisputable advantage as the distributions of all calculated ratios are generally revealing a lower standard deviation, e. g. of σ = 0.67 for the 1:1 sample vs. σ = 2.12 for ProRata, a fact that is also indicated by the 95% confidence interval: sum quantification: $[-1.56;\ 0.94]$, ProRata: $[-4.00;\ 5.07]$.

8.1.3 Accuracy of the elution peak quantification

Regarding the application of the elution peak quantification approach, it must be considered that the algorithm can be configured in different ways. In a nutshell, there are two opposing effects that have to be taken into account: on the one hand, the attainable number of quantified proteins, and on the other hand, the accuracy and quality of the calculated ratios. To evaluate

Table 8.6 – This table demonstrates the impact on the quantification results regarding the 1:1 sample, if the regression coefficient r is used for filtering. For this purpose, the isotopic similarity $d_{S\times T}$ is set to a fixed value of 0.9, and the signal-to-noise threshold S/N to 2.0. For a detailed description of the column headers please refer to Table 8.1.

				Elution peak quantification - regression coefficient					
$d_{S\times T}$	r	S/N	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
0.9	**0.9**	2.0	0	**-0.22**	0.45	-0.24	-0.95;0.56	3457	501
0.9	**0.8**	2.0	0	**-0.23**	0.49	-0.24	-1.22;0.66	4077	549
0.9	**0.7**	2.0	0	**-0.23**	0.56	-0.24	-1.38;0.82	4444	568
0.9	**0.6**	2.0	0	**-0.23**	0.60	-0.24	-1.47;0.94	4694	586
0.9	**0.5**	2.0	0	**-0.24**	0.66	-0.24	-1.57;1.07	4870	592
0.9	**0.4**	2.0	0	**-0.24**	0.70	-0.25	-1.71;1.17	5008	601

Table 8.7 – This table shows the impact on the quantification results regarding the 1:1 sample, if the signal-to-noise threshold S/N is used for filtering. In this case, the isotopic similarity $d_{\mathbf{S\times T}}$ is set to a fixed value of 0.9, and the regression coefficient r to 0.6. For a detailed description of the column headers please refer to Table 8.1.

				Elution peak quantification - signal-to-noise					
$d_{\mathbf{S\times T}}$	r	S/N	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
0.9	0.4	4.0	0	**-0.26**	0.61	-0.25	-1.49;0.80	1758	366
0.9	0.4	3.0	0	**-0.25**	0.67	-0.25	-1.59;1.04	3302	505
0.9	0.4	2.0	0	**-0.24**	0.70	-0.25	-1.71;1.17	5008	601

the impact of different settings on the quantification results, at first, the 1:1 sample was investigated in detail for a selected set of parameters. Adapted to the characteristics of the experiment, the accuracy of the isotopic distribution calculation was set to 0.1 m/z, and a tolerance value ε of 0.01 Da was configured. While for each peptide a retention time of 60 seconds before and after the identifying mass spectrum was analyzed, the concrete detection of the elution peak refers to the CWT-based approach.

Different settings of the isotopic similarity $d_{\mathbf{S\times T}}$, the regression coefficient r, and the signal-to-noise threshold S/N were successively evaluated. The impact of each of these parameters is summarized in the Tables 8.5, 8.6, and 8.7. In addition, a comprehensive parameter comparison can be found in the appendix in section B.2.1. It can be observed that the isotopic distribution similarity has almost no influence on the quantification result, neither on the number of quantified peptides, nor on the accuracy of the ratio (Table 8.5). This presumably indicates that—at least in this experiment—the influence of any disturbance variables on the mass spectrometry signal was comparably low. In addition, it may be assumed that the peptide identifications were mostly accurate and contained only a small number of false

Table 8.8 – Using the following parameters, the application of the elution peak quantification approach on the five benchmark datasets provided by the University of Greifswald yielded the results shown in this table. The configuration was as follows: accuracy of the isotopic distribution calculation: 0.1 m/z, tolerance value ε = 0.01 Da, investigated retention time interval for each peptide: 60 seconds before and after the identifying mass spectrum, CWT-based peak detection, isotopic similarity $d_{\mathbf{S\times T}} > 0.9$, regression coefficient $r > 0.6$, signal-to-noise threshold $S/N > 3.0$. For a detailed description of the column headers please refer to Table 8.1.

		Elution peak quantification					
^{14}**N**/^{15}**N Ratio**	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
1:1	0	**-0.23**	0.57	-0.24	-1.34;0.8	3144	492
1:2	-1	**-1.26**	0.66	-1.29	-2.3;0.17	2993	485
2:1	1	**0.68**	0.57	0.72	-0.65;1.7	3714	490
10:1	3.32	**2.83**	0.9	2.93	0.4;3.94	2150	393
1:10	-3.32	**-3.77**	1.41	-4.02	-5.50;0.61	1829	407

positive hits. Filtering by the regression coefficient[2] revealed in contrast a major effect on the quantification results (Table 8.6). Figuratively, this value specifies the degree of similarity between the form of the elution peaks of the light and the heavy peptide, and thus, allows to filter out measurements in which one peptide variant is missing. On the other hand, noise in the mass spectra, even if it occurs only in a few time points, may result in imperfect peak matches and therefore lead to the incorrect rejection of a calculated ratio. Increasing the filter threshold from $r > 0.4$ to $r > 0.9$ yields (independent of the S/N value) approximately 30% less peptides, which then accounts for a decrease of about 20% in the number of quantified proteins. On the opposite, however, the standard deviation is decreasing by approximately 33%.

A similar effect can be observed if the S/N value is taken into account for filtering (Table 8.7). Here, an increase of the S/N value from 2.0 to 4.0 allows to further improve the accuracy of the results: dependent on other parameter settings, in particular of the r threshold (see section B.2.1 for further details) the standard deviation can be decreased by up to 18%. However, at the same time, the number of quantified peptides is more than halved, and in the end, about 40% less proteins can be quantified.

Taking together the results of the parameter evaluation, the elution peak quantification algorithm was applied on all five benchmark datasets provided by the University of Greifswald. In this case, the isotopic similarity $d_{\mathbf{S}\times\mathbf{T}}$ was set to 0.9, a regression coefficient of at least $r > 0.6$ was demanded, and the signal-to-noise threshold was set to $S/N > 3.0$. The results are displayed in Table 8.8, and reveal that taking the elution time of each peptide into account allows to significantly improve the accuracy of protein quantification. Especially in comparison to the sum quantification approach, the variance of the calculated abundance ratios could be further decreased.

A closer look at the calculated abundance ratios suggests that all measurements are affected by a small but distinct deviation, which resulted in a shift of approximately $M = -0.25$. This deviation might have been introduced during the experiment, possibly during the preparation of samples and, moreover, emphasizes the necessity to include a normalization step in the analysis of quantitative proteomics data that allows to compensate for this error in measurement.

8.2 Protein mixtures – unlabeled vs. partially labeled

In the frame of this work, the idea of the elution peak quantification algorithm was extended to allow the comparison of the abundances of two differentially labeled peptides, i. e. a partially-labeled peptide and its fully-labeled or fully-unlabeled counterpart. This algorithm has proven its applicability in pulse chase experiments but also for the quantification of

[2] Strictly speaking, due to the perpendicular regression, it is not the regression coefficient but the correlation coefficient that is used for filtering by this algorithm.

Table 8.9 – This table shows the results of the application of the pulse chase quantification on six datasets provided by colleagues at the University of Bochum. In compliance with all previous tables $< M >$ denotes the expected mean value based on the given ratio, $\bar{M}$ the mean value of all calculated peptide abundance ratios together with their standard deviation σ, $\tilde{M}$ the median of all ratios, and $\bar{M} \pm 0.95$ the 95%-confidence interval. The last two columns denote the overall number of calculated peptide abundance ratios (#peptides) and the number of proteins these peptides account for (#proteins). In addition, the column entitled $< Ape >$ contains the expected incorporation rate of ^{15}N, $\bar{Ape}$ the mean incorporation rate averaged over all quantified proteins with the corresponding standard deviation σ, $\tilde{Ape}$ the median incorporation rate, and $\bar{Ape} \pm 0.95$ the 95%-confidence interval of all estimated rates. Further configuration details of the algorithm are described in the text.

		Pulse chase quantification						
^{14}N/^{15}N Ratio	**^{15}N Incorporation**	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
1:1	45%	0	**0.46**	0.51	0.41	-0.29;1.36	402	160
1:6	45%	-2.585	**-2.26**	0.39	-2.28	-2.90;-1.30	120	54
6:1	45%	2.585	**3.12**	0.80	3.22	0.79;4.24	239	117
1:1	70%	0	**0.66**	0.60	0.65	-0.48;1.71	459	191
1:6	70%	-2.585	**-2.21**	0.54	-2.21	-3.29;-1.14	154	73
6:1	70%	2.585	**3.22**	0.96	3.40	1.44;4.28	235	122
		$< Ape >$	$\bar{Ape}$	σ	$\tilde{Ape}$	$\bar{Ape} \pm 0.95$		
1:1	45%	0.45	**0.45**	0.07	0.44	0.4;0.56		
1:6	45%	0.45	**0.44**	0.02	0.44	0.4;0.48		
6:1	45%	0.45	**0.49**	0.13	0.44	0.4;1.0		
1:1	70%	0.70	**0.69**	0.05	0.68	0.62;0.84		
1:6	70%	0.70	**0.68**	0.02	0.68	0.64;0.72		
6:1	70%	0.70	**0.70**	0.1	0.68	0.43;0.9		

protein samples extracted from cells or organisms, in which the full incorporation of a stable isotope label is hardly achievable, for example, with regard to higher eukaryotes.

The evaluation of the functionality of the pulse chase quantification algorithm required to investigate not only whether sample mixtures can be accurately quantified but also in how far incorporation rates can be correctly estimated. Trötschel and colleagues at the University of Bochum provided benchmark datasets to perform this evaluation. They prepared samples of *Corynebacterium glutamicum* with different incorporation rates of stable nitrogen isotopes (see Trötschel* et al. 2012 for further details). In total, six datasets were created, in which samples were mixed in ratios of 1:1, 1:6, and 6:1 combining each one sample with natural abundances of nitrogen isotopes and one sample having been labeled with either 45% or 70% ^{15}N. The raw mass spectra were imported into QuPE together with the corresponding protein identifications, which had been generated using Thermo™'s software ProteomeDiscoverer and the Sequest™ algorithm.

8.2.1 Accuracy of the pulse chase quantification

The application of the pulse chase quantification algorithm on the six benchmark datasets led to the results shown in Table 8.9. The algorithm was configured with the following parameters: accuracy of the isotopic distribution calculation: 0.1 m/z, tolerance value $\varepsilon = 0.01$ Da, investigated retention time interval for each peptide: 60 seconds before and after the identifying mass spectrum, CWT-based peak detection, isotopic similarity $d_{\mathbf{S}\times\mathbf{T}} > 0.9$, regression coefficient $r > 0.6$, signal-to-noise threshold $S/N > 3.0$. The potential incorporation rate to be determined by the algorithm for each protein was restricted to a value of 0.4, at the lower bound, and 0.98 at the upper bound.

It can be observed that for all six datasets the calculated incorporation rates match the employed enrichments of ^{15}N to a high degree. Despite a small but systematic bias of about $M = 0.5$, which has presumably been introduced during sample preparation, the calculated abundance values adequately reflect the true ratios of the data and verify the applicability of the implemented procedure to gain valid abundance ratios of a partially-labeled peptide in relation to an – in this case – fully unlabeled peptide.

8.3 Protein quantification: final considerations

The methods described in this chapter targeted the quantification of metabolically stable-isotope labeled proteins. It could firstly be shown that the provided implementations allow to gain accurate and reliable quantification results, and constitute an improvement in terms of the quality of achievable results. Furthermore, an algorithm was introduced that has proven successful in the quantification of proteins having an *a priori* unknown and variable number of stable isotopes incorporated. Although not demonstrated in this performance evaluation, quantification methods implemented within the frame of this work were also successfully applied to proteins labeled with the SILAC-approach (see section 3.3.1.2), which were yielded not only using an ESI mass spectrometer but also by a MALDI-TOF instrument.

Chapter 9

A workflow for the analysis of quantitative proteomics data

Based on the QuPE system and its flexible and extensible tool and job concept, a workflow was devised and implemented allowing the comprehensive analysis of quantitative proteomics data and offering experimenters the possibility to reveal proteins that play a key role in the biological processes under investigation. The derivation of this approach is explained in detail in Albaum et al. (2011b) by means of three case studies on quantitative proteomics datasets provided by Hahne et al. (2010), Otto et al. (2010), and Haußmann et al. (2009).

9.1 Case studies

9.1.1 Experimental setups

The first experiment, in the following designated experiment A, constitutes a study on the wildtype strain of *Bacillus subtilis* and its adaption to salt stress. Hahne et al. (2010) treated growing cells with unnaturally high concentrations of NaCl and investigated the process immediately before the stress, and 10, 30, 60, and 120 minutes afterwards. A second experiment (B) was conducted by Otto et al. (2010) and targeted the soil bacterium *Bacillus subtilis* again. In this experiment the effects of glucose starvation were investigated not only in the proteome, but also the transcriptome, and the metabolome. Time series data was collected during the bacteria's growth, in the exponential phase, at the transition to stationary phase,

and 30, 60, and 120 minutes afterwards. The experimental procedure in terms of sample preparation and labeling was similar in both experiments with three biological replicates being investigated in each of them. These samples were grown in normal medium and in a medium in which ammonium sulphate and L-tryptophan were replaced by ^{15}N-labeled variants. After harvesting, the samples were mixed in equal amounts and measured on an LTQ Orbitrap XL (Thermo Scientific™) coupled to a nanoAcquity UPLC (Waters™). In contrast to the original experiment of Hahne et al., in which different cell fractions were analyzed, this case study only takes the membrane fraction of the proteome samples into account. This fraction, however, also includes a large amount of cytosolic proteins (> 70%). In total, 60 runs were performed on a mass spectrometer for experiment A resulting in an equal amount of raw data files. The course of action being taken was similar in experiment B, in which only the cytosolic fraction was investigated. This was, however, impressive in itself as it consists of 292 individual data files.

The third experiment (C) has been specifically selected to evaluate the workflow on a comparatively smaller dataset. The original experiment profiled the physiological adaption of *Corynebacterium glutamicum* to the two carbon sources benzoate and glucose. Therefore, a comprehensive MudPIT experiment was performed on three biological replicates and different cell fractions. Yet in this case study, only one replicate of the so called predigest fraction was used, summing up to 22 LC-MS/MS runs (LTQ Orbitrap, Thermo Scientific™).

9.1.2 Protein identification

For all three experiments the raw data files were firstly transformed into the mzXML format using the tool 'ReAdW' (Keller et al. 2002; Nesvizhskii et al. 2003), and then uploaded into the QuPE system. In the cases of experiments A and B, mass spectra were then searched using Mascot™ against a database that consists of the completely annotated genome of *Bacillus subtilis* and an equally-sized set of randomized amino acid sequences to facilitate the calculation of false discovery rates. In addition, a number of common laboratory contaminants (The Global Proteome Machine Organization 2011) were included in the database. "Obviously, these proteins were not subject to the labeling, but some showed high signal-to-noise values for the unlabeled peptide. We kept these—actually senseless—proteins in our analysis as they provide a good example for measurements having a high variance. Due to a label swap (control ^{15}N, experiment ^{14}N) in one of the samples not only very high but also very low ratios were obtained" (Albaum et al. 2011b, p.18). Other search parameters were adjusted as follows: the peptide tolerance was set to 10 ppm, the ms/ms tolerance to 1000 mmu; up to two missed cleavage sites were allowed, and oxidation of methionine was configured as a variable modification. As already used in the quantification evaluation (see section 8.1), a potential one Dalton shift was set up for arginine and lysine to account for situations in which the mass spectrometer had not automatically selected the monoisotopic peak of a ^{15}N-labeled precursor for analysis but instead a neighboring peak. Only those hits were taken into further consideration that had, firstly, a score above Mascot™'s own significance threshold ($p < 0.05$), and secondly, a false discovery rate q of less than 0.05, which was estimated based on the

decoy database (see section 7.4.2). For each spectrum only the best-scoring hit was kept so that, in summary, 173,044 peptides remained for experiment A, and 620,305 peptides for experiment B. These in turn accounted for 1,445 and 2,472 proteins, respectively.

In case of experiment C, protein identifications, provided by Haußmann et al., were directly imported into the QuPE system. Originally, the Sequest™ search engine was utilized to compare all mass spectra against a *Corynebacterium glutamicum* database. Following the search, filter criteria such as score thresholds were adjusted in such a way that in the end a false discovery rate of less than 0.01 had to be accepted. The final list included 12,870 peptides representing 712 proteins.

9.1.3 Protein quantification

Protein quantification was performed as described in section 7.3.2. In experiment A, the accuracy of the isotopic distribution calculation was set to 0.01 m/z, a tolerance value ε of 0.2 Da was utilized, and for each identified peptide a retention time interval of 30 seconds before and after the corresponding mass spectrum was investigated. For peak detection the CWT-based algorithm was used; an isotopic similarity $d_{\mathbf{S}\times\mathbf{T}} > 0.8$, a regression coefficient $r > 0.6$, and a signal-to-noise threshold $S/N > 3.0$ were used for filtering. For experiments B and C these parameters were slightly modified. Here, an accuracy value of 0.1 m/z for the isotopic distribution calculation, an increased retention time interval of 60 seconds, a regression coefficient $r > 0.4$, and a signal-to-noise threshold of $S/N > 2.0$ were configured. For experiment B, in addition, a smaller tolerance value of $\varepsilon = 0.1$ was utilized.

In summary, 58,895 peptides could be quantified for experiment A, 180,913 for experiment B, and 3,699 for experiment C. These in turn accounted for 1,285, 2,321, and 589 proteins.

9.2 Detection of differentially regulated proteins

Asking which proteins are differentially regulated regarding one or more selected experimental conditions is probably the most frequently posed question in any quantitative proteomics experiment. In chapter 6 the analysis of variance (ANOVA) was introduced as a common method to answer this question, but also its prerequisites that, in case of infringement, require the use of methods such as the Kruskal-Wallis rank sum test. Based on the three case studies, the applicability of these methods on quantitative proteomics data was investigated, and it was evaluated in how far these methods produce congruent results in terms of the detection of the same proteins as significantly differentially regulated.

A statistical test such as the ANOVA demands the formulation of a model that (sufficiently) accurately describes the experiment's data. In case of the time series experiments A and B, in which each protein can be characterized by a vector $\mathbf{x} = \{x_i, i = 1, \ldots, N\}$ of relative

abundance values and a vector **t** that assigns each value x_i to a fixed time point t_i, this fixed effects model can be put forward by the equation

$$y = \mathbf{x} \sim \mathbf{t} \tag{9.1}$$

The same model can, in principle, be used to describe the data of experiment C, yet instead of a vector **t** of time points, in this case, the factor carbon source ($\mathbf{c} = \{c_i, i = 1, \ldots, N\}$), has to be defined to link one of the two growth conditions benzoate and glucose to each value x_i. In this connection, it has to be noted that "in view of the limited number of biological replicates for all three experiments, statistical tests were performed on every peptide measurement, i. e. each abundance ratio determined by a ^{15}N-labeled/unlabeled peptide pair was considered as an independent measurement of the protein's quantity" (Albaum et al. 2011b, p.3).

At first, the ANOVA was applied on the data of each experiment. In addition, the prerequisites of this statistical test, namely, the homogeneity of the error variances and the Gaussian distribution of all error components were examined using the Fligner-Killeen as well as the Shapiro-Wilks test (see section 6.1.2 for further details). Secondly, the Kruskal-Wallis test was investigated and the outcomes of both tests were compared. The significance level was *a priori* set to $\alpha = 0.05$ for all tests. Comprehensive results can be found in Albaum et al. (2011b) as well as online in the QuPE system.

