Statistical considerations for the design and interpretation of proteomics experiments

Bram Burger

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2021

UNIVERSITY OF BERGEN

Statistical considerations for the design and interpretation of proteomics experiments

Bram Burger

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 10.06.2021

Year: 2021

Title: Statistical considerations for the design and interpretation of proteomics experiments

Name: Bram Burger

Print: Skipnes Kommunikasjon / University of Bergen

Scientific Environment

The work has been carried out at the Computational Biology Unit (CBU), Institute of Informatics, Faculty of Mathematics and Natural Sciences, and at The Proteomics Unit at the University of Bergen (PROBE), Department of Biomedicine. Dr. Harald Barsnes has been the main supervisor. Dr. Marc Vaudel and Prof. Frode S. Berven have been co-supervisors.

The candidate was associated with NORBIS, the National Research School in Bioinformatics, Biostatistics and Systems Biology. In addition to the research, the PhD fellowship included 25% teaching duties spread throughout the four years.

Acknowledgements

To begin by stating the obvious: while it is only my name on the cover, this would not have been possible without the support and kindness of those around me.

I would firstly like to thank Harald for all the stimulating discussions, the support, and hikes to beautiful places I would otherwise never have seen. I would also like to thank Marc and Frode for the support and discussions from a slightly more distant viewpoint. The people in the Barsnes group, and at CBU and PROBE for the stimulating conversations and kind environment. Especially the great team of engineers at PROBE for all the conversations about the machines and the chemistry and biology. Yehia and Francisco for being wonderful and always able to put a smile on my face. And Olav for also being the kindest landlord, making these trying times much easier than it could have been. And finally, my family and friends for the endless moral support.

Thanks everyone!

Abstract

The methods to study proteins are continuously improving, making it possible to identify and study increasingly more proteins. It is important to keep studying and improving the way experiments are designed, performed, and interpreted. This thesis concerns statistical considerations for the design and interpretation of proteomics experiments using mass spectrometry. Mainly the focus is on protein and pathway interactions and networks, extending an existing workflow to additionally identify very small peptides and single amino acids, and effectively making balanced batches in a fixed cohort.

The biological functions of proteins are determined by their interactions with other molecules. While most proteins have a limited number of specific interaction partners, the whole system of these interactions is part of what defines organisms. Interactions between proteins, and with other molecules, can be grouped in different types of reactions. A set of actions and reactions leading to a specific outcome is a pathway, which defines a, more or less specific, process. Results from statistical analyses of protein abundances are often put in a biological context by way of mapping proteins to these pathways, i.e. by pathway analysis.

The interactions between proteins and the definitions of pathways are collected in pathway databases, based on available literature or computational inference. Due to the central role these databases play in the interpretation of proteomics experiments, they can directly influence the outcome of pathway analysis algorithms. We investigated the structure and evolution of one of the manually curated pathway databases, by using network statistics focusing on the connectivity of the annotated proteins and the hierarchical nature of the pathways. Additionally, we hope to aid in improving the understanding of the underlying basis for pathway analysis results and give practical advice for their interpretation.

Several pathways involve the digestion of proteins. This is done not only to extract energy from food, but also in order to modify proteins by proteolytic cleavage or to recycle proteins so their amino acids can be reused in other proteins. Additionally, small peptides and single amino acids can have functions of their own and play an important role in maintaining homeostasis. Differential abundance of endogenous (i.e. naturally occurring) peptides can thus potentially be used as an indication of possible dysfunctional proteases, peptidases, or proteolytic pathways.

Using mass spectrometry only molecules that can attract a charge will be possible to detect. Many amino acids and very small peptides are not able to attract a charge under standard conditions in mass spectrometry. Additionally, singly charged molecules are often ignored as they are likely to be contamination. While there are kits to specifically detect single amino acids and some small peptides using mass spectrometry, these are not easily incorporated into standard workflows. In a proof-of-concept experiment we used isobaric tags in a standard peptidomics workflow. As the tags acquire a charge themselves, this makes it possible to detect the single amino acids and very small peptides when singly charged molecules are also selected for fragmentation. The identification of the isobaric tag in the MS^2 spectra serves as a rudimentary check that the molecule is not a contaminant. This simple addition to a standard workflow, combined with a naive mass-based identification approach, showed that with a limited amount of extra work a large amount of single amino acids and small endogenous peptides could be identified, which could lead to potential novel insights and understanding.

Processing of samples in biomedical experiments generally proceeds sequentially. Due to time and technical limitations, the number of samples that can be processed at once is often limited to less than the number of available samples, which means that at least some steps of the experiment will have to be processed in different batches. It then becomes important to make sure that the batching process does not introduce confounders. For smaller experiments this is feasible to do by hand for expert scientists, but quickly becomes cumbersome and time-consuming for large and imbalanced cohort set-ups.

To aid in the process of designing batch allocations that are as balanced as possible, given the available samples, we have developed a fast and intuitive heuristic algorithm, which can be applied to single-variable model where the treatment variable is nominal. This automated procedure can free researchers to focus on other aspects of the experiment, and provides a marked improvement over a more naive random allocation procedure.

Contents

List of Tables

List of Figures

1 Proteomics

This thesis concerns statistical considerations when designing proteomics experiments using mass spectrometry. First, a short introduction to the main biological concepts important for the included papers will be provided.

1.1 Proteins

Living cellular organisms share the same fundamental building blocks and machinery for basic functions. Some species consist of a single cell, e.g. bacteria, while others consist of many cells, e.g. humans consist of more than 10^{13} cells, though start out as a single cell (Alberts et al., 2015). The largest constituent of a cell, after water, is the proteins, the molecules that are responsible for most of the biochemical activity. For example, enzymes are proteins with catalytic activity, while other proteins function as signal receptors, transporters, or structural elements. Proteins perform their functions by interacting with other proteins and other molecules. The proteins that make up the functioning of these systems are essential for a better understanding of the processes in the cell, and the diversity between cells and organisms (Twyman, 2014).

The proteins themselves are built up from amino acids, of which 20 are directly coded for in the genetic material (Table 1.1). Each amino acid consists of a carbon atom with a hydrogen atom, a carboxyl group (the C-terminal), an amino group (the N-terminal), and a side chain (Figure 1.1). The side chain defines the amino acid (Alberts et al., 2015; Nelson and Cox, 2017). When two or more amino acids bind to each other, they form a peptide. In this process the N-terminal of one amino acid connects to the C-terminal of the other amino acid, and a molecule of water is released (Figure 1.2). A peptide consisting of two amino acids is also called a dipeptide, tripeptides consist of three amino acids, and so on. In general, a peptide containing many amino

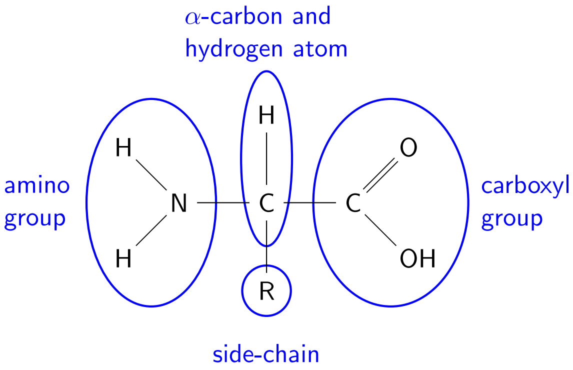


Figure 1.1: Basic structure of amino acids. Each amino acid consists of an α-carbon with a hydrogen atom, an amino group (H_2N), a carboxyl group and a side chain. The side chain is specific for each amino acid.

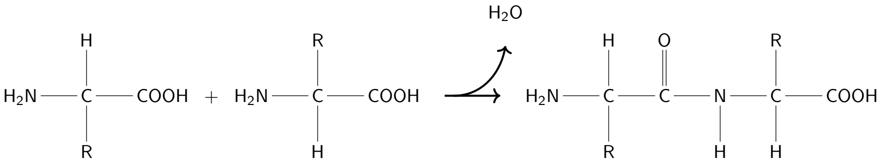


Figure 1.2: Peptide bond. When an amino acid is added to another amino acid (or a chain of amino acids), the N-terminal of the new amino acid connects to the C-terminal of the amino acid (chain). In this process one water molecule is released.

acids is called a polypeptide, or a polypeptide chain, which mass can be calculated as the sum of the residual masses of the constituent amino acids plus a water molecule. A protein consists of one or more such polypeptide chains.

Proteins are translated from ribonucleic acid (RNA), which itself is transcribed from deoxyribonucleic acid (DNA) (Alberts et al., 2015). Going from DNA (the genome) to RNA (transcriptome) to proteins (proteome) from an omics point of view, each step is an order of magnitude more complex. There are roughly 2×10^4 human genes, which are transcribed into about 10^5 transcripts. The translation of an mRNA transcript gives the amino acid sequence of a protein, which is the base form of the protein. Due to the flexible backbone and the different

Table 1.1: Names and key properties of the 20 amino acids directly coded for in DNA. Masses are residual masses. The mass of a peptide is the sum of the amino acid residues plus H_2O. Hydropathy index taken from Kyte and Doolittle (1982).

Amino acid	Abbreviation		Formula (Residual)	Hydropathy index	Monoisotopic Mass (Da)
	3-Letter	1-Letter			
Isoleucine	Ile	I	$C_6H_{11}NO$	4.5	113.084 06
Valine	Val	V	C_5H_9NO	4.2	99.068 41
Leucine	Leu	L	$C_6H_{11}NO$	3.8	113.084 06
Phenylalanine	Phe	F	C_9H_9NO	2.8	147.068 41
Cysteine	Cys	C	C_3H_5NOS	2.5	103.009 19
Methionine	Met	M	C_5H_9NOS	1.9	131.040 49
Alanine	Ala	A	C_3H_5NO	1.8	71.037 11
Glycine	Gly	G	C_2H_3NO	−0.4	57.021 46
Threonine	Thr	T	$C_4H_7NO_2$	−0.7	101.047 68
Tryptophan	Trp	W	$C_{11}H_{10}N_2O$	−0.9	186.079 31
Serine	Ser	S	$C_3H_5NO_2$	−0.8	87.032 03
Tyrosine	Tyr	Y	$C_9H_9NO_2$	−1.3	163.063 33
Proline	Pro	P	C_5H_7NO	−1.6	97.052 76
Histidine	His	H	$C_6H_7N_3O$	−3.2	137.058 91
Glutamic acid	Glu	E	$C_5H_7NO_3$	−3.5	129.042 59
Glutamine	Gln	Q	$C_5H_8N_2O_2$	−3.5	128.058 58
Aspartic acid	Asp	D	$C_4H_5NO_3$	−3.5	115.026 94
Asparagine	Asn	N	$C_4H_6N_2O_2$	−3.5	114.042 93
Lysine	Lys	K	$C_6H_{12}N_2O$	−3.9	128.094 96
Arginine	Arg	R	$C_6H_{12}N_4O$	−4.5	156.101 11

properties of the amino acids, the polypeptide chain folds itself into a particular shape (conformation), partly defined by the interactions between the amino acids themselves, and also due to the interaction between hydrophobic and hydrophilic amino acids and their surroundings.

The main function of a protein is defined by the order of its constituent amino acids (Alberts et al., 2015). Each protein can have many different forms, called proteoforms, which can each have different functions. The different proteoforms can arise in the process of producing the protein from DNA, i.e. when transcribing DNA to RNA, by alternative splicing, or when translating mRNA into a polypeptide, or after the protein has been produced. The latter modifications are called post-translational modifications (PTMs). Common modifications are changes to a standard amino acid, e.g. oxidation, methylation, or glycosylation. A different type of modification is where part of the protein is cut-off. These can lead to a different conformation, switch proteins on or off, and change the function of the protein. Proteins can have multiple PTMs at the same time, and in different combinations, with the total number of proteoforms in a cell being on the order of 10^6 (Aebersold et al., 2018; Eidhammer et al., 2013).

The biological functions of a protein are determined by its interactions with other molecules, e.g. antibodies interact with viruses or bacteria, and other proteins interact with small molecules. Similarly, when a protein binds to other proteins, they form a protein complex. While the type of interactions and the interaction partners of proteins can vary widely, the interactions of a specific protein are usually limited to a small set of specific interaction partners (Alberts et al., 2015). The (inter)actions of proteins are part of a large system trying to maintain homeostasis. Series of actions/reactions within a specified system, or with a specified goal, are called pathways (Jassal et al., 2020; Kanehisa et al., 2016).

As a pathway can be defined as a set of actions/reactions leading to a specific outcome, the smallest pathways can be defined as a single reaction. These pathways can be connected to other pathways, which generate its inputs or use its outputs. Together these pathways may then form a larger, more general pathway. The largest pathways, e.g the immune system, are very general and contain different sub-systems that might work independent of each other. Still, all analytes relate to all other analytes and a perturbance in one part of the system can have consequences in other parts of the system by following the paths of the reactions ‘downstream’, so-called pathway cascades (e.g., Benchoula et al., 2021; Katoh and Katoh, 2007).

1.2 Proteomics and peptidomics

All proteins together form the proteome, and the systematic characterisation of the proteome, under a specific set of conditions, is called proteomics (Nelson and Cox, 2017). This undertaking can be approached from different, complementary angles. The most basic aspect of the proteome is the examination of the constituent proteins. This can be charting all proteins that exist in a species, but also more narrowly the proteins present in e.g. a specific cell type, tissue section, body fluid, or cell compartment. When proteins are identified, another aspect of the proteome is the characterisation of the properties of the different proteins. This can range from defining the amino acid sequence and associated PTMs, to estimate protein abundances, identifying the three-dimensional structure, and finding interaction partners in species and cellular components. A different aspect of the proteome, once the proteins in a sample are identified, is the quantification of the detected proteins.

Very small proteins and naturally occurring shorter amino acid chains can also have specific roles in the cell, e.g. cytokines, or antimicrobial agents, however, such bioactive peptides can usually not be predicted from the genome (Schrader, 2018). Other peptides are products of larger proteins, the result of (specific or unspecific) proteolytic degradation. Note that there is no clear distinction between proteins and peptides, but when a distinction is made it is usually with a size or weight cut-off around 10–15 kDa. The endogenous peptides present in a sample are referred to as the peptidome, the study of which is called peptidomics.

A standard part of proteomics protocols, for reasons that will be explained in Chapter 2, is to cut the proteins into smaller peptides before analysing them. First, proteins are denatured, turning them into unfolded polypeptide chains that a protease can subsequently cut into smaller peptides. With a highly specific protease, the obtained peptides can be predicted from the protein amino acid sequence. Trypsin, the most used protease for proteomics studies, cuts after arginine and lysine, unless followed by proline. Both arginine and lysine are common amino acids, so each protein is cleaved into many small pieces suitable for mass spectrometry analysis (Vandermarliere et al., 2013).

1.3 Biomarkers

The goal of biomedical experiments is often to gain insight into how biological processes work, or are altered by interventions or diseases. One way of doing this is to search for indications in the body that can tell us the state of the organism. These indications are called biomarkers and are defined by Biomarkers Definitions Working Group (2001) as

> A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (p. 91)

As a simple example, body temperature gives an indication of fever. Similarly, there are biomarkers used to diagnose rheumatoid arthritis which can be detected by adding specific antibodies to blood samples (Rose et al., 1948).

The development of a biomarker is a long process, requiring development steps in both research and clinical settings (Simon, 2008). One way of classifying biomarkers is by their intended use: (i) diagnostic biomarkers are used to diagnose, or rule out, a certain disease, (ii) prognostic biomarkers are used to predict how a disease would progress with or without treatment, and (iii) predictive biomarkers are used to differentiate between those benefiting from an intervention and those who would not. Predictive biomarkers are especially relevant for cancer treatments, as those are often toxic, expensive, and only marginally effective (Ankeny et al., 2018; Janes et al., 2015).

The four performance metrics commonly used to assess biomarkers are sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV) (Simon, 2015). These are easiest to define for diagnostic markers, though can be defined analogously for the other markers. PPV is the probability that the subject has the disease given a positive marker, while NPV is the opposite: the probability that the subject does not have the disease given a negative marker. Sensitivity is the probability that the marker is positive for subjects with the disease, and, similarly, specificity is the probability that the marker is negative for subjects who do not have the disease. Thus, when a subject receives a positive result on a diagnostic biomarker, this does not necessarily mean that the person definitely has the disease. Instead, basing calculations solely on the result from the biomarker, it is possible to calculate the probability of having the disease, given a positive test on the marker with Bayes' formula as (Gelman et al., 2013)

$$\begin{aligned} P\left(disease|marker = true\right) = \\ \frac{P\left(marker = true|disease\right) \times P\left(disease\right)}{P\left(marker = true\right)} \end{aligned} \tag{1.1}$$

where $P\left(marker = true|disease\right)$ is the sensitivity of the test, and $P\left(disease\right)$ the disease prevalence. $P\left(marker = true\right)$ is the probability of getting a positive result on the biomarker, which can happen in two ways: either when having the disease $(sensitivity \times prevalence)$, or when not having the disease $((1 - specificity) \times (1 - prevalence))$, thus $P\left(marker = true\right) = (sensitivity \times prevalence) + (1 - specificity) \times (1 - prevalence)$. In reality, there will usually be more indications of possibly having the disease, which should be weighed along with the evidence from the biomarker. Additionally, multiple biomarkers can be used in a panel, to strengthen the sensitivity, or to rule out diseases with similar characteristics. Although in those cases, correlations between different markers should be taken into account.

