Quantitative analysis of mass spectrometry proteomics data

Software for improved life science

Teleman, Johan

2016

Quantitative analysis of mass spectrometry proteomics data

Software for improved life science

Johan Teleman

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Engineering, Lund University, Sweden.

To be defended in the Lundmark lecture hall, Sölvegatan 27, Lund

Friday May 27th at 09:15.

Faculty opponent

Prof. Michael MacCoss

Dep. of Genome Sciences, University of Washington, Seattle, USA

LUND UNIVERSITY	**Document name:** Doctoral dissertation
Dep. of Immunotechnology Ideon Medicon Village, Bld. 406 SE-22381 Lund Sweden	**Date of issue:** 2016-05-27
Johan Teleman	**Sponsoring organization:**

Title and subtitle:
Quantitative analysis of mass spectrometry proteomics data - software for improved life science

Abstract:
The rapid advances in life science, including the sequencing of the human genome and numerous other techiques, has given an extraordinary ability to aquire data on biological systems and human disease. Even so, drug development costs are higher than ever, while the rate of new approved treatments is historically low. A potential explanation to this discrepancy might be the difficulty of understanding the biology underlying the acquired data; the difficulty to refine the data to useful knowledge through interpretation. In this thesis the refinement of the complex data from mass spectrometry proteomics is studied. A number of new algorithms and programs are presented and demonstrated to provide increased analytical ability over previously suggested alternatives. With the higher goal of increasing the mass spectrometry laboratory scientific output, pragmatic studies were also performed, to create new set on compression algorithms for reduced storage requirement of mass spectrometry data, and also to characterize instrument stability. The final components of this thesis are the discussion of the technical and instrumental weaknesses associated with the currently employed mass spectrometry proteomics methodology, and the discussion of current lacking academical software quality and the reasons thereof. As a whole, the primary algorithms, the enabling technology, and the weakness discussions all aim to improve the current capability to perform mass spectrometry proteomics. As this technology is crucial to understand the main functional components of biology, proteins, this quest should allow better and higher quality life science data, and ultimately increase the chances of developing new treatments or diagnostics.

Keywords: Computational proteomics, mass spectrometry, algorithms, data complexity, life science, bioinformatics

Classification system and/or index terms (if any):

Supplementary bibliographical information:		**Language:** English
ISSN and key title:		**ISBN:** 978-91-7623-818-9
Recipient's notes:	**Number of pages:**	**Price:**
	Security classification:	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature ______________________________ Date ____________

Quantitative analysis of mass spectrometry proteomics data

Software for improved life science

Johan Teleman

LUND UNIVERSITY

Cover by Henrik Teleman

Faculty | Department | Lund University

ISBN 978-91-7623-818-9
ISBN 978-91-7623-819-6 (pdf)

Printed in Sweden by Media-Tryck, Lund University
Lund 2016

to Kikki, Stina and Klara

Contents

Original papers

I. Automated selected reaction monitoring software for accurate label-free protein quantification

Johan Teleman, Christofer Karlsson, Sofia Waldemarson, Karin Hansson, Peter James, Johan Malmström, Fredrik Levander
Journal of Proteome Research, **2012**, 11 (7), 3766-3773
doi: 10.1021/pr300256x

II. Automated quality control system for LC-SRM setups

Johan Teleman, Sofia Waldemarson, Johan Malmström, Fredrik Levander
Journal of Proteomics, **2013**, 95, 77-83
doi: 10.1016/j.jprot.2013.03.029

III. Numerical compression schemes for proteomics mass spectrometry data

Johan Teleman, Andrew W Dowsey, Faviel F Gonzalez-Galarza, Simon Perkins, Brian Pratt, Hannes L Röst, Lars Malmström, Johan Malmström, Andrew R Jones, Eric W Deutsch, Fredrik Levander
Molecular & Cellular Proteomics, **2014**, 13 (6), 1537-1542
doi: 10.1074/mcp.O114.037879

IV. Dinosaur: refined open source peptide MS feature detector

Johan Teleman, Aakash Chawade, Marianne Sandin, Fredrik Levander and Johan Malmström
Submitted

V. DIANA − algorithmic improvements for analysis of data-independent acquisition MS data

Johan Teleman, Hannes L Röst, George Rosenberger, Uwe Schmitt, Lars Malmström, Johan Malmström, Fredrik Levander
Bioinformatics, **2015**, 31 (4), 555-562
doi: 10.1093/bioinformatics/btu686

VI. Assay generation for Data Independent Acquisition mass spectrometry

Johan Teleman and Simon Hauri, Johan Malmström
Manuscript

My contributions to the papers

I. Implemented and devised algorithm, performed data analysis and evaluation. Co-wrote manuscript.

II. Designed study, implemented software and visualizations, and co-wrote manuscript.

III. Designed compression schemes and performed data analysis and evaluation. Coordinating manuscript author.

IV. Designed study, designed and implemented algorithms, and performed data analysis and evaluation. Coordinating manuscript author.

V. Participated in study design, devised and implemented DIANA algorithms, and performed data analysis and evaluation. Assisted in PyProphet implementation. Coordinating manuscript author.

VI. Co-designed study and devised algorithms with S. Hauri. Implemented algorithms. Coordinating manuscript author.

Excluded publications

Data processing has major impact on the outcome of quantitative label-free LC-MS analysis
Aakash Chawade, Marianne Sandin, Johan Teleman, Johan Malmström, Fredrik Levander
Journal of Proteome Research, **2015**, 14 (2), 676-687
doi: 10.1021/pr500665j

The representation of selected-reaction monitoring data in the mzQuantML data standard
Da Qi, Craig Lawless, Johan Teleman, Fredrik Levander, Stephen W. Holman, Simon Hubbard, Andrew R. Jones
Proteomics, **2015**, 15 (15), 2592–2596
doi: 10.1002/pmic.201400281

Data processing methods and quality control strategies for label-free LC–MS protein quantification
Marianne Sandin, Johan Teleman, Johan Malmström, Fredrik Levander
Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, **2014**, 1844 (1), 29-41
doi: 10.1016/j.bbapap.2013.03.026

Analysis of bacterial surface interactions with mass spectrometry-based proteomics
Christofer Karlsson, Johan Teleman, and Johan Malmström
Methods in Molecular Biology, Bacterial Pathogenesis: Methods & Protocols, **2016**, Chapter 3
Accepted

Baccus - a novel way of estimating bacterial load using mass spectrometry
Ola Kilsgård, Johan Teleman, Erik Malmtröm, Johan Malmström
Manuscript

The normal variation of cancer associated proteins in blood plasma
Markus Stenemo, Johan Teleman, Martin Sjöström, Gabriel Grubb, Erik Malmström, Johan Malmström, Emma Niméus
Submitted

Abbreviations and dictionary

DIA	data-independent acquisition, used in place of the proprietary term SWATH, instead of the clumsy SWATH-like
MS	mass spectrometry
LC	liquid chromatography
MS1	mass spectrum of intact analytes
MS2	mass spectrum of analyte fragments
DDA	data-dependent acquisition
SRM	selected reaction monitoring
PRM	parallel reaction monitoring
ETD	electron transfer dissociation
CID	collision induced dissociation
HCD	higher energy collision dissociation
ESI	electro-spray ionization
MALDI	matrix assisted laser desorption/ionization
TOF	time-of-flight mass analyzer
SWATH	sequential window acquisition of all theoretical spectra
iRT	a normalized retention time scale
DIANA	just a name
PTM	post-translational modification
DNA	deoxiribonucleic acid, encodes genes
RNA	ribonucleic acid, encodes transcripts

Foreword

The work presented in this thesis is the result of several pleasant years of attempting to apply previous skills, knowledge and intuition in a for me completely unknown domain: bottom-up mass spectrometry based proteomics. Coming from a masters degree in engineering physics, I am convinced that mass spectrometers are essentially simple and understandable machines, that have had the bad fortune of being very useful in the study of something vast, complex and chaotic: molecular biology. Further, mass spectrometers, no matter how accomplished and technically advanced, cannot answer questions about molecular biology. They merely limit the plausible subset of possible molecular mechanisms; interpretation and understanding will always remain the challenge of the scientist.

My aim with this body of text is to highlight and communicate my current understanding of all aspects of computational proteomics, not as the analytical chemist or the pragmatic physician, but as the physics engineer and computer scientist that I label myself as. With this in mind, there should be ample amounts of misunderstandings for the informed reader to point out, given the writers tendency to draw early conclusions in areas where there is little experience. Still, some of the greatest revelations come with questions from an unexpected angle, and I therefore hope there is such a questionable angle hidden somewhere.

Chapter 1: Setting the Stage

The universe is. Building from this most basic meta-physical observation, science tends to narrow down to increasingly specific statements. Of the infinite such statements available, life science focuses on the ones concerning life, namely matter that is organized into fat-enclosed µm-sized volumes, filled with DNA and other organic molecules. Cells, as these volumes are called, are the building blocks of all living things including plants and animals and humans, but also the minuscule things that are bacteria and amoeba.

Human society has benefited greatly from life science. Through history many discoveries in breeding, medicine and brewing have been made that can now be beautifully explained using life science. But also later, using actual molecular biology techniques of studying specific types of organic molecules, great feats have been accomplished for the human good. Tuberculosis and Polio vaccination programs and small pox eradication have saved millions of lives, the fast development of HIV medicines has completely changed life expectancy and spread of this disease, and cancer has transformed from being utterly mortal to having several treatment options, at least for the most common cancer subtypes.

With the great feats of past life science in mind, it must be stated that progress in the medical field seems to have stalled, or at least become increasingly difficult. Although the world population continues to show a wide array of diseases in need of treatment[1,2], the rate of approved new drugs is exceedingly low[3,4], while companies are reporting greater costs than ever for drug development[5–7]. It would seem that the low-hanging fruit in terms of treatable diseases and prognostic markers have been discovered, while safety requirements for population-wide approval have increased[4,8].

Apart from new treatment drugs, much money and brainpower is invested in developing biomarkers – essentially measurable patient properties that are informative regarding the patient health or disease status[9]. Biomarkers can be used to stratify patients into treatment or prognosis groups, or simply mark for the presence of disease. For example, the presence of the BRCA1 gene in breast cancer tumors could be used to select patients for a specific treatment that is not as efficient in patients lacking this mutation[10]. Of further importance in biomarker discovery is the method of biomarker measurement, specifically in terms of patient inconvenience, why human plasma is a frequent target for biomarker discovery studies[9,11,12].

Progress has been very fast in life science in the last decades. Following the monumental completion of the sequencing of the human genome in 2001 after more than a decade of work and at an estimated cost of $3 billion[13–15], the cost of genome sequencing has reduced by 1 million times[16], and much has been learned of human mutations and their relation to different diseases. Yet, genomes do not necessarily explain the mechanisms of any observed biological conditions, or offer new methods of pushing the biological system in a beneficial direction. From only the DNA we have so far no capability to simulate the entire cell to predict its behavior, let along multi-cell systems[17,18]. The next targets for study are therefore the downstream parts of the canonical cell production chain, the translated genes that are in transit for protein expression (RNA) and the proteins themselves.

Some general principles of the cell are very useful to understand the contents of the work described in this thesis. In short, cells exist because they are self-replicating: given reasonable conditions they will make more copies of themselves. Two primary things are needed to copy oneself: a blueprint, and a making-mechanism. The DNA is the blueprint in cells, and the making-mechanism is a series of steps centered on the molecular structure called the ribosome. Essentially, the DNA is a long chain composed of 4 types of nucleic acids. Parts of the DNA, *genes*, are copied and transported to the ribosomes for translation into the primary components of the cell, the *proteins* (Fig. 1.1). Proteins are also long chains, but composed of amino acids rather than nucleic acids, and folded into complex 3-dimensional structures that are critical for the protein function. Proteins carry out most high-order functionality in cells, like propulsion, control of the intra-cellular molecular composition, release of signaling molecules, and tracking of whether to reproduce or gather resources.