In experiment A, the ANOVA revealed 73 proteins as significantly differentially regulated regarding the factor time. However, of these 73 proteins 15 showed inhomogeneous variances, and 29 had non-Gaussian distributions of their residuals. Taking all premises of the ANOVA into account, therefore, only 38 proteins can be regarded as significantly differentially regulated. The subsequently performed Kruskal-Wallis tests found 64 proteins with statistically significant deviations in their abundance values. Interestingly, the number of proteins being congruent between both methods was rather low (18) if only the 38 proteins that fulfilled all criteria were compared. If, however, the prerequisites of the ANOVA were elided, in total more than 80% of the proteins were found as significantly differentially regulated by both methods.

The results were similar in experiment B, although in this case the number of proteins declared significantly differentially regulated by the ANOVA as well as the Kruskal-Wallis test was comparably high with an accordance of more than 90%. Yet the number of proteins found significantly differentially regulated by both methods was rather impressive with 386 proteins by the ANOVA and 493 by the Kruskal-Wallis test. The application of the Fligner-Killeen test to investigate inhomogeneous error variances led to the rejection of only 30 of the 386 proteins. The Shapiro-Wilks test, however, found a violation of the Gaussian distribution in 325 cases so that in summary only 61 proteins could strictly spoken be regarded as differentially regulated.

In experiment C only 17 proteins were found significantly differentially regulated by the ANOVA. At least in this experiment, no further observation was rejected by the Fligner-Killeen test, and only one protein had to be sorted out after the distribution of the residuals had been investigated. In comparison, the application of the Kruskal-Wallis test yielded 10

proteins as differentially regulated, which were—without any exception—also detected by the ANOVA.

To complete the picture and compare the outcome of both methods on the whole, the resulting lists of p-values for all proteins from the ANOVA as well as the Kruskal-Wallis test were set into relation by means of Spearman's rank correlation coefficient (Spearman 1904). In all cases the coefficient was approximately at $r = 0.8$ ($r = 0.829$ for experiment A, $r = 0.837$ for experiment B, and $r = 0.778$ for experiment C) so that, in summary, and following Cohen's rating of $r \geq 0.5$ as a strong correlation (Cohen 1988), a large degree of similarity between both methods was found.

9.3 Identification of co-regulated proteins

With regard to an experiment, researchers are often interested in finding groups of proteins that are characterized by similar abundance ratios. These proteins might underlie a common regulation, for example, in reaction to changing environmental conditions or different growth states of a cell culture. In chapter 6 various algorithmic approaches have been introduced to conduct this clustering task. Purpose and scope of the following evaluation is, firstly, to determine in how far the outcomes of different cluster algorithms applied on real-world quantitative proteomics datasets show similarities and/or differences—does the choice of a specific cluster algorithm influence the clustering result? The evaluation includes nine different cluster algorithms, namely, K-means, Neuralgas, fuzzy C-means, as well as hierarchical cluster analysis (HCA) using, on the one hand, Euclidean distances and Single-, Complete-, Average-, and Ward-linkage, and on the other hand, both Pearson's uncentered and centered correlation coefficient in combination with Average-linkage. Secondly, the clustering results themselves are evaluated in terms of both computational and biological significance. Therefore, a number of cluster validity measures are employed.

In contrast to the problem of detecting differentially regulated proteins, which is best solved by taking the variability of peptide measurements into account, it is advisable to reduce the complexity of the data in order to tackle the challenge of cluster analysis, and hence, to combine all abundance ratios that were measured for different replicates of one protein per condition. Apart from the median and the trimmed mean, the arithmetic mean is a commonly used summary statistic for this purpose. At this point of analysis, one may furthermore decide to discard those protein measurements for which not at least a specific number of replicates is available. In the end, a matrix of abundance values over all proteins and conditions provides the input for any cluster algorithm. It has, however, to be noted that due to the aforementioned filtering but also due to errors in measurements and for other reasons values might be missing in this matrix. This demands the application of missing value replacement strategies, e. g. by using the protein's average abundance value over all conditions. In a stringent way these proteins may also be completely discarded. Striving to achieve utmost accurate results, in the present study this strategy has been followed so that the matrix for experiment A included 188 proteins, for experiment B 935, and for experiment

C 196 proteins. At least two replicate measurements per protein and condition were required, which were then aggregated using the arithmetic mean.

9.3.1 Similarities and differences between cluster algorithms

To estimate the similarity and/or differences between different algorithms, a pairwise comparison of clustering results was conducted. Therefore, each algorithm was subsequently applied on the three experimental datasets to produce clustering results having cluster numbers in the range from two to 50 for the experiments A and C, and up to 100 for experiment B to reflect the comparatively larger dataset size of this experiment. Subsequently, for each algorithm clustering results with identical cluster numbers were compared to each other using the adjusted Rand index (see section 6.2.6 for further details). Afterwards, the mean index value for each pair of algorithms was calculated, which is displayed in form of a heatmap in Figure 9.1.

At first sight, a strong degree of similarity between K-means and Neuralgas becomes apparent with adjusted Rand index values greater than at least 0.45 (A/B: $R > 0.45$, C: $R > 0.6$). This is, however, not too surprising as the authors of Neuralgas claim the algorithm to be an extension of the K-means approach (cf. 6.2.4.2; Martinetz et al. 1993)[1]. In addition, a comparably high similarity (up to $R > 0.6$) can be found between these two algorithms and HCA using Ward-linkage and Euclidean distances, and in case of experiment C—albeit to a smaller extent—also to fuzzy C-means, Complete- and Average-linkage in combination with Euclidean distances. Interestingly, some clustering results show a considerable degree of similarity in some but not all experiments. This is, for example, the case for the results of HCA using Complete- and Average-Linkage with Euclidean distances in experiments A and C with an index value $R > 0.45$, which is only at $R = 0.25$ for experiment B. On the contrary, the two cluster algorithms using correlation-based distances (HCA using Average-linkage) and, for the majority of cases, Single-linkage using Euclidean distances yield entirely unique outputs, that do not compare to the results of other cluster algorithms. "In summary, the results of this comparison [...] demonstrate that the choice for a cluster algorithm is not arbitrary but instead strongly influences the outcome" (Albaum et al. 2011b, p.10).

9.3.2 Computational and biological significance of clustering results

Given the outcomes of different cluster algorithms, a decision must be taken whether the achieved results provide a 'good' solution for the problem to be solved, namely, the identification of co-regulated proteins. But, at this point, there remains one central question: what are the characteristics of a 'good' solution? Looking at the clustering shown in Figure 9.2 it becomes evident at first sight that the herein used method, that is HCA using Single-linkage

[1] This was also analyzed in detail in (Albaum et al. 2011b) by repeatedly comparing equally-sized K-means to Neuralgas clustering results invoked on the same dataset. Here, the Neuralgas approach outperformed K-means showing considerably less variance in its results.

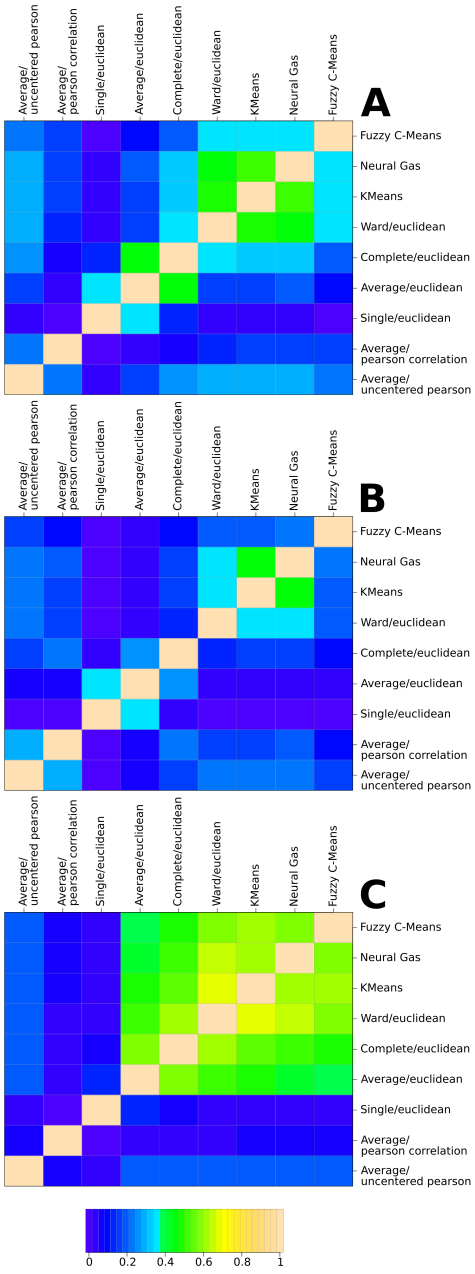


Figure 9.1 – A pairwise degree of similarity between different clustering results was estimated using the adjusted Rand index. In each case, the two compared clustering results had identical cluster numbers but were produced by two different algorithms. For each experiment, the Rand index was calculated for each cluster number in the range from two to 50 (experiment A and C), or 100 (experiment B). A heatmap visualization was chosen to display the mean Rand index for each pair of algorithms. It is evident that the three cluster algorithms HCA using Ward-linkage, K-means, and Neuralgas produce highly similar results. Only in the third experiment (C), these results were also similar to fuzzy C-means, Complete- and Average-linkage. It can, furthermore, be observed for two of the three experiments, namely A and B, that there is a slight similarity between Single- and Average-linkage (with Euclidean distances).

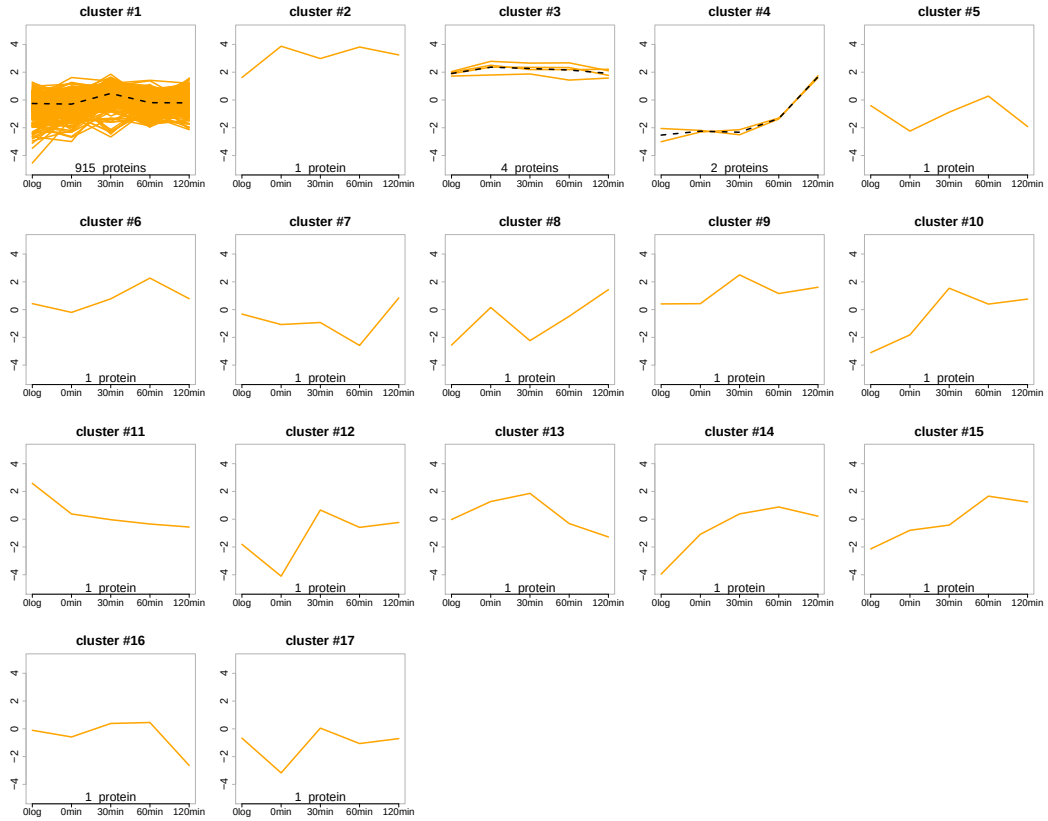


Figure 9.2 – This figure prominently illustrates a possible property of a clustering termed connectedness. It shows the results of a hierarchical cluster analysis (HCA) using Euclidean distances and Single-linkage. The algorithm clearly tends to group all proteins into one cluster that reveal only a slight similarity. It seems obvious that this method is not very practical to be used within the context of proteomics data analysis (each orange line represents a protein, the dashed black line indicates a cluster's prototype).

and Euclidean distances, does not provide a meaningful solution to the problem. All proteins that show a slight similarity are grouped together while only those proteins having exceptional abundance values are found in individual clusters. To computationally assess the significance of a clustering, a number of quality measures have been proposed, which are based on the input data and criteria inherent to a clustering (cf. 6.2.5). Two prominent characterizations to describe the structure of a clustering are, for example, connectedness and compactness (Handl et al. 2005). These are, however, opposing properties, which are at best represented by the two hierarchical cluster methods Single- and Complete-linkage. In the end, it is particularly this oppositeness which exposes the impossibility to formulate universal criteria to describe an optimal—let alone the best—clustering of a dataset.

Worse still, the search for an optimal cluster solution affects not only the choice of a specific cluster algorithm but also the determination of the 'true' number of clusters of a given dataset. In case of hierarchical cluster analysis, a simple but sufficient approach to gain this

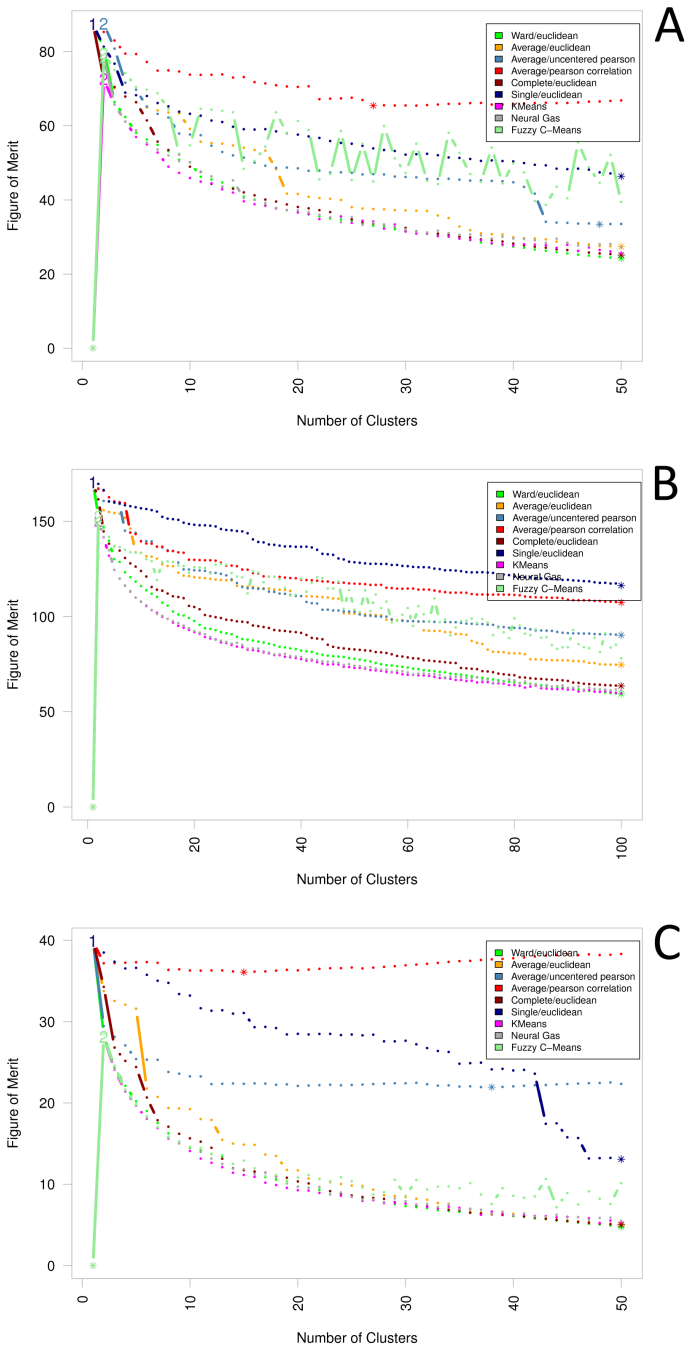


Figure 9.3 – The Figure of Merit estimates the predictive power of a cluster algorithm. Following this index, HCA using Ward-linkage with Euclidean distances, K-means, and Neuralgas belong to the most competitive algorithms for the clustering of quantitative proteomics data, while for example HCA using Single-linkage and Euclidean distances produces considerably less reliable results.

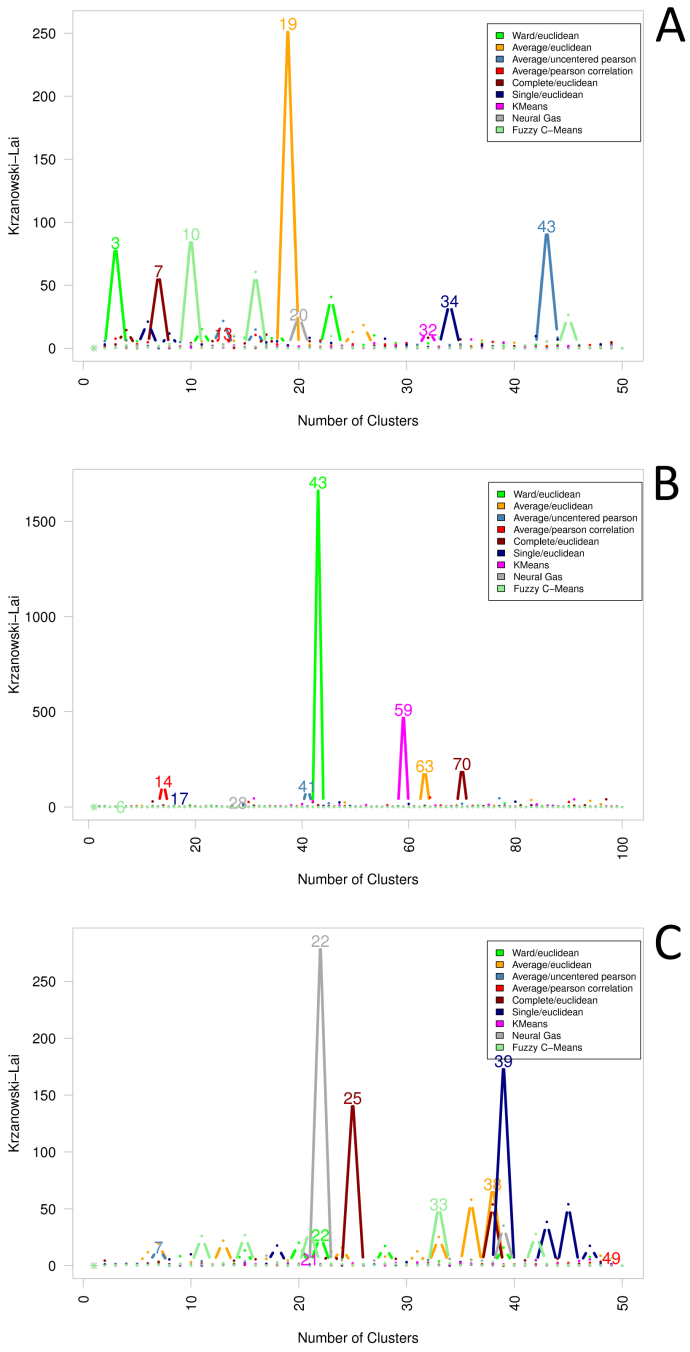


Figure 9.4 – The cluster index of Krzanowski and Lai showed both from a biological as well as from a computational point of view meaningful cluster numbers: for the data of experiment A, there were found cluster numbers between 3 for Ward/Euclidean—here a second local maximum was found at 23 clusters—and 43 for Average/Uncentered Pearson as the true clustering of the data; for experiment B, between 14 (Average/Pearson correlation) and 70 (Complete/Euclidean), with a protruding 43-cluster solution applying Ward/Euclidean; and for experiment C, e. g. at 22 clusters for Ward/Euclidean.