Some disease-specific genetic biomarkers are known, e.g. for breast cancer the oestrogen and progesterone receptors and the human epidermal growth factor receptor 1, which can be measured at gene, transcript, or proteome level. The search for novel biomarkers using omics platforms, however, still faces several challenges (Nicolini et al., 2018). For example, many markers based on multiple genes/proteins are not reproducible in other datasets, or with different platforms, and suffer from instability when reprocessed on the same dataset (Begley and

Ioannidis, 2015; Goh and Wong, 2016b; Venet et al., 2011; Wang et al., 2017).

1.4 Diversity of the proteome

Compared to the genome, the proteome shows dramatically more variability over the course of a day and over the course of a lifetime (Plumel et al., 2019; Robin et al., 2020; Robles et al., 2014; Tsuji et al., 2007; Wang et al., 2018). Monogenic diseases are easy to find when it is known what to look for (e.g. maturity onset diabetes of the young (McDonald and Ellard, 2013), cystic fibrosis (O'Sullivan and Freedman, 2009), haemophilia A and B (Dolan et al., 2018)), though the consequences are visible as non- or dis-functioning proteins. More complex diseases that do not have a single genetic component might be possible (or easier) to understand in terms of protein regulations (Zaghlool et al., 2021).

The diversity of the proteome manifests itself at several levels. Firstly, proteins can exist in multiple different proteoforms. Secondly, proteins can be present in different proportions or abundances, as they are affected by processes inside and outside the body. Finally, some proteins exist in some species or cell-types but not in others, as evidenced by e.g. species specific databases (The UniProt Consortium, 2019) and the human protein atlas (Uhlén et al., 2015). This high variability of the proteome poses major challenges in terms of data analysis. Changes in protein abundance can arise due to differential regulation due to, e.g., disease state, but can also come from inherent differences between subjects, or subtle, unknown biases in sample collection. Disease-driven changes of the proteome thus have to be larger than the inherent variability between the subjects, and/or have to be known in advance and targeted specifically.

Different body fluids and tissues also have different characteristics. For example, blood, brain, liver, and cerebrospinal fluid (CSF) have different properties and functions, and thus a different proteome. To be able to properly analyse the different sample types, the differences need to be accounted for in sample preparation and data analysis. For example, in blood, the protein serum albumin is so abundant that it dwarfs the abundance of other proteins (Pietrowska et al., 2019; Tomascova et al., 2019). Furthermore, platelet proteomics is special in the sense that platelets do not have a nucleus (Smith et al., 2013), but display different behaviour depending on their surroundings (Clemetson and Clemetson, 2013). Another interesting example is urine, for both proteomics and metabolomics, as it is one of the easiest accessible sources of waste-products of bodily processes (Palanski et al., 2021; Wu and Gao, 2015).

A body fluid that is especially relevant for this thesis is CSF. It contains waste-products of brain processes, and is thus potentially very interesting for the investigation of neurological disorders (Hansson et al., 2017). The CSF proteome contains both proteins produced in the central nervous system, used by the brain, and also proteins that have breached the blood-brain barrier. In addition, it is one of the places where proteins are broken down into smaller peptides and single amino acids, for reuse by the body. Both the proteome and the peptidome can give complementary indications concerning the state of the organism, such as aberrant brain processes or protein degradation.

2 Mass spectrometry-based proteomics

The analysis of the protein/peptide samples is done on specialised instrumentation. Mass spectrometry (MS), coupled to liquid chromatography (LC) is the most common method used for the identification and quantification of the proteins (or peptides) present in a sample, by measuring the mass to charge ratio of proteins/peptides. Here follows a short introduction to the instrumentation and techniques relevant to the papers included in this thesis.

2.1 Instrumentation

The most common way to identify and quantify proteins in a sample is by liquid chromatography coupled to mass spectrometry (LC-MS). A mass spectrometer is in its essence a large, very precise scale for measuring small charged molecules. In its simplest representation, a mass spectrometer consists of three elements: an ionisation source, a mass analyser, and an ion detector. To reduce the number of particles entering the mass spectrometer simultaneously, the sample content is first separated by liquid chromatography (LC), which feeds directly into the mass spectrometer's ionisation source.

This section mainly relies on the mass spectrometry textbook from Gross (2017), which, for ease of reading, will not be cited after every paragraph. Where additional information is taken from other sources, these are cited specifically.

2.1.1 Liquid Chromatography

Liquid chromatography consists of a solid and a mobile phase, sorting the molecules in the sample based on their differences in interaction with the two phases. The most common type of chromatography used in mass spectrometry is reverse phase high-performance liquid chromatography (HPLC), which separates molecules based on hydrophobicity (Twyman, 2014). Beads with hydrophobic ligands, packed in a column, form the stationary phase. The mobile phase, a polar solvent containing the dissolved sample, is loaded onto the column entirely and forced under pressure through the solid phase. Molecules in the sample interact with both the solid and the mobile phase, such that molecules with a certain hydrophobicity stick to the solid phase, while other molecules stay in the liquid phase and are carried through. The interaction between the molecules and the solid phase is regulated by the concentration of an organic modifier in the elution buffer, controlling the strength with which molecules are adsorbed by the solid phase. By gradually increasing this concentration, an increasing amount of molecules can pass through the column, such that the weakest hydrophobic interactions are released first. In addition, retention in the column tends to increase with molecular mass, resulting in an additional mass-dependent separation.

2.1.2 Ionisation

Next, the sample has to be made available to the mass spectrometer by ionising the molecules. The two most common ways of ionising biomolecules are matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionisation (ESI). In MALDI, the sample is smeared over a metal plate and mixed with a matrix. A laser pulse irradiates part of the sample, vaporising it and ionising the molecules, which can then be accellerated into the mass spectrometer by an electric field. This method produces mainly singly charged ions. ESI on the other hand can be directly coupled to the HPLC system, injecting the molecules that come out of the HPLC system into the ioniser through a capillary, resulting in a fine spray of charged droplets. As the solvent evaporates, the droplets shrink and are eventually ripped apart until only charged molecules remain. This ionisation method produces molecules with a stochastic number of ions/charges, which are subsequently guided into the high vacuum chamber of the mass spectrometer by an electric field.

2.1.3 Mass analysers

Mass analysers select ions based on their mass over charge ratio (m/z). The simplest mass analysers are linear quadrupole mass analysers, which consist of four rod electrodes, placed in a square configuration, in line with the direction of the ions. Two rods at opposite sides form a pair, and the two pairs of rods are held at the same potential but with opposite polarity. Molecules passing through the mass analyser oscillate between the rods according to their m/z and the potential applied to the electrodes. Thus, by changing the potential, one can select the range of m/z which will have a stable trajectory, i.e. will pass through the mass analyser without hitting a rod or escaping through the gaps.

By placing electrodes with slightly higher potential than the rod electrodes at the in- and the outlet of the quadrupole, a linear ion trap is created. Setting the voltage at zero at the inlet and higher than the rod electrodes at the outlet lets molecules into the quadrupole. After the molecules have entered the mass analyser the inlet can be closed again, by applying the same voltage as at the outlet, before the smallest and fastest molecules can escape, thus trapping the ions. By lowering the voltage at the outlet to something in between the voltage of the rod electrodes and the inlet electrode, molecules are pushed out of the quadrupole through the outlet. The quadrupole ion trap has, in contrast to the linear quadrupole mass analyser, two hyperbolically shaped electrodes as end caps and a ring electrode replacing two of the linear quadrupole rods, resulting in an electrical field in three instead of only two dimensions.

2.1.4 Detectors

After the selection of molecules, the next step is to detect which, and how many, molecules are present. The conceptually simplest mass detector is the time-of-flight (TOF) device, with the main principle being that, when accelerated in an electric field, molecules with a higher m/z take longer to fly a certain distance than molecules with a lower m/z. By feeding the molecules at the same time into the TOF at a right angle to the direction of flight, all molecules start at the same time and with the same velocity. Recording the time it takes the molecules to fly through the whole tube, allows their m/z to be calculated. Longer flight paths increase the difference in time it takes two molecules of different m/z to reach the detector, thus increasing the resolution of the instrument.

A different approach is taken by the Orbitrap, which is both a mass analyser and ion trap, as well as a detector. It consists of an inner electrode in the shape of a wire with a double-cone/spherical shaped bulge, surrounded by an outer electrode. An electric field is applied between the electrodes, such that molecules fly in a stable orbit around the inner electrode. Due to the shape of the inner electrode, the molecules also fly from one end of the sphere of the inner electrode to the other with a stable frequency depending on the m/z. By splitting the outer electrode into two parts, at the widest part of the inner electrode, the frequencies of the oscillations of all molecules in the trap can be recorded. The m/z and abundances of the molecules can then be calculated and deconvoluted by Fourier transform. To work properly, the molecules must be fed into the Orbitrap at a specific angle, which is usually achieved with a specially (C-)shaped quadrupole, called a C-trap.

2.1.5 Quality control

To monitor the stability of the instrument, quality control samples should be run on a regular basis. In a recent review three types of quality control samples were identified: (i) single protein samples, (ii) cell lysates, and (iii) synthetic peptides (Bittremieux et al., 2018). Single protein samples are cheap and take a relatively small amount of time to analyse, so they can be run often in between experimental samples. These are generally used to quickly assess LC performance, by monitoring the retention time and peak width. Cell lysates are more complex samples, which more closely mimic experimental samples, making it possible to monitor mass spectrometer performance. Typically small amounts of quality control samples are injected, to monitor performance at low abundances. Samples containing synthetic peptides can consist of as many (or few) different peptides as required. As the exact peptide composition is known, these can be used to monitor performance for targeted approaches, and can even be added to experimental samples. For all quality control samples it is important to avoid crosscontamination of experimental samples, e.g. by using (lysates containing) peptides that do not occur in the experimental samples.

The type of quality control sample defines what quality control metrics can be computed. In another recent review several types of quality control metrics were identified (Bittremieux et al., 2017). Firstly, quality control metrics can be calculated for a specific experiment, or to compare different experiments. Intra-experiment metrics indicate how the instrument behaves over the course of a single experiment. These can be, e.g. the mass accuracy of identified spectra, or the evolution of the total ion current, which can show whether something unforeseen happened in a specific run. Inter-experiment metrics, on the other hand, summarise the quality of an experiment in a single number to compare multiple experiments to each other, e.g. to assess the performance of the mass spectrometer over time, which allows to compare runs to each other and can indicate whether conditions between runs have changed.

Secondly, a distinction can be made between metrics derived directly from the raw spectra, or metrics combining the raw spectra with identification results. Identification-free metrics, based only on the raw spectra, assess the whole MS workflow, e.g. number of spectra, or scan rate. Identification-based methods combine the information from raw spectra with identifications, and are thus also dependent on the quality of the identifications. However, as they use more information, these metrics can provide more detailed quality assessments, e.g. sequence coverage for a known sample, or number of identifications of peptides and proteins. Lastly, instrument metrics provide low-level information on different parts of the mass spectrometer. These are not directly related to experimental results, but are important for maintenance.

The type of quality control metric clearly depends on the quality control samples (Bittremieux et al., 2017). For samples containing a single protein digest, it is reasonable to expect the sequence coverage to be stable across runs, while, as there is only a single protein in the digest, the number of identifications is less relevant. In complex quality control samples the identifications themselves may be less interesting due to random fluctuations, and thus the number of identifications may become more relevant than sequence coverage. Similarly, for discovery experiments the number of identifications is relevant, while for targeted experiments sequence coverage is more appropriate.

Table 2.1: Elements and stable isotopes occurring naturally in proteins (de Laeter et al., 2003).

Element	Mass	Abundance (%)
Hydrogen		
^{1}H	1.0078	99.99
^{2}H	2.0141	0.01
Carbon		
^{12}C	12.0000	98.93
^{13}C	13.0034	1.07
Nitrogen		
^{14}N	14.0031	99.64
^{15}N	15.0001	0.36
Oxygen		
^{16}O	15.9949	99.76
^{17}O	16.9991	0.04
^{18}O	17.9992	0.21
Sulphur		
^{32}S	31.9721	94.99
^{33}S	32.9715	0.75
^{34}S	33.9679	4.25
^{36}S	35.9671	0.01

2.1.6 Isotopic distribution

As the mass measurements are done at the scale of m/z, the observed m/z values have to be transformed back into mass measurements. The isotopic distributions of elements can help with this. The chemical elements making up amino acids exist in variants, called isotopes, which have different numbers of neutrons. Apart from the radioactive isotopes, which are outside the scope of this thesis, these are biologically indistinguishable, but, as shown in Table 2.1, have slightly different masses and naturally occur in different abundances. The distribution of the elemental masses (and thus of amino acids, peptides, and proteins), is called the isotopic distribution, and can be used to identify charge states and chemical composition.

The mass of an amino acid consisting only of elements of the light-

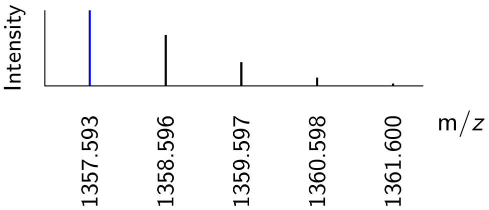
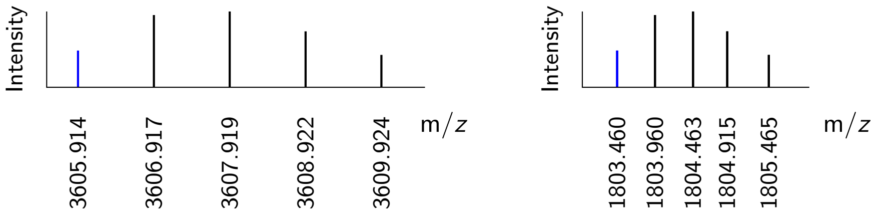


Figure 2.1: Isotopic distribution of a small peptide: MTDTVFSNSSNR. The blue bar represents the monoisotopic peak.



(a) Single charge

(b) Two charges

Figure 2.2: Isotopic distribution of a larger peptide: PQPMPIKKTKPQQPVSEPAAPEQPAPEPKHPAR. The blue bars represent the monoisotopic peak for each charge state. **(a)** Peptide with a single charge. **(b)** Peptide with two charges.

est isotopes is called the monoisotopic mass. Similarly, the peak of a peptide consisting only of elements of the lightest isotopes is called the monoisotopic peak. For small peptides this is also the highest peak, corresponding to the most abundant version of the peptide (Figure 2.1). However, the more copies of an element that are present in the peptide, the smaller the chance that they are all of the lightest isotope, e.g. the chance that a single oxygen atoms is monoisotopic, is 0.9976, while the chance that out of 100 oxygen atoms all are monoisotopic is $0.9976^{100} = 0.7864$. The distance between the isotopic peaks are roughly one for ions with a single charge, 0.5 for ions with two charges (Figure 2.2), and so on.

2.1.7 Tandem MS

When intact peptides are measured, the only available information is the m/z, the isotopic distribution, and, in case of LC-MS, the retention time or elution time through the column. From these measurements it is, unfortunately, not possible to obtain information about the amino acid sequence, given that multiple sequences can have the same m/z, amongst which peptides with the same amino acids in a different order. To get information about the amino acid sequence, tandem mass spectrometry (also referred to as MS^2 or MSMS) has to be performed. Basically, first intact peptides are scanned (precursor scan), which are then dissociated (fragmented) and analysed again (product ion scan). The product ions can be dissociated further and analysed again, to obtain an even finer resolution of the amino acid sequence (MS^3, or more generally MS^{n}). In general experiments, MS^2 and MS^3 are the most common.