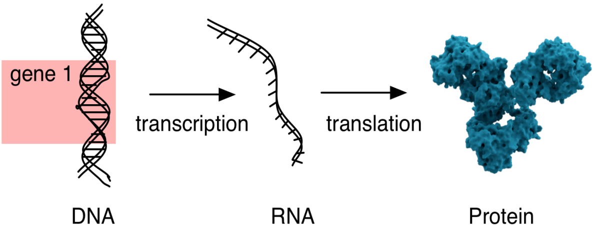


Figure 1.1 - The canonical model of protein expression. A gene in transcribed into an RNA sequence, which is translated into a protein. (The protein 3D image is from "Protein composite" by Thomas Splettstoesser (www.scistyle.com), CC BY-SA 3.0)

Having drawn the basic picture of cell function, I must emphasize that in biology no rule is absolute, and all order has exceptions. In the above canonical model, DNA is translated into RNA that is expressed as proteins (Fig. 1.1). The DNA carries the information to describe all functions of the cell and proteins perform them. However, every word in the last two sentences is only true in the general sense, as a multitude of amendments are necessary to describe actual cell biology (Fig. 1.2). In fact, most genes in the DNA seem to have several possible translations depending on the translator molecules (RNA or protein) that are connected at the time of translation[13,19,20], and some RNA is not expressed as proteins but carries out primary functions by itself[21,22]. To complicate things further, proteins are often modified by some of hundreds of possible post-translational modifications (PTMs)[23,24]. The exact setup of PTMs on a single protein can, but might not, completely change its function or location in the cell[25]. Protein location might in turn completely change the proteins function in the cell[26]. Finally, the same protein might be folded differently, resulting in different function[27–29]. All together, stochastic mutation processes and Darwinian selection, while very successful, seems to have made life highly complex and messy[i].

[i] Divine intervention would probably have been preferable, from an engineering point of view.

The great biological complexity points to a need for large amounts of diverse information to understand biological systems. Luckily, the large-scale study of transcripts and proteins seems to be evolving at paces close to the genome counterpart[30] (Fig. 1.3), and the measurement of the majority of cell transcripts and proteins in simple cell systems is now commonly performed[31,32]. In the work presented here, proteins have been the exclusive targets of interest. The study of the protein contents of a sample is called *proteomics*. A biological sample is at any given time said to contain a

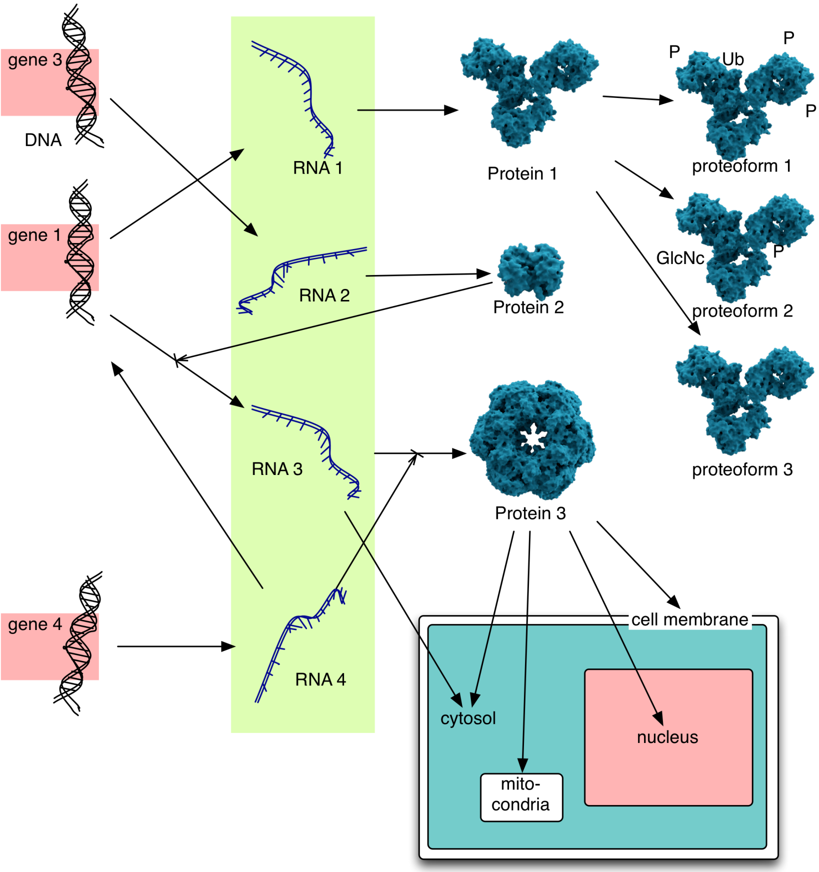


Figure 1.2 - Actual protein expression. In real biology, genes might be translated into several alternative RNA sequences, while some RNA carries out primary functions such as silencing genes or inhibiting protein translation. A single gene-product might turn into different proteoforms by post-translational modifications, with different cellular function. Finally, protein localization in the cell might completely change it's function. (The protein 3D images are from "Protein composite" by Thomas Splettstoesser (www.scistyle.com), CC BY-SA 3.0)

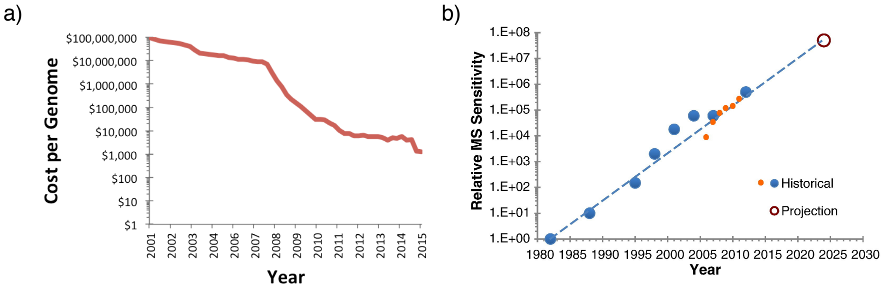


Figure 1.3 - The exponential progress of life science technology capability. a) Cost of sequencing a human genome (data from the NHGRI Genome Sequencing Program[16]). b) Mass spectrometry sensitivity (by NL Anderson, from the historical review *Six decades searching for meaning in the proteome*[30]. Reproduced with permission.)

proteome, which is described by the exact listing of all proteins in the cell, including their quantities, modifications, and sub-cellular location. Proteomics presents some distinct challenges over it's gene and transcript counterparts, in that protein signal cannot be amplified, proteins are present at a large range of concentrations, and that the proteome is constantly adapting to the cellular environment and inner state.

The research described in this thesis is ultimately driven by the desire to measure the proteome to provide insight in biological systems. This can be accomplished by mass spectrometry instrumentation, and my work revolves around automated interpretation and analysis of the sizeable data that mass spectrometers generate when challenged with biological samples. Before this analysis task and my results can be described however, a deeper foundation of the underlying technology is needed.

Mass spectrometry proteomics

Large-scale proteomics can be pursued by several measurement techniques, for example 2-dimensional separation followed by protein staining[33,34], binding by fluorescently labeled antibodies[35,36], or mass spectrometry[37,38]. Even though many considerations are shared between these techniques due to their similar goals of identification and quantification of large numbers of proteins, this work is focused exclusively on mass spectrometry based proteomics. Mass spectrometry proteomics builds upon the vastly increased availability of sequenced genomes, including the human[13,14], the soft ionization techniques electrospray ionization (ESI)[39] and matrix-assisted laser adsorption ionization (MALDI)[40], as well as several years of very rapid instrument performance increases[30].

In this work bottom-up mass spectrometry proteomics has been performed, which is one of the most common varieties of mass spectrometry proteomics. This workflow consists of protein digestion by some enzyme, online separation by hydrophobicity on a reversed phase chromatographic column, and electrospray ionization followed by mass spectrometry.

The mass spectrometer is defined by one singular trait: the ability to separate gas-phase analytes based on their mass over charge ratio (m/z). To accomplish this target analytes are ionized, followed by subjection to highly controlled electro-magnetic fields[41]. In the quadrupole mass analyzer[42,43] and different ion traps[44,45], oscillating fields are utilized to create stable trajectories only for ions at a certain m/z[46]. In the time-of-flight instrument, a constant field is used to reverse ion trajectories[46]. Because heavier ions take longer to deflect, the time of flight in the field can be used to compute the m/z of the ion. In the Fourier transform ion cyclotron resonance (FTICR)[47,48] analyzer and the Orbitrap[49,50], ions are subjected to a 3-dimensional or a mixed cylindrical field that puts them in different orbitals depending on their m/z. The induced current is measured in fixed detectors, where the specific cyclical motion at each m/z will generate signal at a specific frequency. The m/z values can then be obtained by the inverse Fourier transform. Naturally, these mass analyzers come with different strengths and weaknesses, and they all have uses in specific workflows[37,41].

Protein digestion is performed for two main reasons in the MS proteomics workflow. First, proteins vary greatly in size and chemical properties such as ionic charge and isoelectric point, and thus any separation strategy at the protein level is likely to be incompatible with a substantial amount of proteins. Cleaving proteins into peptides increases the chances that at least one or a few peptides of each protein will be detectable, and therefore provide information about the protein. Second, digestion by for example trypsin results in peptides that are very suitable for ionization and mass spectrometric analysis. This suitability comes from the almost guaranteed presence of a lysine or argine amino acid at the C-terminus of the peptide, in combination with a suitable size of 5-36 amino acids[ii]. The basic amino acids attract positive charge, resulting in peptides that are charge 2 or more, meaning that the peptide m/z will lie around 300-1000 m/z, which is a range where mass spectrometers are very capable.

Chromatographic separation is performed because of limitations in electrospray and MS technology. No current instrumentation can concurrently ionize and analyze thousands of analytes. Limitation of the momental complexity is needed to properly ionize as many peptides as possible, and to allow their fragmentation and identification in the mass spectrometer. The retention time information can further be used to identify peptides by comparison to predicted hydrophobicity or previous empirical measurements.

The bottom-up proteomics workflow provides increased probability of protein measurement by distributing risks over peptides, and strong separation of peptides by hydrophobicity and m/z. Still, this is not enough to tackle the full complexity of biological samples, and one additional component is commonly used for even higher specificity.

Gas-phase fragmentation

In the primary form of mass spectrometry, analytes are separated based on m/z and the intensity of each m/z is recorded, resulting in an array of m/z and intensity pairs, that is called a spectrum. A single m/z value though, is generally not specific enough to uniquely determine the corresponding molecule, because of the large amount of possible peptides in biological samples. Modern mass spectrometers are therefore in addition built to dynamically isolate a part of the analyte m/z range, subject this sub-range to fragmentation using some fragmentation technique, and collect a new spectrum of the generated fragment ions[51–53]. As this associates promising m/z values with complementary measurements of their fragment spectra, greatly increased specificity is achieved, due to the exceedingly small

[ii] 90% Confidence interval computed for unique peptides from *in silico* tryptic digest of Human proteome as downloaded from SwissProt / UniProt 2015-05.

likelihood of multiple peptide ions sharing both m/z, retention time and several fragment m/z values[54].

To differentiate between the fundamentally different types of chemical information in the intact peptide spectra and the fragment spectra, the primary intact analyte MS spectrum is called an MS1 spectrum, while MS spectra of analyte fragments are called MS2 spectra.

The analyte of interest in bottom up proteomics is the peptide, which consists of a chain of amino acids (Fig. 1.4). Amino acids all share a backbone molecular motif, and are differentiated by the structure of their side chains. In gas-phase fragmentation, energy is carefully added to the peptide, to enable the dissociation of one bond of the peptides amino acid backbone. This stochastic process can generate many different types of fragments, which are mainly categorized by where the amino acid back-bone breaks, giving rise to a-, b- and c-fragments from the N-terminal end, and x-, y- and z-fragment from the C-terminal end[55,56] (Fig. 1.4). As peptide fragmentation pathways are still not completely understood[57–59], the probabilities of each fragment type and fragmentation site cannot be readily predicted although attempts have been made with some success[60,61]. Importantly, the fragments observed depend heavily not only on peptide length but also on amino acid composition[62–64]. Apart from the main terminal fragments, certain amino acids frequently give rise to internal fragments, and multiple losses such as deamidation can occur to vary the exact chemical composition of the fragment.

Depending on the intentions of the analysis, optimal peptide fragmentation might consist of complete coverage of all the amino acids in the peptide, or

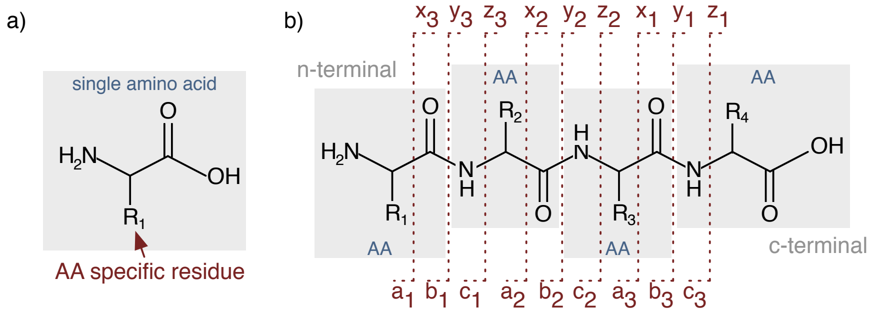


Figure 1.4 - Chemical composition of the amino acid and chains thereof.

of fewer but more distinct fragments. In both cases, fast reactions times and high reproducibility are key properties. In pursuit of these goals many fragmentation techniques have been developed. The most employed fragmentation method, collision induced dissociation (CID)[65,66], works by colliding analyte ions into an innate gas like argon, and primarily produces y- and b-fragments. The related higher-energy collisional dissociation (HCD)[67] generates mostly y-fragments, but also some b-fragments. The two methods have the moderate weakness of not generating complete fragment series, meaning that ambiguities will typically exist regarding the order of some neighboring amino acids. Electron capture dissociation (ECD)[68] and electron transfer dissociation (ETD)[69] are completely different fragmentation methods that enable fragmentation by low-energy electron donation, producing mainly c- and z-type fragments. The main benefit of ECD/ETD fragmentation is the very complete fragment series that is generated, while the drawback of the methods is the longer reaction times needed for the election transfer to occur, which can limit sequencing speed. Even though the exact mechanisms of fragmentation are unknown or unpredictable for a given peptide, publicly available data has been used to establish empirical knowledge about the probabilities of generating each fragment type for the common fragmentation techniques[70–73].