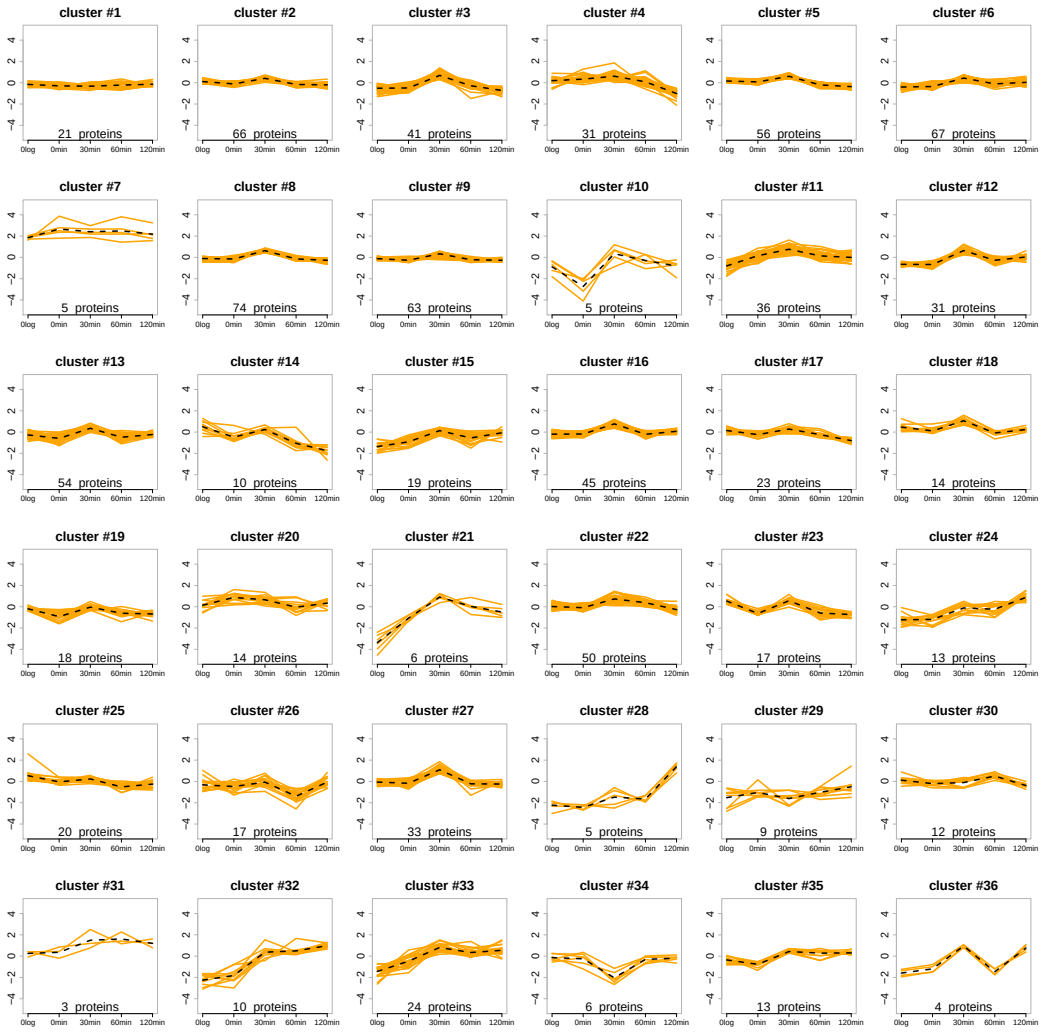


Figure 9.5 – This cluster profile plot demonstrates the property of HCA using Ward's linkage method to form compact clusters as the algorithmic approach attempts to minimize the increase in variance during the iterative cluster process (each orange line represents a protein, the dashed black line indicates a cluster's prototype). On this basis, the method seems to be well suited to identify co-regulated proteins. The cluster solution herein displayed was indicated as optimal using the cluster index proposed by Krzanowski and Lai.

cluster number is to investigate the increase in distance resultant from each (cluster) join operation. Typically, the distance between the first clusters[2] is comparatively small while it is usually surpassingly increased after several iterations. A method, which makes use of this assumption, is to plot the increase in distance against the cluster number, and afterwards search the 'turning point' or 'knee' in this plot.

[2] Please note that, in this consideration, each protein forms its own cluster at the beginning of a calculation.

To assess the significance of a clustering, a number of cluster validity measures have been introduced such as the cluster index of Calinski and Harabasz or the index *I* (see section 6.2.5 for more details). But, in the context of proteomics and in addition to this computational point of view, the significance of a clustering also needs to be seen from a biological perspective. In this case, further information about the proteins in a cluster needs to be taken into account. This can, for example, be a general functional description of a protein or the knowledge that an enzyme participates in a specific metabolic pathway reaction. However, in many experiments, an impeccable functional classification of each protein is usually only available for a subset of all identified proteins, not to mention the typically remarkable high number of hypothetical or putative proteins. An automatic inclusion of gene annotation data in the assessment of cluster solutions can therefore hardly be put into practice. In this evaluation, it is hence the aim to utilize existing cluster validity measures to gain an optimal clustering of the three different quantitative proteomics datasets. The final assessment of a good solution is then based on a manual evaluation, which incorporates additional knowledge of the clustered proteins such as functional annotations.

The Figure of Merit provides assistance in the selection of a cluster algorithm since it aims to estimate the predictive power of an algorithm by means of a bootstrapping approach (cf. 6.2.5.5). According to this cluster index, HCA using Ward-linkage with Euclidean distances, K-means, and Neuralgas clearly outperform the other cluster algorithms in this evaluation. In particular HCA using Single-linkage and Euclidean distances and—at least applied on the data of experiments A and B—fuzzy C-means yield, in contrast, the least reliable results (see Figure 9.3).

Calinski and Harabasz proposed a cluster validity measure that sets the similarity of all proteins within each cluster in relation to the pairwise computed similarities between all clusters (see section 6.2.5.1). This approach, however, tends to favor smaller cluster numbers such as two or three. Even though this result, from a computational point of view, may be well grounded on clear and understandable reasons, it generally characterizes only individual proteins with exceptional patterns of abundance (see Figure B.2). The application of the index to the data of experiment C represents an exception as in this case higher cluster numbers such as 14 for HCA using Complete-linkage and Euclidean distances could be observed. A reason for this might be the lower dimensionality of this dataset with only two conditions, namely glucose and benzoate. The tendency to predict very low cluster numbers as optimal emerges even more clearly for the cluster index called Index-*I* as it can be seen in Figure B.1. This is, however, not surprising since the computation of this index value follows an approach very similar to those of Calinski and Harabasz (see section 6.2.5.2, Figure B.1).

The application of the cluster index of Davies and Bouldin (see section 6.2.5.3 for further details) to the three proteomics datasets did not allow for any meaningful interpretation. In contrast to the other cluster indexes utilized in this study, a local minimum of the index value is said to indicate an optimal cluster solution. Yet in most cases, the index value is constantly decreasing with larger cluster numbers (see Figure B.3). An exception denotes Average-linkage using correlation-based distances as the index value is rather fluctuating in

this case, or for experiment C even steadily increasing. In summary, a clear and consistent statement cannot be formulated using this validity measure.

A pleasant surprise was the cluster index of Krzanowski and Lai (see section 6.2.5.4), which turned out to provide both from a computational and a biological point of view meaningful results (see Figure 9.4). In experiment A, cluster numbers between three for HCA using Ward-linkage and Euclidean distances—with a second local maximum at 23 clusters—and 43 for Average-linkage in combination with Pearson's uncentered correlation coefficient were indicated as optimal. A number of promising cluster solutions were investigated in detail. Here, in particular the 23-cluster solution (Ward/Euclidean) was riddled with interesting biological findings. It consisted of several clusters of proteins sharing a common function. This involves groups of proteins responsible for cell wall biogenesis, metabolism of amino acids, and for motility and chemotaxis. Overall, the findings correspond to the observations made by Hahne et al. (2010) in their original study.

Similarly positive results were achieved for experiment B. Here, the cluster index of Krzanowski and Lai suggested cluster numbers between 14 (HCA using Average-linkage and Pearson's correlation coefficient) and 70 (HCA using Complete-linkage with Euclidean distances). Following the advice of the aforementioned Figure of Merit that indicated HCA using Ward-linkage in combination with Euclidean distances as one of the best performing algorithms, the 43-cluster solution of this method was analyzed in detail. The result is shown in Figure 9.5 and reveals several biologically interesting groups of proteins. It can, in particular, be observed in how far different groups of proteins are regulated during the bacteria's growth phase. While several clusters of proteins, which play e. g. a role in secondary metabolites biosynthesis, transport and metabolism, can be found with decreasing abundance ratios after the cells entered the stationary phase, other clusters consist of proteins that are clearly up-regulated at that time point. In case of experiment C, the range of cluster numbers indicated as optimal ranges between seven for HCA using Average-linkage and Pearson's uncentered correlation coefficient and 38 clusters for Average-linkage with Euclidean distances. Again HCA using Ward-linkage and Euclidean distances was scrutinized, which in this case manifested in a 22-cluster solution. It revealed several ribosomal proteins with no change in their regulation due to the different growth media. In contrast, proteins that fulfill functions in amino acid transport and metabolism as well as energy production are down-regulated during growth on benzoate.

9.4 Proposal of a workflow for the analysis of quantitative proteomics experiments

The aim of this chapter, which summarizes the results of a methodological publication in Proteome Science, was to provide an answer to two of the most frequently posed questions in quantitative proteomics experiments: firstly, which proteins are differentially regulated regarding all investigated experimental conditions, and secondly, whether groups of proteins show similar abundance values, which in turn might indicate that these proteins are

commonly regulated. Certainly, a variety of methods have been introduced that allow to answer similar questions in other fields of PolyOmics such as Microarray data analysis. This, however, neglects the particular characteristics of mass spectrometry-based proteomics data i. e. for example noise in the data due to unrelated background signals in the mass spectra or missing values in the data matrix as peptides could not be correctly quantified nor even identified (Karpievitch et al. 2009). To face these challenges a workflow was derived and evaluated based on three recently published, real-world datasets.

The ANOVA constitutes the most powerful approach (see section 6.1.2) to detect differently regulated proteins. However, it was found that the results of this statistical test, strictly spoken, often had to be discarded as mandatory preconditions were not fulfilled. Especially the requirement that error components are Gaussian-distributed was in many cases not met. "Asking whether ANOVA [...] assumptions are satisfied is not idle curiosity. The assumptions of most mathematical models are always false to a greater or lesser extent. The relevant question is not whether ANOVA assumptions are met exactly, but rather whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions" (Glass et al. 1972, p.237). Since the objective of many experiments is to find all proteins that may be influenced by a stress stimulus, play a role in a specific regulatory mechanism, or could be a potential target for a specific therapeutic agent, one may therefore argue that it is of primary interest to find any possible candidate before a stricter investigation and analysis take place. According to the statistical doctrine, instead of the ANOVA a non-parametric test such as the Kruskal-Wallis rank sum has to be applied. Interestingly, a strong congruence between the results of both tests has been discovered. In conclusion, it can be recommended "to firstly rely on the results of an ANOVA, but secondly, to always take into consideration Kruskal-Wallis. Results should then be compared and further visually investigated using for example Box- and Whisker-plots. In all tests, because of the multiple testing situation, adjustment of computed p-values should take place" (Albaum et al. 2011b, p.16).

When it comes to the identification of groups of potentially co-regulated proteins, cluster analysis is the method of choice. It, however, strongly depends on the utilized algorithm and the utilized measure of validity whether a biologically meaningful cluster solution can be obtained. In this connection, the objective is clearly to find and separate those groups of proteins that reveal an utmost similar pattern of abundance regarding the selected experiment conditions. It can, hence, be stated that HCA using Single-linkage is not applicable for this purpose since this method tends to cluster all proteins together that show only a slight similarity. A good indicator for this non-applicability was also given by the development of the Figure of Merit. Following the evaluation it can be concluded that "if the benefits of a hierarchical cluster analysis are requested, Ward's method has proven a good choice. If there isn't, Neuralgas should be selected, which clearly outperforms the K-means approach, in particular, regarding the reproducibility of its results. The only drawback of this algorithm might be its comparatively high computational complexity, which is, however, negligible taken into consideration today's average computing resources. [...] The most difficult part is the validation of a cluster result to gain the 'true' number of clusters of a dataset. Here, the cluster index of Krzanowski and Lai turned out to produce both computationally as well

as biologically meaningful results. In contrast to other investigated validity measures the index solely relies on the internal compactness of clusters, which seems to correspond to our objective of clustering those proteins that reveal a highly similar pattern of regulation" (Albaum et al. 2011b, p.16-17).

Chapter 10

Discussion and Conclusion

Positive changes have taken place in the field of protein analysis over the last decade. Technical developments and in particular the advent of novel experimental procedures forged ahead and finally provide today's scientist with a comprehensive inventory to scrutinize the biomolecules that correlate the closest to the phenotype of an organism[1]. The most important contribution to this development was certainly made with the invention of the soft ionization methods MALDI and ESI, which in the first place allowed to determine the masses of large biomolecules such as proteins. Also, the laboratory methods targeting the employment of stable isotopes and other mass tags for protein quantification played a crucial role towards the establishment of a high-throughput analysis of an organism's complete proteome. Since these technical and experimental methods have found a firm place in the tool box of modern proteomics researchers, there is undoubtedly a strong need for computational assistance in processing and in particular evaluating the enormous amounts of data that are accumulating in such mass spectrometry-based quantitative proteomics experiments.

10.1 The rich internet application QuPE

In this work, the concept of a software application for quantitative proteomics experiments was devised and put into practice. This envisaged a platform, firstly, to manage all data and meta data related to these experiments, and secondly, to ease the development and

[1] One may of course argue as well that it is the metabolome that best reflects any differences in the phenotype of two organisms.

integration of novel analysis methods. Based on the capabilities of the system a variety of new methods have been designed and implemented starting from procedures for the assessment of protein identifications, to optimized but also novel algorithms for protein quantification, to the first-time derivation of a workflow for the multivariate statistical analysis of quantitative proteomics experiments.

The system has already been used successfully in a number of national and international projects. Currently, more than 50 users are registered as members of QuPE projects via the GPMS: the unique project and user management system employed at the CeBiTec. The list of application cases comprises scientific collaborations with different institutions, as for example in the frame of the BMBF-funded QuantPro initiative with workgroups at the universities of Bochum and Greifswald [grant 0313812], and the list of organisms that is worked with covers all domains and ranges from *Burkholderia cenocepacia* to *Xanthomonas campestris pv. campestris*, from *Arabidopsis thaliana* to *Triticum aestivum*.

QuPE constitutes the first rich internet application to provide data management capabilities as well as analysis functionality for protein identification, quantification, and in particular statistical evaluation from any location in the world via a standard web browser. This advantage of QuPE, which is best expressed by the concept of 'Software as a Service' (SaaS, Mell and Grance 2010), has led to cooperations, *inter alia*, with the Heart and Diabetes Center in Bad Oeynhausen, the University College Cork in Ireland, and the Palacký University in the Czech Republic. It is one of the most characteristic features of the system that also locally dispersed users, once they have uploaded their data, can start and continue their analysis in a collaborative way, whenever an internet connection is available. Using the language of advertisement, QuPE, so to say, enables quantitative proteomics 'in the cloud'.

The first version of QuPE (under the working title 'ProSE') was used productively by members of Bielefeld University beginning at the end of the year 2008. Since then, the system has evolved towards a comprehensive platform for the storage and analysis of quantitative proteomics data with a plethora of new features being implemented and added to the web interface, the application logic, and the data model. During this time, the foundation of the system has proven successful as new ideas and methods such as advanced quantification algorithms could easily be integrated. Nowadays, the application programming interface provides a very extensive and rich basis for all kinds of proteomics data analyses. The devised data model for mass spectra, for protein identifications from database searches, and especially for analysis results, fulfilled all requirements in terms of performance, scalability, and flexibility. It was, in particular, the decision to rest the design of the system on the Spring framework (Johnson 2003; SpringSource, a division of VMware 2011) that yielded a modular and easy-to-extend architecture. The implemented model-view-controller pattern, which makes use of the Echo web framework (NextApp, Inc. 2011), facilitated a rapid advancement of the user interface and allows for an interactive, desktop-like experience of the software. The realization and deployment of QuPE as a rich internet application has shown to be an economical, reliable, and attractive solution for the development and operation of the system as well as its usage by experimenters.

10.2 Algorithms for protein quantification

A significant part of the work of this thesis was dedicated to the optimization and enhancement of algorithms for the calculation of (relative) abundance values from isotope-labeled protein samples. This started with the implementation of a rather simple single-spectrum based approach, which nevertheless achieves competitive results (see section 8.1.2), and ended with a new method that now allows to compare the abundances of two differentially labeled peptides, i. e. a partially-labeled peptide and its fully-labeled or fully-unlabeled counterpart in a high-throughput manner. Overall, the newly developed algorithms allow to accurately and precisely determine relative abundance values of metabolically stable isotope-labeled data and furthermore represent a significant improvement in terms of quality in comparison to other existing approaches.

The algorithms' key features are the utilization of exact theoretical isotopic distributions to ensure that, on the one hand, the complete set of peaks belonging to a peptide can be used for quantification, but on the other hand, any noise due to errors in measurement or overlapping peptides is omitted. In case liquid chromatography has been employed in the experiment and therefore a peptide's elution can be taken into account, the continuous wavelet transform showed the best performance to accurately predict the elution peak of a peptide. In contrast to other methods, as for example a simple top-down approach (see section 7.3.2) that searches for the apex of a peak and its ascending and descending flanks, the impact of instrumental errors is minimized and, moreover, the application of a smoothing filter such as the Savitzky-Golay filter is in general not necessary. A further feature of the algorithm is the calculation of relative abundance values based on linear regression instead of setting the area under the two XICs into relation. This approach is similar to those used in the Tool RelEx (see section 4.3.2), yet vertical offsets have been replaced by perpendicular offsets to allow for uncertainties in the intensity measurements of both the labeled and the unlabeled peptide.

A new algorithm was developed for so called pulse chase experiments. The basic idea is to replace the growth medium of an organism at a distinct time point, and thereby, introduce a new and, in particular, differentially stable isotope-labeled nutrient. The incorporation of the added mass tags can afterwards be followed in newly synthesized proteins or, similarly, their loss due to protein degradation. Colleagues at the university of Bochum conducted an experiment in which *Corynebacterium glutamicum* was transfered from minimal medium with either $^{15}NH_4Cl$ as nitrogen source or ^{13}C-labeled glucose as sole carbon source to normal growth medium. The challenge of this pulse chase approach that bases on metabolic labeling with ^{15}N or ^{13}C is to find out the current ratio of heavy to light isotopes of each peptide at any point of time over the course of the experiment. With protein half-lives in the range of a few minutes up to hours (Belle et al. 2006; Maier et al. 2011), it can be expected that after the switch of the medium all proteins with a high turnover will quickly be found with a decreasing number of incorporated heavy stable isotopes. In contrast, it is very likely to observe those proteins which have a slower rate of turnover with a higher isotope enrichment. Certainly, the incorporation can be different for each individual peptide. The correct determination

of the current incorporation rate is, moreover, exacerbated by two factors: first of all, an unknown number of incorporated stable isotopes also leads to an unknown mass shift, and second, partly-labeled peptides reveal a complex isotopic distribution, in which the most abundant peak is typically not the first peak.