Tandem mass spectrometry is performed in two ways: tandem-in-space or tandem-in-time. Tandem-in-time mass spectrometry performs the discrete steps (ion selection, activation, and product ion analysis) in the same mass analyser, sequentially in time, while with tandem-in-space mass spectrometry each level of mass spectrometry is performed in separate mass analysers. Tandem-in-space mass spectrometry is done, for example, in triple-quadrupoles, which have three quadrupoles: the first and third act as mass analysers, while the middle quadrupole (or hexa-/octopole) acts as a collision cell. The collision cell is filled with a noble (inert) gas. As the molecules move through the collision cell, they collide with the gas and break into smaller pieces. This is called collision induced dissociation (CID). Every next level would require another collision cell and mass analyser, so this is really only feasible up to MS^3. Higher-order tandem MS is also possible by exchanging the last quadrupole with a linear ion trap, combining tandem-in-space (for MS^2) with tandem-in-time (for MS^3) mass spectrometry. Similarly, the Orbitrap does not have the option to perform tandem MS by itself, but needs a dedicated ion trap for ion isolation and dissociation, prior to analysis of fragments in the Orbitrap. To achieve better fragmentation, the molecules can be (further) dissociated in the C-trap, or passed through the C-trap into a collision cell and then fed back into

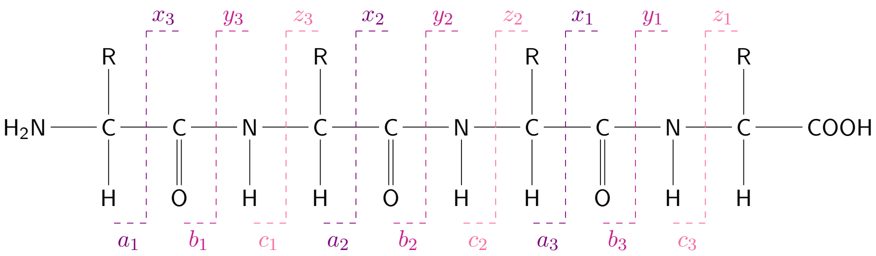


Figure 2.3: Fragment ion definition for the fragments observed in tandem mass spectrometry.

the C-trap to be transferred into the Orbitrap. This is called higher-energy C-trap dissociation, or, as the dissociation generally happens outside the C-trap, higher-energy collisional dissociation (HCD).

Usually the aim is to break each peptide once, leading to peptides with the original N-terminus and peptides with the C-terminus, but no peptides with neither or both. Breaks between two amino acids in the peptide sequence can occur in three places: at the C-terminal side of the α-carbon of the 'left-hand' amino acid, at the N-terminal side of the α-carbon of the 'right-hand' amino acid, or at the bond between the two amino acids. The resulting peptide fragments are named, for ease of reference, according to the break point from N- to C-terminus: a, b, and c ions for the fragment ions containing the N-terminus, and x, y, and z ions for those containing the C-terminus. Additionally, they are numbered according to the number of amino acid side chains they contain (Figure 2.3).

Fragment spectra only show intensity vs m/z, thus for proper identification of fragment ions, the charge state needs to be taken into account. Furthermore, the same fragments can be present in the spectrum with different charge states, adding additional complexity to the fragment spectra. To make interpretation simpler, charge deconvolution, using the isotopic distribution, is applied to spectra with multiply charged fragments, calculating the singly charged m/z for multiply charged fragments. This is mainly a challenge for ionisation with ESI, which generally leads to multiply charged ions, while MALDI mainly generates singly charged ions.

2.1.8 Mass spectrum files

In the process of peptide analysis and fragmentation, packets of ionised molecules are sent into the detector one by one, recording the m/z and the intensity of the signal at each m/z, resulting in a raw mass spectrum for each of these packets. The number of mass spectra from the analysis of a single sample easily runs in the thousands, leading to data in the order of several GB for a single sample.

A mass spectrum is usually displayed as a diagram with m/z on the horizontal axis and the intensities on the vertical axis. Due to the nature of detectors and the reality in which the molecules exist, the m/z is not measured with infinite precision. The peaks are thus not straight lines, but rather stretch over a small m/z range, dependent on the resolution of the instrument. With low-resolution instruments from a couple of decades ago, nearby peaks regularly merged with each other, making interpretation very difficult. For further processing of the spectra it is necessary to summarise the small m/z ranges into peaks at a single m/z, generating a peak list (Eidhammer et al., 2013).

The generation of this peak list from the raw mass spectra is not trivial: even when there are no molecules in the detector, the detector still 'detects' a certain amount of background noise. This baseline measurement needs to be removed from all intensities, but is not the same across the whole experiment. Furthermore, summarising the intensities of a peak across a m/z range into a single intensity can be done in several ways (Eidhammer et al., 2013). The simplest way is to just pick the intensity at the apex of the curve. Alternatively, a model can be used to approximate the curve, e.g. by fitting a quadratic curve through the apex and two points on either side of the apex, or by fitting a Gaussian (or Laurentian) curve through the apex. From these estimated curves the area under the curve is calculated to be the intensity of the measurement.

2.2 Interpreting mass spectra

After obtaining the peak lists, peptides have to be identified and quantified. Peptide identification is done on the basis of the fragment spectra, while the quantification is usually based on the data in MS^1 spectra. An important exception is experiments with isobaric labels, which also perform quantification on the fragment spectra. When the peptides have been identified, they have to be mapped to the proteins they came from, referred to as the protein inference problem, and subsequently protein abundances have to be calculated. The protein identifications and abundances then form the basis for further statistical analysis.

2.2.1 Peptide identification

The generated peak list is used to identify the peptide(s) present in the mass spectrum. This can be done by comparing the spectra to fragment spectra in a pre-generated database, to a spectral library, or by *de novo* sequencing. To make a database with theoretical fragment spectra, protein sequences are downloaded (e.g. from UniProt (The UniProt Consortium, 2019) or RefSeq (O'Leary et al., 2016)) and from these protein sequences a decoy database is generated. The usual strategies to generate the decoy proteins are to either reverse the protein sequences, or to generate random protein sequences (Edwards, 2017; Elias and Gygi, 2007; Moosa et al., 2020). The protein sequences are then digested *in silico* (algorithmically, on the computer) with the same rules as the enzyme used in the experiment. As in the sample processing, the enzyme will usually miss some cleavage sites and the *in silico* digestion is thus performed allowing for a small number of missed cleavages. The resulting peptides are then fragmented, leading to a series of expected peak positions on the scale of m/z. These peak positions can then be compared to the peaks found in the experimental spectra, and scores for the compatibility can be computed to estimate the level of certainty concerning matching theoretical peptides. By using both a relevant protein database and a decoy database it is possible to estimate probabilities of chance matches (Edwards, 2017).

Especially for complex organisms, and including multiple PTMs, the database can become extremely large. In addition, the height of the fragment peaks is often not taken into account, while it does contain valuable information (Silva et al., 2019). An alternative for theoretical fragment library search is to create a spectral library from experimental fragment spectra. First, a number of runs is done, identifying peptides in the spectra as before. The identified experimental spectra then form the spectral library against which the fragment spectra from subsequent runs are compared. This vastly reduces the size of the database to compare against, as only peptides (with PTMs) that have been identified before are included in the spectral library. Additionally, peak heights can be used for more accurate scoring of the spectrum comparisons (Craig et al., 2005; Shiferaw et al., 2020). Finally, a recent approach is to generate a spectral library from a sequence database using instrument specific peptide detectability predictions (Yang et al., 2020).

When mapping against a protein database, one is only able to identify peptides that can come from proteins in that database. For well annotated species this can be the main driver for peptide identification, but for species lacking annotation many peptides are likely to be missed or identified incorrectly. To identify peptides from fragment spectra without an annotated database, one can perform *de novo* sequencing. This is done by inferring the amino acid sequence by looking at the distance between different peaks. Depending on the dissociation strategies some fragment ion pairs are more likely than others (Gross, 2017). Additional evidence that a peak in the spectrum comes from a fragment ion (instead of e.g. background noise) presents itself when peaks for the other fragmentation points are present. By mapping the distance between two neighbouring ion peaks of the same series, e.g. two y-ions, to amino acid masses, one can infer which amino acid is present at a particular point in the sequence (Figure 2.3). Repeating this process for all peaks can in many cases lead to identification of most amino acids in the peptide, and their sequence, for high-quality fragment spectra, as in Figure 2.4. *De novo* sequencing can also be performed in addition to spectrum matching, to attempt to identify unidentified experimental spectra. The generated peptide sequence(s) can additionally be compared to the database.

Both for spectrum matching and *de novo* sequencing the incorporation of possible PTMs can dramatically increase the search space (Verheggen et al., 2016). Fixed modifications, i.e. modifications that are assumed to affect all relevant amino acids, simply replace those amino acids with their modified counterpart when building the database. A standard fixed modification for mass spectrometry analyses is carbamidomethylation of cysteine. When generating the peptide database the cysteine residues are all replaced by their modified counterparts, leaving

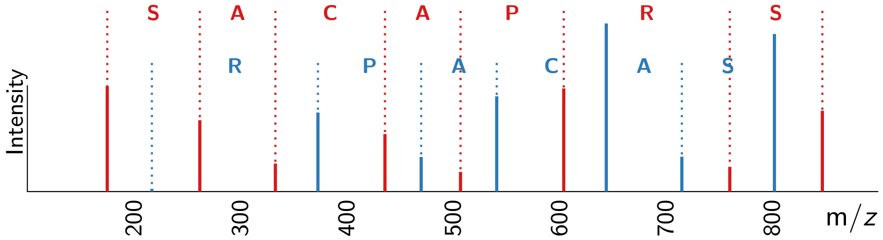


Figure 2.4: Simulated fragment spectrum with *de novo* annotation for the peptide `ESRPACASR`. Fragments are singly charged, peak heights are predicted from the singly charged peptide with MS^{2}PIP (Gabriels et al., 2019). Blue: b-ions, red: y-ions. Distances between peaks correspond to the annotated amino acid residue masses.

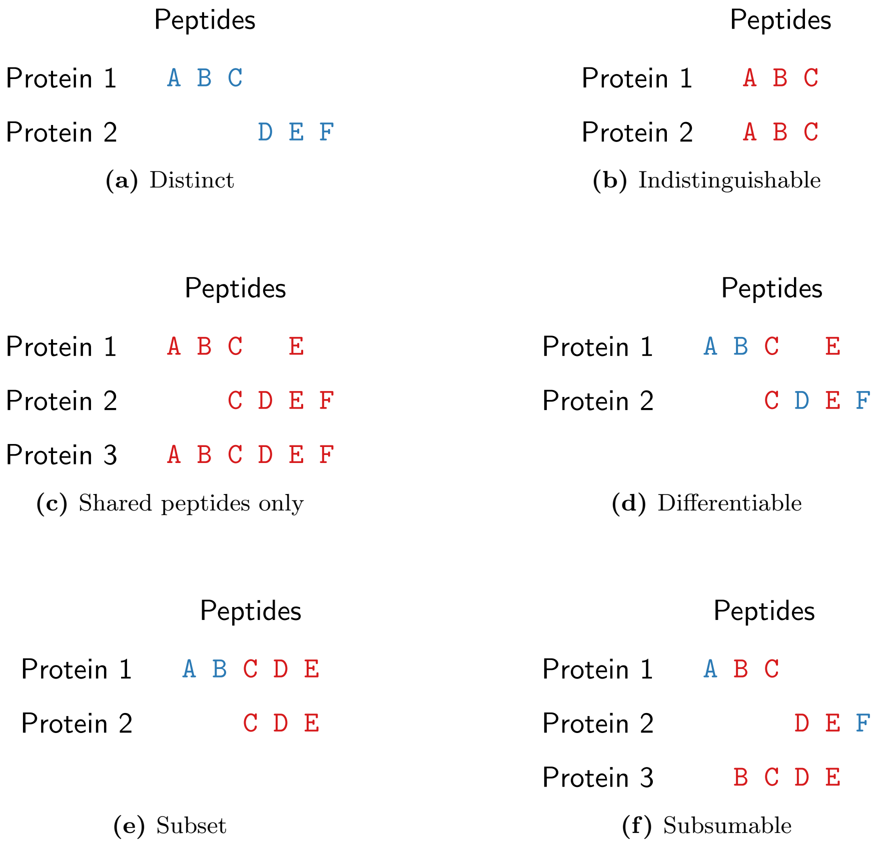
the database the same size. Variable modifications, on the other hand, are modifications that might or might not be present on the amino acid residues, e.g. oxidation of methionine and phosphorylations. These are usually of much lower abundance, and for each of the relevant residues it would have to be checked whether the modified or unmodified residue fits, thus dramatically increasing the size of the database and the search space for *de novo* sequencing.

2.2.2 From peptide to protein abundance

With peptides identified, the next steps are to infer the proteins from the peptides, and, subsequently or simultaneously, calculate protein abundances. Unfortunately, neither step is trivial, and many approaches are available.

Protein inference To determine which protein a peptide came from, the only information available is the amino acid sequences of the peptide and the proteins. Unfortunately, there are many peptides that can come from multiple proteins. This makes it difficult to determine which proteins are present in the sample, based only on the identified peptides. This challenge, mapping peptides to proteins, is called the protein inference problem.

Given a set of identified peptides and a set of proteins from which



(a) Distinct **(b)** Indistinguishable

(c) Shared peptides only **(d)** Differentiable

(e) Subset **(f)** Subsumable

Figure 2.5: Different scenarios for protein identification. Unique peptides are blue (as in **(a)**), shared peptides red (as in **(b)**).

the peptides could have come from, there are six general ways peptides can map to proteins. Nesvizhskii and Aebersold (2005) developed a nomenclature to distinguish these scenarios, and named them distinct, indistinguishable, shared peptides only, differentiable, subset, and subsumable (Figure 2.5).

The simplest scenario is where the proteins have no shared peptides. Each peptide belongs to only one protein, so it is clear that both proteins occur in the sample and which peptide comes from which protein. These proteins are called distinct proteins (Figure 2.5a). The opposite happens when all the peptides from two proteins are shared by those two proteins (Figure 2.5b). None of the peptides is unique to one of the proteins, so either one or both of the proteins is present in the sample, but it is impossible to know which. When there are more than two proteins this generalises to a group of proteins that is identified by shared peptides only (Figure 2.5c). In this set there is no peptide that is unique to one protein, though any pair of proteins might not share all their proteins with each other.

If instead the proteins share some peptides, but both have at least one unique peptide, the proteins are differentiable and both are present in the sample (Figure 2.5d). This is in contrast to the situation in which all peptides in the set belong to one protein, and a subset of these peptides belongs to the other protein, but this second protein does not have a unique protein. This second protein is then called a subset protein (Figure 2.5e). A similar scenario is the one with a subsumable protein (Figure 2.5f). This occurs when a protein shares peptides with two other proteins, which both have at least one unique peptide and only share peptides with the first protein, but not with each other. While the last two proteins are definitely present in the sample, it is unclear whether the subsumable protein is.

The reasons why non-unique peptides occur are numerous. For example, as a gene can produce many different protein isoforms, these are likely to share peptides. Proteins that originate from alternative splice forms would only have distinct peptides for those peptides that contain a splice site. Furthermore, proteins with (different) PTMs are indistinguishable apart from the peptides where the modification(s) exist. Similarly, cleavage of a signal or transit peptide might not be observable unless the cleaved peptide is observed. Ideally, proteins that are indistinguishable should be presented as protein groups, although that might be challenging for subset and subsumable proteins.

Protein quantification Given the different ways the peptides belonging to a protein can be shared with other proteins, there are many ways peptide abundances can be summarised into protein abundances. Limiting the abundance calculation to unique peptides, might ignore most of the peptides that could belong to a protein. In this simple case, where all peptides are certain to come from the same protein, the mean or median of the peptide abundances can be used as a protein abundance. Exclusively using this strategy for all proteins will exclude many peptides and potential proteins.

The challenge becomes more complex for sets of proteins that are identified by shared peptides only. Taking the example in Figure 2.5c, the relative abundances of the three proteins could be approached as a constraint satisfaction problem. However, the peptide abundances are not perfect, and the number of proteins with shared peptides is usually much larger than what is depicted in the examples. Especially taking into account different proteoforms, the summarising of the peptide abundances into one protein abundance is an incredible challenge (Plubell et al., 2021).

As there are many different ways to decide which proteins are (not) present in the sample in the above situations, different algorithms can give different sets of identified proteins and different protein abundances. Typically, protein inference and quantification is done in an integrated software pipeline, collecting tools for each step in one interface. Even so, oftentimes the different steps in the protein inference and quantification process are treated as distinct, and only point estimates for quantifications are given which hides variability in peptide abundances and subsequent uncertainty of the protein quantifications. Triqler/Quandenser (The and Käll, 2019, 2020) recently proposed an alternative, taking uncertainty of all previous stages into account when estimating protein abundances and outputting posterior distributions rather than point estimates.

2.3 Analytical approaches

Depending on the goals of the experiment, different types of analyses might be appropriate. Here the focus is on quantitative proteomics using mass spectrometry. Other types of analysis, such as which proteins interact with each other or what the three-dimensional structure of a protein is, are performed with different experimental set-ups and instruments and will not be discussed here. The most common goals of quantitative proteomics studies are to find the effect of an intervention, or differences between populations. Another general distinction is between targeted or discovery experiments. Furthermore, within discovery experiments the main distinction is between data-dependent aquisisiton (DDA) and data-independent aquisition (DIA). Additionally, all these experiments can be performed with various different labels or as label-free (Eidhammer et al., 2013).

2.3.1 Targeted vs discovery

A usual sequence of proteomics experiments starts with a discovery study, to get a rough idea of the proteins present in the samples and to generate a preliminary list of proteins that might be interesting to study in the given situation. This is then followed by a targeted experiment where the focus is specifically on the peptides or proteins in that list. By only targeting a relatively small number of proteins/peptides, more accurate and stable results can be achieved at the cost of vastly reduced coverage.