Optimizing information gain: different flavors of mass spectrometry

Because of the complex complexion of biological samples, no current mass-spectrometric method is capable of performing the ideal theoretical feat of identifying and quantifying all peptides that can be proteolytically derived from the sample. Therefore a number of mass spectrometry methods have been devised to optimize the sensitivity, specificity, parallelism, reproducibility or accuracy of quantification of the experiment performed[37] (Fig. 1.5).

The classical MS method for cataloguing a sample of unknown composition is data-dependent acquisition (DDA) MS, also nicknamed shotgun MS[74,75].

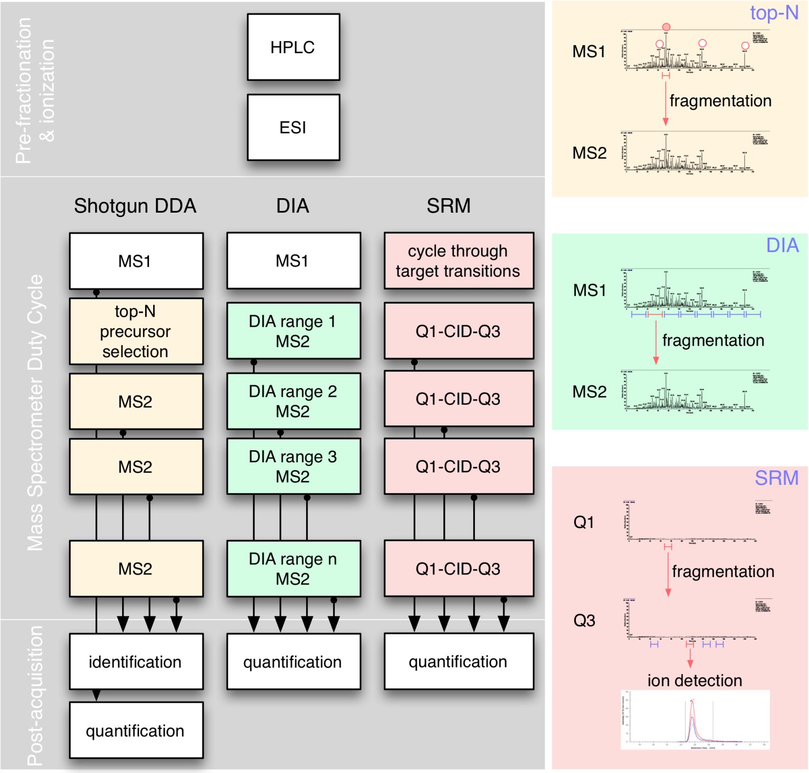


Figure 1.5 - Representative mass spectrometry proteomics workflows. All starting with liquid chromatography (HPLC) and electro spray ionization (ESI), the workflows diverge on the mass spectrometer side. In shotgun MS, the top-N measured MS1 peaks are selected for fragmentation. In DIA, the whole MS1 range is fragmented in fixed subsets, and in SRM, predefined precursor and fragment m/z values are used to hopefully only measure peptide ions of interest.

Here, the goal is to identify as many peptides as possible, sacrificing sensitivity and reproducibility for high parallelism. The method works by performing initial MS1 scans to characterize the composition of the currently injected ions in terms of their m/z and abundances. Acting upon this scan, the instrument selects the most abundant m/z values. Each such m/z precursor is isolated, and fragmented using for example HCD fragmentation, and scanned. These MS2 scans are performed sequentially, and the number of peptides that can be identified is heavily dependent on the speed of MS2 scans, with modern mass spectrometers reaching 20 MS2 scans/s[76,77]. To not repeatedly scan the same analyte ion, instruments typically use a dynamic exclusion list; blacklisting m/z values from further MS2 scanning for a short period after being subjected to a MS2 scan.

On the opposing end of the scale, selection reaction monitoring[78–80] (SRM) and parallel reaction monitoring[81] (PRM) provide high sensitivity, specificity, reproducibility and accuracy of quantification while limiting the measured targets to a predefined set of analyte ions (Fig. 1.5). Here the data-driven fragmentation and MS2 scanning of precursors is replaced with a pre-defined list of precursors of interest, which are measured continuously or in a scheduled fashion. In SRM, a triple-quadrupole instrument is employed, using the first quadrupole for precursor filtering, the second for CID fragmentation, and the third for fragment filtering, after which the signal intensity is measured using an ion detector. Because of the double filtering step, operation of the SRM workflow requires assays: prior information on the analytes to be measured. Assay information includes the precursor mass and charge, the collision energy to use, the expected retention time, and the fragments to be measured and their charges. The instrument uses this information to constantly measure m/z channels of interest, making it highly sensitive. Assays are slightly simplified in PRM, because the third quadrupole is replaced with some scanning mass analyzer like the Orbitrap, and therefore all fragments within a broad range will be measured and no fragment pre-selection needs to be made.

In-between the parallel shotgun MS and the sensitive SRM/PRM are a number of compromising methods such as data-independent acquisition (DIA), where the benefits of both regimes are pursued for high general performance[82–86]. When involving DIA, this thesis describes work using the DIA method SWATH, as popularized by Gillet et. al.[87] (Fig. 1.5). Here a scanning mass spectrometer performs an MS1 scan, followed by isolation, fragmentation and scanning of predefined subsets of the precursor m/z range. In the original formulation the precursor range 400-1200 m/z is scanned and then divided into 32 bins spanning 25 m/z. Each bin is isolated, fragmented and scanned in sequence. Because of the large amount of MS2 scans performed, SWATH cycle times on current instruments are long compared to chromatographic peak-widths in HPLC, as several MS2 spectra are needed to characterize the chromatographic peak[88]. Interestingly, too sharp chromatography could negatively affect the ability to analyze this data. Initially only the fast TOF instruments were capable of SWATH analysis, but quadrupole-Orbitrap instruments are also emerging as capable of SWATH-like[iii] analysis[88]. In this work I will simply used the acronym DIA when referring to this type of analysis using any instrument.

Aims of this thesis and results provided

We study life science to improve and prolong life quality for those affected by disease, and to achieve others goals such as food production or waste water cleaning. To understand life we now have very strong capabilities to map genes and measure their transcripts, but for true understanding of biological systems we need to measure the first order actors, the proteins. Because of the proteomes great complexity this poses a great problem, even though substantial progress has been made by for example mass spectrometry proteomics.

We are still not capable of measuring complete proteomes, partly because of primary limitations in the measuring techniques, but also because of difficulties in efficiently interpreting the large datasets generated by mass spectrometers when exposed to complex biological samples.

In this thesis, I will describe the research I have been performing during the last five years. The overarching goal of this work has been to improve upon the computational tools available across the board of mass spectrometry based proteomics techniques, to solidify technical advances on the instrument side, and to improve data quality and allow better conclusions at the biological and medical level. The pursuit of this goal has resulted in 6 scientific papers, which constitute the main contribution of my work. These papers describe a new tool and algorithm for data analysis of SRM data (Paper I), a workflow for standardized quality control of SRM instruments (Paper II), numerical compression algorithms for mass spectrometry data (Paper III), an algorithm and tool for detection of peptide isotopic envelopes (Paper IV), an algorithm and tool for analysis of DIA data (Paper V), and finally a workflow to generate assays for DIA analysis (Paper VI).

[iii] 'SWATH' is regrettably a proprietary term that can only be used for the application of the technique on Q-TOF instruments. The term *SWATH-like* has been suggested for the use with other instruments.

To put the scientific papers in context, the papers involving enabling technologies (Paper II, III and VI) are summarized along with a discussion of mass spectrometry proteomics data in Chapter 2. The papers presenting primary analysis algorithm of MS data (Paper I, IV and V) are described in Chapter 3, together with a general discussion of the common solutions employed by the field for such analysis tasks.

Apart from the main research described in the 6 papers, some subjects and insights have been synthesized that did not fit into scientific publications. I believe these conceptual results should be considered secondary contributions of this thesis work, and they are presented in Chapter 4 and 5. From implementation and usage of computational proteomics software, I have come to identify problems with academic software programming and software quality, and these issues and how to potentially resolve them are discussed in Chapter 4, in the interest of improving software quality in proteomics. Finally, in Chapter 5, I attempt to identify what I believe to be the key instrumental and measurement-technical weaknesses of the currently used mass spectrometry workflow when considering the higher goals of life science and the biological complexity. These weaknesses are included to raise awareness and possibly contribute to their eventual removal.

In the final chapter of this thesis, the full contributions of the presented papers and conceptual insights are summarized and evaluated for relevance. An outlook is provided on possible directions for future research building on the presented contributions, with the continuing goal of improving proteomics data quality to provide better and more powerful life science.

Chapter 2: Too Much Information

We need to characterize mass spectrometry proteomics data to properly understand it and eventually create efficient algorithms for its analysis. Data can be described by amount and the tightly coupled redundancy, but also in terms of data complexity. In this chapter, I will also try to explain the instrumental causes of variation in the different properties of the data, which will highlight some difficulties that are encountered when designing analysis algorithms. Mass spectrometry proteomics data comes in many shapes depending on the used acquisition method (Fig. 2.1). In addition, an important type of meta-data required for SRM and targeted DIA analysis is the MS assay (Fig. 2.2), which is also presented here. Finally, this chapter ends by description of the studies I have performed to describe and compress MS data, and to generate MS assays (Paper II, III and VI).

Data redundancy

Mass spectrometry proteomics generates quite substantial amounts of data, even though talking about *big data*[89,90] is stretching the buzzword too far. A typical proteomics mass spectrometer at the time of writing might generate in the order of 1 GB data/h when in production[iv], and a typical MS proteomics study could encompass between a few GB to a few TB raw data. Even though these data-sizes are by no means exceptional in size, they still require

[iv] Mass spectrometers are apparently highly complicated machines that exhibit ingenius ability to break in creative and costly ways including internal fires and unknown goo covering the transfer tube.

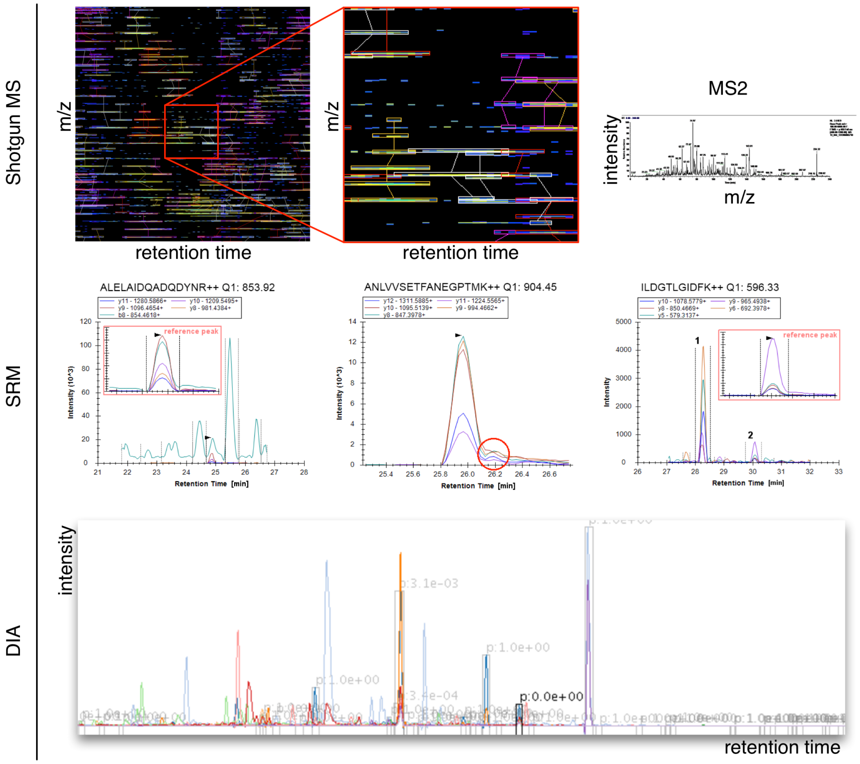


Figure 2.1 - Examples of raw data from shotgun MS, SRM and DIA

dedicated computational hardware for efficient production analysis, and are no longer manageable on personal laptops.