Several software tools have been introduced for the quantification of metabolically-labeled protein samples as for example ASAPRatio, ProRata, Census, and QN (see section 4.3). All of these tools have in common that they can be used to calculate relative abundance ratios from a mixture of two samples with defined enrichments of stable isotopes, in general, containing both unlabeled as well as fully-labeled proteins. The pulse chase approach, however, demands an algorithm that is able to quantify sample mixtures in which one peptide is only found partly-labeled, and moreover, at an unknown rate of enrichment. It was not long ago that Gouw et al. (2010, p.16) noted that "in these cases, the lack of suitable software hampers data processing so far". Rao et al. (2008) addressed this problem by investigating all peptides which had been identified in their experiment. They developed a model to estimate the average m/z ranges based on the number of N-atoms in each peptide, and finally performed their quantification based on these estimations. Cargile et al. (2004) utilized a Poisson distribution model to predict the isotopic distribution patterns of labeled isotopes. Yet the drawback of these approaches is the requirement for manual estimation of parameters, and the necessity for time-consuming preprocessing of the data. They are therefore hardly to be used for high-throughput data analysis. The software tool QuantiSpec (Haegler et al. 2009, cf. 4.3.6) denotes one of the first automated approaches for the relative quantification of one partially-labeled and one fully-labeled or -unlabeled protein sample. While intended for MALDI-TOF data, the tool does not support LC-MS/MS experiments, and thus cannot refer to the temporal information that could be gained through the elution of peptides. Very recently, Guan et al. (2011) were the first to propose a pipeline for the calculation of protein turnover rates from ^{15}N-labeled samples in a high-throughput manner. Price et al. (2010) employed this approach to calculate protein turnover rates for over 2,500 proteins from three different tissues of mice. In this experiment, the animals were fed with a diet of ^{15}N-labeled algae. The algorithm has, however, one drawback which hampers its unrestricted application in other experiments. As protein identifications are transfered from the first time point to any subsequently recorded sample, the procedure places high demands on the individual samples—it is necessary to ensure that no retention time drifts occur for all peptides within the different samples.

The pulse chase quantification algorithm realized within QuPE constitutes an approach for LC-MS/MS data that is capable of calculating protein turnover rates in a high-throughput manner from any metabolically stable isotope-labeled sample, in which a mixture of proteins is found fully-labeled or -unlabeled as well as partially-labeled to an unknown extent. The algorithm, in particular, does not require any preconditions to be fulfilled. Apart from the initial pulse chase experiment conducted at Bochum University, the new quantification algorithm has been used by Grasse et al. (2011) for the investigation of a novel dimeric photosystem in a mutant strain of *Thermosynechococcus elongatus*. The algorithm has also proven its applicability for the quantification of protein samples from organisms for which the full incorporation of a stable isotope is hardly to achieve. This can have economic reasons

as the provision of an appropriate diet, e. g. to feed higher eukaryotes such as mice, is without any doubt laborious and time-consuming (Gouw et al. 2010; Zhang et al. 2011). But there may be other difficulties that prevent a complete incorporation. In an experiment to determine the qualification of different strains of the green algae *Chlamydomonas reinhardtii* for the production of biofuels, members of Olaf Kruse's workgroup at Bielefeld University faced the problem of finding—unintentionally—only partially-labeled protein samples. Using the quantification method developed within this work, they were yet able to successfully calculate relative protein abundance values.

The runtime of the pulse chase quantification algorithm is mainly characterized by three steps of the procedure: firstly, the calculation of theoretical isotopic distributions; secondly, the extraction of ion chromatograms; and thirdly, the final calculation of relative abundance ratios from these XICs. In the current implementation of the algorithm, a polynomial approach is utilized for the isotopic distribution calculation. The effort of this approach is acceptable as biomolecules consist, in general, only of the five elements carbon, hydrogen, nitrogen, oxygen, and sulfur (abbrev. CHNOPS). Given a peptide has each n_E atoms of an element $E \in [C, H, N, O, P, S]$, which in turn has I different isotopes E_i, $i \in [1, \ldots, I]$ the peptide's isotopic distribution can therefore be computed by expanding the following term:

$$(C_1 + C_2)^{n_C} \cdot (H_1 + H_2)^{n_H} \cdot (N_1 + N_2)^{n_N} \cdot (O_1 + \ldots + O_3)^{n_O} \cdot (S_1 + \ldots + S_4)^{n_S} \qquad (10.1)$$

To reduce the computational complexity of this calculation it is in general advisable, to combine single peaks if their difference in mass falls below a certain threshold. In QuPE, this value is configurable and thereby allows to attain an optimal balance between exactness and computational efficiency. Nevertheless, the runtime of the algorithm could be further improved in the future if methods that utilize Fourier transform are taken into account as proposed by Rockwood and Orden (1996), Cossio (2010), Sperling et al. (2008), or the dynamic programming-based approach described by Snider (2007). The extraction of ion chromatograms is undoubtedly the most demanding factor contributing to the computational costs of the algorithm. This step of the procedure requires the retrieval of mass spectra from the database for each peptide. As the amount of data can easily comprise several kilobytes per spectrum, it is not only the running time that is affected but also the memory consumption. Optimizations of the algorithm therefore targeted the amount of required memory, e. g. by flushing the Hibernate session cache after a specific number of processed spectra. To give an example, the running time for the elution peak quantification algorithm applied on a comparatively large experiment that contained over 1,400 proteins with almost 1,000,000 individual peptides was at approximately 43.5 hours on the CeBiTec compute cluster (averaged over four runs, 2x Quad-Core Intel Xeon™ E5640, 48GB RAM). It took less than two hours to compute the isotopic distributions for all peptides, while most of the time was spend for the creation of XICs. The final computation of relative abundance ratios then lasted only about 35 minutes.

10.3 A workflow for the analysis of quantitative proteomics experiments

Mass tags have become an established technique to gain an understanding of regulation at the protein level, and also the software tools to perform protein quantification have reached a high degree of quality (although it was very recently expressed that "there is still room for improvement of quantitation algorithms", Arsova et al. 2011, p.9). The end product of these tools and algorithms is usually a list of identified peptides together with their (relative) abundance values, which account for varying environmental conditions or different growth states of an organism (cf. sections 4.5, 5.1.4, Kumar and Mann 2009). "At this point, data will be ready for various statistical analyses" (Becker and Bern 2011, p.178), yet one has to denote that, "proteomics researchers are somehow left out in the cold, since existing software solutions [as listed above] lack support of advanced data analysis" (Albaum et al. 2009a, p.3129).

The provision of user-friendly and conceivable statistical analysis methods is, however, only 'half the battle'—moreover, it needed to be elucidated which statistical analysis strategy promises success for stable isotope-labeled proteomics data, and allows to draw accurate and valid conclusions from the data. The two central questions posed in a multitude of quantitative proteomics experiments are, firstly, which proteins are differentially regulated regarding the selected experimental conditions, and secondly, whether there are groups of proteins that show similar abundance ratios and thus might have a similar turnover. To answer these questions, a comprehensive evaluation was conducted within the scope of this work taking into account three real-world datasets from recently published experiments. This finally led to the derivation of a workflow for quantitative proteomics data analysis, which has been described in detail in the previous chapter.

Different statistical analysis methods were evaluated regarding their suitability to identify up- or down-regulated proteins in multivariate experimental data. In the same manner, cluster algorithms were investigated and their outcomes compared to each other in order to determine the method that best fits to this type of data. The evaluation assessed not only the cluster algorithms itself but also their validation to obtain the optimal number of clusters for a specific dataset. In this connection, the inclusion of external information such as COG functional categories turned out to be a key element to gain meaningful clusterings, both from a biological and a computational point of view.

10.4 Further developments of the QuPE system

Overall, the QuPE system has proven to be highly extensible. This could not only be demonstrated with the implemented algorithms for protein quantification and the provided range of statistical analysis functionality but also with minor extensions and enhancements of the system, such as the recently added support for spot coordinates on 2D-gels. This was realized in close cooperation with the Institute of Plant Genetics at the Leibniz Universität Hannover.

Rode et al. (2011) developed GelMap as a web-based tool for the storage and representation

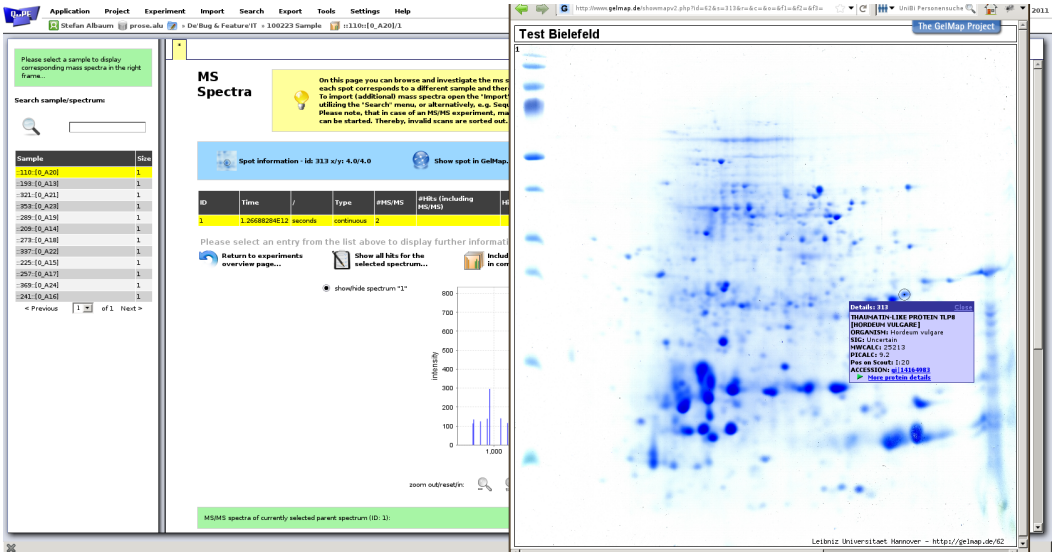


Figure 10.1 – This screenshot demonstrates the recently added support for 2D-gels in QuPE via the connection with the web-based software tool GelMap. This was made possible by a cooperation with the Leibniz Universität Hannover. After an experiment has been linked to a GelMap project, and spot coordinates have been added to each imported sample, it is possible to directly link a mass spectrum and the respective protein identification to the original spot position on a 2D-gel.

of 2D-gel electrophoresis results. In the context of QuPE, spot coordinates can be assigned to each imported sample, e. g. via import of an Excel spreadsheet, and linked to a GelMap project. This is illustrated in the screenshot shown in Figure 10.1. A researcher is thereby able to directly link protein identification and mass spectra to the respective position of the protein on an underlying 2D-gel.

Although QuPE has reached a high level of functionality, applicability and stability, it is, of course, still work in progress as developments in the field of proteomics in terms of novel laboratory and technical methods are continuously moving forward, and thus pose new challenges for data management, integration, and analysis. An important problem in mass spectrometry-based proteomics, these days, concerns the investigation of post-translational protein modifications (PTMs). Protein phosphorylation sites are, for example, of particular interest since these may result in a conformational change which in turn may lead to the protein's activation or inactivation. While the well-known and established search engines such as Mascot™or Sequest™do support database searches that take PTMs into account, the inclusion of more than a few potential modifications, in general, drastically increases the computational complexity of the search. To tackle this issue alternative approaches need to be evaluated and, subsequently, integrated into the QuPE system. The search engine InsPecT (Tanner et al. 2005) is a promising candidate as it has specifically been designed to address the identification of post-translational protein modifications (PTMs).

While protein quantification methods have reached a high level of quality and accuracy, it is conceivable that the amount of quantifiable proteins can be further increased if the protein quantification does not primarily rely on an *a priori* conducted protein identification step. Instead, the feature space which is spanned by the three domains 'retention time', 'mass to charge ratio', and 'signal intensity' could be searched independently for pairs of labeled and unlabeled proteins. A similar approach was introduced with the software MAXQUANT (see section 4.3.7). This can, however, only be applied to SILAC-labeled data. No approach has yet been conceived for stable isotope labeled data, for example, based on heavy nitrogen. Due to the nonuniform and complex peak pattern of proteins labeled in such a way, this certainly represents a comparatively far more difficult task. While there would be the benefit of an increased number of quantified peptides, it remains questionable whether these abundance values can afterwards also be assigned to a specific protein—this would need to be analyzed.

10.5 Final remarks

"Mass spectrometry (MS)-based proteomics has significantly contributed to the development of systems biology, a new paradigm for the life sciences in which biological processes are addressed in terms of dynamic networks of interacting molecular networks" (Sabidó et al. 2011, p.1). In this spirit, it is in particular the utilization of metabolic labeling approaches in combination with the pulse chase approach that allows to gain detailed insights into the processes that are responsible for the amounts of proteins in a cell—protein synthesis as well as degradation. QuPE constitutes a comprehensive platform for the analysis of these quantitative proteomics experiments, especially of metabolic stable isotope labeling approaches. Due to its extensible nature, the system can easily be extended to cope with future developments in this field of research.

Appendix

Appendix A

Implementation of the QuPE system – additional information

In this section of the appendix, additional information regarding the implementation of the QuPE system is provided.

A.1 Isotopic Distribution Calculation

Calculation of isotopic distributions of amino acids is derived from an open source program named 'Isotopic Pattern Calculator' (Nolting 2008). Given the elemental composition of a peptide, the algorithm computes relative peak intensities with a user-defined accuracy. In addition, an intensity threshold may be set to omit peaks that contribute only an irrelevantly small amount to the overall intensity of a molecule.

```
INPUT: elemental composition
INPUT: accuracy, e. g. 0.01
INPUT: charge state of peptide
INPUT: minimal peak intensity

// the variable peaks defines a mapping of masses on intensity values
DEFINE peaks = Mapping[mass,intensity]

// initialization of peaks variable
SET peaks[0.0] = 1.0
```

```
// loop over all elements possibly occurring in biomolecules...
// i.e. carbon, hydrogen, nitrogen, oxygen, sulfide
FOR EACH element OF CHNOS DO
  // ...for each occurring atom...
  FOR EACH atom OF element DO
    // ...adjust each previous peak...
    FOR EACH peak OF peaks DO
      // ...taking each isotope and its probability into account...
      FOR EACH isotope OF element DO

        DEFINE peakMass = peak.mass
        DEFINE peakIntensity = peak[mass]

        DEFINE newPeakMass = ROUND((peakMass + isotope.mass) / accuracy) * accuracy

        DEFINE newPeakIntensity = peakIntensity * isotope.frequency

        IF peaks[newPeakMass] != null THEN
          peaks[newPeakMass] = peakIntensity + newPeakIntensity
        ELSE
          peaks[newPeakMass] = newPeakIntensity
        FI
      OD
    OD
  OD
OD

// finally, 'translate' to relative peak intensities
DEFINE maxIntensity = MAX OF ALL intensities OF peaks
DEFINE finalPeaks = Mapping[mass,intensity]

FOR EACH peak OF peaks DO
  DEFINE peakMass = peak.mass
  DEFINE peakIntensity = peak[mass]

  DEFINE relativeIntensity = peakIntensity / maxIntensity * 100

  // peaks below a user-defined threshold are omitted
  IF relativeIntensity > threshold THEN
    finalPeaks[peakMass / charge] = relativeIntensity
  FI
OD
```

Appendix B

Performance and accuracy of protein quantification – additional information

In this section of the appendix, additional information regarding the performance and accuracy of algorithms for the quantification of isotope-labeled protein samples is given.

B.1 Reference measurements – additional information

The universities of Bielefeld and Greifswald provides benchmark datasets for the evaluation of algorithms for the quantification of proteins. In the following, tool configurations and further information used for the processing of these datasets can be found.