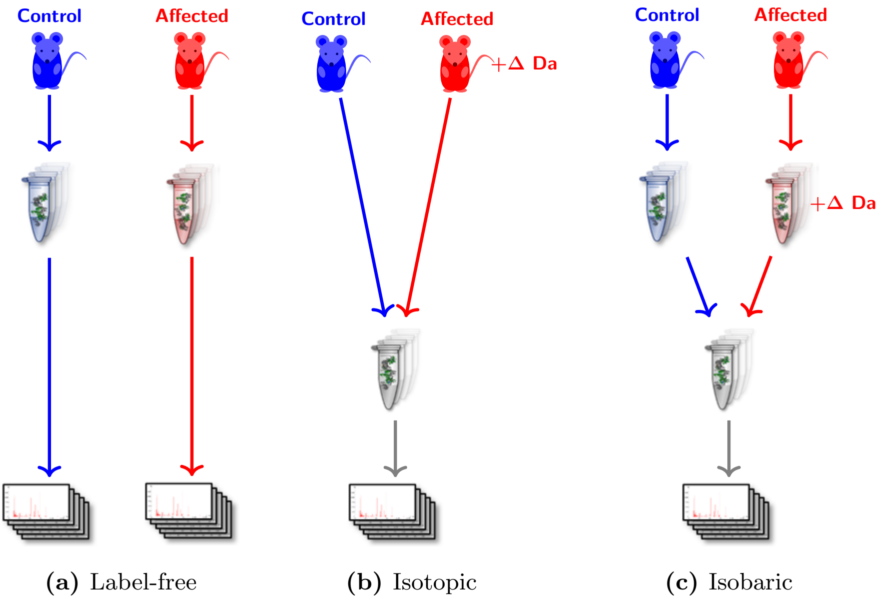
In discovery studies, the goal is to screen for differences between two (or more) populations. These can be subjects which have e.g. received a certain drug or have a certain disease. The first questions are which proteins can be found, and which are differentially expressed between the populations. As these are screening experiments, the protein abundances found in these experiments are not as accurate as one would like for definitive answers. Rather, the goal is to select a small subset of the proteins, e.g. the top differentially abundant proteins, for further study. This selection is then augmented by other proteins that are interesting, e.g. because of previous research or because they have roles in the same, potentially relevant, pathways.

In targeted experiments only a small number of proteins/peptides are quantified, e.g. the resulting list of proteins from the discovery experiment. If the proteins have unique peptides (or peptide fragments) after digestion, these can be selected and isolated by the mass analyser. As only a relatively small number of peptides are targeted, this gives the opportunity to perform absolute quantification using synthetic peptides. These peptides are assembled with a specific mass difference, which is achieved by synthetically generating the peptide with one amino acid replaced by a heavy version containing stable (heavy) isotopes of carbon, nitrogen, and hydrogen (i.e. ^{13}C instead of ^{12}C, ^{15}N instead of ^{14}N, and ^{2}H (deuterium) instead of ^{1}H elements). As the amount of the synthetic (heavy) peptide(s) in the sample is known, and the same across samples, this can be used as a reference to calculate the actual amount of the peptide observed, even when there are small differences in the acquisition performance between samples (Kirkpatrick et al., 2005).

2.3.2 Labelled vs label-free

For discovery experiments there is a choice between labelled or label-free analysis. In label-free analysis, the peptides/proteins are loaded onto the mass spectrometer one sample at a time without any labels added to the peptides/proteins in the sample (Figure 2.6a). Comparing proteins and peptides between samples thus has added variance due to differences in sample acquisition. By using labels, one can directly compare peptide abundances between samples avoiding the variability between different runs. Different labels have different properties and use-cases, but the main idea is the same. The labels are added to the different samples (e.g. iTRAQ, TMT), cell cultures (e.g. SILAC), or are fed to animals in the form of special food. The labelled samples are then combined into a single pooled sample, with the number of samples that can be differentiated depending on the number of distinct labels.

Isotopic labelling is performed by using specific food, in the case of animals, or growth medium, in the case of cell lines. For clarity, the examples here concern only animals and food, though the same concepts apply to cell lines and growth medium. The isotopic labelling is done by using food with monoisotopic amino acids for one animal, and food with stable heavy isotopes of one of the amino acids for the other animal (Figure 2.6b). For example, when the food contains arginine with ^{13}C, the mass difference between heavy and light peptides is calculated as the number of carbon atoms in arginine residue (see Table 1.1), times the number of arginines in the peptide, times the mass difference between ^{13}C and ^{12}C (see Table 2.1). A peptide containing one arginine would thus present a mass difference of just over 6 Da. The mass difference between the heavy and light peptides depends on the number of modified amino acids in the peptide, and peptides not



(a) Label-free (b) Isotopic (c) Isobaric

Figure 2.6: Different types of labelling in proteomics experiments. See main text for details.

containing any modified amino acid are indistinguishable between the two samples.

Isobaric labels are instead attached to each of the peptides in the sample after digestion (Figure 2.6c). Sample preparation is mostly as in label-free studies, except that near the end of sample preparation a sample-specific label is added to each sample, after which the samples are combined into one pooled sample. The labels are designed with three distinct parts: (i) one chemical group that can bind to a part of the peptides, e.g. an amine specific group that attaches to the peptide N-terminus, (ii) a reporter ion, defined by its specific mass to differentiate it from the other labels in the same set, which can attract a charge and is preferentially dissociated at a specific chemical bond so it can be observed in fragment spectra, (iii) and a balancer group which makes sure the masses of all the different labels are the same. This way the relative abundances of the peptide in the different samples are visible in the fragment mass spectra.

Thus, isotopic labels allow to avoid the variance between samples introduced during the experiment, while isobaric labelling allows to avoid adding variance after digestion. In label-free workflows, however, the full experimental variance is added to the biological variance.

2.3.3 Tandem mass spectra acquisition

Ionised molecules are processed in small groups, which are collected during a short period of time. The m/z of the molecules in such a group are analysed (the precursor scan), and subsequently a decision is made which molecules (which m/z ranges) are fragmented for MS^2. In the most common acquisition strategy, data dependent acquisition (DDA), the m/z ranges corresponding to the highest MS^1 intensities are selected to be analysed further with MS^2. For each of these m/z ranges, the peptides are directed into the collision chamber, such that each fragment spectrum is likely to contain the fragments of a single precursor ion. To prevent the repeated selection of very high abundant peptides dynamic exclusion of m/z ranges can be used.

The selection of a m/z range corresponding to peaks in the MS^1 spectrum ideally leads to the selection of a single peptide species. In that case the MS^2 fragment spectra contain fragment peaks originating from a single peptide. However, often a peak can consist of two or more coeluting peptides of the same m/z, leading to MS^2 fragment spectra contain peaks originating from a multiple peptides. These chimeric spectra have to be deconvoluted in order to identify all present peptides (Dorfer et al., 2018).

As only the most intense precursor ions are selected for MS^2, many of the low abundant peptides will not be fragmented, and subsequently not identified as identification is done based on the fragment spectra. In different samples the top most abundant peptides in each precursor scan might also be slightly different, leading to some peptides getting selected in one sample, but not in another. Longer elution gradients spread the molecules more and can alleviate this problem to a certain extent, at the cost of longer run-times per sample.

An alternative is instead to fragment all the molecules in the precursor scan, termed data independent acquisition (DIA). The selection window for which molecules to fragment in each round is much wider than in DDA, leading to fragment spectra containing fragments from multiple, potentially many, precursor ions. Interpretation of the spectra is therefore much more challenging than in DDA (Yang et al., 2020). An additional advantage is that the sample volume needed for DIA is less and the throughput is higher. Both DDA and DIA are used for label-free analyses, however, as in DIA multiple precursor ions can be fragmented together, this method is not suitable for labelled experiments, e.g. the intensities of isobaric tags would be a combination of the intensities of all selected precursors.

For targeted analysis the main approaches are selected reaction monitoring (SRM) and parallel reaction monitoring (PRM). In both SRM and PRM, first a peptide precursor is selected in MS^1, subsequently a complete fragment scan is performed in PRM, while in SRM only a select number of fragments is selected (Ankney et al., 2018). This additionally allows the use of synthetic peptides for absolute quantification.

2.4 Analysis of intact and digested proteins

A major distinction in proteomics approaches is the difference between top-down and bottom-up proteomics. In addition, peptidomics shares some aspects of both proteomics techniques, and presents its own challenges and opportunities.

2.4.1 Top-down proteomics

When intact proteins are loaded onto the mass spectrometer, this is called top-down proteomics. By analysing and fragmenting the intact proteins, the protein inference problem is avoided, additionally making it possible to identify different proteoforms created by e.g. PTMs and endogenous proteolysis. In case the proteins are not denatured, even protein aspects such as folding and complexes can be analysed (Brown et al., 2020). Fractionation and fragmentation of large molecules is more challenging than for smaller peptides, and reduces the sensitivity of the measurements. With standard LC-MS only around 100 proteoforms can be identified, with limited coverage above 30 kDa (Brown et al., 2020). Longer columns or different fractionation strategies are needed to lower the sample complexity and increase the sensitivity. The mass spectra of fragmented intact proteins contain many different product ions, which leads to a low signal to noise ratio, making interpretation of the spectra challenging. Low abundant proteins and proteoforms are thus hard to find in top-down proteomics (Brown et al., 2020; Cupp-Sutton and Wu, 2020; Schaffer et al., 2019).

2.4.2 Bottom-up proteomics

In bottom-up proteomics one starts by denaturing the proteins, breaking the noncovalent bonds and converting them into their primary structure, a flexible polypeptide chain, which is then digested into smaller peptides by a protease. The most common protease is trypsin, which cuts only after (at the C-terminal side of) arginine and lysine, unless followed by a proline, leading to peptides usually consisting of less than 25 residues (Eidhammer et al., 2013; Twyman, 2014). These peptides can attract a limited number of charges, and can produce fewer different fragment ions, leading to easier to interpret fragment spectra with higher signal to noise ratios. Mapping the peptides to proteins (the protein inference problem), and recombining the peptide intensities into protein intensities present significant challenges. Additionally, a proteoform with modifications on multiple peptides is impossible to distinguish from individual proteoforms with a single modification (Schaffer et al., 2019).

2.4.3 Peptidomics

Peptidomics has aspects of both bottom-up and top-down proteomics, as intact endogenous peptides are loaded onto the mass spectrometer. Interest is usually in peptides of less than 15 kDa, including disease-specific small protein fragments and signalling molecules. The identification of the peptide necessarily has to rely on fragment spectra, as amino acid sequences of unknown peptides are not known. The identification therefore combines database searching and *de novo* sequencing. PTMs influence the biological function of peptides, and thus also need to be taken into account. Furthermore, the terminals are not (necessarily) cleaving sites for a known enzyme, dramatically increasing the search space (Schrader, 2018).

2.5 Small peptides and single amino acids

In general proteomics and peptidomics experiments, peptides with only a single charge are ignored as these are likely to be contaminants. However, this workflow ignores small peptides that attract at most one charge, possibly discarding valuable information. Recently these very small molecules have become a focus of research for different diseases and bodily fluids (Fonteh et al., 2007; Martelli et al., 2014).

While there exist several kits that allow the specific targeting of single amino acids and some small peptides (e.g. EZ:FAAST (Badawy et al., 2008) and aTRAQ (Held et al., 2011)), these require special sample handling and are thus not easily combined with standard proteomics workflows. Metabolomics approaches such as calculating the fragmentation tree of the molecules are resource intensive and can be overkill for the analysis of a set of compounds with known structures (Böcker and Rasche, 2008). The number of charges a peptide can attract depends on the pH and the constituent amino acids. As isobaric labels need to be able to attract a charge, all labelled peptides (including labelled single amino acids) can attract at least one charge. When these singly charged peptides are fragmented in MS^2 they can potentially be separated from contaminants due to the isobaric tags. For small enough peptides and a limited number of PTMs, the possible fragments ion masses can be calculated, and mapped to the fragment spectra. Thus, even very small peptides that would normally not be able to attract a charge become available for analysis by using isobaric labels and also fragmenting the singly charged peptides.

This novel approach is further explored in *Paper II*.

3 Statistical analysis of proteomics data

When protein abundances have been quantified, the next step is usually to look for differences between the populations of interest. Proteomics studies are often split into two parts: first a discovery experiment, in which as many proteins as possible are identified and quantified, which serves to find a set of ‘interesting’ proteins that will be studied more closely in a second experiment, often called a verification or validation experiment. This second experiment should be performed with new samples, and focuses on only a relatively small set of proteins, usually employing a targeted method and absolute quantification. Finally, the results of the statistical analysis have to be put in a biological context, which is often done with pathway analysis.

3.1 Comparison of protein abundances

In complex samples, the protein abundances are usually assumed log-normally distributed, which practically means that the errors are assumed to arise from a multiplicative process. By applying a log-transformation, the relationship between the variances becomes additive, making standard least squares approaches to estimation applicable. The choice of base for the logarithm is usually chosen pragmatically to aid interpretation. As differences in protein abundances between groups are usually given in fold changes, the most commonly used base is two.

3.1.1 Differences between two groups

With quantified protein abundances, a natural first step is to perform a per protein analysis to look at differences between the populations for each protein separately. The simplest way to compare normally distributed continuous values, such as protein abundances, between two populations (or treatment groups) is with a two-sample t-test. As the protein abundances (for a single protein) are not the same for each subject in each of the two treatment groups, the variability of the measurements within each treatment group should be taken into account. Thus, the test statistic is defined as the difference between the group means divided by the weighted average of the sample standard deviations. When the variation within each group can be assumed to be (roughly) equal, the standard deviations can be pooled, and the test statistic is defined as (Kutner et al., 2005)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}} \tag{3.1}$$

with $\bar{x}_1$ and $\bar{x}_2$ the observed sample means, d_0 the mean difference under the null hypothesis, and s_p the pooled sample standard deviation

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \tag{3.2}$$

The null and alternative hypotheses associated with the test statistic are

$$H_0 \quad : \mu_1 - \mu_2 = d_0 \tag{3.3}$$

$$H_1 \quad : \mu_1 - \mu_2 \neq d_0 \tag{3.4}$$

Usually d_0 is set to 0 to test whether there is ‘no difference’ between the two means. To test the null hypothesis the observed t is compared to the appropriate cumulative density from the t-distribution with the appropriate degrees of freedom. Degrees of freedom indicate the number of values that are free to vary with the given test statistic. Here the comparison is between two means, each of which ‘costs’ one degree of freedom, the null distribution thus being $t_{n_1+n_2-2}$. The cumulative density for a two-sided test (as shown above) is taken at $\alpha/2$, where the usual cut-off for α is 0.05.

When one of the treatment groups is likely to be more variable than the other it is not appropriate to pool the standard deviations. This can happen, for example when a healthy population is compared to a population containing subjects with diverse disease progression states. The denominator of the t-statistic is still the weighted mean of the sample variances $\sqrt{s_1^2/n_1 + s_2^2/n_2}$, but the degrees of freedom are now approximated by

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1 - 1) + \left(s_2^2/n_2\right)^2/(n_2 - 1)} \tag{3.5}$$

Large differences between the two test statistics only occur when there are substantial differences in sample sizes and/or standard deviations (Boneau, 1960; Lumley et al., 2002; Moser and Stevens, 1992).

In case of severe departures from normality it is more appropriate to use a rank-based test, such as the Wilcoxon rank-sum test, also known as the Mann-Whitney U test. With this test the ranks of the abundances of each group are summed and used as test-statistic, and the null distribution is generated by enumerating all possible rank sums, given the sample sizes. The choice of test should ideally be made as part of the design of the experiment.

To test for normality when the data has been collected presents several problems. Firstly, a test of normality, such as the Shapiro-Wilk test, has very low power with small sample sizes. With large sample sizes, approximate normality is achieved due to the central limit theorem, which ensures that the t-test is appropriate. However, with large sample sizes small departures from normality can be identified, rejecting normality at a level where the t-test is still appropriate. Secondly, departures from normality can present as skewness (one tail longer than the other) and/or as kurtosis (longer or shorter tails). Of these, only skewness has a detrimental effect on the t-statistic when the skew is in opposite directions in the two groups (Boneau, 1960; Box, 1953; Lumley et al., 2002).

3.1.2 Multiple testing correction

With per-protein analyses, p-values for the individual proteins should not be viewed in isolation of each other. The two common ways of correcting for multiple testing are by limiting the family-wise error rate (FWER), or by limiting the false discovery rate (FDR). FWER limits the probability of one falsely rejected hypothesis (i.e. the chance of one false positive) to a specific level, while FDR limits the fraction of the expected number of falsely rejected hypotheses amongst all rejected hypotheses.

In discovery studies the interest is in finding a limited set of proteins for further study, as false discoveries can be filtered out again later, as long as there are not too many of them. It is more important not to have many false negatives, proteins that are important for the difference between the treatment groups but are discarded for further analysis. FWER would thus be too strict for this setting, and FDR the better choice (Hastie et al., 2009).