Even though generating respectable data sizes, one should remember that MS proteomics data is highly redundant. Studies have estimated the expressed proteome of tissues or cell types to consist of roughly 10 000 gene products[91–93], commonly identified by about 100 000 peptides[91,93,94]. Even adding the permutations of post-translational modifications[23,24] and splice variants[95,96], a reasonable guess[v] at the number of unique identifiable molecules in a bottom-up proteomics sample might be 10e6. Listing the quantities of these would constitute the optimal representation of the

[v] kudos, Enrico Fermi

quantitative information in a sample, and would uncompressed only use tens of MB. Apart from high-level arguments, there is also local redundancy because of the sequential scanning that creates mass spectra. This local redundancy can, and should, be reduced by compression to simplify data handling and lower storage requirement, and to potentially even speed up computations[97,98].

Data complexity

Containing millions of unique molecules, life science samples are commonly considered complex. How this *sample complexity* translates into MS proteomics *data complexity* is however non-trivial. In general, increasing the number of molecules subjected to analysis will increase the number of detectable and measured analytes, but the increase is non-linear and starts to saturate already at low sample complexity (Fig. 2.2). This is largely unexplored in complex proteomics samples, but ion competition should play a large part in limiting the observable molecules to only the ones most suitable for ionization[99–101]. Additional molecules of lower ionizability might

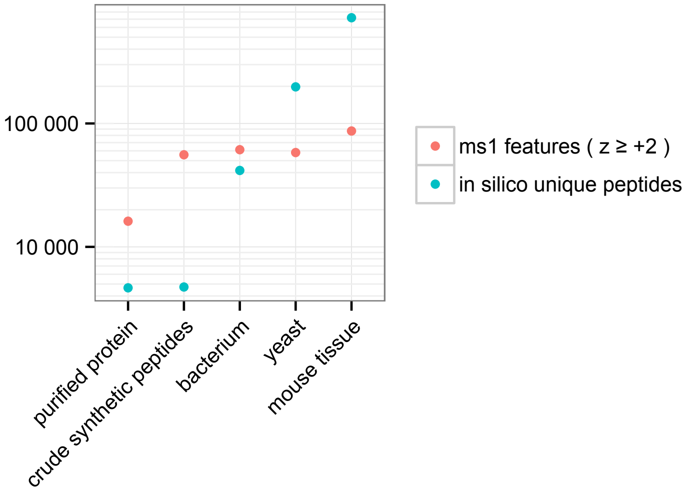


Figure 2.2 - Sample complexity saturation: detected MS1 features vs theoretical sample complexity. A Thermo Scientific Q-Exactive Plus interfaced with a 1000 bar Easy LC was used to measure the data for this figure.

not be observable at all, depending on their relative abundances. Another critical aspect limiting the MS-perceived complexity of samples is the limited number of ions that can be trapped concurrently. Most ion trap space will be filled by high abundant species, making rare species too few to be detectable. In summary, it would seem as if the current instrument setup is saturating at around 100 000 detectable species - regardless of sample complexity.

Human blood is also related to data complexity. Blood plasma samples are one of the major targets of life science, and can readily be collected with low patient discomfort to provide systemic information of the patient health status. Unfortunately, plasma is also very challenging to analyze, and proteomics analyses of plasma detect substantially smaller numbers of proteins than other body samples like biopsies. This challenge originates in the enormous abundance range of plasma analytes, commonly cited as 10 orders of magnitude[11], with the top 14 most abundant proteins constituting 98% of the molecules[11]. In individual eukaryotic cells protein abundances range 6 orders of magnitude[31,102], with a much more even distribution.

Variability in MS data

In order to design effective algorithms for analysis of MS proteomics data, we need to understand how the processes and measurement technologies generating this data work. In particular, we need to know how analytes of interest behave under observation and what deviations from this behavior that should be expected. I will therefore elaborate on the sources of variability that are inherent to the used MS proteomics methodology. In short this amounts to variability in chromatographic elution time, signal intensity, mass accuracy and peptide fragmentation.

One thing that is commonly considered less accurate is the hydrophobicity-based retention time of peptides in nano-flow liquid chromatography[103–105] (Paper II). Linear shift can be seen because of differences in the exact configuration assembly of the current chromatography lines and also because of different column quality. However, even more troublesome are the non-linear effects that occur with too high sample concentrations, where the most populated part of the hydrophobicity scale might be slowed down[106].

Signal intensity will depend on multiple aspects, such as instrument contamination and ion suppression[107–111]. As the instrument is used, molecules will randomly attach to the electrospray tip or the walls of the mass spectrometer inlet. These contaminations will build up to eventually disturb ion flows and reduce signal intensity. Ion suppression will interestingly also depend on the chromatographic elution and sample composition. Because co-eluting peptides in complex samples are competing for ions, slight shifts in the momental composition of peptides might change the experienced ionization efficiency for all these peptides. Finally, low abundant ions will be measured by few individual molecules, and whether it is 3 or 5 molecules that hit the detector will have a big impact on the quantification of that ion.

Mass accuracy is in general limited by the finite control of injected ions. In quadrupoles the m/z width of the transmission window is limited to roughly 1 m/z since smaller windows severely reduce transmission[43,112]. For Orbitraps, limitations are the simultaneous injection of all ion trap ions, as well as the frequency resolution that is limited by the sampling speed[113]. These errors are typically normally distributed, but trickier is the m/z blurring that occurs due to space-charge effects when to large populations of ions are present in a small volume. This effect is caused by the electrostatic repulsion of same-charged particles, and causes ions to come slightly of trajectory as the force of electrostatic repulsion becomes comparable with the induced force of the controlling fields.

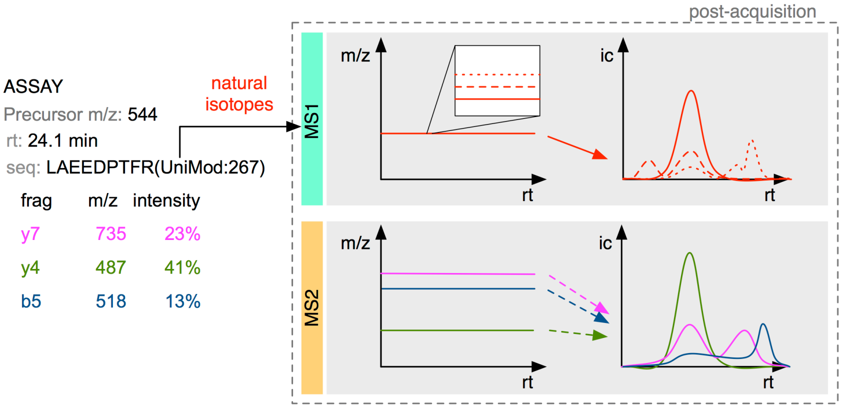
Fragmentation of peptide ions is a probabilistic process regardless of fragmentation method, where addition of energy to the ion allows breakage through one of many pathways. For individual ions the fragmentation pathway chosen depends on the position of an eventual mobile proton[57], which cannot be known but has some probabilistic distribution over the peptide. Further influence could come from kinetic, rotational and vibrational energies of the whole ion[114], as well as the concentration of collision gas or electron donors[115]. For calibrated experiments and for large populations of ions however, fragmentation seems to be highly conserved[64] (Paper II).

As in all scientific disciplines based on empirical measurement, it is apparent that MS proteomics must take into account small inaccuracies and unknowns in the data. In practice a substantial part of algorithm development circles around estimating the size of the expected variation in all dimensions, and constructing appropriate analysis steps to optimally separate inaccurately measured molecules from completely different molecules. In this task it is sometimes helpful with more detailed knowledge of the exact data pattern that a specific molecule will cause, and the generation of such MS assays is a key step in targeted MS and targeted analysis workflows.

MS assays: leveraging what we already know

Many useful mass spectrometry workflows rely on the use of previously empirically observed behavior of analytes and peptides to re-identify them in the current data. One can refer to the information used to identify an analyte as an assay (Fig. 2.3). The MS proteomics assay to target one specific peptide ion typically contains, apart for the peptide sequence and the ion charge, the expected retention time of the peptide and information on how it fragments under the used fragmentation conditions. This fragmentation information contains the most common fragments, denoted by ordinal and fragment type, as well as the relative intensities at which they were observed.

Acquisition of MS assays can be done in multiple ways, depending on the target analytical platform. For SRM, acquisition of assays involves major effort. Suitable peptides first need to be selected based of empirical data from some other MS platform[116–118] or by *in silico* prediction[119–125]. Then synthetic peptides are ordered, followed by several iterations of fragment selection on the target triple-quadrupole[117,126]. This selection process is needed because



frag	m/z	intensity
y7	735	23%
y4	487	41%
b5	518	13%

Figure 2.3 - The MS assay retains empirical information of the observed peptide ion charge state, chromatographic retention time, and most intense fragment ions, including their relative intensity. This information is used trace the peptide ion online in SRM experiments, or post-acquisition in DIA experiments.

peptide fragmentation can vary substantially between HCD and CID instruments[64], and also many proteins of interest simply have no or few peptides that are detectable by shotgun MS. For DIA, assay generation is greatly simplified, as the same instrument is capable of shotgun MS as well as DIA analysis, and assays can thus be derived directly from shotgun data from the same instrument[127]. Still, critical assays should be validated by synthetic peptides to ensure full confidence.

Contributions to MS data description, compression and MS assay generation

As a secondary effect of our aim to improve computational proteomics tools, several studies have been performed to create enabling technology to understand MS data, to simplify MS data handling, and to generate MS assays. I summarize these scientific contributions here, while all details are described in the respective scientific paper.

The operation of the quite expensive mass spectrometers requires a lot of maintenance and instrument performance constantly fluctuates because of the issues with keeping the hundreds of parameters in the instrumentation calibrated. In such an environment it is highly useful to keep track of instrument performance to determine when the instrument condition is suitable for analysis of rare or costly samples. In Paper II we investigated the magnitude of uncontrollable effects[vi] in SRM. Here, a quality control sample was injected repeatedly 400 times over the course of 6 months, while following key parameters such as peptide retention time, signal intensity and fragmentation pattern. In this study, we find peptide retention times to vary in a mostly linear fashion, corroborating the idea of post-acquisition retention time normalization[128,129]. We also find that ion intensities are internally consistent, although they might vary from run-to-run, especially over long periods.

[vi] That is, effects of phenomena that were not forseen and can not be easily compensated by meta-data.

In the interest of reducing MS data size by reducing local redundancy, we developed a numerical compression package called MS-Numpress, consisting of 3 near-lossless numerical compression schemes (Paper III). This need arose from the voluminous data generated by DIA on the TripleTOF instrument, where data in the uncompressed text-based mzML-format could require 46 GB of storage per injection, compared to 3 GB in the binary vendor format. The compression schemes leverage log transforms, fixed-point arithmetics, and m/z and retention time data smoothness to store MS data in a more compressed representation. These techniques and compression ideas have all been used from the onset of computers, but have to my knowledge not been employed on MS data previously. Using an extensive set of 10 different benchmarking files of different samples and measured using different instruments, we demonstrate the speed and compression ratios of these compression schemes when embedded in the MS proteomics community standard data format mzML.

The process of extracting assays for DIA analysis from shotgun MS results is still quite unexplored, although it might seem very straightforward[127,130]. Essentially the MS2 spectra underlying shotgun MS peptide ion identifications are queried for quantitative information on how the peptide ion fragments, and this data is used to define MS assays for the DIA targeted analysis. We investigated assay extraction for DIA in Paper VI, with a focus on an upcoming instrument for DIA analysis, the quadrupole-Orbitrap[88]. Most importantly, we discover that assay quality plays a major role in this workflow, and suggest improvements for making high-quality libraries that increase the number of successful quantifications by 14-36% when employed in a DIA analysis, compared to low quality libraries of the same size. Embedded in this insight is the need to evaluate assay libraries in terms of DIA performance in addition to merely the number of assays.

Our findings regarding DIA assay extraction offer a way to increase quality of assay libraries essentially without any additional costs or experimental complications. As assays represent a fixed resource to be used in many further studies, improvements to assay quality can exhibit positive influence over a large number projects and injections.