B.1.1 Configuration of the tool ProRata

The configuration of the tool ProRata (Pan et al. 2006) was essentially used as provided by the authors, respectively, in the downloaded software package (Version 1.0):

```
<?xml version="1.0"?>
<CONFIG version="1.0" >
    <SIC_EXTRACTION>
        <MS_FILE_TYPE>mzXML</MS_FILE_TYPE>
        <ID_FILE_TYPE>DTASelect</ID_FILE_TYPE>
        <RETENTION_TIME_INTERVAL>
            <MINUTES_BEFORE_MS2>2</MINUTES_BEFORE_MS2>
            <MINUTES_AFTER_MS2>2</MINUTES_AFTER_MS2>
            <MINUTES_BETWEEN_DUPLICATE_MS2>2</MINUTES_BETWEEN_DUPLICATE_MS2>
        </RETENTION_TIME_INTERVAL>
```

```
        <MASS_TO_CHARGE_INTERVAL>
            <PLUS_MZ_ERROR> 0.5 </PLUS_MZ_ERROR>
            <MINUS_MZ_ERROR> 0.5 </MINUS_MZ_ERROR>
            <ISOTOPIC_ENVELOP_CUTOFF> 0.1 </ISOTOPIC_ENVELOP_CUTOFF>
        </MASS_TO_CHARGE_INTERVAL>
        <ATOM_ISOTOPIC_COMPOSITION>
            <C>
                <MASS_DA> 12.000000,  13.003355 </MASS_DA>
                <NATURAL> 0.9893,   0.0107    </NATURAL>
                <ENRICHED>  0.02,   0.98    </ENRICHED>
            </C>
            <H>
                <MASS_DA> 1.007825, 2.014102  </MASS_DA>
                <NATURAL> 0.999885, 0.000115  </NATURAL>
                <ENRICHED>  0.02,   0.98    </ENRICHED>
            </H>
            <O>
                <MASS_DA> 15.994915,  16.999132,  17.999160 </MASS_DA>
                <NATURAL> 0.99757,  0.00038,  0.00205   </NATURAL>
                <ENRICHED>  0.02,   0.0,    0.98    </ENRICHED>
            </O>
            <N>
                <MASS_DA> 14.003074,  15.000109 </MASS_DA>
                <NATURAL> 0.99632,  0.00368   </NATURAL>
                <ENRICHED>  0.02,   0.98    </ENRICHED>
            </N>
            <P>
                <MASS_DA> 30.973762 </MASS_DA>
                <NATURAL> 1.0   </NATURAL>
                <ENRICHED>  1.0   </ENRICHED>
            </P>
            <S>
                <MASS_DA> 31.972071,  32.971459,  33.967867,  35.967081 </MASS_DA>
                <NATURAL> 0.9493,   0.0076,   0.0429,   0.0002    </NATURAL>
                <ENRICHED>  0.02,   0.0,    0.98,   0.0   </ENRICHED>
            </S>
        </ATOM_ISOTOPIC_COMPOSITION>
        <RESIDUE_ATOMIC_COMPOSITION>
            <ISOTOPOLOGUE name="N14" >
                <R> NTerm,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> CTerm,  0,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> L,  6,  11, 1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> A,  3,  5,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> S,  3,  5,  2,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> G,  2,  3,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> V,  5,  9,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> E,  5,  7,  3,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> K,  6,  12, 1,  2,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> I,  6,  11, 1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> T,  4,  7,  2,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> D,  4,  5,  3,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> R,  6,  12, 1,  4,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> P,  5,  7,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> N,  4,  6,  2,  2,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> F,  9,  9,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> Q,  5,  8,  2,  2,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> Y,  9,  9,  2,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> M,  5,  9,  1,  1,  0,  1,  0,  0,  0,  0,  0,  0 </R>
                <R> H,  6,  7,  1,  3,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> C,  3,  5,  1,  1,  0,  1,  0,  0,  0,  0,  0,  0 </R>
                <R> W,  11, 10, 1,  2,  0,  0,  0,  0,  0,  0,  0,  0 </R>
    <R> *,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0 </R>
    <R> #,  2,  3,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0 </R>
            </ISOTOPOLOGUE>
            <ISOTOPOLOGUE name="N15" >
                <R> NTerm,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> CTerm,  0,  1,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0 </R>
                <R> L,  6,  11, 1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> A,  3,  5,  1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> S,  3,  5,  2,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> G,  2,  3,  1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> V,  5,  9,  1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> E,  5,  7,  3,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> K,  6,  12, 1,  0,  0,  0,  0,  0,  0,  2,  0,  0 </R>
                <R> I,  6,  11, 1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> T,  4,  7,  2,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> D,  4,  5,  3,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> R,  6,  12, 1,  0,  0,  0,  0,  0,  0,  4,  0,  0 </R>
                <R> P,  5,  7,  1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> N,  4,  6,  2,  0,  0,  0,  0,  0,  0,  2,  0,  0 </R>
                <R> F,  9,  9,  1,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> Q,  5,  8,  2,  0,  0,  0,  0,  0,  0,  2,  0,  0 </R>
                <R> Y,  9,  9,  2,  0,  0,  0,  0,  0,  0,  1,  0,  0 </R>
                <R> M,  5,  9,  1,  0,  0,  1,  0,  0,  0,  1,  0,  0 </R>
```

```
                <R> H, 6, 7, 1, 0, 0, 0, 0, 0, 3, 0, 0 </R>
                <R> C, 3, 5, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0 </R>
                <R> W, 11, 10, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0 </R>
                <R> *, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 </R>
    <R> #, 2, 3, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0 </R>
            </ISOTOPOLOGUE>
        </RESIDUE_ATOMIC_COMPOSITION>
    </SIC_EXTRACTION>
    <PEPTIDE_QUANTIFICATION>
        <PEAK_DETECTION>
            <CHROMATOGRAM_SMOOTHING>
                <ORDER>2</ORDER>
                <WINDOW_SIZE>7</WINDOW_SIZE>
            </CHROMATOGRAM_SMOOTHING>
            <PEAK_SHIFT>
                <LEFT>0</LEFT>
                <RIGHT>0</RIGHT>
            </PEAK_SHIFT>
        </PEAK_DETECTION>
        <ABUNDANCE_RATIO>
            <NUMERATOR_ISOTOPOLOGUE>N14</NUMERATOR_ISOTOPOLOGUE>
            <DENOMINATOR_ISOTOPOLOGUE>N15</DENOMINATOR_ISOTOPOLOGUE>
        </ABUNDANCE_RATIO>
        <LOG2_RATIO>
            <MINIMUM>-10</MINIMUM>
            <MAXIMUM>10</MAXIMUM>
        </LOG2_RATIO>
        <LOG2_SNR_CUTOFF>1</LOG2_SNR_CUTOFF>
        <REMOVE_AMBIGUOUS_PEPTIDES>true</REMOVE_AMBIGUOUS_PEPTIDES>
    </PEPTIDE_QUANTIFICATION>
    <PROTEIN_QUANTIFICATION>
        <MIN_PEPTIDE_NUMBER>2</MIN_PEPTIDE_NUMBER>
        <MAX_CI_WIDTH>5</MAX_CI_WIDTH>
        <MAX_LOG2_SNR>4</MAX_LOG2_SNR>
        <LOG2_RATIO>
            <MINIMUM>-5</MINIMUM>
            <MAXIMUM>5</MAXIMUM>
        </LOG2_RATIO>
        <LOG2_RATIO_DISCRETIZATION>0.1</LOG2_RATIO_DISCRETIZATION>
        <FASTA_FILE>bs.faa</FASTA_FILE>
        <STANDARD_DEVIATION>
            <SLOPE>-0.288</SLOPE>
            <INTERCEPT>1.305</INTERCEPT>
        </STANDARD_DEVIATION>
        <MEAN>
            <SLOPE>1.2</SLOPE>
            <INTERCEPT>0</INTERCEPT>
        </MEAN>
        <SMOOTHING_PROBABILITY_SPACE>0.15</SMOOTHING_PROBABILITY_SPACE>
    </PROTEIN_QUANTIFICATION>
</CONFIG>
```

B.1.2 Configuration of the tool Census

As in the case of ProRata, the configuration of the tool Census (Park et al. 2008) was essentially used as provided by the authors, respectively, in the downloaded software package (Version 1.33):

```
<?xml version="1.0" encoding="UTF-8"?>
<config>
  <label_type labeling="true">
    <name>sample</name>
    <name>reference</name>
  </label_type>
  <params>
    <scan_type>MS</scan_type>
    <extract_method>1</extract_method>
    <mass_accuracy unit="mz">0.3</mass_accuracy>
    <enrich>0.98</enrich>
    <max_win>50</max_win>
  </params>
  <element_comp>
    <each_sample>
      <residue name="A">
        <ele_C>3</ele_C>
        <ele_H>5</ele_H>
        <ele_O>1</ele_O>
        <ele_N>1</ele_N>
        <ele_S>0</ele_S>
        <ele_P>0</ele_P>
        <ele_15N>0</ele_15N>
        <ele_2H>0</ele_2H>
        <ele_13C>0</ele_13C>
      </residue>
      <residue name="C">
        <ele_C>5</ele_C>
        <ele_H>8</ele_H>
        <ele_O>2</ele_O>
        <ele_N>2</ele_N>
        <ele_S>1</ele_S>
```

```
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="D">
      <ele_C>4</ele_C>
      <ele_H>5</ele_H>
      <ele_O>3</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="E">
      <ele_C>5</ele_C>
      <ele_H>7</ele_H>
      <ele_O>3</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="F">
      <ele_C>9</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="G">
      <ele_C>2</ele_C>
      <ele_H>3</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="H">
      <ele_C>6</ele_C>
      <ele_H>7</ele_H>
      <ele_O>1</ele_O>
      <ele_N>3</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="I">
      <ele_C>6</ele_C>
      <ele_H>11</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="K">
      <ele_C>6</ele_C>
      <ele_H>12</ele_H>
      <ele_O>1</ele_O>
      <ele_N>2</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="L">
      <ele_C>6</ele_C>
      <ele_H>11</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="M">
      <ele_C>5</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>1</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="N">
      <ele_C>4</ele_C>
      <ele_H>6</ele_H>
      <ele_O>2</ele_O>
      <ele_N>2</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="P">
      <ele_C>5</ele_C>
      <ele_H>7</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="Q">
      <ele_C>5</ele_C>
      <ele_H>8</ele_H>
      <ele_O>2</ele_O>
      <ele_N>2</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="R">
      <ele_C>6</ele_C>
      <ele_H>12</ele_H>
      <ele_O>1</ele_O>
      <ele_N>4</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="S">
      <ele_C>3</ele_C>
      <ele_H>5</ele_H>
      <ele_O>2</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="T">
      <ele_C>4</ele_C>
      <ele_H>7</ele_H>
      <ele_O>2</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
```

```
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="V">
      <ele_C>5</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="W">
      <ele_C>11</ele_C>
      <ele_H>10</ele_H>
      <ele_O>1</ele_O>
      <ele_N>2</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="Y">
      <ele_C>9</ele_C>
      <ele_H>9</ele_H>
      <ele_O>2</ele_O>
      <ele_N>1</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="NTERM">
      <ele_C>0</ele_C>
      <ele_H>1</ele_H>
      <ele_O>0</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="CTERM">
      <ele_C>0</ele_C>
      <ele_H>1</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="*">
      <ele_C>0</ele_C>
      <ele_H>0</ele_H>
      <ele_O>0</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="#">
      <ele_C>0</ele_C>
      <ele_H>0</ele_H>
      <ele_O>0</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="@">
      <ele_C>0</ele_C>
      <ele_H>0</ele_H>
      <ele_O>0</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
  </each_sample>
  <each_sample>
    <residue name="A">
      <ele_C>3</ele_C>
      <ele_H>5</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="C">
      <ele_C>5</ele_C>
      <ele_H>8</ele_H>
      <ele_O>2</ele_O>
      <ele_N>1</ele_N>
      <ele_S>1</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="D">
      <ele_C>4</ele_C>
      <ele_H>5</ele_H>
      <ele_O>3</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="E">
      <ele_C>5</ele_C>
      <ele_H>7</ele_H>
      <ele_O>3</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="F">
      <ele_C>9</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="G">
      <ele_C>2</ele_C>
      <ele_H>3</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="H">
      <ele_C>6</ele_C>
      <ele_H>7</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>3</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
```

```
    </residue>
    <residue name="I">
      <ele_C>6</ele_C>
      <ele_H>11</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="K">
      <ele_C>6</ele_C>
      <ele_H>12</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>2</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="L">
      <ele_C>6</ele_C>
      <ele_H>11</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="M">
      <ele_C>5</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>1</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="N">
      <ele_C>4</ele_C>
      <ele_H>6</ele_H>
      <ele_O>2</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>2</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="P">
      <ele_C>5</ele_C>
      <ele_H>7</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="Q">
      <ele_C>5</ele_C>
      <ele_H>8</ele_H>
      <ele_O>2</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>2</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="R">
      <ele_C>6</ele_C>
      <ele_H>12</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>4</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="S">
      <ele_C>3</ele_C>
      <ele_H>5</ele_H>
      <ele_O>2</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="T">
      <ele_C>4</ele_C>
      <ele_H>7</ele_H>
      <ele_O>2</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="V">
      <ele_C>5</ele_C>
      <ele_H>9</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="W">
      <ele_C>11</ele_C>
      <ele_H>10</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>2</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="Y">
      <ele_C>9</ele_C>
      <ele_H>9</ele_H>
      <ele_O>2</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>1</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="NTERM">
      <ele_C>0</ele_C>
      <ele_H>1</ele_H>
      <ele_O>0</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="CTERM">
      <ele_C>0</ele_C>
      <ele_H>1</ele_H>
      <ele_O>1</ele_O>
      <ele_N>0</ele_N>
      <ele_S>0</ele_S>
      <ele_P>0</ele_P>
      <ele_15N>0</ele_15N>
      <ele_2H>0</ele_2H>
      <ele_13C>0</ele_13C>
    </residue>
    <residue name="*">
      <ele_C>0</ele_C>
```

```
        <ele_H>0</ele_H>
        <ele_O>0</ele_O>
        <ele_N>0</ele_N>
        <ele_S>0</ele_S>
        <ele_P>0</ele_P>
        <ele_15N>0</ele_15N>
        <ele_2H>0</ele_2H>
        <ele_13C>0</ele_13C>
      </residue>
      <residue name="#">
        <ele_C>0</ele_C>
        <ele_H>0</ele_H>
        <ele_O>0</ele_O>
        <ele_N>0</ele_N>
        <ele_S>0</ele_S>
        <ele_P>0</ele_P>
        <ele_15N>0</ele_15N>
        <ele_2H>0</ele_2H>
        <ele_13C>0</ele_13C>
      </residue>
      <residue name="@">
        <ele_C>0</ele_C>
        <ele_H>0</ele_H>
        <ele_O>0</ele_O>
        <ele_N>0</ele_N>
        <ele_S>0</ele_S>
        <ele_P>0</ele_P>
        <ele_15N>0</ele_15N>
        <ele_2H>0</ele_2H>
        <ele_13C>0</ele_13C>
      </residue>
    </each_sample>
  </element_comp>
</config>
```

Results were afterwards exported using the following parameters:

```
Determination Factor : 0.5
Outlier pValue : 0.1
Filter Fragment Ions on MS/MS pValue : true
Correction Factor Value : 0.0
allNoneLowerBound : 0.1
allNoneUpperBound : 10.0
allNoneCompositeScore : 0.95
Unique Peptide only : false
```

B.2 Evaluation of implemented quantification algorithms – additional information

This section of the appendix provides additional information regarding the evaluation of own quantification algorithms (see chapter 8 for further details).

B.2.1 Accuracy of the elution peak quantification – parameter evaluation

An evaluation was performed to investigate the impact of different parameters on the quantification results achievable with the elution peak quantification algorithm (see sections 7.3.2 and 8.1.3). Therefore, the 1:1 sample provided by colleagues at the University of Greifswald was analyzed in detail. The following settings were used in accordance with the characteristics of this experiment: accuracy of the isotopic distribution calculation: 0.1 m/z, tolerance value ε = 0.01 Da, investigated retention time interval for each peptide: 60 seconds before and after the identifying mass spectrum, CWT-based peak detection. The results of varying settings of the isotopic similarity $d_{\mathbf{S}\times\mathbf{T}}$, the regression coefficient r, and the signal-to-noise threshold S/N are shown in the following table. Here, $< M >$ denotes the expected mean value based on the given ratio, thus, in this case 0. The column entitled $\bar{M}$ shows the mean value of all calculated peptide abundance ratios together with their standard deviation σ. The median is given in column $\tilde{M}$, while the sixth column contains the 95%-confidence interval. The last two columns denote the overall number of calculated peptide abundance ratios (#peptides) and the number of proteins these peptides account for (#proteins).

				Elution peak quantification - parameter evaluation					
$d_{S\times T}$	r	S/N	$< M >$	$\bar{M}$	σ	$\tilde{M}$	$\bar{M} \pm 0.95$	#peptides	#proteins
0.9	0.9	4.0	0	**-0.22**	0.37	-0.23	-0.77;0.40	1414	313
0.9	0.9	3.0	0	**-0.22**	0.42	-0.24	-0.85;0.46	2457	421
0.9	0.9	2.0	0	**-0.22**	0.45	-0.24	-0.95;0.56	3457	501
0.9	0.8	4.0	0	**-0.24**	0.40	-0.24	-0.92;0.42	1561	335
0.9	0.8	3.0	0	**-0.23**	0.45	-0.24	-1.00;0.55	2816	457
0.9	0.8	2.0	0	**-0.23**	0.49	-0.24	-1.22;0.66	4077	549
0.9	0.7	4.0	0	**-0.24**	0.47	-0.24	-1.04;0.51	1636	346
0.9	0.7	3.0	0	**-0.23**	0.52	-0.24	-1.18;0.68	3007	474
0.9	0.7	2.0	0	**-0.23**	0.56	-0.24	-1.38;0.82	4444	568
0.9	0.6	4.0	0	**-0.24**	0.53	-0.24	-1.20;0.71	1694	358
0.9	0.6	3.0	0	**-0.24**	0.57	-0.23	-1.34;0.80	3144	492
0.9	0.6	2.0	0	**-0.23**	0.60	-0.24	-1.47;0.94	4694	586
0.9	0.5	4.0	0	**-0.25**	0.59	-0.24	-1.42;0.78	1735	364
0.9	0.5	3.0	0	**-0.24**	0.63	-0.24	-1.49;0.94	3237	499
0.9	0.5	2.0	0	**-0.24**	0.66	-0.24	-1.57;1.07	4870	592
0.9	0.4	4.0	0	**-0.26**	0.61	-0.25	-1.49;0.80	1758	366
0.9	0.4	3.0	0	**-0.25**	0.67	-0.25	-1.59;1.04	3302	505
0.9	0.4	2.0	0	**-0.24**	0.70	-0.25	-1.71;1.17	5008	601
0.95	0.6	4.0	0	**-0.24**	0.53	-0.24	-1.21;0.71	1692	358
0.95	0.6	3.0	0	**-0.23**	0.57	-0.24	-1.33;0.80	3138	491
0.95	0.6	2.0	0	**-0.23**	0.60	-0.24	-1.45;0.94	4683	584
0.8	0.6	3.0	0	**-0.23**	0.57	-0.24	-1.34;0.80	3145	492
0.8	0.6	2.0	0	**-0.23**	0.60	-0.24	-1.47;0.94	4695	586
0.8	0.5	3.0	0	**-0.24**	0.63	-0.24	-1.49;0.94	3239	499
0.8	0.5	2.0	0	**-0.24**	0.66	-0.24	-1.57;1.07	4872	592
0.8	0.4	3.0	0	**-0.24**	0.67	-0.25	-1.59;1.03	3304	505
0.8	0.4	2.0	0	**-0.24**	0.70	-0.25	-1.71;1.17	5010	601

B.3 Analysis of quantitative proteomics data – additional information

This section of the appendix includes additional analysis results regarding the three case studies and the development of a workflow for the analysis of quantitative proteomics data (see chapter 9 for further details).

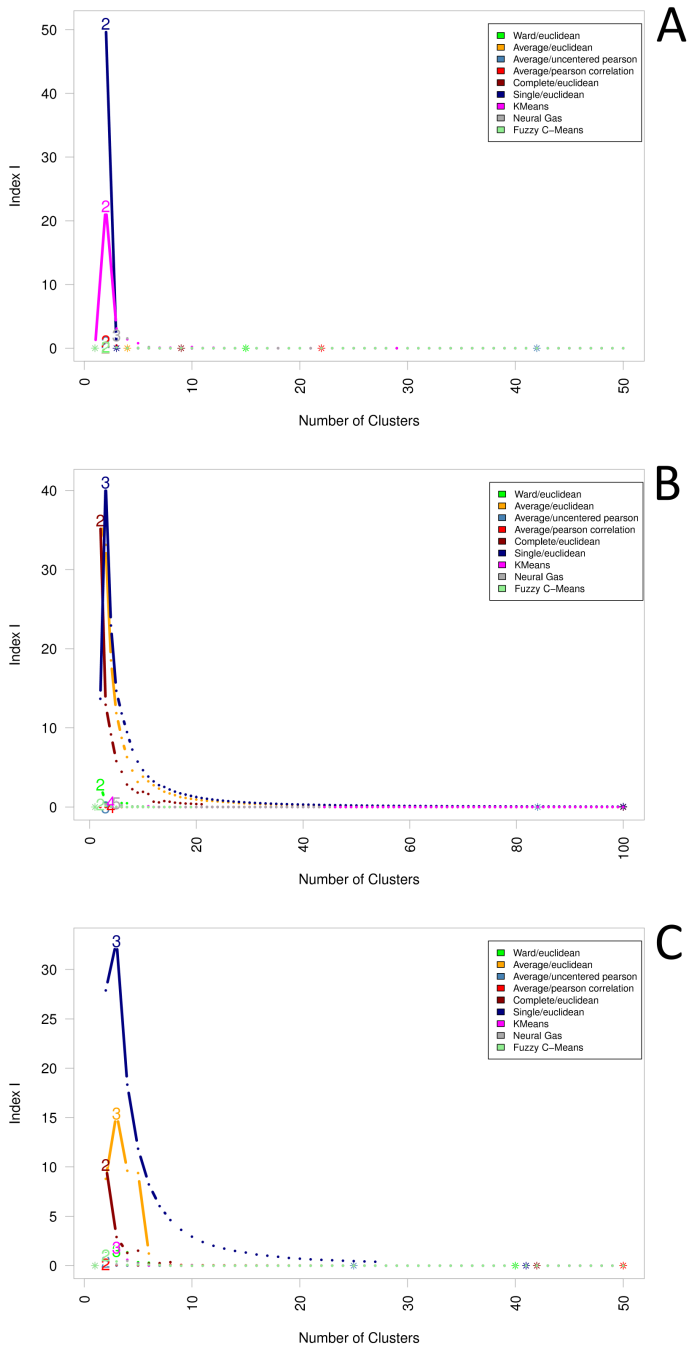


Figure B.1 – The cluster index "Index I" tends to favor smaller cluster numbers between two and three clusters. From a computational point of view this is clearly a good result. Unfortunately, from a biological point this does not allow any meaningful interpretation of the data. In general, these small clusterings only characterize individual outliers, while the rest of the clusters are found with a high number of cluster members having everything clustered together that reveals only a slight similarity.