The FDR estimation procedure that is most used in proteomics is the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) which starts from an ordered list of the m p-values (ascending) and rejects all hypotheses $i : p_i < \frac{i}{m}q$, with m the total number of hypotheses (proteins) and q the required FDR level. Instead of calculating cut-off values for each protein, it is possible to calculate adjusted p-values as $p_i\frac{i}{m} = p_i^*$. Hypotheses where $p_i^* < q$ are then rejected.

One of the assumptions underlying most FDR estimation procedures is that the p-values of all the hypotheses follow a mixed distribution, with the p-values of the non-differentially abundant proteins following a uniform distribution between 0 and 1, and the p-values of the other proteins concentrated near 0. This assumption can be tested by e.g. generating a histogram from the p-values or using calibration plots as described in Gianetto et al. (2016).

As the FDR is a step-wise procedure, selecting proteins from the list of rejected hypotheses invalidates the FDR, i.e. the fraction of false positives can be higher or lower than the chosen level. This is, however, common practice in proteomics discovery studies (Schwämmle et al., 2020). In Goeman and Solari (2011), a different approach is taken, which is especially relevant for these settings. Here, a closed testing procedure is used to calculate the expected number of false discoveries for every subset of hypotheses simultaneously. Thus, when a researcher has performed the per-protein analysis, selects some proteins based on FDR and some proteins based on their expert knowledge, the expected number of false discoveries can be found by applying this procedure.

3.1.3 Testing multiple variables

If there are additional variables that have to be taken into account, such as batch- or sex-effects, the preceding tests are not appropriate. Ignoring batch-effects, or other variables that can influence the difference between the treatment groups, can lead to incorrect inferences (Box et al., 2005). Instead a (linear) model should be employed to estimate the effect of all variables under consideration. For example, when there is a binary treatment variable (e.g. *Placebo* vs *Treatment*) and two batches, a model can be defined as

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 B_i + \varepsilon_i \tag{3.6}$$

with Y_i the protein abundance of protein i, T_i and B_i indicator variables for the treatment and batch respectively, β_0 the intercept, β_1 and β_2 the estimated effect of the treatment and the batch respectively, and ε_i the deviation of subject i from their expected value given the model. ε_i is thus a random error with mean 0 and variance σ^2 (Kutner et al., 2005).

When the interaction between the treatment variable and another variable is also of interest, interpretation of the results becomes more involved. Say there is a binary treatment variable (e.g. *Placebo* vs *Treatment*) and two sexes (*Female* and *Male*), and the interaction between the treatment and sex is expected to be of interest. A model can then be defined, similarly as above, as

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 T_i S_i + \varepsilon_i \tag{3.7}$$

The effect of the treatment is now partly the marginal treatment effect β_1 and partly the interaction effect β_3. This makes (semi-)automated selection of 'interesting' proteins for the discovery experiments difficult, but could aid with interpretation of validation experiments.

As the test statistic is defined by the difference between the groups divided by the spread of the data, proteins with small fold change and small variance can appear highly significant. To stabilise the variance, pooling variances from all proteins and pulling the variances of the individual proteins towards this pooled variance, an empirical Bayes approach such as LIMMA can be used (Smyth, 2004).

3.1.4 Missing values

Often many of the quantified proteins contain missing values for some of the subjects. There are generally two ways missing values can be handled in a statistical analysis: either variables or subjects with missing values are removed from the analysis, or the missing values are imputed. Removing variables or subjects can lead to biased inferences, and should not be done without thorough justification. Imputation has to be based on the type of missingness and allow for a quantification of the uncertainty of the imputations (Molenberghs et al., 2015; van Buuren, 2018).

One of the simplest imputation methods is to replace the missing values by the median of the observed values, or half of the lowest observed value. By imputing a single value, the assumption is made that the imputed value is the only possible true value, reducing the overall variance in the data. Moreover, as usually none of the protein abundances are exactly alike, it would actually be highly unlikely that all missing values are exactly the same.

One way of taking the likely location and variability of the data into account is to sample imputed values from a normal distribution based on the mean and variance of the observed protein abundances. The Perseus software takes this approach, assuming values to be missing due to low abundance (Tyanova et al., 2016). This requires a strong assumption on the expected location and variability of the missing values. In addition, as each missing value is imputed by a single imputed value, the uncertainty of each imputation is not carried forward into downstream analysis.

Instead, the analysis of missing values should be either part of the downstream data analysis (i.e. joint modelling), or each missing value should be imputed multiple times, such that the uncertainty of the imputation can be taken into account (Molenberghs et al., 2015; van Buuren, 2018).

The mechanism by which values go missing can be categorised into three main categories: missing completely at random, missing at random, and missing not at random (Table 3.1). Censored variables are included in the missing not at random category, but are interesting in and of themselves. Thorough treatment of the handling of miss-

Table 3.1: Mechanisms for missing values. MCAR: Missing completely at random. MAR: Missing at random. MNAR: Missing not at random. Censored values are also MNAR.

Mechanism	Missingness dependency
MCAR	Independent of both observed and unobserved variables
MAR	Dependent only on fully observed variables
MNAR	Dependent on unobserved variables
Censored	Dependent on missing values themselves

ing values in different fields and manners can be found in handbooks on missing values such as Molenberghs et al. (2015) and van Buuren (2018), which provided the basis for the following descriptions.

When values are missing completely at random (MCAR), the missingness (whether an observation is observed or not) is independent of both observed and unobserved variables. Thus the missing values can be ignored (i.e. the subjects can be left out of the analysis) without biasing the analysis. Unfortunately, this hardly ever happens.

When values are missing at random (MAR), the missingness is dependent only on fully observed variables. Only when all the variables that influence the missingness are included in the analytical model can the missingness be ignored. The assumption that the missingness depends only on some fully observed variables is impossible to verify, thus in specific cases it will need to be substantiated with convincing arguments about the nature of the data.

When the missingness depends on unobserved variables, the data is not anymore ‘at random’. A special case of this presents itself when the missingness depends on the missing values themselves. In a proteomics context this could, for example, be peptides in data-dependent acquisition that for some samples have too low abundance to be selected for fragmentation. When this happens for all peptides of a protein, the protein in question can be missing for the subject. For these last two cases, the data are missing not at random (MNAR), and the missingness will either have to be modelled explicitly, or some bias in the inference will have to be accepted.

The missingness mechanism should be leading for the imputation method used. For MCAR and MAR the imputation strategies can be more or less general, depending only on the observed values and (context-specific) dependencies between the variables. Contrariwise, the imputation strategies have to be context-specific for MNAR, depending on the reason why the values go missing. In a recent review on imputation strategies for label-free proteomics experiments (Lazar et al., 2016), it is concluded that: (i) both MCAR and MNAR values occur in proteomics data, in varying proportions; (ii) in absence of knowledge about the nature of the missing values they show it is most appropriate to consider them as MCAR/MAR; and (iii) when the missingness mechanism is known, on the other hand, it is better to use a strategy dedicated to the type of missingness in the data.

As the data from the mass spectrometer is in the form of peptide abundances, if no imputation is done at the peptide level, there is an implicit assumption about the nature of the missingness mechanism when deriving the protein abundance from (partially missing) peptide abundances. Therefore, the imputation should ideally be performed at the peptide level. It is also important to note that most current software for imputation presents an imputed dataset without propagating the uncertainty around the imputed values. Imputed values are thus presented at the same level of certainty as observed values, which can be potentially misleading. Instead, with multiple imputation or joint modelling of the missing values and protein quantities/group differences the uncertainty about the missing values can be propagated to protein quantifications and/or test statistics for differential abundances (Goeminne et al., 2020; Lazar et al., 2016; Schwämmle et al., 2020; The and Käll, 2019).

3.1.5 Handling outliers

Observed values that fall far outside the range of the other observed values are called outliers. These form a challenge similar to that of missing values, as they can have biological reasons, in which case they are valid values and care should be taken how to handle them, or technical reasons, in which case they are invalid values that should be discarded. Often it is not possible to find the reason why a value is an outlier. In

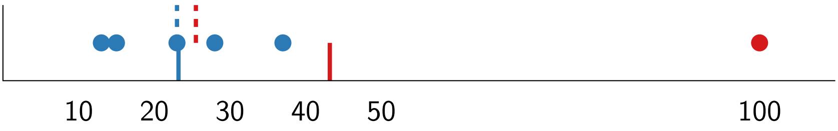


Figure 3.1: The effect of an outlier on the mean and median of a sample. Blue dots: main sample. Red dot: outlier. Solid blue line: mean excluding outlier. Solid red line: mean including outlier. Dashed blue line: median excluding outlier. Dashed red line: median including outlier.

information that is coded manually, coding errors can occur which are easier to spot, e.g. a length in inches where there should be centimetres. In computationally generated measurements as from a mass spectrometer, or in the quantification of proteins from their peptide values, it is less obvious how to spot the nature of outliers.

A single outlier can greatly influence the mean of a set of values. When comparisons between groups are based on the mean abundance values, aberrant values can thus have a big influence on the results of the analysis. To give an extreme example, when the observed values are between 10 and 40, adding only a single outlier, say 100, the mean already migrates outside the range of the original values (Figure 3.1, solid lines). The median is the value that is both higher and lower than 50% of the observed values. If instead of the mean, one would perform inference based on the median, the influence of the outlying value is much reduced (Figure 3.1, dashed lines). Thus, the median is a measure of location that is robust against outliers. Similar statistics exist for variance (e.g. median absolute deviation, interquartile range), and for other statistics in general, e.g. rank-based statistics are robust to outliers by design. The downside of robust statistics is that they and their null-distributions are usually harder to compute.

3.1.6 Multivariate modelling and dimension reduction

When the number of subjects is smaller than the number of proteins, the usual regression model with proteins as parameters is not identifiable. Thus, penalised regression or other methods should be employed (Hastie et al., 2009). The two most common penalised regression methods are ridge regression and the least absolute shrinkage and selection operator (lasso), with elastic net being a combination of the two. These methods impose a penalty on the size of the parameters, shrinking them towards zero and towards each other. Where ridge regression shrinks parameter values towards zero (but not at zero), the lasso does. Unfortunately, the lasso has problems with highly correlated features, which are common in proteomics data. By using elastic net, groups of correlated features can be selected (or rejected) together. When the interest is in specific, pre-defined groups of proteins (such as protein complexes or pathways), a method such as (sparse) group lasso can be used.

Different machine learning techniques, such as e.g. random forests or neural networks can also be used (see e.g. Hastie et al. (2009) for an overview). These are, however, less easily interpretable as it is unclear what the relation of the proteins are to each other and to the outcome of the final decision rule. All the methods described in this section require the optimisation of hyperparameters: extra variables that define, e.g. the level of the penalty (for penalised regression), the depth and number of trees (for random forests), or the depth of, and number of nodes in, neural networks. To optimise these hyperparameters, models need to be fitted multiple times with different values for the hyperparameters to compare their performance. This can be done by splitting the data set into a training-set and a test-set. For each value of the hyperparameters, the model is fitted on the training-set, and the performance of the model with the given hyperparameter values is subsequently tested on the test-set, with the best performing hyperparameter values chosen for the final model. The data set thus needs to be large enough that both training- and test-set are of sufficient size to fit a decent model and to give a meaningful performance evaluation, respectively. One thing to be careful about is that the different treatment groups should be present in sufficient quantities in both training- and test-set, and cross-validation uses the data more efficiently (Hastie et al., 2009).

A different approach is orthogonal projections to latent structures (OPLS) (Trygg and Wold, 2002). In this method, linear combinations of all variables are generated that explain the maximum variance in the protein abundance data in the direction of the variables of interest. Additionally, variance orthogonal to (i.e. not correlated with) the variance of interest is modelled, allowing for the analysis of both the differences of interest and, separately, where additional noise in the data might come from. A natural comparison can be made with principal component analysis (PCA), which generates linear combinations of variables in the direction of the most variation in the data. This is however an unsupervised method, in which an extra analysis is necessary to map the principal components to either variables of interest or systematic noise.

3.1.7 Peptide-centric and protein complex-based approaches

Several novel approaches apply a statistical model to peptide abundances instead of the aggregated protein abundances. This both avoids the loss of uncertainty of the protein abundance estimation, combining abundance estimation and differential abundance testing. The approach in Goeminne et al. (2016) is to fit a robust ridge regression on the peptide abundances, while in The and Käll (2019) a Bayesian approach is used.

A per-protein analysis assumes that the protein abundances are independent of each other. Later in the analysis, when performing pathway analysis, a major assumption is that proteins work together and the abundance of proteins in active pathways should rather be correlated. This can be seen clearly in enzyme inhibitors, which can inhibit the production of some other enzyme. Similarly, proteins work together with other proteins in protein complexes. When proteins are in the same complex they could be assumed to have correlated abundances. For example, several ways of performing statistical analysis on the protein complexes instead of the individual proteins are descibed in Goh and Wong (2016a). This approach can find additional low-abundant proteins that might be interesting for further analysis that would otherwise be rejected.

3.2 Pathway analysis

It is common practice to first perform a per-protein analysis, followed by a mapping of the selected proteins to pathways. This way the per-protein analysis can hopefully be put into their biological context, and additional potentially interesting proteins can be included in follow-up analyses.

As pathways are series of reactions, where the output of one reaction is the input to the next reaction, they can be modelled as directed networks of reactions. Elements that are part of reactions can be proteins, metabolites, RNA, and other molecules, from here on collectively referred to as molecules. The reactions can be seen as the edges between molecules (e.g. KEGG (Kanehisa et al., 2016)), or as nodes with molecules connected to them (e.g. Reactome (Jassal et al., 2020)).

Looking at the complex nature of protein interactions and the pathways proteins are active in, a structural way of analysing the pathways in which differentially abundant proteins participate can help with interpreting the mapping of proteins to pathways. As proteins can participate in multiple pathways, and pathways have different sizes (numbers of constituent proteins), the main idea is to calculate whether in a pathway there are more proteins found differentially abundant than expected by chance. What is expected by chance (and how to calculate this) is mostly where the different pathway analysis methods differ.

In Khatri et al. (2012) three generations of pathway analysis methods are identified: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology (PT). ORA starts with a list of differentially abundant proteins. Then, for each pathway, the number of proteins in that list is compared to a background list of proteins, e.g. all the identified proteins. A Fisher exact test (or similar) is then used to test whether the proteins in the pathway are over-represented in the input list of proteins. This only uses the number of differentially abundant proteins, and thus ignores both the differential abundance and the variability of the proteins in the list. Moreover, by only using a subset of the identified proteins a (more or less arbitrary) cut-off has to be made between proteins that are included and those that are deemed irrelevant. The strict cut-off thus created can lead to a false dichotomy between those proteins that are marginally included and those that are marginally excluded. Finally, both the proteins and the pathways are treated as independent from each other, which is paradoxical as proteins in a pathway are unlikely to act independently from each other. Similarly, pathways are connected to, and nested in, each other, and are thus not independent of each other.

The main idea behind FCS is that both big changes in protein abundance and many small changes within a pathway are important. Starting from the per-protein statistics of all identified proteins, these are then transformed into a score for each pathway, which can depend on e.g. the per-protein statistics, pathway-size, and correlations between different protein abundances. Statistical significance can then be computed by permuting class labels and comparing the proteins within each pathway, or by permuting protein labels and comparing the proteins in the pathway to proteins not in the pathway. With this approach arbitrary thresholds are avoided, and the level of differential abundance is used at least to some extent. However, the pathways are still considered as independent from each other, just as with ORA.

PT methods follow the same steps as FCS, but make use of the topology of the pathways when calculating the per-protein statistics. The topology of a pathway is concerned with which proteins have roles in the same reaction and what those respective roles are. If *Protein A* is a catalyst of *Reaction R*, and *Protein B* is the output, the relation between the two proteins would be different to when *Protein A* is an inhibitor instead. Similarly, if *Protein A* and *Protein B* are in the same reaction, their scores can be expected to correlate more than when they are three reactions separated from each other. Elements like this can potentially be used to compute more relevant scores for the pathways. However, connections between pathways are still difficult to account for with this approach.

3.3 Protein networks

3.3.1 Analysing network structure

Investigating the network structure of pathway databases can be done at the level of the network or the level of the constituent proteins. At the most general level, the network consists of a number of nodes (the proteins) and edges between them (protein interactions). A group of nodes which are connected to each other by edges is called a connected component (Harary, 1969). The nodes do not have to be directly connected, i.e. if *Node A* is connected to *Node B*, and *Node B* is connected to *Node C*, and there is no edge directly between *Node A* and *Node C*, this is still one connected component. Additionally, the size of the connected component can be defined by its diameter: between each pair of nodes in the connected component there is a shortest path, the smallest number of edges connecting the two nodes. The longest of all these shortest paths is the diameter of the connected component.