Wrap-up on MS data

To summarize the chapter, mass spectrometry proteomics data is highly complex, although the measured complexity does not fully reflect the actual sample complexity. The discrepancy causes the data complexity to saturate, and likely comes from instrument limitations. Also due to instrument limits, variability can be expected in peptide elution time, signal intensity, mass accuracy and fragmentation. This variability needs to be considered by a successful computational proteomics tool to correctly separate true signals from noise. Finally, MS proteomics data is shown to be highly redundant. It turns out however, that reducing this redundancy actually is the same task as the primary task of analyzing the data: to find a more compact representation of the information we are interested in, the quantities of the peptides and proteins.

Chapter 3: Algorithms for Quantification in Raw MS Data

Modern MS proteomics is not possible without the use of custom analysis algorithms. From no less than 5 independent reports of intact peptide mass matching algorithms in 1993[131–135], mass spectrometry turned to automated analysis of MS2 spectra. The techniques of MS2 peptide interpretation[51,52] were adapted to the first modern MS2 search program by Eng in 1994, using a translated genome database to correlate all measured MS2 spectra with the theoretical spectra of all database peptides, and then select the best correlating peptide for each spectrum[136]. Many analysis improvements have been described since, and in particular novel MS workflows have further motivated new algorithms.

In this chapter a number of common analysis steps are identified that are frequently performed to quantify data, regardless of originating mass spectrometry variant. These steps are preprocessing and peak detection, scoring, machine learning for target identification, and finally statistical confidence calculation. Following this general review, some specific considerations for analysis of shotgun MS, SRM and DIA data will be presented. Finally, I will summarize the 3 algorithms that we have developed for analysis of SRM, shotgun MS data, and DIA (Paper I, VI and V).

Common themes of computational proteomics algorithms

A number of themes are common to the analysis of almost every MS workflow. First, some form of preprocessing is performed to decrease the size of the raw MS data. This can be spectral centroiding, deconvolution and de-charging, or peak detection in targeted analyses. Second, data are scored using one or multiple scores that are expected to differentiate the correct peptide ions from everything else. For example the probability of getting the observed degree of matching between the measured fragment pattern and the assay fragment pattern, retention time deviation from the assay retention time, or correlation between spectral peaks and theoretical fragments. Third, decoys are employed to normalize the scoring scale to the null hypothesis of not measuring the currently considered analyte. This phase also frequently includes some semi-supervised learning strategy to linearly combine scores and other parameters in an attempt to maximize analysis sensitivity. Fourth, the final decoy score distribution is used to calculate statistical confidence measures, such as the p-value, the q-value, and the posterior error probability[137].

Preprocessing of MS data in proteomics is perhaps not the most exciting task, but several studies have nonetheless bravely touched upon the subject, albeit mostly in the more mature shotgun proteomics field[138–143]. Spectra are typically centroided using either quadratic- or normal-matching to the most intense peaks or local maxima, or by template matching to some wavelet. The matching of full isotopic envelope templates to spectra seems to have fallen out of favor with the increasing prevalence of high-resolution spectra though, presumably because of too high computational costs. When performed, de-isotoping and de-charging is often done by simple heuristic rules where the typical isotope mass shift and considered charge states are compared to some user-defined threshold[144], with some notable exceptions[141,142]. It is likely that some statistical method could remove several parameters here and slightly improve performance, analogous to the feature detection algorithm of MaxQuant[145] and Dinosaur (Paper IV). In the targeted data format, assay traces and hills are smoothed by a wavelet transform, sliding window means or Savitzky-Golay smoothing[146]. For assay traces the initial preprocessing is followed by peak detection, meaning that improved performance might be gained by direct template matching on the raw data.

In scoring MS proteomics data, chemical and physical properties of the target analytes are utilized to separate them from other analytes and noise (Fig. 3.1). Peptide-like molecules are separated from other molecules by correlating isotopic envelopes to that of the theoretical averagine[147] at the same mass. Precursor and fragment intensity profiles can be correlated to each other to leverage their shared elution characteristics. An important score is the deviation from the expected retention time, which differentiates the target from other molecules. However, both in shotgun and targeted analysis the

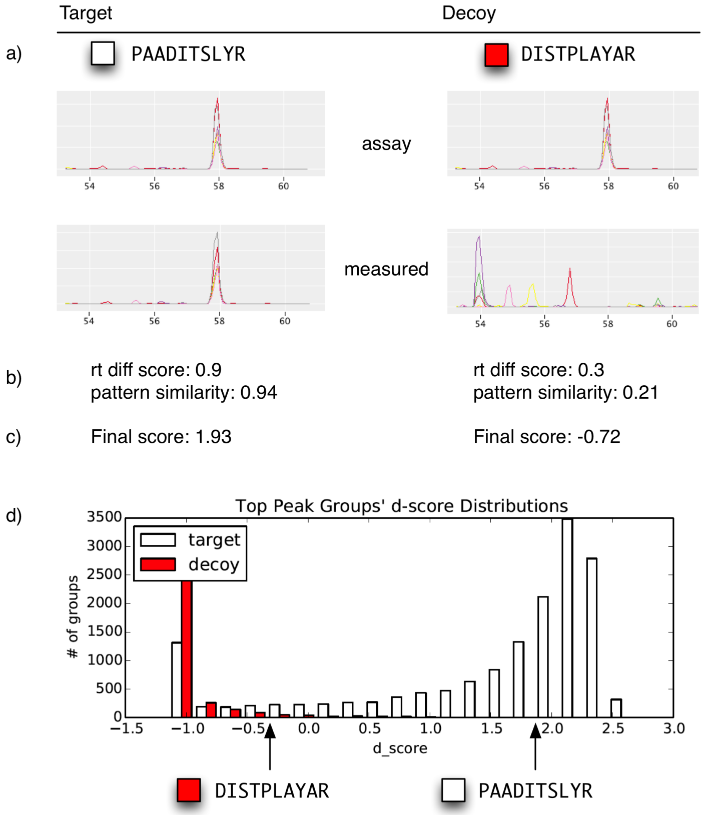


Figure 3.1 - Decoys, scoring and statistical analysis of MS proteomics data. a) Target peptide amino acids are permuted to generate decoys, while retaining amino acids specific to the used enzyme. b) The targets and decoys are scored by a number of orthogonal scores, c) that are combined into one final score, either by some fixed weighting or by an iterative optimizing approach. d) The decoy score distribution is assumed to approximate scores of targets that have not been detected, and is used to compute p-values for targets.

doubtlessly most powerful score is the matching of experimental and target fragment m/z values. In the targeted setting the known relative intensities of fragment ions is also utilized for additional power[148]. Without this knowledge, shotgun analysis can use the statistical distributions of fragment m/z errors[149] or prevalence of different fragment ion types[150].

Although highly powerful, scores merely prioritize the results internally. The use of one or multiple scores is therefore not enough to determine the absolute confidence in the analysis results. A statistical measure is needed to properly judge the validity of a result, meaning some model of a null hypothesis is necessary. In proteomics the decoy strategy is employed for this purpose[151] (Fig. 3.1). Here, the target database or assay library is permuted in a way that ensures no permuted peptides/assays will be present in the actual data. The scores that result from analyzing the data based on the decoys are assumed to represent the score-distribution of false associations between peptide/assay and data. Under this assumption, scores from targets can be tested against the decoy score distribution to attain p-values describing the likelihood of targets being false associations.

Several studies have shown that multiple scores can be combined for increased analytical power, including scores from different analysis tools[152–155]. Such a combination is usually a linear weighted sum of the sub-scores, where an iterative process selects the weights. The decoy null score distribution is used to specify a set of high-confidence target results, which are in turn used to update the weights of the sub-score combinator[156]. With the new weights, decoy and target scores are recomputed, giving new p-values and a new set of high-confidence targets. This procedure is repeated until saturation or detection of overfitting. Finally, the number of false matches in the result set can be estimated by utilizing the by definition uniform distribution of false matches[157].

Because of the large amount of hypotheses tested in proteomics, using the raw target p-values for assigning confidence is of limited value due to the considerations of multiple hypothesis testing. The preferred measure of statistical confidence in proteomics is rather the false discovery rate (FDR), which denotes an estimate of the percentage of false positives in the reported results. Alternatively, the posterior error probability denotes the probability of an individual result being false. In practice, the FDR, which is computed for a set of results, is achieved by computation of q-values[158], that for any score represents the lowest fraction of false positives that can be included while also including this score in the results.

Having listed some common steps of proteomics quantification algorithms, it must be noted that countless variants and ingenious solutions have been omitted for brevity, including specific techniques for MS2 identification, de novo sequencing, PTM identification and localization, label-free retention time alignment and isotopic/isobaric labeling quantification. To frame the new MS data algorithms presented in this thesis however, I need to elaborate on a few specific subjects relating to analysis and quantification of SRM, shotgun MS, and DIA data.

Analysis of shotgun MS, SRM and DIA data

As the above computational steps are applied to specific MS workflows, characteristics of the target workflow will raise particular difficulties to overcome in a successful algorithm. To understand the significance of the research presented in Paper I, IV and V on primary MS proteomics analysis algorithms, some specific difficulties in analysis of shotgun MS, SRM and DIA data must be outlined.

While MS proteomics has been obsessed with protein identifications, the importance of accurate quantification is increasingly recognized. In low-resolution shotgun MS, where analyte isotopes are not resolved, spectral counting is a suitable method for pseudo-quantification of proteins. Here the number of peptide identifications of a protein is used as an abundance proxy. This works well for abundant proteins and is easily computed from a list of identifications, but the method is sharply limited by the very finite amount of identifications per injection, which are spread over all identified proteins. As changed protein abundance will affect both the number of peptides identified and the number of identifications per peptide, spectral counting suffers from non-linear response, and since few identifications support most proteins, spectral counting has low accuracy for medium- and low-abundant proteins[159,160]. Because of its limitations, one could argue that spectral counts should only be employed with low-resolution data or when aggregated over groups of proteins such as pathways, where the drawbacks of spectral counting are mitigated by increased numbers of proteins. Studies using spectral counts are none the less continuously published, even though instruments typically have high resolution, and several quantification alternatives exist[161,162], like isobaric labeling[163], label-free MS1 quantification[164], and metabolic labeling[165].

For more accurate label free MS quantification, ion counts are used rather than spectral counts, either on precursor or fragment level. In SRM and DIA fragment ions are used because of their higher specificity, while shotgun MS uses precursor ions since the MS2 event might occur far from the chromatographic apex of the peptide, and thus poorly represent the peptide abundance. Precursor based quantification has been hindered though by the complexity of MS1 data, which complicates feature detection, as well as the difficulty of correctly matching features to the MS2 spectra and matching features between samples. Still, precursor based quantification of shotgun data has been demonstrated to work well in a large number of studies[166–169].

In SRM, the relatively small amount of data has so far limited the number of developed analytical tools compared to shotgun MS. The overwhelmingly most used tool is Skyline[170], which relies heavily on manual curation through a well-implemented graphical user interface. One clear challenge in SRM data analysis, compared to analysis of data from scanning MS instruments, is the lack of noise data that can be used to model the null hypothesis. Reiter et. al. took the pragmatic approach of solving this for their tool mProphet by purposefully measuring decoy assays on the instrument[148]. While statistically ideal, this has the drawback of further reducing the already limited SRM multiplexity. In the database solution DDB[128], Malmström et. al. used the shotgun inspired approach of modeling the null distribution of assay-data matches by scoring all assays against all data. This approach allows full SRM throughput, but comes with the requirement of a substantial MS assay database to compute accurate null distributions.

Finally, the analysis of DIA data has only lately seen a number of algorithms appear. In essence, two strategies are pursued for this purpose, targeted analysis and shotgun-like analysis. In the initially proposed targeted analysis, MS assays are used to *in silico*-trace the target fragments and precursor so that some SRM-like analysis can be employed. This was initially described by Gillet et. al.[87] and implemented by Röst et. al. in the OpenSWATH algorithm[171]. In the shotgun-like analysis, different strategies are employed to deconvolute the multiplexed DIA MS2 spectra into simpler versions, to be analyzed by conventional shotgun analysis and identification pipelines[172]. Here MS1 features might be used and correlated with MS2 traces to separate MS2 information from different analytes[172]. Because of the youth of the DIA method, it can be expected that multiple algorithmic advances remain to be discovered, even though the pioneering solutions clearly already utilize several key wisdoms from shotgun MS and SRM algorithms.

Contributions to the analysis of shotgun MS, SRM and DIA data

With the explicit goal of improving data quality from MS proteomics workflows to allow better and more accurate biological conclusions, we have developed 3 algorithms for identification and quantification of peptide ions in MS data from SRM, shotgun MS, and DIA (Paper I, VI and V). As with the enabling research studies, the results are summarized here, and described in detail in the respective scientific paper.