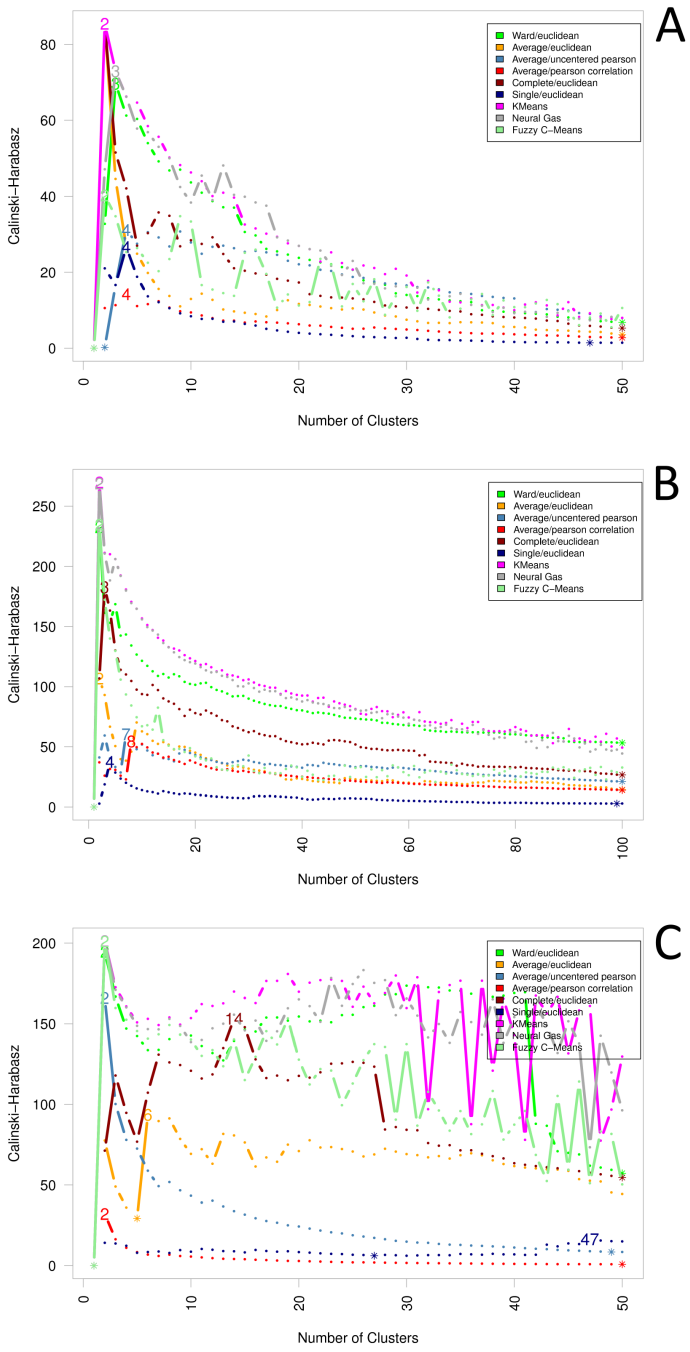


Figure B.2 – Similar to the "Index I" the cluster index of Calinski and Harabasz tends to favor smaller cluster numbers between three and four clusters. In the same manner, the applicability with respect to the biological question also remains questionable.

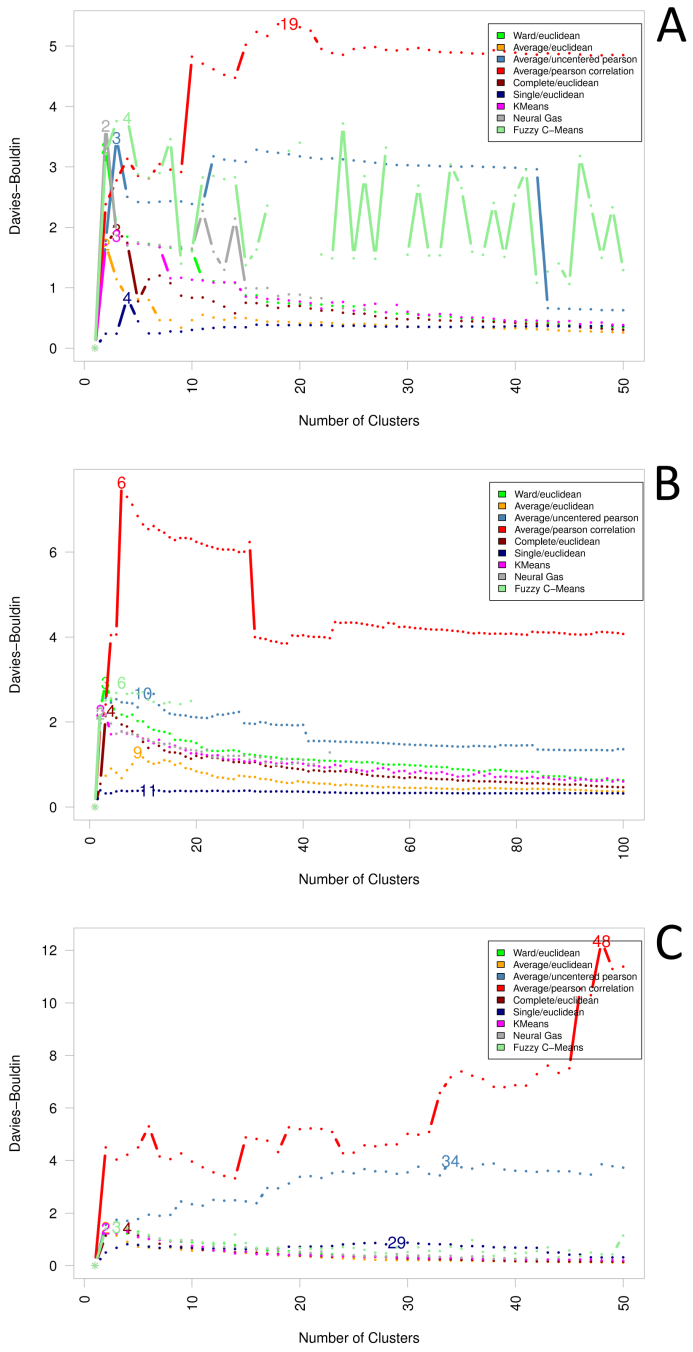


Figure B.3 – Instead of simply proposing a cluster index, Davies and Bouldin formulated a general framework for the evaluation of the outcomes of cluster algorithms. In contrast to other indexes, an optimal cluster solution is indicated by the minimal calculated index value. For instance, for the two cluster algorithms K-means and Neuralgas a local minimum can be located around the 30-cluster solution. A general interpretation of this index, however, seems to be difficult due to a strong tendency towards constantly decreasing index values with regard to large cluster numbers.

Glossary

2D-electrophoresis Two-dimensional electrophoresis—protein separation.

amino acids Small molecules that each consist of an amine group, a carboxylic acid group, and a variable side chain together bond to a central C-atom. Currently, 22 amino acids are known that constitute the building blocks of proteins all differing in size, form and charge.

CID Collision-induced dissociation refers to the utilization of a collision gas such as nitrogen to gain a defined fragmentation of a (peptide) ion. The masses of these fragments give hint to the peptide's amino acid structure.

corpuscles The physicist and Nobel prize-winner Thomson gave this name to the particles he discovered. Nowadays, these are better known as electrons.

DIGE Differential gel electrophoresis—samples are each labeled with a fluorescence dye such as Cy3 and Cy5. After 2D-electrophoresis, a scan process with different wavelengths in analogy to the utilized dyes then allows to visualize the differences in protein expression.

DNA Deoxyribonucleic acid is a macro molecule consisting of nucleotides and carrying the genetic information of every known living organism.

EIC In contrast to the total ion current (TIC), in the extracted ion chromatogram (also XIC) only the summed intensities of one distinct m/z value (or a small range of values) are used to reconstruct the elution of a specific analyte from a number of subsequently

recorded mass spectra. The EIC of a peptide, also termed its elution peak, provides a measure of the peptide's abundance.

ESI Electrospray ionization is a soft ionization technique that allows the generation of ions directly from dissolved molecules.

FWHM Full width at half maximum height; the resolution power R of a mass spectrometer is, for example, calculated as the ratio of a peak's mass to the peak's width at half maximum.

IEF Isoelectric focussing—under the influence of an electric field any mixture of zwitterionic compounds (ampholytes) rearranges itself in such a way that each individual compound of the mixture shifts to that position in this pH-gradient where it has a net charge of zero. If a protein is added to this gradient it also moves to the position where it is not electrically charged, its so called isoelectric point.

ion trap An ion trap is a type of mass analyzer that consists of a ring electrode as well as as two hyperboloid-shaped end cap electrodes. An electric field at the ring electrode forces incoming protein ions to traverse on a circular path—the ions are trapped. At a particular voltage the trajectory of all ions having a specific m/z value get unstable, and the ions are, figuratively speaking, thrown out of the trap..

isoelectric point With their amine and carboxylic acid group amino acids are able to react both as an acid and a base. There exists a certain pH level, where an amino acid is not electrically charged. Termed the isoelectric point this is unique for each kind of amino acid.

LC Liquid chromatography aims to separate a mixture of proteins by their specific properties such as size, charge or hydrophobicity.

LC-MS/MS The combination of liquid chromatography and tandem mass spectrometry.

m/z Mass to charge ratio—resultant from any mass spectrometry analysis is a list of peaks, each described by its intensity and mass to charge ratio.

MALDI Matrix-assisted laser desorption/ionization belongs to the group of soft ionization techniques. The sample and an UV-absorbing matrix compound are co-crystallized on a plate. Irradiation with an UV-laser results in matrix vaporization and sample ions movement into gas phase.

MIS MS/MS ion search utilizes the fragmentation pattern resultant from collision-induced dissociation to identify the amino acid structure of a peptide.

mRNA During the process of transcription the genes of the DNA are rewritten into a single-stranded so called messenger ribonucleic acid (mRNA).

MS Mass spectrometry allows to determine the molecular weight of molecules.

MudPIT Multi dimensional protein identification technology.

PMF Peptide mass fingerprinting takes advantage of the fact that a proteolytic enzyme such as trypsin cleaves a protein (or amino acid sequence) at specific positions. The enzyme produces defined fragments of a protein, so to speak its fingerprint, which allows for a precise protein identification.

ppm Parts per million—a measure often used to denote the accuracy of a mass spectrometer with which the mass of an ion has been determined.

precursor ion In tandem mass spectrometry, the precursor ion refers to the molecule that is subjected to dissociation. Generally, the term is used for any ion before its reaction e. g. with another molecule to form a particular product. The precursor ion is sometimes also called parent ion.

PTM The term post-translational modification summarizes all types of chemical modifications of a protein that occur after the protein has been assembled at the ribosomes. Since PTMs may influence the structure and function of a protein, they constitute an additional level of regulation in a cell.

quadrupole A quadrupole mass analyzer consists of four metal rods arranged in parallel. Utilizing an applied voltage on the metal rods an electric field is established that allows only those ions to pass the analyzer that have a certain m/z value.

RNA Ribonucleic acid is a macro molecule similar to DNA containing the sugar ribose instead of deoxyribose. It plays an important role in protein synthesis.

SDS-page Polyacrylamide gel electrophoresis based on sodium dodecyl sulfate (SDS).

TIC The total ion current denotes the sum of all intensities recorded across the full mass range of a spectrum. Given a number of subsequently recorded mass spectra, the TIC chromatogram provides an overview of the total intensities, and thereby the amounts of analytes, detected over time.

TOF Time-of-flight mass spectrometry denotes a type of analyzer. Masses are determined by measuring the time an ionized particle takes until it hits a detector.

tRNA Type of RNA—transfer RNA is an amino acid carrying helper molecule involved in translation. Each molecule is characterized by an anticodon sequence of 3 bases that matches with a corresponding mRNA sequence.

XIC see EIC.

XML The extensible markup language is a standardized language to store hierarchically structured data in a text file. A mayor advantage of the format is its handling simplicity due a range of available application programming interfaces.

zwitterion A molecule with at least two oppositely charged functional groups. Despite positively as well as negatively charged atoms the overall molecule is electrically neutral. Amino acids are a well-known example of twitterions in solids or polar solutions, e. g. in water.

Bibliography

Albaum, S. P., B. Linke, S. Jaenicke, J. Blom, N. Kessler, S. Juenemann, and A. Goesmann (2011). "Tools for Genome and Post-Genome Data Analysis Developed by the Technology Platform Bioinformatics (Poster abstract)". In: *5th European Conference on Prokaryotic and Fungal Genomics*. Göttingen, Germany.

Albaum, S. P., H. Neuweger, S. Lange, D. Mertens, K. Runte, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2008). "ProSE – "Software as a Service" for Quantitative Proteomics (Poster abstract)". In: *Human Proteome Organisation 7th Annual World Congress*. Amsterdam, Netherlands.

Albaum, S. P., H. Neuweger, B. Fränzel, S. Lange, D. Mertens, C. Trötschel, D. Wolters, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2009). Qupe–a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics* 25.23, pp. 3128–3134.

Albaum, S. P., H. Hahne, A. Otto, U. Haußmann, D. Becher, A. Poetsch, A. Goesmann, and T. W. Nattkemper (2011). A guide through the computational analysis of isotope-labeled mass spectrometry-based quantitative proteomics data: an application study. *Proteome Science* 9.1, p. 30.

Albaum, S., H. Neuweger, S. Lange, D. Mertens, J. Kalinowski, T. W. Nattkemper, and A. Goesmann (2009). "ProSE – a Rich Internet Application to securely Store, Organise, and Analyse Quantitative Proteomics Experiments (Poster abstract)". In: *17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference on Computational Biology*. Stockholm, Schweden.

Alberts, B. (2003). DNA replication and recombination. *Nature* 421.6921, pp. 431–435.

Alici, A. (2007). "Comprehensive Proteomics of Sorangium cellulosum So ce56". dissertation. Faculty of Biology: Bielefeld University.

Allaire, J. (2002). *Macromedia Flash MX - A next-generation rich client*. Tech. rep. Macromedia white paper.

Anderson, L. and J. Seilhamer (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18.3-4, pp. 533–537.

Anderson, N. L. and N. G. Anderson (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* 19.11, pp. 1853–1861.

Andreev, V. P., L. Li, T. Rejtar, Q. Li, J. G. Ferry, and B. L. Karger (2006). New algorithm for 15N/14N quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. *Journal of Proteome Research* 5.8, pp. 2039–2045.

Araki, T. (1992). An analysis of the effect of changes in growth temperature on proteolysis in vivo in the psychrophilic bacterium Vibrio sp. strain ANT-300. *Journal of General Microbiology* 138.10, pp. 2075–2082.

Arsova, B., S. Kierszniowska, and W. X. Schulze (2011). The use of heavy nitrogen in quantitative proteomics experiments in plants. *Trends in Plant Science*.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25.1, pp. 25–29.

Aston, F. W. (1922). *Mass Spectra and Isotopes—Nobel lecture in chemistry*. `http://www.nobel.se/chemistry/laureates/1922/aston-lecture.pdf`.

Avery, O. T., C. M. Macleod, and M. McCarty (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *Journal of Experimental Medicine* 79.2, pp. 137–158.

Bacher, J. (1996). *Clusteranalyse*. 2nd. Oldenbourg.

Bantscheff, M., M. Schirle, G. Sweetman, J. Rick, and B. Kuster (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* 389.4, pp. 1017–1031.

Bartels, D., S. Kespohl, S. Albaum, T. Drüke, A. Goesmann, J. Herold, O. Kaiser, A. Pühler, F. Pfeiffer, G. Raddatz, J. Stoye, F. Meyer, and S. C. Schuster (2005). BACCardI–a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. *Bioinformatics* 21.7, pp. 853–859.

Becker, C. H. and M. Bern (2011). Recent developments in quantitative proteomics. *Mutation Research* 722.2, pp. 171–182.

Belle, A., A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea (2006). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* 103.35, pp. 13004–13009.

Benölken, M. (2007). "JO-JavaO2DBI: Entwicklung eines Code Generators zur Erzeugung eines O2DBI kompatiblen Java Persistenz Layers". MA thesis. Bielefeld University.

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

Beynon, R. J. (2005). The dynamics of the proteome: strategies for measuring protein turnover on a proteome-wide scale. *Briefings in functional genomics & proteomics* 3.4, pp. 382–390.

Blom, J., S. P. Albaum, D. Doppmeier, A. Pühler, F.-J. Vorhölter, M. Zakrzewski, and A. Goesmann (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10, p. 154.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. 6. Auflage. Springer.

Bradshaw, R. A. (2005). Revised draft guidelines for proteomic data publication. *Molecular and Cellular Proteomics* 4.9, pp. 1223–1225.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 29.4, pp. 365–371.

Breukelen, B. van, H. W. P. van den Toorn, M. M. Drugan, and A. J. R. Heck (2009). StatQuant: a post-quantification analysis toolbox for improving quantitative mass spectrometry. *Bioinformatics* 25.11, pp. 1472–1473.

Broberg, P. (2012). *SAGx: Statistical Analysis of the GeneChip*. R package version 1.26.0.

Cabo-Bilbao, A., S. Spinelli, B. Sot, J. Agirre, A. E. Mechaly, A. Muga, and D. M. A. Guérin (2006). Crystal structure of the temperature-sensitive and allosteric-defective chaperonin GroELE461K. *Journal of Structural Biology* 155.3, pp. 482–492.

Calinski, R. B. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics* 3, pp. 1–27.

Cargile, B. J., J. L. Bundy, A. M. Grunden, and J. L. Stephenson (2004). Synthesis/degradation ratio mass spectrometry for measuring relative dynamic protein turnover. *Analytical Chemistry* 76.1, pp. 86–97.

Chair for computeroriented statistics and data analysis (2008). *rJava - Low-level R to Java interface*. `http://rosuda.org/rJava/`.

Chernushevich, I. V., A. V. Loboda, and B. A. Thomson (2001). An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry* 36.8, pp. 849–865.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum, XXI, 567 S.

Comisarow, M. B. and A. G. Marshall (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters* 25.2, pp. 282–283.

Cormack, R. (1971). A Review of Classification. *Journal of the Royal Statistical Society (Series A)* 134.3, pp. 321–367.

Cossio, J. F. de (2010). Computation of the isotopic distribution in two dimensions. *Analytical Chemistry* 82.15, pp. 6726–6729.

Côté, R. G., P. Jones, R. Apweiler, and H. Hermjakob (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 7, p. 97.

Coursey, J. S., D. J. Schwab, N. I. o. S. R. A. Dragoset, and M. Technology Gaithersburg (2005). *Atomic Weights and Isotopic Compositions (version 2.4.1)*. `http://physics.nist.gov/Comp`.

Cox, J. and M. Mann (2007). Is proteomics the new genomics? *Cell* 130.3, pp. 395–398.

— (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26.12, pp. 1367–1372.

Craig, R. and R. C. Beavis (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20.9, pp. 1466–1467.

Crawley, M. J. (2007). *Statistics - An Introduction using R*. Wiley.

Creasy, D. M. and J. S. Cottrell (2004). Unimod: Protein modifications for mass spectrometry. *Proteomics* 4.6, pp. 1534–1536.

Davies, D. L. and D. W. Bouldin (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1, pp. 224–227.

Desiere, F., E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M. G. Katze, K. A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J. A. Ranish, D. J. Rawlings, L. E. Samelson, Y. Shiio, J. D. Watts, B. Wollscheid, M. E. Wright, W. Yan, L. Yang, E. C. Yi, H. Zhang, and R. Aebersold (2005). Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology* 6.1, R9.

Dimitriadou, E. (2009). *cclust: Convex Clustering Methods and Clustering Indexes*. R package version 0.6-16.

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel (2011). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6.