In a pathway network, the nodes are reactions with inputs and output, and possibly catalysts and/or regulators, where the output of one reaction can form the input to another reaction. This can also be seen as a directed network, with edges going from input/catalyst/regulator to output. The number of edges of a node is called the degree of the node, with in-degree (the number of edges going to the node) and out-degree (the number of edges exiting the node) defined for directed networks (Harary, 1969). A straightforward way of looking at the importance of a node is to look just at the degree of the node. However, this only gives an indication of the immediate neighbourhood of the protein, which is a rather limited perspective given the interdependent nature of the reactions. Calculating the average number of steps in the network between a protein and all other proteins in the network can act as a proxy of how influential a protein could be across the network (Valente and Foreman, 1998).

Proteins often perform their functions together with other proteins in protein complexes. All proteins in a complex have to be present for the complex to be able to perform its function. The conceptual opposite is proteins that are alternatives for each other in a reaction or a complex. A set of such proteins can be annotated as an entity set. Then only one of these has to be present for the reaction to be possible. This is important for the pathway analysis, as finding multiple proteins in an entity set might not be of as much interest as finding all proteins in a complex, and makes the interpretation of the network more challenging.

The analysis of the structure of pathway databases is further explored in *Paper I*.

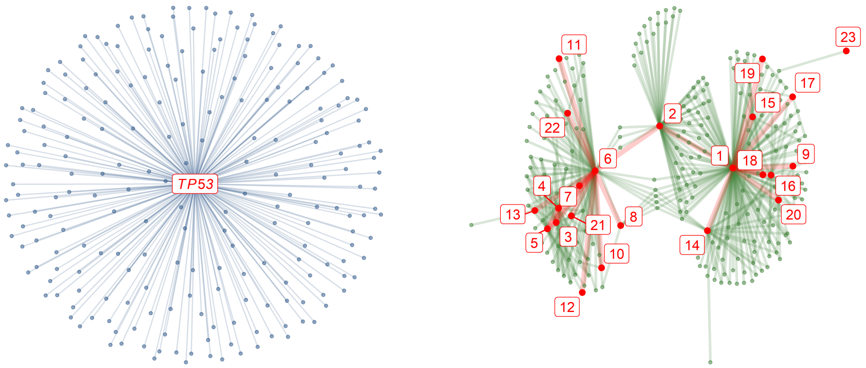
3.3.2 Proteoform- vs gene-centric networks

Pathway databases are designed with specific goals and use-cases in mind. For example, KEGG (Kanehisa et al., 2016) originated with a focus on genes and metabolites, while Reactome (Jassal et al., 2020) is centred around reactions involving metabolites and proteins. Aggregating the interactions of a protein into its parent gene combines the interactions of multiple proteins that originate from that gene, making a protein-centric database necessarily more specific than a gene-centric database. Similarly, proteoforms and PTMs are important for specifying the function of a protein. These are increasingly commonly identified in samples, but only to a lesser extent annotated in pathway databases.

Where proteoform-level information is annotated in the pathway database, it indicates minimal requirements on the status of a protein to perform a reaction, e.g. phosphorylation at a given site. By creating a proteoform-centric pathway network it becomes clear that the proteoforms make the reactions more specific. Thus, analysing data at the level of proteoforms leads to less irrelevant pathway hits. Additionally, it shows that there are interactions between different proteoforms of the same protein, which are lost at the protein (or gene) level.

Both these aspects can be illustrated with the interactions of annotated proteoforms of the protein *cellular tumor antigen p53* (P04637). At the gene level, there are 220 interactions between the *TP53* gene and other genes (Figure 3.2a). Splitting the genes into their corresponding proteoforms, the network expands to 23 annotated proteoforms originating from the *TP53* gene, with 24 interactions between two proteoforms originating from this gene, and 390 interactions with 227 proteoforms originating from other genes (Figure 3.2b).

While it is not always possible to identify the precise protein from a gene, or proteoform of a protein, with the current methods, where these can be identified they can possibly make the interpretation of the experimental results more specific. In *Additional Paper I* we developed a tool to query the Reactome pathway database at different levels of specificity, allowing a more fine-grained interaction and pathway retrieval depending on the level of the protein identifications in the experiment.



(a) Gene-centric network

(b) Proteoform-centric network

Figure 3.2: Proteoform-centric vs gene-centric network representation of the protein *cellular tumor antigen p53* (P04637). **(a)** Gene-centric network, with other genes depicted as small blue nodes. **(b)** Proteoform-centric network, with the proteoforms of the *p53* protein numbered in red nodes, and proteoforms of other proteins as small green nodes. Red edges indicate interactions between *p53* protein isoforms, green edges interactions between *p53* protein isoforms and isoforms from other proteins. Figure adapted from *Additional Paper I* (Hernández Sánchez et al., 2019).

4 Design of proteomics experiments

When designing a comparative proteomics experiment there are multiple aspects to take into account. As a general rule, the better the experiment is planned, the more straightforward the analysis of the data. The first step is deciding what treatments (or disease states, etc.) to compare and which other variables are likely to influence the outcome measurement(s) studied. Next, the type of samples should be appropriate to answer the research question, and subjects need to be recruited or cell lines grown. Finally, to be able to answer the research question, a sufficient number of samples should be available, and the sample processing order should not introduce confounders (Box et al., 2005).

4.1 Sample populations

Different research questions require different types of samples to be answered efficiently. Samples can be taken from a single subject, or from an environment, e.g. soil samples. The latter consist of multiple species, so-called metaproteomics, and are outside the scope of this thesis. On the other end of the spectrum there is single-cell analysis, in which the proteome of each cell in the sample is analysed separately. This technique is becoming increasingly popular (e.g. Specht et al., 2021), though the bulk of proteomics analyses is still done on populations of cells.

One of the main differences for single subject samples is between biological samples and cell lines, which are grown in a laboratory. By applying a treatment to the cells, one can see what happens to the treated vs. the untreated cell line. If the treatment would do nothing at all, the treated and untreated cell lines should be the same. Biological samples are taken from biological entities, such as humans or other animals, which differ on more than just the treatment. This biological variation needs to be taken into account when modelling the effect of treatments, and is vital for extrapolating to individuals outside the study.

When using samples from subjects, one of the first questions concerning the design of the experiment is whether to make comparisons between groups of subjects or comparisons within subjects in different states. When the interest is in differences between disease states, the only option is often to recruit subjects who do have the disease and subjects who do not have the disease. This results in average differences between subjects in different disease states. In diseases where the disease manifests itself in only a part of a tissue, as in specific cancers, it is possible to take a ‘healthy’ and a ‘disease’ sample of the same tissue type from the same subject. Then, the difference between the tissues from the same subject should only differ in terms of the disease, while for the whole population the variation between subjects can still be modelled.

Similarly, when analysing the effect of a drug, a common strategy is to perform a cross-over trial or a longitudinal study. In a cross-over trial, subjects receive two or more treatments at different time points and one can compare how each subject reacts to both treatments. To avoid biases, subjects should receive the treatments in a random order and there should be sufficient time between the treatments.

When interest is in the progression of a disease or the recovery trajectory of subjects, measuring protein abundances at multiple time points in a longitudinal study can show differences in the trajectories of the subjects. There are generally two ways to approach a longitudinal study, which is to either process samples as they come in, or to process all samples together after collecting all samples (Mertens, 2017). The former has the advantage of avoiding storage effects, but introduces sample processing days or batches as a confounder for the time-dependent effect studied. However, the latter approach introduces sample storage time as a perfect confounder, in which case care should be taken that the interest is in proteins that are stable when stored over longer time periods (Gast et al., 2009; Meier et al., 2020; Tsuchida et al., 2018; Tworoger and Hankinson, 2006).

4.2 Power and balance

When the types of samples, the type of analysis, and the research question is decided upon, a statistical model should naturally present itself. This model should account for the research design, and should make it possible to answer the research question. Once a preliminary model is chosen, it is advisable to calculate whether the number of available samples is enough to estimate the parameters of interest with satisfactory precision. Alternatively, the number of samples needed for this can be calcualted.

The traditional way of calculating the required number of samples is by performing a power analysis, which is done by calculating the Type I (false positive) and Type II (false negative) error probabilities based on pre-specified effect size, variance and sample size (Oberg and Vitek, 2009; Riter et al., 2005). Type I and 1 − Type II error rates are also called the size and power of a test. However, given the interdependent nature of protein abundances, the complex models and expected high prevalence of missing values, and the subsequent multiple testing correction, it is not always clear how to perform or interpret classical power analyses in omics settings. While sample size calculations for omics studies have been proposed (e.g. Müller et al., 2004; Tibshirani, 2006), these all rely either on pilot data or known/expected test statistic distributions. In a typical discovery study, however, this might not be available. With potentially limited population sizes, and the high cost of acquiring and processing samples, the discovery phase of the experiment is often seen as a kind of pilot study for the validation phase. Expected variances and effect sizes might also be hard to estimate from related previous studies, when these are done with different protocols, using different instruments, and/or different software pipelines.

It is therefore easier to perform a power analysis in the validation phase. Then, one can use the effect size and variance estimates from the discovery phase to estimate what an appropriate sample size would be for the proteins/peptides included in the validation phase.

An alternative to power analysis is an evaluation in terms of Type S (sign) and Type M (magnitude) errors (Gelman and Carlin, 2014). These give an indication concerning how likely it is that the magnitude of the effect is far from the true effect size (a Type M error), and how likely the true effect is actually in the opposite direction (a Type S error). For all these statistics concerning the design of the experiment, i.e. both for power calculations and Type S- and Type M-errors, it is important to remember that the initial effect size and variance estimates on which these calculations are based should come from realistic external estimates. Both published literature and the ‘significant’ proteins from the current study are likely to overestimate effect sizes.

The numbers of samples that are recruited can always, for various reasons, be lower than what was aimed for. Ideally, the number of subjects recruited is (nearly) balanced in terms of the characteristics important for the research question, i.e. the variables in the analytical model. When there are many more samples in one group compared to another the estimation of group differences can become challenging. Balance has certain advantages as parameter estimation is optimal (i.e. has the least variance) when sample sizes are equal, for a given total sample size.

In general the more samples the better, though there are some notable exceptions. Firstly, the larger the sample size, the better parameters such as the mean or group differences can be estimated. However, at some point the sample size is large enough that the exact minimal meaningful difference between the groups can be observed, and increasing the sample size(s) after that is a waste of effort and materials. Similarly, when there is a large imbalance between sample sizes, adding more samples to the larger group might not meaningfully make an impact on the power of the test while adding more samples to the smaller group would.

It might happen that for some groups the number of subjects recruited is much too low. In that case it might make sense to reevaluate the statistical model and perhaps drop a variable one is less interested in. Of course, this is only possible if the proposed model still makes sense without the dropped variable.

4.3 Batching and sample processing order

When the samples are collected, the processing of the samples has to be prepared. The number of batches and how the treatment groups are divided into them should be deliberated in the design phase of the study. Ideally all samples can be processed as a single batch. When this is not possible, care should be taken to avoid the introduction of confounders when dividing samples over batches.

The number of samples that can be processed at once is usually limited by the time it takes to process one sample and/or limitations of equipment. As samples generally have to be stored under a strict set of circumstances, e.g. long term storage at −80°C, and start to degrade when defrosted, the number of samples that can be processed after defrosting can be limited (Tsuchida et al., 2018; Tworoger and Hankinson, 2006). Similary the number of samples that can be processed on e.g. a single 96-well plate is limited to 96, and with 10-plex TMT the maximum number of samples in one pool is nine or ten, depending on the inclusion of a shared reference sample. The need for batch processing is thus commonplace.

A batch is a set of samples that is treated (almost) exactly the same way. Say one set of samples (Set *A*) is processed on one day, and another set of samples (Set *B*) is processed the next day, these would be two different batches. When, in addition, both sets of samples are processed by two different instruments (Instrument *1* and Instrument *2*), this will lead to four batches: *A1*, *A2*, *B1*, and *B2*. However, if all samples of Set *A* are processed on one instrument, and all samples of Set *B* on the other instrument, there are still only two batches: *A1*, and *B2*. The batch effects introduced by the days of processing and those introduced by the different instruments are confounded, and thus combined into a single batch effect. Each type of batch (here *instrument* and *processing day*) has to be added to the analytical model as a factor variable. If the batch variable is not included in the model, the batch effect would at best, i.e. when orthogonal to the treatments, increase variances, but more likely also influence the estimates of the treatment effects.

When creating batches, the main goal is to make each batch like its own small experiment. Two general rules of thumb can aid in this process:

1. *All batches should be as similar to each other as possible.*
2. *Each batch should contain as many different treatments as possible.*

It is easy to make all batches exactly the same when each treatment has the same number of subjects, the number of subjects per batch equals the number of treatments, and the number of batches equals the number of subjects per treatment. In that case each batch contains one subject of each treatment. Where it is not possible for the batches to be exactly the same, the similarity of the batches should be seen in terms of overlap between the treatments represented in each batch. The number of subjects in each treatment should thus ideally be representative of their relative frequency in the complete experiment. When the number of treatments is larger than the batch size, each batch should contain as many different treatments as possible.

For specific settings, where interest is only in a certain subset of treatment contrasts, other strategies should be followed. A typical example is where two current treatments (T_a and T_b) are compared to a single new treatment (T_n), and the main goal is to compare the new treatment to each of the current treatments. Comparisons between the two current treatments would not change the evaluation of the new treatment and are consequently not of interest. Batches would then mainly (or only) combine samples from the new treatment with the current treatments, rather than combining samples from the current treatments.

Even with only a single batch one can usually process only one sample at a time, both in the laboratory and on the instrument that analyses the samples. This can introduce a time-dependent effect, which means that one should also consciously decide on a processing order of the samples. The most appropriate strategy is to use block randomisation to decide upon the order of the treatments, followed by a random allocation of subjects to slots according to their treatment (Figure 4.1). While it is important to minimise time-dependent effects using block randomisation, run-order effects on LCMS are usually unavoidable, making it essential to afterwards check the run-order effect. This also underlines the importance of reference samples, which should be included in each batch and run at set intervals in the run order (Surowiec et al., 2018). Further details on block randomisation can be found in *Additional Paper II*.

While for smaller cohorts and standard and/or balanced settings it is feasible to find (near-)optimal batch allocations by hand, this can

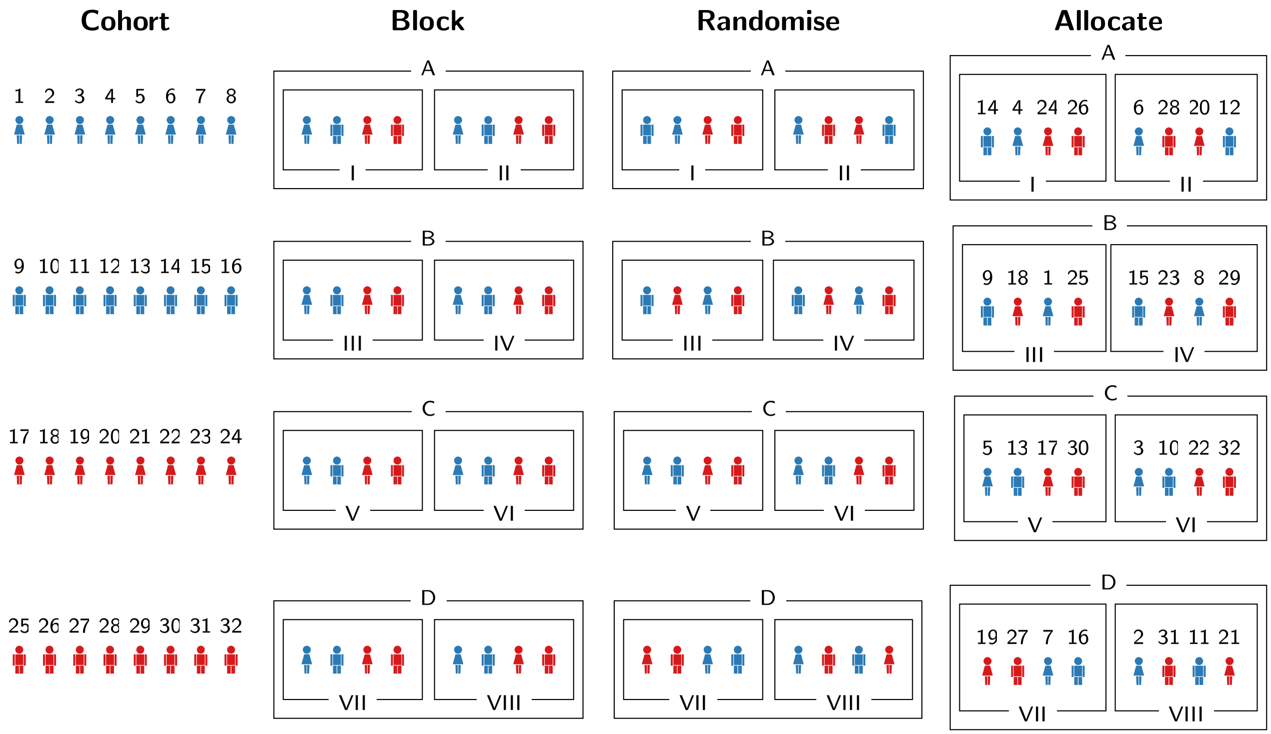


Figure 4.1: Batch allocation and block randomisation. The cohort shows the recruited subjects and their identifier. Each batch contains two blocks with one subject of each group. In each block the order of the groups is randomised. Subjects are then randomly allocated to a spot for their group.

quickly become challenging and time-consuming. An automated procedure for allocating subjects into batches can aid researchers, freeing them to concentrate on other aspects of the experiment. Therefore we devised an algorithm to allocate subjects into batches, based on the principles stated above, where blocking is based on only a single variable, and all treatment contrasts are considered equally important. Further details on the algorithm and its performance can be found in *Paper III*.