In Paper I, we developed the Anubis algorithm for SRM data analysis. This algorithm contributes 3 novel concepts to targeted data analysis. First, a new peak detection procedure is designed that detects chromatographic peaks directly by the expected ratios between fragments in the assay. This procedure utilizes the very stable fragmentation pattern of peptide ions, meaning that it is very robust to chromatographic or signal intensity aberrations. The drawback of this procedure might be difficulties in differentiating differently modified peptide versions when the measured fragments do not capture the modification site. Second, the constructed fragment ratio model is used to detect interference in individual data points and fragments, and to correct such interference based on the expected fragmentation pattern. This interference correction is discussed further below. Third, we propose computational resampling of the assay traces as a novel way to create a null model for evaluating the statistical quality of the reported assay results. The main properties that distinguish true assay-peak-matches are the correlation of the measured fragments and the agreement between their relative intensities and that of the assay. A random permutation of each channel will break these properties in the resampled data, and thus simulate noise with similar frequencies and intensity in each channel, but lacking the composite pattern. This method of null hypothesis estimation avoids the reduction of instrument throughput of mProphet[148] and also avoids the DDB need for large assay databases[128], thus providing an attractive alternative for null hypothesis modeling.

Paper IV revolves around creating a robust algorithm and tool for detecting precursor ions and their isotopic envelopes in MS1 data. Precursor ion isotopic envelopes are commonly called features, and the task of specifying the exact retention time, intensity profiles and isotopes of all features is called feature detection. The algorithm described in Paper IV, implemented in the tool Dinosaur, provides an improved method of feature detection. Dinosaur is shown to find a larger proportion of expected features, while maintaining quantification accuracy and feature quality when compared to 4 alternative feature detection tools. We also demonstrate faster runtimes compared to other algorithms, and include a visual quality control strategy. The primary use of features is quantification of peptide ions, but features can also enable identification of chimeric MS2 spectra - meaning MS2 spectra where multiple peptide ions have been fragmented together. Although a majority of spectra are estimated to be chimeric to some degree[94], this has been largely ignored, discarding up to 35% of identifiable peptides[144,173–176]. In summary, improved feature detection can potentially aid several MS proteomics applications, making the Dinosaur results a highly attractive outcome.

The final scientific contribution to MS data analysis in this thesis is a targeted analysis tool for DIA called DIANA. DIANA explores concepts from Anubis, focusing on detecting expected fragment ratios in the assay trace. However, the algorithm was completely reworked because of the increased noisiness of DIA data compared to SRM, which stems from the wider precursor isolation window. The major conceptual novelty of DIANA lies in the procedure to statistically estimate the significance of measuring a found number of data points matching the expected fragment ratio for each fragment, and then combining these probabilities in a statistically sound way. Benchmarking DIANA we show levels of sensitivity and specificity similar to the only alternative algorithm, OpenSWATH, but with a different spread of detected assays, providing a complementary option for DIA analysis. In the development of DIANA a joint effort was also made to create the tool PyProphet, reimplementing the semi-supervised learning module of mProphet, but with drastically improved computational performance.

The targeted MS technologies, both SRM and targeted analysis of DIA data, share the trait of explicitly looking for multiple fragments with a known relative pattern. This pattern is highly conserved for a single peptide on the same instrument[64] (Paper II). In fact, the information offered by each additionally added fragment grows increasingly smaller, as the intensity of the *n:th* fragment is highly predictable from the known fragmentation pattern and the *n-1* previous fragments. We have therefore in both Anubis and DIANA employed a method of interference correction, where internal consistency of the known fragmentation pattern is checked at each trace data point. By this method, fragment trace data points that are of higher intensity than expected can be identified and adjusted by the internally correct fragments. This might sound like a dangerous path to tread; when intensities are adjusted to "expected" values we are very close to a self-fulfilling algorithm. Indeed, caution is advised, and this correction is only employed for deviations of more than two-fold above the expected value. Empirical data show interference corrected quantities to be very close to raw quantities, but display higher correlation with the theoretical levels in a dilution series (Paper V). I have also manually checked the interference correction of hundreds of trace peaks, and verified that the corrected signals appear valid.

The performed projects on algorithms for MS data analysis were selected and prioritized based on a combination of need and opportunity. Anubis was fuelled by a need for large scale SRM quantification with low manual intervention. With the Anubis concepts in place, targeted analysis of DIA data was a natural next application that initially appeared straightforward, as no automated tool was yet published on the topic. Dinosaur finally was driven by frustration of the first few feature detection tools tested for an MS1 quantification workflow. Together these 3 tools can assist life scientists in achieving higher quality measurements across the palette of mass spectrometry proteomics methods, from broad discovery experiments to focused biomarker or pathway validation studies.

Wrap-up on algorithms

In the grand scheme of MS proteomics, computational tools have come very far in the last decades. One could argue that analysis of discovery MS data is essentially a solved problem, but considering the recent renaissance of chimeric spectral analysis and the gains from it[144,173–176] (Paper IV & VI), there still appears to be undiscovered improvements to make[177]. The continuous adaptation of algorithms to new instrumentation and MS techniques will also uphold a steady need for computational proteomics development.

One dimension that has not been utilized in any of the studies presented in this thesis is the integrative analysis of raw-data from multiple samples. Such analysis can yield substantial statistical strength, but I would advice caution. While the target signal should be reproduced, so will most of the noise, as most *noise* is simply signal from other molecules! Further caution is also advised upon later statistical testing for quantitative differences based on these results, because for such a test samples usually are assumed to be independent.

Chapter 4: On Algorithm Implementation

When one set of lenses is employed, computational proteomics can be viewed as just another application domain to apply computer science and software programming, along social media applications, toaster controllers and word processors. By this viewpoint, it appears as if the computational proteomics field is very ample at devising complicated algorithms[178], but quite poor at providing robust implementations of these. A fragile implementation might reveal itself by unnecessarily slow execution time, unexplained failures to analyze apparently valid input, or inability to handle standard data formats. There are naturally many counter-examples to this rather pessimistic picture, particularly the most used programs and pipelines display higher implementation quality, but when considering published algorithms without weighting for popularity of the tool I maintain the position that proteomics software holds a generally low level of implementation quality.

With the omnipresent need of software in all modern MS proteomics labs, I perceive software quality to be a hidden problem, partly for computational proteomics scientists, but mostly for biological and medical applications. In this chapter I will seek explanations to the experienced poor proteomics software quality by comparing academic and commercial software development. I will also describe some selected programming techniques that I have found useful for improving software quality, and finally make a proposal or wish for how global MS software quality might be improved.

Academic and commercial software development

It is easy to list differences in software development strategies between the commercial and academic spheres. Pair programming is the default mode of operation at many software development companies, along with test-driven programming, agile development when suitable, automated daily tests and builds, and other quality-oriented techniques. Meanwhile, a lone Ph.D. student (like myself) or post-doc programs the typical proteomics tool, while lacking large-scale programming experience and maybe also formal education in software programming. The comparison falters though; the money involved in academic software is orders of magnitude less than that of the commercial actors. Still, we should expect companies competing for customers to settle on processes that deliver the greatest quality software at the lowest cost. Application of these processes should therefore be even more urgent with the lesser funding available in academia.

Development resilience is the second major difference between the academic and commercial software practices. In computational proteomics the primary unit of merit, the publication, motivates focused development up until publication, followed by low-priority development post-publication as the next project is pursued. This is a poor strategy to achieve high-quality software; very few users will test the implementation pre-publication, when development is focused, while development is low-priority post-publication during the main exposition of the software to users. By contrast, companies commonly employ both alpha and beta test phases pre-launch, and maintain dedicated teams of bug-fixers during the entire lifetime of a product.

To finally demonstrate the difference in software quality that can be achieved with better development techniques and experience, consider the computational proteomics software with the hands down highest implementation quality: Skyline[170]. Skyline is backed by a team of developers with decades of accumulated industry experience, funded by grants and instrument vendor support, and presumably measures success in downloads and program launches rather than publications.

Programming techniques for improved software quality

In the following section, programming techniques will be discussed that have proven useful in the software development backing this thesis. They have been subjectively selected based on their frequency of use, and their ability to produce high quality software in short development times.

Code packaging. One extremely important stepping-stone for fast software development is code reuse[179,180]. This can be performed by factorizing code with similar purpose into libraries, that are versioned and can be embedded or used from multiple applications and other libraries[181,182]. For example, code related to modeling molecules and amino acids for computing their exact masses could be suitably packaged into one library, but should not be co-packaged with code for parsing command line arguments[vii]. Although this might be perceived as trivial, code packaging is complicated, and commonly boils down to minimizing risks. When source code errors are detected, and they always are[183], the bug fix in a library will affect all depending libraries and programs, potentially causing a massive need for down-stream fixes. Therefore the most general libraries that are frequently used benefit from being the most well tested and well defined, and from not containing any unnecessary functionality.

Error messages. The sign of mature software is software that communicates with its user[184]. When an error occurs, especially when program input is violating program assumptions, error messages benefit from being infinitely precise in describing the problem[185]. Achieving this is relatively easy from the programmers view, and frequently turns programs from unusable into merely slightly annoying.

Data standards. The spread of data standards is the greatest software quality achievement in computational proteomics this far. Use them if possible and applicable.

Premature optimization. All optimization should be performed after bottlenecks have been pinpointed by measuring elapsed computation time,

[vii] Beware of the library called *util*.

and the program as a whole was deemed to be too slow[186]. However, program execution time must be evaluated based on user needs for real input data, not on programmer needs for test data.

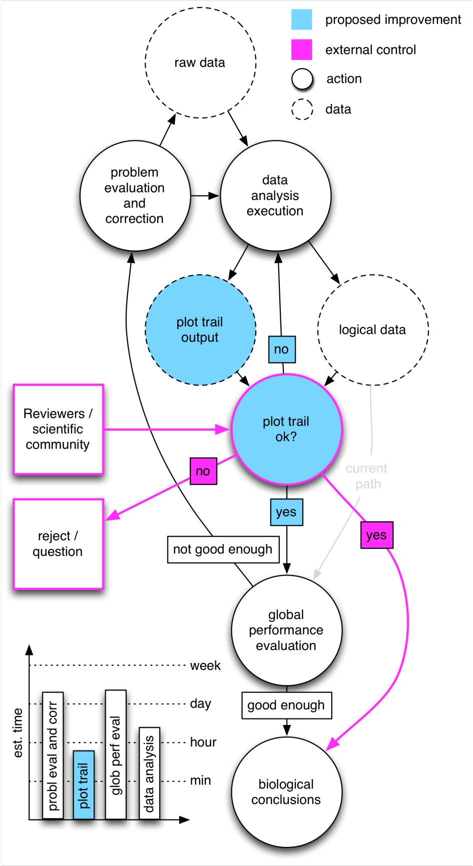


Figure 4.1 - Conceptual overview of the plot trail quality control functionality

Sort-merge. One recurrent task in this work has been the merging of two lists based on proximity in a shared property. The lists could for example contain MS1 features and MS2 identifications. Although the initial solution of comparing units in the first list with all units in the second is easy to program, sorting both lists and walking through them simultaneously will asymptotically speed up performance, if the data is reasonably sparse.

Reactive parallelization. For larger computational tasks, currently meaning several gigabytes of input data, it is worthwhile to provide parallel implementations for faster and scalable operation. This can be accomplished without introducing overly complicated concurrent and stochastic bugs by using some reactive programming technique[187] that allows function computation in an asynchronous way without introducing massive threading overhead. Using reactive parallelization, logical subunits of the data, like individual features or DIA assay traces, can be isolated and processed as soon as computational cycles are available. I have employed this strategy in Paper I, IV and V.

Analysis transparency and traceability. Proteomics datasets are constantly growing, since long passing the point where complete manual analysis is feasible. Current workflows therefore use highly powerful automated computational analysis tools. Herein lies a hidden potential drawback however: scientists could suffer reduced understanding of the analytical and computational steps actually performed. When data analysis is treated like a black box of unknown internal workings, the ability to determine if the data analysis is performing as intended might be compromised. We have focused some work at preventing this, and developed a visual quality control strategy that we call *the plot trail* (Fig. 4.1). The plot trail borrows its goal from economical auditing, meaning to minimize the risk of any large mistakes having been made. Because the proteomics data volume is too large for complete manual control, we need to subsample it. In the plot trail strategy, the results of each computation step are therefore randomly subsampled, and this subsample is visually plotted and written to disk for retrospective control. This achieves a two-fold goal. First, the program user is allowed suitably limited but highly relevant information of the details of algorithm execution, and might therefore detect sub-standard performance before heavy investment is made in down-stream analysis and interpretation of the results (Fig. 1.4). Second, the program user is subtly given understanding of the key elements of the algorithm used, and might become better at predicting effectiveness of the program or at localizing errors.

Wrap-up on implementation

Although perhaps a slightly unconventional topic in a Ph.D. thesis on computational proteomics, it is my sincere opinion that software implementation is important, and should receive more focus in the scientific world. It is important even though it's barely *publishable*, because it saves you and everyone else time, money and mindboggling bugs in the long run. It is the explicit hope that the described programming techniques could be adapted by others to help them improve their software quality, and that they in turn would share experience and techniques so that computational proteomics as a field might improve.