Doherty, M. K., D. E. Hammond, M. J. Clague, S. J. Gaskell, and R. J. Beynon (2009). Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *Journal of Proteome Research* 8.1, pp. 104–112.

Dondrup, M., S. P. Albaum, T. Griebel, K. Henckel, S. Jünemann, T. Kahlke, C. K. Kleindt, H. Küster, B. Linke, D. Mertens, V. Mittard-Runte, H. Neuweger, K. J. Runte, A. Tauch, F. Tille, A. Pühler, and A. Goesmann (2009). EMMA 2–a MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics* 10, p. 50.

Du, P., W. A. Kibbe, and S. M. Lin (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22.17, pp. 2059–2065.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2000). *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*. Tech. rep. 578.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* 2.12, pp. 919–929.

Elias, J. E. and S. P. Gygi (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4.3, pp. 207–214.

Ellison, S. L. R., V. J. Barwick, and T. J. D. Farrant (2009). *A Practical Statistics for the Analytical Scientist: A Bench Guide 2nd Edition. A Bench Guide*. 2nd. The Royal Society of Chemistry.

Eng, J. K., A. L. McCormack, and J. R. Y. III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5.11, pp. 976–989.

Everitt, B. S., S. Landau, and M. Leese (2001). *Cluster Analysis*. Fourth Edition. Arnold.

Falconer, I. (1987). Corpuscles, Electrons and Cathode Rays: J. J. Thomson and the 'Discovery of the Electron'. *The British Journal for the History of Science* 20.3, pp. 241–276.

Farrah, T., E. W. Deutsch, and R. Aebersold (2011). Using the Human Plasma PeptideAtlas to study human plasma proteins. *Methods in molecular biology* 728, pp. 349–374.

Feng, H.-T., N. S. C. Wong, L. C. Sim, L. Wati, Y. Ho, and M. M. Lee (2010). Rapid characterization of high/low producer CHO cells using matrix-assisted laser desorption/ionization time-of-flight. *Rapid Communications in Mass Spectrometry* 24.9, pp. 1226–1230.

Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Society* 52, pp. 399–433.

Fligner, M. A. and T. J. Killeen (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of The American Statistical Association* 71.353, pp. 210–213.

Forgy, E. (1965). Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications. *Biometrics* 21, pp. 768–769.

Frank, E., M. S. Kessler, M. D. Filiou, Y. Zhang, G. Maccarrone, S. Reckow, M. Bunck, H. Heumann, C. W. Turck, R. Landgraf, and B. Hambsch (2009). Stable isotope metabolic labeling with a novel N-enriched bacteria diet for improved proteomic analyses of mouse models for psychopathologies. *PLoS One* 4.11, e7821.

Fränzel, B. (2010). "Targeting Integral Membrane Proteins in Quantitative Proteomics for Medical Applications and Biotechnological Issues by Corynebacterium glutamicum and Escherichia coli". dissertation. Faculty of Chemistry and Biochemistry: Ruhr-Universität Bochum.

Fränzel, B., C. Trötschel, C. Rückert, J. Kalinowski, A. Poetsch, and D. A. Wolters (2010). Adaptation of Corynebacterium glutamicum to salt-stress conditions. *Proteomics* 10.3, pp. 445–457.

Fränzel, B., A. Poetsch, C. Trötschel, M. Persicke, J. Kalinowski, and D. A. Wolters (2010). Quantitative proteomic overview on the Corynebacterium glutamicuml-lysine producing strain DM1730. *Journal of Proteomics*.

Garavelli, J. S. (2003). The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Research* 31.1, pp. 499–501.

Gärdén, P., R. Alm, and J. Häkkinen (2005). PROTEIOS: an open source proteomics initiative. *Bioinformatics* 21.9, pp. 2085–2087.

Garrett, J. J. (2005). *Ajax: A New Approach to Web Applications*. `http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications`.

Garwood, K., T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. P. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton (2004). PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* 5, p. 68.

Gau, R. (2008). "Bericht zum Projektmodul - Entwicklung eines lernfähigen Laboratory Information Management Systems". MA thesis. Bielefeld University.

Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research* 3.5, pp. 958–964.

Gerber, S. A., J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America* 100.12, pp. 6940–6945.

Glass, G. V., P. D. Peckham, and J. R. Sanders (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 42.3, pp. 237–288.

Goesmann, A., B. Linke, O. Rupp, L. Krause, D. Bartels, M. Dondrup, A. C. McHardy, A. Wilke, A. Pühler, and F. Meyer (2003). Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. *Journal of Biotechnology* 106.2-3, pp. 157–167.

Google (2011). *Google Web Toolkit*. `http://code.google.com/intl/de/webtoolkit/`.

Gordon, A. D. (1987). A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General)* 150.2, pp. 119–137.

Gottesman, S., E. Roche, Y. Zhou, and R. T. Sauer (1998). The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes & Development* 12.9, pp. 1338–1347.

Gouw, J. W., J. Krijgsveld, and A. J. R. Heck (2010). Quantitative proteomics by metabolic labeling of model organisms. *Molecular and Cellular Proteomics* 9.1, pp. 11–24.

Grasse, N., F. Mamedov, K. Becker, S. Styring, M. Rögner, and M. M. Nowaczyk (2011). Role of Novel Dimeric Photosystem II (PSII)-Psb27 Protein Complex in PSII Repair. *Journal of Biological Chemistry* 286.34, pp. 29548–29555.

Griffin, N. M., J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature Biotechnology* 28.1, pp. 83–89.

Guan, S., J. C. Price, S. B. Prusiner, S. Ghaemmaghami, and A. L. Burlingame (2011). A data processing pipeline for Mammalian proteome dynamics studies using stable isotope metabolic labeling. *Molecular and Cellular Proteomics* 10.12, p. M111.010728.

Gudgin, M., M. Hadley, N. Mendelsohn, J.-J. Moreau, H. F. Nielsen, A. Karmarkar, and Y. Lafon (2011). *SOAP Version 1.2*. `http://www.w3.org/TR/soap12-part1/`.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* 17.10, pp. 994–999.

Haegler, K., N. S. Mueller, G. Maccarrone, E. Hunyadi-Gulyas, C. Webhofer, M. D. Filiou, Y. Zhang, and C. W. Turck (2009). QuantiSpec–Quantitative mass spectrometry data analysis of (15)N-metabolically labeled proteins. *Journal of Proteomics* 71.6, pp. 601–608.

Hahne, H., U. Mäder, A. Otto, F. Bonn, L. Steil, E. Bremer, M. Hecker, and D. Becher (2010). A comprehensive proteomics and transcriptomics analysis of Bacillus subtilis salt stress adaptation. *Journal of Bacteriology* 192.3, pp. 870–882.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2002). Cluster Validity Methods: Part I & II. *SIGMOD Rec.* 31.2, pp. 40–45.

Höller, J. (2005). *Offenheit ist der Grund für den Erfolg von Spring*. `http:\www.entwickler.de/php/029881`.

Hamdan, M. and P. G. Righetti (2005). *Proteomics today - Protein assessment and biomarkers using mass spectrometry, 2D electrophoreses, and microarray technology*. Hoboken, NJ: Wiley-Interscience, XVII, 426 S. : Ill., graph. Darst.

Hamilton, G. (1997). *JavaBeans™API specification, Version 1.01A*. Tech. rep. Sun Microsystems.

Han, D. K., J. Eng, H. Zhou, and R. Aebersold (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology* 19.10, pp. 946–951.

Handl, J., J. Knowles, and D. Kell (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, pp. 3201–3212.

Hansen, M. E. and J. Smedsgaard (2004). A new matching algorithm for high resolution mass spectra. *Journal of the American Society for Mass Spectrometry* 15.8, pp. 1173–1180.

Hartigan, J. A. and M. A. Wong (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108.

Hartler, J., G. G. Thallinger, G. Stocker, A. Sturn, T. R. Burkard, E. Körner, R. Rader, A. Schmidt, K. Mechtler, and Z. Trajanoski (2007). MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics* 8, p. 197.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Haußmann, U. and A. Poetsch (in-press). Global proteome survey of protocatechuate- and glucose-grown Corynebacterium glutamicum reveals multiple physiological differences. *Journal of Proteomics*.

Haußmann, U., S.-W. Qi, D. Wolters, M. Rögner, S.-J. Liu, and A. Poetsch (2009). Physiological adaptation of Corynebacterium glutamicum to benzoate as alternative carbon source - a membrane proteome-centric view. *Proteomics* 9.14, pp. 3635–3651.

Hendrickson, E. L., Q. Xia, T. Wang, J. A. Leigh, and M. Hackett (2006). Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *The Analyst* 131.12, pp. 1335–1341.

Hennig, C. (2010). *fpc: Flexible procedures for clustering*. R package version 2.0-3.

Herman, C., D. Thévenet, P. Bouloc, G. C. Walker, and R. D'Ari (1998). Degradation of carboxy-terminal-tagged cytoplasmic proteins by the Escherichia coli protease HflB (FtsH). *Genes & Development* 12.9, pp. 1348–1355.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75.4, pp. 800–802.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics* 6.2, pp. 65–70.

Hoopmann, M. R., G. L. Finney, and M. J. MacCoss (2007). High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Analytical Chemistry* 79.15, pp. 5620–5632.

Hubel, D. H. and T. N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160, pp. 106–154.

Huber, L. A. (2003). Is proteomics heading in the wrong direction? *Nature Reviews Molecular Cell Biology* 4.1, pp. 74–80.

Hubert, L. and P. Arabie (1985). Comparing Partitions. *Journal of Classification* 2, pp. 193–218.

Hufnagel, P. and R. Rabus (2006). Mass spectrometric identification of proteins in complex post-genomic projects. Soluble proteins of the metabolically versatile, denitrifying 'Aromatoleum' sp. strain EbN1. *Journal of Molecular Microbiology and Biotechnology* 11.1-2, pp. 53–81.

James, P., M. Quadroni, E. Carafoli, and G. Gonnet (1993). Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications* 195.1, pp. 58–64.

Jayapal, K. P., S. Sui, R. J. Philp, Y.-J. Kok, M. G. S. Yap, T. J. Griffin, and W.-S. Hu (2010). Multitagging proteomic strategy to estimate protein turnover rates in dynamic systems. *Journal of Proteome Research* 9.5, pp. 2087–2097.

JBoss Inc., R. H. (2011). *Hibernate*. `http://www.hibernate.org`.

JFree.org (2011). *JFreeChart*. `http://www.jfree.org/jfreechart/`.

Jmol Entwicklerteam (2010). *Jmol: an open-source Java viewer for chemical structures in 3D*. `http://www.jmol.org/`.

Johnson, R. (2003). *Expert One-on-One J2EE Design and Development*. Wiley Publishing, Inc.

Kalinowski, J., B. Bathe, D. Bartels, N. Bischoff, M. Bott, A. Burkovski, N. Dusch, L. Eggeling, B. J. Eikmanns, L. Gaigalat, A. Goesmann, M. Hartmann, K. Huthmacher, R. Krämer, B. Linke, A. C. McHardy, F. Meyer, B. Möckel, W. Pfefferle, A. Pühler, D. A. Rey, C. Rückert, O. Rupp, H. Sahm, V. F. Wendisch, I. Wiegräbe, and A. Tauch (2003). The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *Journal of Biotechnology* 104.1-3, pp. 5–25.

Kaltschmidt, E. and H. G. Wittmann (1970). Ribosomal proteins. VII. Two-dimensional polyacrylamide gel electrophoresis for fingerprinting of ribosomal proteins. *Analytical Biochemistry* 36.2, pp. 401–412.

Kanehisa, M. and S. Goto (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28.1, pp. 27–30.

Karas, M., D. Bachmann, U. Bahr, and F. Hillenkamp (1987). Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes* 78, pp. 53–68.

Karpievitch, Y., J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W.-J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney (2009). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* 25.16, pp. 2028–2034.

Köster, C. and A. Holle (1999). "A new intelligent annotation procedure: SNAP". In: *ASMS 1999, Dallas, TX, USA, Poster MPA 003*.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data*. Wiley.

Kebarle, P. and U. H. Verkerk (2009). Electrospray: from ions in solution to ions in the gas phase, what we know now. *Mass Spectrometry Reviews* 28.6, pp. 898–917.

Keller, A., A. I. Nesvizhskii, E. Kolker, and R. Aebersold (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry* 74.20, pp. 5383–5392.

Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 26.3, pp. 231–243.

Knippers, R. (2001). *Molekulare Genetik*. 8. Auflage. Thieme.

Koch, M. (2008). "Interactive visualization and gene selection tool based on webservices". BA thesis. Bielefeld University.

Koehler, W., G. A. Schachtel, and P. Voleske (2002). *Biostatistik*. 3. Auflage. Springer.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the Institute of Electrical and Electronics Engineers* 78.9, pp. 1464–1480.

Krijgsveld, J., R. F. Ketting, T. Mahmoudi, J. Johansen, M. Artal-Sanz, C. P. Verrijzer, R. H. A. Plasterk, and A. J. R. Heck (2003). Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nature Biotechnology* 21.8, pp. 927–931.

Krzanowski, W. J. and Y. T. Lai (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics* 44, pp. 23–34.

Kumar, C. and M. Mann (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Letters*.

Lawrence, E. O. and M. S. Livingston (1931). A method for producing high speed hydrogen ions without the use of high voltages. *Physical review* 37 (12), p. 1707.

Levander, F., J. Hakkinen, G. Vincic, O. Mansson, and K. Warell (2009). The Proteios Software Environment - An extensible multi-user platform for management and analysis of proteomics data. *Journal of Proteome Research*.

Levin, Y. (2011). The role of statistical power analysis in quantitative proteomics. *Proteomics* 11.12, pp. 2565–2567.

Li, X. J., H. Zhang, A. Ranish, and R. Aebersold (2003). Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. *Analytical Chemistry* 75.23, pp. 6648–6657.

Lies, M. and M. R. Maurizi (2008). Turnover of endogenous SsrA-tagged proteins mediated by ATP-dependent proteases in Escherichia coli. *Journal of Biological Chemistry* 283.34, pp. 22918–22929.

Lilley, K. S. and P. Dupree (2006). Methods of quantitative proteomics and their application to plant organelle characterization. *Journal of Experimental Botany* 57.7, pp. 1493–1499.

Link, A. J., J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology* 17.7, pp. 676–682.

Linke, B. (2002). "O2DBI II – ein Persistenz-Layer für Perl-Objekte". MA thesis. Bielefeld University.

Liu, H., R. G. Sadygov, and J. R. Yates (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry* 76.14, pp. 4193–4201.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28.2, pp. 129–137.

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology* 25.1, pp. 117–124.

Lucas, A. and S. Jasson (2006). Using amap and ctc Packages for Huge Clustering. *R News* 6.5, pp. 58–60.

MacCoss, M. J., C. C. Wu, H. Liu, R. Sadygov, and J. R. Yates (2003). A Correlation Algorithm for the Automated Quantitative Analysis of Shotgun Proteomics Data. *Analytical Chemistry* 75.24, pp. 6912–6921.

MacQueen, J. (1965). "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. L. Cam and J. Neyman. Vol. 1. University of California Pr., pp. 281–297.

Maier, T., A. Schmidt, M. Güell, S. Kühner, A.-C. Gavin, R. Aebersold, and L. Serrano (2011). Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology* 7, p. 511.

Makarov (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry* 72.6, pp. 1156–1162.

Mallick, P. and B. Kuster (2010). Proteomics: a pragmatic perspective. *Nature Biotechnology* 28.7, pp. 695–709.

Mann, M. and M. Wilm (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry* 66.24, pp. 4390–4399.

Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, and R. Apweiler (2005). PRIDE: the proteomics identifications database. *Proteomics* 5.13, pp. 3537–3545.

Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch (2010). mzML - a Community Standard for Mass Spectrometry Data. *Molecular and Cellular Proteomics* 10, R110.000133.

Martinetz, T. M., S. Berkovich, and K. J. Schulten (1993). Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks* 4.4, pp. 558–569.

Martinsen, D. P. and B.-H. Song (1985). Computer applications in mass spectral interpretation: A recent review. *Mass Spectrometry Reviews* 4.4, pp. 461–490.

Matthiesen, R., ed. (2007). *Mass spectrometry data analysis in proteomics*. Vol. 367. Methods in molecular biology. Humana Press.

Matthiesen, R. (2007). Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* 7.16, pp. 2815–2832.

Maulik, U. and S. Bandyopadhyay (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12, pp. 1650–1654.

McCarty, M. (2003). Discovering genes are made of DNA. *Nature* 421.6921, p. 406.

McCormack, A. L., D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates (1997). Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Analytical Chemistry* 69.4, pp. 767–776.

McLuckey, S. A., G. J. V. Berkel, D. E. Goeringer, and G. L. Glish (1994). Ion trap mass spectrometry. Using high-pressure ionization. *Analytical Chemistry* 66.14, 737A–743A.

McQuitty, L. L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educational and Psychological Measurement* 26, pp. 825–831.

Mell, P. and T. Grance (2010). *The NIST Definition of Cloud Computing*. Tech. rep. Version 15. National Institute of Standards and Technology, Information Technology Laboratory.

Mertens, D. (2008). "Global quantitative proteomics by stable isotope labeling and tandem mass spectrometry". MA thesis. Bielefeld University.

Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, and A. Pühler (2003). GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research* 31.8, pp. 2187–2195.

Microsoft (2011). *Microsoft SQL Server*. `http://www.microsoft.com/sqlserver`.

— (2011). *The .NET Framework*. `http://www.microsoft.com/germany/net/`.

Milligan, G. W. and M. C. Cooper (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* 50.2, pp. 159–179.

Milligan, G. (1979). Ultrametric Hierarchical Clustering Algorithms. *Psychometrika* 44, pp. 343–246.

Moore, R. E., M. K. Young, and T. D. Lee (2002). Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry* 13.4, pp. 378–386.

Mueller, L. N., M.-Y. Brusniak, D. R. Mani, and R. Aebersold (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of Proteome Research* 7.1, pp. 51–61.

Mulder, G. J. (1839). Ueber die Zusammensetzung einiger thierischen Substanzen. *Journal für Praktische Chemie* 16.1, pp. 129–152.

Narayana, N., S. L. Ginell, I. M. Russu, and H. M. Berman (1991). Crystal and molecular structure of a DNA fragment: d(CGTGAATTCACG). *Biochemistry* 30.18, pp. 4449–4455.

Nelson, E. K., B. Piehler, J. Eckels, A. Rauch, M. Bellew, P. Hussey, S. Ramsay, C. Nathe, K. Lum, K. Krouse, D. Stearns, B. Connolly, T. Skillman, and M. Igra (2011). LabKey Server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics* 12, p. 71.

Nesvizhskii, A. I., A. Keller, E. Kolker, and R. Aebersold (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry* 75.17, pp. 4646–4658.

Nesvizhskii, A. I., O. Vitek, and R. Aebersold (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* 4.10, pp. 787–797.

Netterwald, J. (2007). Got MudPIT? *Drug Discovery & Development* 7.1, G4–G8.

Neubauer, G., A. Gottschalk, P. Fabrizio, B. Séraphin, R. Lührmann, and M. Mann (1997). Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 94.2, pp. 385–390.

Neuroth, H., A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth, eds. (2009). *Nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.0, Juni 2009. Boizenburg : Hülsbusch; Göttingen : Univ.-Verl. Göttingen.

Neuweger, H., J. Baumbach, S. Albaum, T. Bekel, M. Dondrup, A. T. Hüser, J. Kalinowski, S. Oehm, A. Pühler, S. Rahmann, J. Weile, and A. Goesmann (2007). CoryneCenter – an online resource for the integrated analysis of corynebacterial genome and transcriptome data. *BMC Systems Biology* 1, p. 55.