5 Papers

5.1 Paper I: Interpreting the structure of pathways

To optimally interpret pathway analysis results, it is important to understand the structure and organisation of the underlying database. Using network statistics we investigated the evolution of the connectivity of the Reactome pathway knowledgebase, and described the hierarchical and interconnected nature of pathways. Taking into account the structure of the protein network when performing pathway analyses can make the analysis more specific, but also more complicated.

This is further described in *Paper I*:

Burger B., Hernández Sánchez L.F., Lereim R.R., Barsnes H., Vaudel M. Analyzing the structure of pathways and its influence on the interpretation of biomedical proteomics data sets. *Journal of Proteome Research*, 2018, 17, 11, 3801–3809.

5.2 Paper II: Detecting small endogenous peptides

In general peptidomic and proteomic workflows, peptides (and single amino acids) that cannot attract a charge are structurally missed in the analysis. We performed a proof-of-concept experiment extending an existing workflow, allowing us to identify single amino acids and small endogenous peptides in human cerebrospinal fluid, using a basic mass-based identification approach. Further development of identification and quantification strategies for these molecules might provide a wealth of additional information, and potentially biomarkers for neurological diseases, using isobaric tags and targeting singly charged molecules.

This is further described in *Paper II*:

Burger B.§, Lereim R.R.§, Berven F.S., Barsnes H. Detecting single amino acids and small peptides by combining isobaric tags and peptidomics. *European Journal of Mass Spectrometry (Chichester)*, 2019, 25, 6, 451–456.

§ These authors contributed equally.

5.3 Paper III: Automated blocking and batching

In biomedical experiments, batches are generally designed by hand, which becomes cumbersome for challenging cohort set-ups. We have developed a fast and intuitive heuristic algorithm to generate balanced allocations of samples to batches, which can be applied to any single-variable model where the treatment variable is nominal. We show that this algorithm provides a marked improvement over random allocations and yields an optimal solution for small cohorts (up to 10^8 allocation possibilities).

This is further described in *Paper III*:

Burger B., Vaudel, M., Barsnes, H. Automated blocking and batching of biomedical experiments with sequential processing. *Unpublished preliminary results.*

6 Additional work

Following is a list of papers describing additional work performed during the period when the PhD was carried out, but which are not included as part of the thesis.

Additional Paper I:

Hernández Sánchez L.F., **Burger B.**, Horro C., Fabregat A., Johansson S., Njølstad P.R., Barsnes H., Hermjakob H., Vaudel M. PathwayMatcher: proteoform-centric network construction enables fine-granularity multiomics pathway mapping. *Gigascience*, 2019, 8, 8, giz088.

Additional Paper II:

Burger B., Vaudel M., Barsnes H. Importance of block randomisation when designing proteomics experiments. *Journal of Proteome Research*, 2021, 20, 1, 122–128.

7 Discussion

This thesis mainly focuses on issues concerning the design and analysis of proteomics experiments. A number of these topics warrant further discussion. First, additional challenges in performing and interpreting pathway analysis will be addressed. These challenges unavoidably arise from the complex biological entities and processes they describe and from the way the databases are built up and maintained. Next, opportunities for expanding proteomics experiments beyond single studies and single laboratories will be discussed. This includes combining multiple studies in meta-analyses, but also the combination of multiple omics platforms to create a systems biology perspective of the relevant differences between populations. Lastly, challenges and opportunities of multidisciplinary research are covered, with a focus on the need for clear and tailored communication.

7.1 Limits of pathway analysis

The main challenges of performing pathway analysis and interpreting the results were already discussed in Chapter 3.2. However, there are more general challenges to pathway analysis, most notably concerning the differences between species and the constant growth of the knowledge base.

Species- and disease-specific pathways When choosing protein and pathway databases, the reference database should be aligned with the species under investigation, e.g. if a study is done in mice, the protein and pathway database should be specific for mice. Even though mouse models are often used to study what would happen in humans, there are differences in proteins, and pathways and processes in mice cannot directly be mapped to human pathways. This does not mean that mouse models are not relevant or important, but the mapping from mouse to human is a separate step that requires expert knowledge about the relevant differences (Mohammed et al., 2020).

Similarly, disease-specific pathways are specific for the given disease. Overlaps with other disease-mechanisms can be very interesting, but as the pathway analysis algorithms will try to map to anything that is in the database, the 'most relevant' pathways in the database might not be the most relevant when interest is in a less exhaustively annotated disease. The disease pathways included in the database should thus be specific for the disease under consideration, possibly including related diseases, to avoid drawing erroneous conclusions.

Database curation Manually curated databases update slower, but the content should be more trustworthy than that in computationally derived databases. Additionally, there are commercial pathway databases, e.g. Ingenuity Pathway Analysis (Qiagen), which are curated, but where the curation process is not necessarily known to the user. In terms of species-specific pathways and reactions, for example, the Reactome pathway database is curated for human interactions, but 'inferred from humans' for all other species (Jassal et al., 2020). Additionally, some pathways are better annotated than others, and annotations of pathways can differ between databases (Müller et al., 2011). Part of the differences in levels of annotation between pathways may lie in sociological biases. Curation efforts seem to be focused mainly on proteins that are already well annotated, and the proteins they interact with, biasing knowledge generation towards areas where there already is extensive annotation (Rolland et al., 2014). It can be hypothesised that the same is true for resources and efforts directed towards different diseases (see, e.g., the biology and diease oriented branch of the Human Proteome Project Aebersold et al., 2013). Thus, a database should be selected which has the required level of curation relevant to the research question.

Pathway specificity An additional challenge is that one can only find those interactions and pathways that are annotated in the database. Pathway databases are constantly expanding to incorporate new knowledge. Results of different experiments analysed with different (versions of) databases might not be directly comparable, and reanalysing the data with a different (newer) version of the database can lead to different results. Using an up-to-date version of the database is thus highly recommended (Wadi et al., 2016).

Ideally all the known proteins and their different proteoforms are functionally annotated in the pathway database, with all of their known interactions. However, many proteins and proteoforms are not annotated in pathway databases. Similarly, in the protein inference step, when it is not possible to uniquely identify proteins from the identified peptides, protein groups are constructed. These are a collection of proteins that the peptides can be mapped to. For pathway analysis these protein groups are usually reduced to a single protein identifier, dropping the other possible proteins from the analysis. When, for a proteoform, only information at the protein or gene level is available, the information is mapped back to the more general level by the pathway database. These generalisations should be transparent for the user, as they can influence the results and relevance of the pathway analysis.

7.2 Moving beyond a single study

To study the differences in experimental procedures, there have been several projects where the same samples have been sent to different laboratories. The variability in these projects comes from different protocols, researchers, equipment, instruments, and software (Bell et al., 2009; Collins et al., 2017; Tabb et al., 2010). In multi-centre studies this variability comes in addition to the biological variability, while a question for single-centre studies becomes how reproducible the results are in other laboratories, and with other protocols and instruments.

Due to limitations in the number of available resources, large studies (e.g. Ellis et al., 2013; Riley et al., 2011) are often not possible. For example, in studies concerning rare diseases the number of possible samples to recruit can be very low. However, for wide applicability there is only so much that can learned from a single experiment.

Combining independent studies Multiple relevant studies can potentially be combined by performing a meta-analysis and/or systematic review, even when they have different protocols and/or instruments and are relatively small themselves. There are many different ways of performing a meta-analysis, but generally the results from each study are extracted and a weighted average is calculated, incorporating the uncertainties from the individual studies. When individual patient information is available, as is usually the case with proteomics studies in the form of publicly available raw data files, this can be used instead to combine different studies (a one-stage approach), or to recalculate study specific results (a two-stage approach) (Debray et al., 2015). While this is common practice in drug development (see e.g. `cochrane.org`), it is less commonly done in proteomics (e.g. Dupae et al., 2014; Rehiman et al., 2020; Srisawat et al., 2017).

Combining different omics platforms Another approach is to perform multiple omics studies with the same samples, usually done by splitting each sample into two (or more) and processing each fraction separately for each platform. Different classes of biomolecules can provide complementary information, increasing the amount of knowledge one can gain from a single experiment. Furthermore, proteins are generated from mRNA, which in turn is transcribed from DNA by proteins. In general, cellular homeostasis depends on an interplay of DNA, RNA, proteins, and metabolites, thus, for a complete understanding of cellular processes, ideally the interplay of all biomolecules is studied (Veenstra, 2021). When sample collection is difficult or expensive, e.g. when taken during an invasive procedure or when studying rare diseases, it is especially important to maximise the knowledge gained from each sample. If different classes of biomolecules can be extracted from the same sample, without needing drastically more material, a multi-omics approach could be an option (Blum et al., 2018; Hörmann et al., 2019).

The study of nucleic acids, i.e. DNA and RNA, generally requires different technologies compared to the study of amino acids and metabolites. Even though the latter two are both commonly processed by mass spectrometry, the sample processing and computational analysis differ substantially. When both proteins and metabolites are extracted from the same sample, sample preparation has to be done in a way that preserves the biological information contained in both classes of biomolecules (Blum et al., 2018). Many protocols for both proteomics and metabolomics exist, depending on sample type, object of the experiment, and possible type of proteins and metabolites targeted (e.g. Lygirou et al., 2014; Micic, 2016; Santamaría and Fernández-Irigoyen, 2019; Walker, 2005). As metabolites, usually including very small peptides, generally have smaller mass than proteins, the sample can be split based on mass, with both fractions subsequently treated and run in a metabolomics- and proteomics-specific way. Results can then be combined by performing pathway analysis based on both protein and metabolite interactions (Blum et al., 2018; Xu et al., 2020; Zhang et al., 2018).

7.3 Communication in multidisciplinary research

Proteomics experiments are inherently multidisciplinary endeavours. First and foremost, there is the biological and biochemical knowledge concerning the research question and the processing of the samples in the laboratory and on the instrument. Given the enormous amount of data that is generated for each sample, bioinformatics methods are necessary to process the data into a form that can hopefully be used to answer the research question(s). Appropriate statistical methods are needed to analyse the resulting data, and the interpretation of the results falls again in the domain of the biologist. Communication across these fields is thus essential.

Given that the vocabulary in these fields differ, the concepts that are basic for one research area might not be so in another field. It is thus important for researchers to have at least a basic understanding of the other fields and be able to communicate at a level that is understandable across different backgrounds. This can be challenging, and requires a lot of patience, but gets easier with practice and experience. Establishing a good working relationship with the researchers in the other fields (and those in your own) is therefore vital for the development as an effective scientist.

References

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., Ge, Y., Gunawardena, J., Hendrickson, R. C., Hergenrother, P. J., Huber, C. G., Ivanov, A. R., Jensen, O. N., Jewett, M. C., Kelleher, N. L., Kiessling, L. L., Krogan, N. J., Larsen, M. R., Loo, J. A., Loo, R. R. O., Lundberg, E., MacCoss, M. J., Mallick, P., Mootha, V. K., Mrksich, M., Muir, T. W., Patrie, S. M., Pesavento, J. J., Pitteri, S. J., Rodriguez, H., Saghatelian, A., Sandoval, W., Schlüter, H., Sechi, S., Slavoff, S. A., Smith, L. M., Snyder, M. P., Thomas, P. M., Uhlén, M., Eyk, J. E. V., Vidal, M., Walt, D. R., White, F. M., Williams, E. R., Wohlschlager, T., Wysocki, V. H., Yates, N. A., Young, N. L., and Zhang, B. (2018). How many human proteoforms are there? *Nature Chemical Biology*, 14(3):206–214.

Aebersold, R., Bader, G. D., Edwards, A. M., van Eyk, J. E., Kussmann, M., Qin, J., and Omenn, G. S. (2013). The biology/disease-driven human proteome project (B/D-HPP): Enabling protein research for the life sciences community. *Journal of Proteome Research*, 12(1):23–27.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2015). *Molecular Biology of the Cell*. Garland Science, New York, NY, 6th edition.

Ankeny, J. S., Labadie, B., Luke, J., Hsueh, E., Messina, J., and Zager, J. S. (2018). Review of diagnostic, prognostic, and predictive biomarkers in melanoma. *Clinical & Experimental Metastasis*, 35(5-6):487–493.

Ankney, J. A., Muneer, A., and Chen, X. (2018). Relative and absolute quantitation in mass spectrometry–based proteomics. *Annual Review of Analytical Chemistry*, 11(1):49–77.

Badawy, A. A.-B., Morgan, C. J., and Turner, J. A. (2008). Application of the phenomenex EZ:faast™ amino acid analysis kit for rapid gas-chromatographic determination of concentrations of plasma tryptophan and its brain uptake competitors. *Amino Acids*, 34(4):587–596.

Begley, C. G. and Ioannidis, J. P. A. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–126.

Bell, A. W., , Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J. J. M., and HUPO Test Sample Working Group (2009). A HUPO test sample study reveals common problems in mass spectrometry–based proteomics. *Nature Methods*, 6(6):423–430.

Benchoula, K., Parhar, I. S., and Wong, E. H. (2021). The crosstalk of hedgehog, PI3K and Wnt pathways in diabetes. *Archives of Biochemistry and Biophysics*, 698:108743.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95.

Bittremieux, W., Tabb, D. L., Impens, F., Staes, A., Timmerman, E., Martens, L., and Laukens, K. (2018). Quality control in mass spectrometry-based proteomics. *Mass Spectrometry Reviews*, 37(5):697–711.

Bittremieux, W., Valkenborg, D., Martens, L., and Laukens, K. (2017). Computational quality control tools for mass spectrometry proteomics. *Proteomics*, 17(3-4):1600159.

Blum, B. C., Mousavi, F., and Emili, A. (2018). Single-platform ‘multi-omic’ profiling: unified mass spectrometry and computational workflows for integrative proteomics–metabolomics analysis. *Molecular Omics*, 14(5):307–319.

Böcker, S. and Rasche, F. (2008). Towards *de novo* identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24(16):i49–i55.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1):49–64.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40(3-4):318–335.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2nd edition.

Brown, K. A., Melby, J. A., Roberts, D. S., and Ge, Y. (2020). Top-down proteomics: challenges, innovations, and applications in basic and clinical research. *Expert Review of Proteomics*, 17(10):719–733.

Clemetson, K. J. and Clemetson, J. M. (2013). Platelet receptors. In Michelson, A. D., editor, *Platelets*, pages 169 – 194. Academic Press, Elsevier, London, UK, 3rd edition.

Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R. L., Ohtsuki, S., Schlapbach, R., Selevsek, N., Thomas, S. N., Tzeng, S.-C., Zhang, H., and Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, 8(1):291.

Craig, R., Cortens, J. P., and Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Communications in Mass Spectrometry*, 19(13):1844–1850.

Cupp-Sutton, K. A. and Wu, S. (2020). High-throughput quantitative top-down proteomics. *Molecular Omics*, 16(2):91–99.

de Laeter, J. R., Böhlke, J. K., De Bièvre, P., Hidaka, H., Peiser, H. S., Rosman, K. J. R., and Taylor, P. D. P. (2003). Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure and Applied Chemistry*, 75(6):683–800.

Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., Reitsma, J. B., and on behalf of the GetReal methods review group (2015). Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Research Synthesis Methods*, 6(4):293–309.

Dolan, G., Benson, G., Duffy, A., Hermans, C., Jiménez-Yuste, V., Lambert, T., Ljung, R., Morfini, M., and Zupančić Šalek, S. (2018). Haemophilia B: Where are we now and what does the future hold? *Blood Reviews*, 32(1):52–60.

Dorfer, V., Maltsev, S., Winkler, S., and Mechtler, K. (2018). CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. *Journal of Proteome Research*, 17(8):2581–2589.

Dupae, J., Bohler, S., Noben, J.-P., Carpentier, S., Vangronsveld, J., and Cuypers, A. (2014). Problems inherent to a meta-analysis of proteomics data: A case study on the plants' response to Cd in different cultivation conditions. *Journal of Proteomics*, 108:30–54.

Edwards, N. J. (2017). Protein identification from tandem mass spectra by database searching. In Wu, C., Arighi, C., and Ross, K., editors, *Protein Bioinformatics*, volume 1558 of *Methods in Molecular Biology*, pages 357–380. Humana, New York, NY.

Eidhammer, I., Barsnes, H., Eide, G. E., and Martens, L. (2013). *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry*. Wiley, Chichester, UK.

Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214.

Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., Townsend, R. R., Kinsinger, C., Mesri, M., Rodriguez, H., and Liebler, D. C. (2013). Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. *Cancer Discovery*, 3(10):1108–1112.

Fonteh, A. N., Harrington, R. J., Tsai, A., Liao, P., and Harrington, M. G. (2007). Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. *Amino Acids*, 32(2):213–224.

Gabriels, R., Martens, L., and Degroeve, S. (2019). Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research*, 47(W1):W295–W299.

Gast, M.-C. W., van Gils, C. H., Wessels, L. F. A., Harris, N., Bonfrer, J. M. G., Rutgers, E. J. Th., Schellens, J. H. M., and Beijnen, J. H. (2009). Influence of sample storage duration on serum protein profiles assessed by surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF MS). *Clinical Chemistry and Laboratory Medicine*, 47(6):694–705.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Texts in Statistical Science Series. CRC, Boca Raton, FL, 3rd edition.

Gianetto, Q. G., Combes, F., Ramus, C., Bruley, C., Couté, Y., and Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics*, 16(1):29–32.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.

Goeminne, L. J. E., Gevaert, K., and Clement, L. (2016). Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. *Molecular & Cellular Proteomics*, 15(2):657–668.

Goeminne, L. J. E., Sticker, A., Martens, L., Gevaert, K., and Clement, L. (2020). MSqRob takes the missing hurdle: Uniting intensity- and count-based proteomics. *Analytical Chemistry*, 92(9):6278–6287.

Goh, W. W. B. and Wong, L. (2016a). Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. *Journal of proteome research*, 15(9):3167–3179.

Goh, W. W. B. and Wong, L. (2016b). Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology*, 14(05):1650029.

Gross, J. H. (2017). *Mass Spectrometry: A Textbook.* Springer, Cham, 3rd edition.

Hansson, K. T., Skillbäck, T., Pernevik, E., Kern, S., Portelius, E., Höglund, K., Brinkmalm, G., Holmén-Larsson, J., Blennow, K., Zetterberg, H., and Gobom, J. (2017). Expanding the cerebrospinal fluid endopeptidome. *Proteomics*, 17(5):1600384.

Harary, F. (1969). *Graph Theory.* Addison-Wesley, Reading, MA.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, New York, NY.

Held, P. K., White, L., and Pasquali, M. (2011). Quantitative urine amino acid analysis using liquid chromatography tandem mass spectrometry and aTRAQ® reagents. *Journal of Chromatography B*, 879(26):2695–2703.

Hernández Sánchez, L. F., Burger, B., Horro, C., Fabregat, A., Johansson, S., Njølstad, P. R., Barsnes, H., Hermjakob, H., and Vaudel, M. (2019). PathwayMatcher: proteoform-centric network construction enables fine-granularity multiomics pathway mapping. *GigaScience*, 8(8):giz088.

Hörmann, P., Barkovits, K., Marcus, K., and Hiller, K. (2019). Co-extraction for metabolomics and proteomics from a single CSF sample. In Santamaría, E. and Fernández-Irigoyen, J., editors, *Cerebrospinal Fluid (CSF) Proteomics*, volume 2044 of *Methods in Molecular Biology*, pages 337–342. Humana, New York, NY.

Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., and Heagerty, P. J. (2015). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute*, 107(8):djv157.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, 48:D498–D503.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

Katoh, M. and Katoh, M. (2007). WNT signaling pathway and stem cell signaling network. *Clinical Cancer Research*, 13(14):4042–4045.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375.

Kirkpatrick, D. S., Gerber, S. A., and Gygi, S. P. (2005). The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods*, 35(3):265–273.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill, New York, NY, 5th edition.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132.

Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, 15(4):1116–1125.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169.

Lygirou, V., Makridakis, M., and Vlahou, A. (2014). Biological sample collection for clinical proteomics: Existing SOPs. In Vlahou, A. and Makridakis, M., editors, *Clinical Proteomics*, volume 1243 of *Methods in Molecular Biology (Methods and Protocols)*, pages 3–27. Humana Press, New York, NY.

Martelli, C., Iavarone, F., Vincenzoni, F., Cabras, T., Manconi, B., Desiderio, C., Messana, I., and Castagnola, M. (2014). Top-down peptidomics of bodily fluids. *Peptidomics*, 1(1):47–64.

McDonald, T. J. and Ellard, S. (2013). Maturity onset diabetes of the young: identification and diagnosis. *Annals of Clinical Biochemistry*, 50(5):403–415.

Meier, L., Carlson, R., Neßler, J., and Tipold, A. (2020). Stability of canine and feline cerebrospinal fluid samples regarding total cell count and cell populations stored in "TransFix®/EDTA CSF sample storage tubes". *BMC Veterinary Research*, 16(1):487.

Mertens, B. J. A. (2017). Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies. In Datta, S. and Mertens, B. J. A., editors, *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, Frontiers in Probability and the Statistical Sciences, pages 1–21. Springer, Cham.

Micic, M., editor (2016). *Sample Preparation Techniques for Soil, Plant, and Animal Samples.* Humana, New York, NY.

Mohammed, B. M., Monroe, D. M., and Gailani, D. (2020). Mouse models of hemostasis. *Platelets*, 31(4):417–422.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G., editors (2015). *Handbook of Missing Data Methodology.* Handbooks of Modern Statistical Methods. CRC, Boca Raton, FL.

Moosa, J. M., Guan, S., Moran, M. F., and Ma, B. (2020). Repeat-preserving decoy database for false discovery rate estimation in peptide identification. *Journal of Proteome Research*, 19(3):1029–1036.

Moser, B. K. and Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician*, 46(1):19–21.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001.

Müller, T., Schrötter, A., Loosse, C., Helling, S., Stephan, C., Ahrens, M., Uszkoreit, J., Eisenacher, M., Meyer, H. E., and Marcus, K. (2011). Sense and nonsense of pathway analysis software in proteomics. *Journal of Proteome Research*, 10(12):5398–5408.

Nelson, D. L. and Cox, M. M. (2017). *Lehninger Principles of Biochemistry.* W. H. Freeman, New York, NY, 7th edition.

Nesvizhskii, A. I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data the protein inference problem. *Molecular & Cellular Proteomics*, 4(10):1419–1440.

Nicolini, A., Ferrari, P., and Duffy, M. J. (2018). Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Seminars in Cancer Biology*, 52:56–73.

Oberg, A. L. and Vitek, O. (2009). Statistical design of quantitative mass spectrometry-based proteomic experiments. *Journal of Proteome Research*, 8(5):2144–2156.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.

O'Sullivan, B. P. and Freedman, S. D. (2009). Cystic fibrosis. *The Lancet*, 373(9678):1891–1904.

Palanski, B. A., Weng, N., Zhang, L., Hilmer, A. J., Fall, L. A., Swaminathan, K., Jabri, B., Sousa, C., Fernandez-Becker, N. Q., Khosla, C., and Elias, J. E. (2021). An efficient urine peptidomics workflow identifies chemically defined dietary gluten peptides from patients with celiac disease. *bioRxiv*.

Pietrowska, M., Wlosowicz, A., Gawin, M., and Widlak, P. (2019). MS-based proteomic analysis of serum and plasma: Problem of high abundant components and lights and shadows of albumin removal. In Capelo-Martínez, J. L., editor, *Emerging Sample Treatments in Proteomics*, volume 1073 of *Advances in Experimental Medicine and Biology*, pages 57–76. Springer, Cham.

Plubell, D. L., Käll, L., Webb-Robertson, B.-J., Bramer, L., Ives, A., Kelleher, N. L., Smith, L. M., Montine, T. J., Wu, C. C., and MacCoss, M. J. (2021). Can we put humpty dumpty back together again? what does protein quantification mean in bottom-up proteomics? *bioRxiv*.

Plumel, M., Dumont, S., Maes, P., Sandu, C., Felder-Schmittbuhl, M.-P., Challet, E., and Bertile, F. (2019). Circadian analysis of the mouse cerebellum proteome. *International Journal of Molecular Sciences*, 20(8):1852.

Rehiman, S. H., Lim, S. M., Neoh, C. F., Majeed, A. B. A., Chin, A.-V., Tan, M. P., Kamaruzzaman, S. B., and Ramasamy, K. (2020). Proteomics as a reliable approach for discovery of blood-based Alzheimer's disease biomarkers: A systematic review and meta-analysis. *Ageing Research Reviews*, 60:101066.

Riley, C. P., Zhang, X., Nakshatri, H., Schneider, B., Regnier, F. E., Adamec, J., and Buck, C. (2011). A large, consistent plasma proteomics data set from prospectively collected breast cancer patient and healthy volunteer samples. *Journal of Translational Medicine*, 9(1):80.

Riter, L. S., Vitek, O., Gooding, K. M., Hodge, B. D., and Julian, Jr, R. K. (2005). Statistical design of experiments as a tool in mass spectrometry. *Journal of Mass Spectrometry*, 40(5):565–579.

Robin, J. D., Jacome Burbano, M.-S., Peng, H., Croce, O., Thomas, J. L., Laberthonniere, C., Renault, V., Lototska, L., Pousse, M., Tessier, F., Bauwens, S., Leong, W., Sacconi, S., Schaeffer, L., Magdinier, F., Ye, J., and Gilson, E. (2020). Mitochondrial function in skeletal myofibers is controlled by a TRF2-SIRT3 axis over lifetime. *Aging Cell*, 19(3):e13097.

Robles, M. S., Cox, J., and Mann, M. (2014). In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genetics*, 10(1):e1004047.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A.-L., Iakoucheva, L. M., Aloy, P., Rivas, J. D. L., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., and Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Rose, H. M., Ragan, C., Pearce, E., and Lipman, M. O. (1948). Differential agglutination of normal and sensitized sheep erythrocytes by sera of patients with rheumatoid arthritis. *Proceedings of the Society for Experimental Biology and Medicine*, 68(1):1–6.

Santamaría, E. and Fernández-Irigoyen, J., editors (2019). *Cerebrospinal Fluid (CSF) Proteomics.* Humana, New York, NY.

Schaffer, L. V., Millikin, R. J., Miller, R. M., Anderson, L. C., Fellers, R. T., Ge, Y., Kelleher, N. L., LeDuc, R. D., Liu, X., Payne, S. H., Sun, L., Thomas, P. M., Tucholski, T., Wang, Z., Wu, S., Wu, Z., Yu, D., Shortreed, M. R., and Smith, L. M. (2019). Identification and quantification of proteoforms by mass spectrometry. *Proteomics*, 19(10):1800361.

Schrader, M. (2018). Origins, technological development, and applications of peptidomics. In Schrader, M. and Fricker, L., editors, *Peptidomics*, volume 1719 of *Methods in Molecular Biology*, pages 3–39. Humana, New York, NY.

Schwämmle, V., Hagensen, C. E., Rogowska-Wrzesinska, A., and Jensen, O. N. (2020). PolySTest: Robust statistical testing of proteomics data with missing values improves detection of biologically relevant features. *Molecular & Cellular Proteomics*, 19(8):1396–1408.

Shiferaw, G. A., Vandermarliere, E., Hulstaert, N., Gabriels, R., Martens, L., and Volders, P.-J. (2020). COSS: A fast and user-friendly tool for spectral library searching. *Journal of Proteome Research*, 19(7):2786–2793.

Silva, A. S. C., Bouwmeester, R., Martens, L., and Degroeve, S. (2019). Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics*, 35(24):5243–5248.

Simon, R. (2008). Lost in translation: Problems and pitfalls in translating laboratory observations to clinical utility. *European Journal of Cancer*, 44(18):2707–2713.

Simon, R. (2015). Sensitivity, specificity, PPV, and NPV for predictive biomarkers. *Journal of the National Cancer Institute*, 107(8):djv153.

Smith, M. C., Schwertz, H., Zimmerman, G. A., and Weyrich, A. S. (2013). The platelet proteome. In Michelson, A. D., editor, *Platelets*, pages 103 – 116. Academic Press, Elsevier, London, UK, 3rd edition.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., Kharchenko, P., Koller, A., and Slavov, N. (2021). Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, 22(50).

Srisawat, K., Shepherd, S. O., Lisboa, P. J., and Burniston, J. G. (2017). A systematic review and meta-analysis of proteomics literature on the response of human skeletal muscle to obesity/type 2 diabetes mellitus (T2DM) versus exercise training. *Proteomes*, 5(4):30.

Surowiec, I., Johansson, E., Stenlund, H., Rantapää-Dahlqvist, S., Bergström, S., Normark, J., and Trygg, J. (2018). Quantification of run order effect on chromatography - mass spectrometry profiling data. *Journal of Chromatography A*, 1568:229–234.

Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Fisher, S. J., Gibson, B. W., Kinsinger, C. R., Mesri, M., Rodriguez, H., Stein, S. E., Tempst, P., Paulovich, A. G., Liebler, D. C., and Spiegelman, C. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research*, 9(2):761–776.

The, M. and Käll, L. (2019). Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & Cellular Proteomics*, 18(3):561–570.

The, M. and Käll, L. (2020). Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nature Communications*, 11(1):3234.

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515.

Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7(1):106.

Tomascova, A., Lehotsky, J., Kalenska, D., Baranovicova, E., Kaplan, P., and Tatarkova, Z. (2019). A comparison of albumin removal procedures for proteomic analysis of blood plasma. *General Physiology and Biophysics*, 38(4):305–314.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128.

Tsuchida, S., Satoh, M., Umemura, H., Sogawa, K., Takiwaki, M., Ishige, T., Miyabayashi, Y., Iwasawa, Y., Kobayashi, S., Beppu, M., Nishimura, M., Kodera, Y., Matsushita, K., and Nomura, F. (2018). Assessment by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of the effects of preanalytical variables on serum peptidome profiles following long-term sample storage. *Proteomics - Clinical Applications*, 12(3):1700047.

Tsuji, T., Hirota, T., Takemori, N., Komori, N., Yoshitane, H., Fukuda, M., Matsumoto, H., and Fukada, Y. (2007). Circadian proteomics of the mouse retina. *Proteomics*, 7(19):3500–3508.

Tworoger, S. S. and Hankinson, S. E. (2006). Collection, processing, and storage of biological samples in epidemiologic studies: Sex hormones, carotenoids, inflammatory markers, and proteomics as examples. *Cancer Epidemiology Biomarkers & Prevention*, 15(9):1578–1581.

Twyman, R. M. (2014). *Principles of proteomics*. Garland Science, New York, NY, 2nd edition.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9):731–740.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.

Valente, T. W. and Foreman, R. K. (1998). Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105.

van Buuren, S. (2018). *Flexible Imputation of Missing Data.* CRC, Boca Raton, FL, 2nd edition.

Vandermarliere, E., Mueller, M., and Martens, L. (2013). Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrometry Reviews*, 32(6):453–465.

Veenstra, T. D. (2021). Omics in systems biology: Current progress and future outlook. *Proteomics*, 21(3-4):2000235.

Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, 7(10):e1002240.

Verheggen, K., Martens, L., Berven, F. S., Barsnes, H., and Vaudel, M. (2016). Database search engines: Paradigms, challenges and solutions. In Mirzaei, H. and Carrasco, M., editors, *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*, volume 919 of *Advances in Experimental Medicine and Biology*, pages 147–156. Springer, Cham.

Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature Methods*, 13(9):705–706.

Walker, J. M., editor (2005). *The Proteomics Protocols Handbook*. Humana Press, New York, NY.

Wang, W., Sue, A. C.-H., and Goh, W. W. B. (2017). Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discovery Today*, 22(6):912–918.

Wang, Y., Song, L., Liu, M., Ge, R., Zhou, Q., Liu, W., Li, R., Qie, J., Zhen, B., Wang, Y., He, F., Qin, J., and Ding, C. (2018). A proteomics landscape of circadian clock in mouse liver. *Nature Communications*, 9(1):1553.

Wu, J. and Gao, Y. (2015). Physiological conditions can be reflected in human urine proteome and metabolome. *Expert Review of Proteomics*, 12(6):623–636.

Xu, L., Zhao, Q., Luo, J., Ma, W., Jin, Y., Li, C., Hou, Y., Feng, M., Wang, Y., Chen, J., Zhao, J., Zheng, Y., and Yu, D. (2020). Integration of proteomics, lipidomics, and metabolomics reveals novel metabolic mechanisms underlying N, N-dimethylformamide induced hepatotoxicity. *Ecotoxicology and Environmental Safety*, 205:111166.

Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., and Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, 11(1):146.

Zaghlool, S. B., Sharma, S., Molnar, M., Matías-García, P. R., Elhadad, M. A., Waldenberger, M., Peters, A., Rathmann, W., Graumann, J., Gieger, C., Grallert, H., and Suhre, K. (2021). Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nature Communications*, 12(1):1279.

Zhang, Y., Yuan, S., Pu, J., Yang, L., Zhou, X., Liu, L., Jiang, X., Zhang, H., Teng, T., Tian, L., and Xie, P. (2018). Integrated metabolomics and proteomics analysis of hippocampus in a rat model of depression. *Neuroscience*, 371:207–220.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 9788230866023 (print)
9788230845288 (PDF)