With the current motivational structures of computational proteomics scientists, it will be hard to achieve greatly improved software quality. But in essence, software quality is mostly a problem for *software users*, meaning the vast bulk of mass spectrometry proteomics scientists who do not publish algorithms or tools, but rather leverage these to draw biological conclusions. My suggestion to this community is to pool resources globally and create a new executive software development unit. This unit would be funded and mandated with the task of maintaining decent-quality implementations of proven computational proteomics algorithms. A secondary task would be to create new implementations of niche algorithms to make them available for general use. The development team should be composed of a mix of industry and research background developers. In essence, this could be seen as a form of translational computational proteomics, where computational results are refined to a level where they can be applied for biological studies.

Chapter 5: Key Weaknesses in MS Proteomics of Today

Considering once more the goal of providing high quality quantitative data to enable comprehension of biological systems, it must be considered that even todays advanced mass spectrometry might not be the ideal technology. It should be of major importance for proteomics scientists to regularly review the main weaknesses of their platforms, and think about how these could be resolved[viii]. This chapter will be devoted to weaknesses in mass spectrometry proteomics that are not directly studied in the included scientific papers, but that appear as unmovable obstacles that are circumvented time and time again. Four weaknesses have been selected: the decoupled peptides, the unknown specific electro-spray function, the non-use of available ions, and the limited sequencing speed (Fig. 5.1). These weaknesses hinder us from doing label-free absolute quantification, from correctly detecting proteoforms rather than gene-products, and from achieving the speed and sensitivity that is needed to truly match the sample complexities of biological tissues.

[viii] Each proteomics scientist will, in confidence after a few beers, reveal the main drawback and weaknesses of their favorite proteomics technology. There will be some discussion on the order of said drawbacks, but nothing that can't be resolved by another round. I assume that I have already shared one or two with most of the readers at this point, but if I haven't and you continue reading, *you owe me one*.

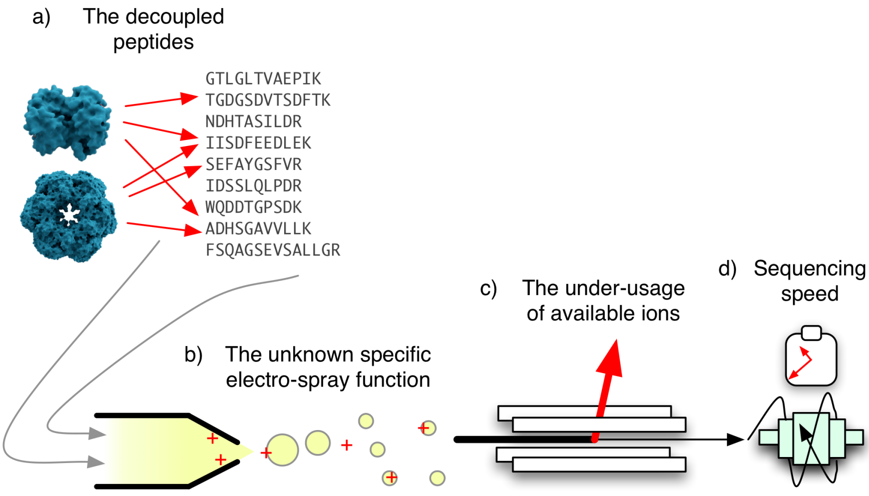


Figure 5.1 - Selected MS proteomics weaknesses. Key instrumental and methodological areas to study are a) peptide decoupling due to digestion, b) the mechanics of electro-spray ionization of thousands of peptides, c) the under-usage of ionized analytes, and d) the ability to sequence these in a timely manner. (The protein 3D images from "Protein composite" by Thomas Splettstoesser (www.scistyle.com), CC BY-SA 3.0)

The decoupled peptides

Unfortunately, the most used MS proteomics workflow, where proteins are digested into peptides prior to MS analysis, bears a weakness that cannot be compensated[ix]. Many refer to this as the protein inference problem[188,189], where the question is: given a set of identified peptides of which some are unique to a theoretical protein and others not, what proteins are in the sample? The inherent problem is that the peptides from a protein are no longer connected during measurement, but are intermingled with peptides from all the other millions of unique proteins. This carries greater implications than simply detecting what *gene* that is expressed and present

[ix] The decoupled peptide problem could alternatively be solved by extensive protein fractionation in for example 2D-gels, at least theoretically. This technique has fallen out of favour though because of the large amount of manual work required.

though, the peptide decoupling also means that we cannot discover or quantify the protein PTM configurations and proteoforms that exist in the sample. Many biological components define their function and state by a specific combination of modifications, located in different parts of its proteins. This information is by design not accessible after digestion, unless modifications are situated locally on the protein.

The only way to avoid peptide decoupling is to not digest proteins, and rather perform mass spectrometry of intact proteins, so called top-down proteomics. Initially ridden with a small ocean of childhood problems, the technique seems to be maturing lately, with one study detecting 1034 gene products and more than 3000 protein isoforms upon top-down analysis of extensively fractionated cell line material[190]. Some issues of MS analysis of whole proteins are the large range of protein sizes, the large number of observed charge states for such big organic molecules, the number of ions needed for complete fragment series, and the lack of algorithms for analysis of top-down data. Nevertheless, top-down proteomics will no doubt become the method of choice once enough advances in chemistry, instrumentation and algorithms have made the technique robust and reliable.

The unknown specific electro-spray function

For systems biology, the biggest weakness of MS based proteomics is the inability to perform native absolute quantification at the peptide level. The weakness can be circumvented with different labeling strategies, by adding isotopes of the target analytes at known concentrations, and assuming linear response. This is however greatly unsatisfactory, as the primary MS strength of massive parallelization is heavily impaired. The key reason that absolute quantification is not possible is the complexities of electro-spray ionization. It is not known in detail how amino acid backbone molecules are ionized during electro spray, and in particular how the degree of ionization depends on the amino acid configuration. The process is clearly deterministic, but the current inability to compute peptide response factors or adjusting for matrix effects makes practical label-free absolute quantification at the peptide level impossible.

Putting costly labeling protocols aside, there is actually a known solution for drastically reducing the effect of the electro-spray response, so that most peptides have very similar responses. By reducing the nano electro-spray flow rate from 300 nl/min to ~1 nl/min, the electro-spray droplet size is greatly reduced, and the ionization efficiency enhanced - resulting in the almost complete ionization of all analytes[191,192]. Initial demonstrations used pump-free chromatography which requires constant expert managing, but later studies have demonstrated multi-spray solutions that manage to achieve small droplets while staying at optimal chromatography speeds[193]. These techniques would be expected to greatly benefit proteomics, and it is not clear why their uptake by mass spectrometry vendors has been so slow. Peptide responses could be expected to vary within an order of magnitude even with these miniaturized droplets however, why a model to explain and predict the electro-spray function would remain valuable.

Solving the electro-spray function will require two main components. First and absolutely crucial is the acquisition of a sizeable learning dataset, containing the measured signal responses of a large number of peptides at exact known concentrations. Second, a powerful framework is needed to represent the function. This could come either from increased understanding of the electro-spray chemistry through a large number of basic well-controlled experiments of increasing complexity, or from modern machine learning approaches.

The acquisition of a large dataset of exactly known peptide composition is non-trivial or expensive. Ordering a pure and accurately quantified synthetic peptide might cost $1000, and because of a need for 10 000 - 100 000 peptides this price tag become prohibitive. Crude synthetic peptides have been utilized as a more affordable alternative, with the trade-off of an approximated 5-fold spread in peptide concentration[125], which clearly limits the maximal achievable algorithm accuracy. Basing the dataset on proteins[194,195] is an attractive alternative to collect peptides at increased rates while maintaining a known concentration of the protein. Here protein digestion needs to be accurately controlled though, as not to introduce unknown variation in the MS input. Further, the protein expression system is very important: it should minimize post-translational modification and protein cleavage. Regardless of synthesis method, the sample peptides should reflect the relevant peptide-space in terms of amino acid compositions and measured peptide responses.

Although the electro-spray function has been provocatively labeled *unknown* here, this is not entirely true. Multitudes of papers describe the ESI droplet formation and their size at different flow rates[196], how droplets are iteratively dispersed[197–201], and what properties of peptides and other chemicals that influence their response factors[107,202–204]. To complement, several machine-learning studies have been completed with the aim of predicting high-responding peptides for inclusion as targets in targeted proteomics workflows[120–124,195]. These display increasing success at this binary classification task, and also indicate peptide properties that influence peptide responses. Still, the knowledge of the mechanisms behind electro-spray ionization has never been scaled above 10 parallel analytes, far from proteomics production samples of millions of analytes. Meanwhile, machine-learning approaches require large data sets, but employ a typically naïve model, where the response factor is guessed to be a function of peptide properties that are summed from the individual amino acids. No machine-learning study has so far been able to analyze the specific sequence of amino acids in the peptides, which likely plays a key part in the ionization, or attempted to predict the actual response factor as opposed to mere classification or ranking of response.

The under-usage of available ions

In all proteomics mass spectrometry, the percentage of ionized analytes that are actually subjected to any mass analysis is minuscule[x]. To support this statement, consider the following estimations. During typical 300 nl/min electrospray on a common instrument[xi] a voltage of ~3 kV is applied, resulting in a current during analysis of a ~200 nA, which translates to roughly 1.2e12 charge units/s. During successful sample acquisition the total ion count measured by the instrument is about 3e9 ions/s. If we assume all ions to be of charge 1, this would entail that about 1 out of 400 electro-sprayed charge units is deposited on a target molecule that enters the instrument and is successfully measured. After entry into the instrument, further ions are lost. For MS1 spectra, ions are typically accumulated in the ion trap for 5 ms, followed by a MS1 scan taking 200 ms, leaving the ion trap unused for 195 out of 200 ms. Even worse, for MS2 spectra a few percent of

[x] Or as amply put by Andrew Krutchinsky: It's like holding a bucket in the Niagara falls.

[xi] A Thermo Scientific Q-Exactive Plus.

the mass range is collected for 5-100 ms, followed by a 100 ms scan event. During MS1 events all ions are discarded 97.5% of the time, and during MS2 events approximately (800 - 2) / 800 = 99.75% of ions are not used, if we naively assume a uniform spread of ions over the m/z range. Therefore, commercial instruments could theoretically gain around a 100-fold sensitivity by better transmission and distribution of charges following ionization, and another 100-fold by better utilization of the ions that do arrive in the instrument. As there are papers claiming ~50% transmission efficiency[205], future improvements here seem plausible.

Scanning instruments all rely on ion traps for collecting populations of ions, which are then subjected as a group to scanning mass analyzers. Drastic improvements to ion trap capacity by several orders of magnitude have been proposed as a means to increase ion usage, but this does not solve the essential problem: how do we analyze such huge amounts of ions in a meaningful way? Already with current ion trap capacities, mass analyzers experience reduced transmission and resolution due to space charge effects, where ions of similar charge push each other outwards because of electrical repulsion, meaning that the m/z - intensity distribution gets blurred. This leads up to the last selected weakness, the MS sequencing speed, which needs to be addressed simultaneously to the under-usage of ions.

Sequencing speed

Sequencing speed used to be the limiting factor in the number of peptides and proteins that could be identified in a proteomics study. However, sequencing speed is lately reaching levels where it is no longer the single key bottleneck of the workflow. For example, ion transmission into the mass spectrometer can be more important[206], as could dynamic range or ion trap capacity. An estimated 16-25 peptide ions/s are injected into the MS by current chromatographic setups[94] (Paper V), which matches closely the ~20 MS2/s of reported high-resolution sequencing speed for two top-of-the-line commercial instrument of today[76,77]. Performing one MS2 per injected peptide ion will most likely not be enough however, even for discovery MS, as suboptimal real-time sampling often necessitates several MS2 scans per peptide. In DIA analyses specifically, increased sequencing speed would allow increased chromatography performance. Targeted analysis of DIA data requires several MS2 over the elution of a peptide, which no mass spectrometer can deliver when interfaced with the best chromatography of today. With increased sequencing speed, sharper DIA chromatography would be possible, giving improved ionization and increased sensitivity in return.

Achieving increases in sequencing speed might not be easy, since instruments are already highly concurrent. In modern instruments, most subcomponents perform their tasks simultaneously in an assembly line fashion. Even so, recent instrument releases seem to always include improvements in sequencing speed, why the practical limit is probably still not reached. In particular, no MS vendor has yet attempted parallel sequencing using more than two mass analyzers[207], which would be a natural way of scaling performance similar to how multi-core computers have improved computation power once processor speeds reached their practical limit around 3 GHz.