Neuweger, H., S. P. Albaum, M. Dondrup, M. Persicke, T. Watt, K. Niehaus, J. Stoye, and A. Goesmann (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics* 24.23, pp. 2726–2732.

Neuweger, H., M. Persicke, S. P. Albaum, T. Bekel, M. Dondrup, A. T. Hüser, J. Winnebald, J. Schneider, J. Kalinowski, and A. Goesmann (2009). Visualizing post genomics data-sets on customized pathway maps by ProMeTra—aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example. *BMC Systems Biology* 3, p. 82.

NextApp, Inc. (2011). *Echo Web Framework*. `http://echo.nextapp.com`.

Nolting, D. (2008). *Isotopic Pattern Calculator*. `http://isotopatcalc.sourceforge.net`.

Object Management Group, I. (2008). *OMG Model Driven Architecture*. `http://www.omg.org/mda/`.

Oda, Y., K. Huang, F. R. Cross, D. Cowburn, and B. T. Chait (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America* 96.12, pp. 6591–6596.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry* 250.10, pp. 4007–4021.

Ong, S.-E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics* 1.5, pp. 376–386.

Oracle (2011). *Java EE Compatibility*. `http://java.sun.com/javaee/overview/compatibility.jsp`.

— (2011). *MySQL*. `http://www.mysql.com`.

— (2011). *Oracle Database*. `http://www.oracle.com`.

— (2011). *Oracle Grid Engine*. `http://www.oracle.com/technetwork/oem/grid-engine-166852.html`.

Orchard, S., H. Hermjakob, and R. Apweiler (2003). The proteomics standards initiative. *Proteomics* 3.7, pp. 1374–1376.

Orchard, S., H. Hermjakob, R. K. Julian, K. Runte, D. Sherman, J. Wojcik, W. Zhu, and R. Apweiler (2004). Common interchange standards for proteomics data: Public availability of tools and schema. *Proteomics* 4.2, pp. 490–491.

Otto, A., J. Bernhardt, H. Meyer, M. Schaffer, F.-A. Herbst, J. Siebourg, U. Mäder, M. Lalk, M. Hecker, and D. Becher (2010). Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis. *Nature Communications* 1, p. 137.

Pan, C., G. Kora, D. L. Tabb, D. A. Pelletier, W. H. McDonald, G. B. Hurst, R. L. Hettich, and N. F. Samatova (2006). Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Analytical Chemistry* 78.20, pp. 7110–7120.

Pappin, D. J., P. Hojrup, and A. J. Bleasby (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* 3.6, pp. 327–332.

Park, S. K., J. D. Venable, T. Xu, and J. R. Yates (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. *Nature Methods* 5.4, pp. 319–322.

Parker, C. M., E. N. Wafula, P. M. C. Swatman, and P. A. Swatman (1994). "Information Systems Research Methods: The Technology Transfer Problem". In: *ACIS 1994 Proceedings*.

Parvin, L., M. C. Galicia, J. M. Gauntt, L. M. Carney, A. B. Nguyen, E. Park, L. Heffernan, and A. Vertes (2005). Electrospray diagnostics by Fourier analysis of current oscillations and fast imaging. *Analytical Chemistry* 77.13, pp. 3908–3915.

Paul, W. and H. Steinwedel (1953). Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift fuer Naturforschung, A: Physical Sciences* 8, pp. 448–450.

Paul, W. and H. Steinwedel (1960). "Apparatus for separating charged particles of different specific charges". Pat. US2939952 (A).

Pearson, H. (2003). DNA: Beyond the double helix. *Nature* 421.6921, pp. 310–312.

— (2008). Biologists initiate plan to map human proteome. *Nature* 452.7190, pp. 920–921.

Pedrioli, P. G. A., J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology* 22.11, pp. 1459–1466.

Peng, J., J. E. Elias, C. C. Thoreen, L. J. Licklider, and S. P. Gygi (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research* 2.1, pp. 43–50.

Perkins, D., D. Pappin, D. Creasy, and J. Cottrell (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20.18, pp. 3551–3567.

Poetsch, A., U. Haussmann, and A. Burkovski (2011). Proteomics of corynebacteria: From biotechnology workhorses to pathogens. *Proteomics* 11.15, pp. 3244–3255.

Polpitiya, A. D., W.-J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. G. Camp, G. A. Anderson, and R. D. Smith (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 24.13, pp. 1556–1558.

PostgreSQL-Team (2011). *PostgreSQL*. `http://www.postgresql.org`.

Pratt, J. M., J. Petty, I. Riba-Garcia, D. H. L. Robertson, S. J. Gaskell, S. G. Oliver, and R. J. Beynon (2002). Dynamics of protein turnover, a missing dimension in proteomics. *Molecular and Cellular Proteomics* 1.8, pp. 579–591.

Price, J. C., S. Guan, A. Burlingame, S. B. Prusiner, and S. Ghaemmaghami (2010). Analysis of proteome dynamics in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* 107.32, pp. 14508–14513.

Proteomics Informatics Standards Group (2011). *mzIdentML*. `http://psidev.info/index.php?q=node/319`.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Ramos, H., P. Shannon, and R. Aebersold (2008). The protein information and property explorer: an easy-to-use, rich-client web application for the management and functional analysis of proteomic data. *Bioinformatics* 24.18, pp. 2110–2111.

Rao, P. K., G. M. Rodriguez, I. Smith, and Q. Li (2008). Protein dynamics in iron-starved Mycobacterium tuberculosis revealed by turnover and abundance measurement using hybrid-linear ion trap-Fourier transform mass spectrometry. *Analytical Chemistry* 80.18, pp. 6860–6869.

Rauch, A., M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman, P. Hussey, M. Igra, B. Maclean, C. W. Lin, A. Detter, R. Fang, V. Faca, P. Gafken, H. Zhang, J. Whiteaker, J. Whitaker, D. States, S. Hanash, A. Paulovich, and M. W. McIntosh (2006). Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *Journal of Proteome Research* 5.1, pp. 112–121.

Rehm, H. (2006). *Der Experimentator - Proteinbiochemie/Proteomics*. Elsevier GmbH Spektrum Akademischer Verlag.

Reidegeld, K. A., M. Eisenacher, M. Kohl, D. Chamrad, G. Körting, M. Blüggel, H. E. Meyer, and C. Stephan (2008). An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. *Proteomics* 8.6, pp. 1129–1137.

Rigbolt, K. T. G., J. T. Vanselow, and B. Blagoev (2011). GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Molecular and Cellular Proteomics* 10.8, O110.007450.

Rocke, D. M. (2004). Design and analysis of experiments with high throughput biological assay data. *Seminars in Cell & Developmental Biology* 15.6, pp. 703–713.

Rockwood, A. L. and S. L. V. Orden (1996). Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry* 68.13, pp. 2027–2030.

Rode, C., M. Senkler, J. Klodmann, T. Winkelmann, and H.-P. Braun (2011). GelMap–a novel software tool for building and presenting proteome reference maps. *Journal of Proteomics* 74.10, pp. 2214–2219.

Roepstorff, P. and J. Fohlman (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry* 11.11, p. 601.

Ross, P. L., 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 (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics* 3.12, pp. 1154–1169.

Sabidó, E., N. Selevsek, and R. Aebersold (2011). Mass spectrometry-based proteomics for systems biology. *Current Opinion in Biotechnology*.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36.8, pp. 1627–1639.

Schäfer, R. (2009). ultrafleXtreme: Redefining MALDI-TOF-TOF Mass Spectrometry Performance. *The Applicationsbook*, pp. 1–2.

Schmidt, A., J. Kellermann, and F. Lottspeich (2005). A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* 5.1, pp. 4–15.

Schröder, S. (2010). "Entwicklung, Implementierung und Optimierung von Verfahren zur quantitativen Analyse von stabilisotop markierten, massenspektrometrischen Protein-Daten". MA thesis. Bielefeld University.

Schuler, G. D., J. A. Epstein, H. Ohkawa, and J. A. Kans (1996). Entrez: molecular biology database and retrieval system. *Methods in Enzymology* 266, pp. 141–162.

Shapiro, S. S. and M. B. Wilk (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52.3/4, pp. 591–611.

Smedley, D., S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk (2009). BioMart–biological queries made easy. *BMC Genomics* 10, p. 22.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, O. B. I. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25.11, pp. 1251–1255.

Sneath, P. H. A. and R. R. Sokal (1973). *Numerical taxonomy - the principles and practice of numerical classification*. Freeman.

Snider, R. K. (2007). Efficient calculation of exact mass isotopic distributions. *Journal of the American Society for Mass Spectrometry* 18.8, pp. 1511–1515.

Sokal, R. R. and C. D. Michener (1958). A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas science bulletin* 38, pp. 1409–1438.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15.1, pp. 72–101.

Sperling, E., A. E. Bunner, M. T. Sykes, and J. R. Williamson (2008). Quantitative analysis of isotope distributions in proteomic mass spectrometry using least-squares Fourier transform convolution. *Analytical Chemistry* 80.13, pp. 4906–4917.

SpringSource, a division of VMware (2011). *Spring Framework*. `http://www.springsource.org`.

Sprinson, D. B. and D. Rittenberg (1949). The rate of interaction of the amino acids of the diet with the tissue proteins. *Journal of Biological Chemistry* 180, pp. 715–726.

Stein, S. E. and D. R. Scott (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* 5, pp. 859–866.

Stephan, C., M. Kohl, M. Turewicz, K. Podwojski, H. E. Meyer, and M. Eisenacher (2010). Using Laboratory Information Management Systems as central part of a proteomics data workflow. *Proteomics* 10.6, pp. 1230–1249.

Storz, G. (2002). An expanding universe of noncoding RNAs. *Science* 296.5571, pp. 1260–1263.

Tanaka, K., H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo (1988). 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, pp. 151–153.

Tang, K., J. S. Page, and R. D. Smith (2004). Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* 15.10, pp. 1416–1423.

Tanner, S., H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna (2005). InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry* 77.14, pp. 4626–4639.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4.41, pp. 1–14.

Taylor, C. F., H. Hermjakob, R. K. Julian, J. S. Garavelli, R. Aebersold, and R. Apweiler (2006). The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS* 10.2, pp. 145–151.

Taylor, C. F., N. W. Paton, K. S. Lilley, P.-A. Binz, R. K. Julian, A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove,

T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, and H. Hermjakob (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* 25.8, pp. 887–893.

Taylor, C. F., P.-A. Binz, R. Aebersold, M. Affolter, R. Barkovich, E. W. Deutsch, D. M. Horn, A. Hühmer, M. Kussmann, K. Lilley, M. Macht, M. Mann, D. Müller, T. A. Neubert, J. Nickson, S. D. Patterson, R. Raso, K. Resing, S. L. Seymour, A. Tsugita, I. Xenarios, R. Zeng, and J. Randall K. Julian (2008). *MIAPE: Mass spectrometry*. `http://www.psidev.info/miape/ms/`. Version 2.24.

Taylor, C. F., D. Field, S.-A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P.-A. Binz, M. Bogue, T. Booth, A. Brazma, R. R. Brinkman, A. M. Clark, E. W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, G. Grimes, J. M. Hancock, N. W. Hardy, H. Hermjakob, R. K. Julian, M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuiper, N. L. Novère, J. Leebens-Mack, S. E. Lewis, P. Lord, A.-M. Mallon, N. Marthandan, H. Masuya, R. McNally, A. Mehrle, N. Morrison, S. Orchard, J. Quackenbush, J. M. Reecy, D. G. Robertson, P. Rocca-Serra, H. Rodriguez, H. Rosenfelder, J. Santoyo-Lopez, R. H. Scheuermann, D. Schober, B. Smith, J. Snape, C. J. Stoeckert, K. Tipton, P. Sterk, A. Untergasser, J. Vandesompele, and S. Wiemann (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* 26.8, pp. 889–896.

The Apache Software Foundation (2011). *Struts*. `http://struts.apache.org`.

— (2011). *Tomcat*. `http://tomcat.apache.org`.

The Global Proteome Machine Organization (2011). *The common Repository of Adventitious Proteins*. `http://www.thegpm.org/crap/index.html`.

The UniProt Consortium (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research* 36.suppl 1, pp. D190–D195.

Tobias, J. W., T. E. Shrader, G. Rocap, and A. Varshavsky (1991). The N-end rule in bacteria. *Science* 254.5036, pp. 1374–1377.

Toepel, J., S. P. Albaum, S. Arvidsson, A. Goesmann, M. la Russa, K. Rogge, and O. Kruse (2011). Construction and evaluation of a whole genome microarray of Chlamydomonas reinhardtii. *BMC Genomics* 12, p. 579.

Trötschel, C., C. Lange, S. Albaum, A. Goesmann, R. Krämer, and K. Marin (2011). "Oxidative Stress Response in Corynebacterium glutamicum – the Proteome in Focus (Poster abstract)". In: *5th European Conference on Prokaryotic and Fungal Genomics*. Göttingen, Germany.

Trötschel*, C., S. P. Albaum*, D. Wolff, S. Schröder, A. Goesmann, T. W. Nattkemper, M. Rögner, and A. Poetsch (2012). Protein turnover quantification in a multi-labeling approach–from data calculation to evaluation. *Molecular and Cellular Proteomics* 11.8. (*contributed equally), pp. 512–526.

Tufte, E. R. (2007). *The visual display of quantitative information*. 2nd ed. Graphics Press.

Unlü, M., M. E. Morgan, and J. S. Minden (1997). Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 18.11, pp. 2071–2077.

Vizcaíno, J. A., R. Côté, F. Reisinger, J. M. Foster, M. Mueller, J. Rameseder, H. Hermjakob, and L. Martens (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 9.18, pp. 4276–4283.

Walesiak, M. and A. Dudek (2012). *clusterSim: Searching for optimal clustering procedure for a data set*. R package version 0.41-5.

Ward Joe H., J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of The American Statistical Association* 58.301, pp. 236–244.

Watson, J. D. and F. H. Crick (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171.4356, pp. 737–738.

Westermeier, R., T. Naven, and H.-R. Höpker (2008). *Proteomics in Practice–A Guide to Successful Experimental Design*. Weinheim: Wiley-VCH.

Westermeyer, M. (2008). "Implementierung von Wizards für ein Laboratory Information Management System in einer quantitativen Proteomanalyse-Plattform". BA thesis. Bielefeld University.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 36.Database issue, pp. D13–D21.

Whitehouse, C. M., R. N. Dreyer, M. Yamashita, and J. B. Fenn (1985). Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry* 57.3, pp. 675–679.

Wilke, A., C. Ruckert, D. Bartels, M. Dondrup, A. Goesmann, A. Huser, S. Kespohl, B. Linke, M. Mahne, A. McHardy, A. Pühler, and F. Meyer (2003). Bioinformatics support for high-throughput proteomics. *Journal of Biotechnology* 106, pp. 147–156.

Wilkins, M. R., J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology & Genetic Engineering Reviews* 13, pp. 19–50.

Wolters, D., M. Washburn, and J. Yates (2001). An automated multidimensional protein identification technology for shotgun proteomic. *Analytical Chemistry* 73.23, pp. 5683–5690.

Yang, C., Z. He, and W. Yu (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics* 10, p. 4.

Yates, J. R., J. K. Eng, A. L. McCormack, and D. Schieltz (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry* 67.8, pp. 1426–1436.

Yergey, J., D. Heller, G. Hansen, R. J. Cotter, and C. Fenselau (1983). Isotopic distributions in mass spectra of large molecules. *Analytical Chemistry* 55.2, pp. 353–356.

Yeung, K., D. Haynor, and W. Ruzzo (2001). Validating clustering for gene expression data. *Bioinformatics* 17, pp. 309–318.

Yin, Y. W. and T. A. Steitz (2004). The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell* 116.3, pp. 393–404.

Yost, R. A. and C. G. Enke (1978). Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society* 100, pp. 2274–2275.

Zhang, N., R. Aebersold, and B. Schwikowski (2002). ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2.10, pp. 1406–1412.

Zhang, Y., C. Webhofer, S. Reckow, M. D. Filiou, G. Maccarrone, and C. W. Turck (2009). A MS data search method for improved 15N-labeled protein identification. *Proteomics* 9.17, pp. 4265–4270.

Zhang, Y., S. Reckow, C. Webhofer, M. Boehme, P. Gormanns, W. M. Egge-Jacobsen, and C. W. Turck (2011). Proteome Scale Turnover Analysis in Live Animals Using Stable Isotope Metabolic Labeling. *Analytical Chemistry*.

Zhu, H., S. Pan, S. Gu, E. M. Bradbury, and X. Chen (2002). Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Communications in Mass Spectrometry* 16.22, pp. 2115–2123.

Zhu, W., J. W. Smith, and C.-M. Huang (2010). Mass spectrometry-based label-free quantitative proteomics. *Journal of Biomedicine & Biotechnology* 2010, p. 840518.

Danksagung

Zunächst möchte ich mich ganz herzlich bei meinen Betreuern Prof. Dr.-Ing. Tim Wilhelm Nattkemper und Dr. Alexander Goesmann bedanken, die mir die Gelegenheit gegeben haben, diese Arbeit zu verwirklichen, mir stets motivierend zur Seite standen, und mich mit unzähligen Anregungen und Ratschlägen unterstützt haben.

Außerdem gilt mein Dank PD Dr. Ansgar Poetsch für die Bereitschaft, diese Arbeit als externer Gutachter zu beurteilen.

Vielen Dank auch an die Systemadministratoren der Bioinformatics Resource Facility für die Bereitstellung und Betreuung von Resourcen und zahlreichen Problemlösungen bei Hard- und Softwareproblemen.

Mein besonderer Dank gebührt darüber hinaus allen Mitgliedern der Arbeitsgruppe Computational Genomics, vor allem meinem ehemaligen Kollegen Dr. Heiko Neuweger für zahlreiche Diskussionen und Anregungen nicht nur zum Thema Massenspektrometrie. Danke für das sehr angenehme Arbeitsklima in dieser Arbeitsgruppe.

Danke auch an Sita Lange, Dominik Mertens, Mirko Westermeyer und Simon Schroeder für Eure Beiträge im Rahmen von Hilfskraftstellen, Bachelor- und Masterarbeiten.

Für die gute Zusammenarbeit danke ich allen Kolleginnen und Kollegen aus den Laboren, insbesondere Dr. Christian Trötschel, Ute Haußmann, Dr. Benjamin Fränzel, Dr. Dirk Wolters, Dr. Andreas Otto, Hannes Hahne, Dr. Dörte Becher, Benjamin Müller, Karin Gorzolka, Anja Bonte, Yaarub Al-Hussuna und Julia Beckmann. Vielen Dank für die Bereitstellung von Messdaten, ohne die viele Methoden in QuPE wohl nur graue Theorie wären, und vielmehr noch für Eure vielfältigen Anregungen, die einen essentiellen Beitrag zur Entwicklung und Optimierung dieser Arbeit geliefert haben.

Ich danke des Weiterem dem Bundesministerium für Bildung und Forschung (BMBF) für die finanzielle Unterstützung.

Vielen Dank auch allen Korrekturleserinnen und -lesern, vor allem meiner Schwester Katrin.

Mein größter Dank gilt schließlich meinen Freunden, meinen Eltern und meiner Familie, und ganz besonders Linda. Ohne Euch wäre diese Arbeit nicht möglich gewesen. Danke für Eure motivationale und seelische Unterstützung, danke dass Ihr immer für mich da gewesen seid.

Gedruckt auf alterungsbeständigem holz- und säurefreiem Papier nach DIN-ISO 9706.

Hiermit erkläre ich, Stefan P. Albaum, die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben. Alle Ausführungen, die ich wörtlich oder sinngemäß aus den Werken anderer Autoren entnommen habe, wurden ausdrücklich als solche gekennzeichnet.

Bielefeld, im April 2012

Stefan P. Albaum