Wrap-up on weaknesses

One thing that is intriguing with mass spectrometry proteomics is that the technology is very powerful and capable of multiplexed measurement of the complex proteome, yet it still lacks some properties that would be considered fundamental in other analytical technologies. The listed weaknesses have in common that they limit the feasibility of interpretation and understanding of the biological mechanisms at play. The ability to measure the absolute concentrations of peptides or relative concentrations of proteins would greatly simplify detection of proteoforms and PTMs. The improved sensitivity and multiplexing from better ion usage and parallel sequencing would increase proteome coverage, in particular for the low abundant proteins and small proteins where data is scarce. Acquiring these abilities should greatly increase the likelihood that mass spectrometry proteomics measurements are interpretable at a biological level, and therefore increased efforts could be warranted in this direction in future research.

Chapter 6: Conclusions and Relevance

With the currently low rate of new drug and biomarker approvals, along with increasing costs of drug development, it is clear that a change is needed if the overall global health and wellbeing should continue to improve. Still, life science has acquired exponentially more powerful tools to acquire genomic, transcriptomic and proteomic data. How can this discrepancy in the amount of available data, and number of new therapeutics be explained? As a last resort, one could argue that we have reached the end of what's possible to achieve in terms of human health - at some point the magnificent machine that is the human body will inevitably fail. I do not believe we have reached this point. Rather, I argue that life science is either asking the wrong questions, or failing to interpret the enormous amounts of data it generates. Indeed, we are living in an age when acquiring data is easy and cheap, while understanding the biological processes that the data reflect is as hard as ever. To make use of the current voluminous data sets, we need to extend our capability of data interpretation in terms of speed, quality and robustness.

This thesis has been aimed at improving the computational tools for mass spectrometry based proteomics, to enable higher quality data for improved ability to draw biological and medical conclusions. In this pursuit I have created enabling technology for better instrument control and better data handling, and algorithms for direct MS data analysis to quantitate peptides. As a secondary effect of these projects, I have come to several conclusions regarding the states of computational software quality and the mass spectrometry technology, which are also presented as results and discussion points.

The work here described is not merely a set of algorithms for analysis of MS proteomics data. It embodies a strategy to tackle key weaknesses in MS methodologies, to mature the technology for greater robustness and performance. I have developed algorithms that I have demonstrated to improve the analytical power of the major MS methodologies DDA, SRM and DIA, in terms of quantitative accuracy (Paper I & V), improved recall (Paper IV), and ease of application (Paper I & IV). However, the primary goal of a computational platform is to maximize the total amount of knowledge it allows the mass spectrometry lab to produce. I have therefore equally focused on speed of execution and standardization (Paper I, III-VI), and practical issues related to instrument performance (Paper II) and data storage (Paper III), so that the mass spectrometry lab as a whole might function efficiently.

The ultimate test of new academic algorithms and programs is whether they are used in later studies. Anubis has been used in a few published studies[208,209], and is a part of additional projects that are still under investigation. I am proud to note that the Numpress algorithms have a quite substantial uptake with support in most maintained proteomics tools and pipelines[210] (Paper III). DIANA is used frequently in our lab, for protein-protein interaction studies and regular quantitative profiling, and coupled to this usage we are replacing the previous MS assay pipeline with the one described in Paper IV to give faster and better performance.

There are of course unlimited possible investigations that can be performed to follow up the presented studies. To name a few, I believe that application of the MS techniques and the quantitative algorithms in protein-protein interaction experiments such as affinity purification MS or cross-linking MS could greatly facilitate biological understanding. There could also be gains in performing online feature detection, or more advanced feature/peak detection in DIA analysis compared to the DIANA implementation. To completely change the game of MS proteomics however, the main instrumental weaknesses need to be tackled, in particular the electro-spray function and the decoupled peptides.

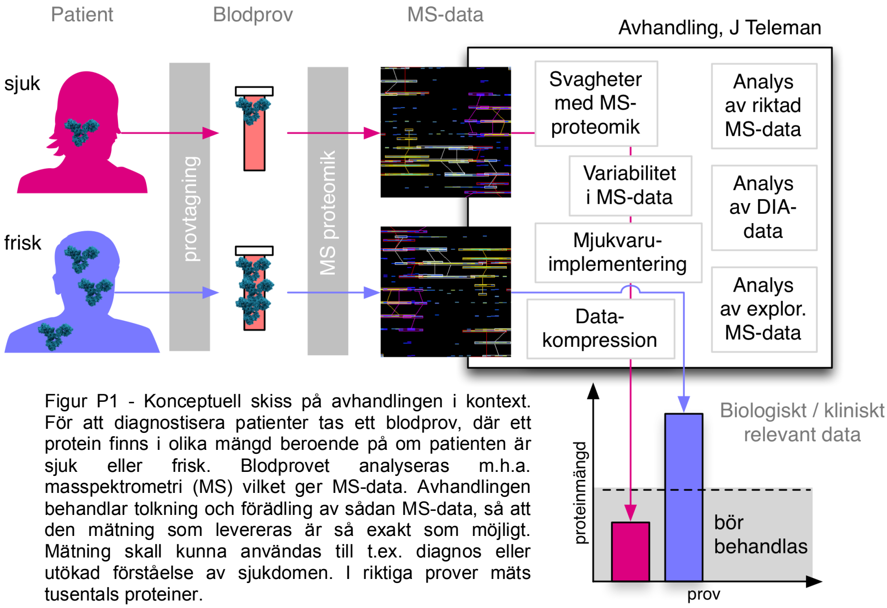
To improve proteomics analytical power and allow deeper understanding of the biological processes underlying life science data, the exposed mass spectrometry proteomics weaknesses and poor proteomics software quality will be key points to address. Therefore the continuous discussion, study and resolution of the instrumental and implementational issues should be a ongoing relevant task for the community, so that the increased technical ability might finally allow comprehensive understanding of the complex biological systems. Although many problems remain to be solved in mass spectrometry based proteomics, we should take heart from the fact that the technique has taken us so far even with such limitations. No doubt resolving these weaknesses will be the goal of substantial effort the coming decades, and then we might finally see systems biology that can model human cells or even tissues, and perhaps fuel continued advancements in life science for better health and human wellbeing.

Populärvetenskaplig sammanfattning

Inom livsvetenskaperna studerar man det som i dagligt tal benämns som levande. Det betyder i praktiken att man studerar celler, de små byggstenar som allt liv består av. Först under de senaste 20 åren har man insett hur otroligt komplicerade celler är, med hjälp av omfattande framsteg inom gensekvensering och andra analytiska verktyg. Hastigheten på den tekniska utvecklingen inom livsvetenskaperna är jämförbar med datorkraftutvecklingen.

I framkanten av dagens forskning i livsvetenskap försöker många forskare förstå och mäta cellens proteiner. Proteinerna är en stor grupp molekyler som utför majoriteten av cellens funktioner, som reproduktion, energiproduktion och kommunikation med andra celler. För att studera cellens proteiner används ett mycket känsligt instrument, masspektrometern, som kan mäta proteinmängden av tusentals proteiner i så lite som 1 μl av ett biologiskt prov. Med masspektrometri kan man bland annat förstå skillnader mellan livshotande infektioner och vanlig halsfluss eller detektera prostatacancer direkt i ett blodprov (Fig. P1), men man kan också t.ex. kvalitetskontrollera livmedelsproduktion.

En svårighet med masspektrometri är att tekniken genererar stora mängder mätdata, eftersom instrumenten är känsliga och celler komplicerade. En modern masspektrometer producerar ungefär 1 GB mätdata i timmen, och för att tolka datan krävs specialiserade algoritmer och datorprogram. Den forskning som jag presenterar här beskriver flera algoritmer för analys av olika typer av masspektrometridata. Vi kan visa att dessa i flera fall ger bättre resultat än de alternativ som tidigare beskrivits, exempelvis med avseende på hur många proteiner som detekteras, eller vilken beräkningstid som krävs.



Figur P1 - Konceptuell skiss på avhandlingen i kontext. För att diagnostisera patienter tas ett blodprov, där ett protein finns i olika mängd beroende på om patienten är sjuk eller frisk. Blodprovet analyseras m.h.a. masspektrometri (MS) vilket ger MS-data. Avhandlingen behandlar tolkning och förädling av sådan MS-data, så att den mätning som levereras är så exakt som möjligt. Mätning skall kunna användas till t.ex. diagnos eller utökad förståelse av sjukdomen. I riktiga prover mäts tusentals proteiner.

Inom masspektrometri skiljer man på explorativ och riktad analys, där den explorativa analysen (discovery MS) antar väldigt lite om provet och kan hitta många proteiner, medan den riktade analysen (SRM) letar efter fördefinierade proteiner men i gengäld mäter dessa med större känslighet och precision. Det finns också hybridmetoder, där man försöker uppnå fördelarna med både den explorativa och riktade metoden. Vi har utvecklat olika program för analys av både explorativ och riktad masspektrometridata, samt för den relativt nya hybridmetoden Data-Independent Acquisition (DIA).

Utöver direkta analysverktyg har vi genomfört tre studier som förenklar hantering och drift av masspektrometerlabb. I den första studien undersöker vi hur skicket på en masspektrometer ändras under ett halvår av vanlig användning. För att genomföra detta utvecklades ett automatiskt verktyg för kvalitetskontroll. Det visar sig att speciellt det första steget av analysen, då provet delas upp baserat på hydrofobicitet, ger något olika resultat beroende på normal användning. Den uppmätta signalen i masspektrometern varierar också, eventuellt beroende på slumpmässig kontaminering. I den andra studien utvecklade vi nya kompressionsalgoritmer som är speciellt anpassade till masspektrometridata för att förenkla hanteringen av den stora mätdatamängden (MS-Numpress). Dessa komprimerar informationen, d.v.s. viker ihop den, så att den tar mindre plats att lagra. Slutligen beskriver den tredje studien ett nytt förbättrat sätt att definiera hur proteiner skall mätas i riktade analysmetoder och hybridmetoder. Vi visar att vi genom att öka kvalitén på våra mätkoordinater (MS assays), kan öka mängden identifierade proteindelar med 30%.

Utöver de vetenskapliga studierna innehåller avhandlingen diskussion kring två ämnen som inte resulterat i publikationer: brister i akademisk mjukvarukvalité, och svagheter med MS-proteomik, d.v.s. mättekniken masspektrometri tillämpat på proteiner. Tyvärr är en stor risk med akademisk mjukvara att den faller i glömska p.g.a. praktiska problem vid användning - akademisk mjukvara utvecklas i allmänhet med mycket mindre finansiellt stöd än kommersiell mjukvara. Detta diskuteras i avhandlingen, liksom eventuella lösningar. I MS-proteomik identifierar jag 4 brister som framtida forskning bör sträva efter att åtgärda för att drastiskt förbättra kvalitén på masspektrometridata och öka möjligheten att dra relevanta biologiska slutsatser från den.

De forskningsresultat och program som vi utvecklat är ämnade att hjälpa livsvetenskaperna - och förhoppningsvis i förlängningen sjukvården - genom att möjliggöra känsligare, snabbare och tillförlitligare analys av proteinnivåerna i biologiska prover. Bättre analyskvalité kan i sin tur bidra till ökad förståelse av biologiska system, eller upptäckten av nya läkemedel eller diagnosmetoder, och även om vägen dit kan verka oändligt lång så hoppas jag att denna avhandling åtminstone går i rätt riktning.

Acknowledgements

One might see very far when standing on the shoulders of giants, and luckily they are big shoulders, to accommodate all the people who helped to get all the way up there. The are certainly those that I have unjustly forgotten to list here, but as a compensation you get the first heartfelt thanks of all.

I have for half a decade been orbiting the twin stars *Johan* and *Fredrik*. The Ever Enthusiastic and The Always Helpful. I will forever remain indebted to you for the skills you taught me, and the knowledge you imparted.

Without the prospect of working under your excellent leadership, I would never have entered science *Johan*. I am more than thankful for the opportunity, and in particular for all the faith I was shown, even though you must have questioned my projects at times.

There are so many things I've learned from you *Fredrik*, who will always see the devil in the details. Especially thanks for prodding me in the right direction during the hairiest moments of some projects, and for the great encouragement.

To all *the IMPs*: you are such an incredible group - please maintain the outstandingly nice atmosphere forever! *Simon:* thanks for deciding to come to Sweden with your crazy hands-on MS skills. That was truely vitalizing, and I'm sure someday BIF and taxlink will rise again! *Ola:* sharing an office with you has been an honour. Your special laid back humour made it feel like home, even though it's just a tincan packed with two sardines. More thanks to *Karlsson:* for the super-solid science that you always present as if it was boring. *Lotta & Anahita:* for shaking my uselessly hibernating Finnish. It didn't wake. *Sjöholm, Emma, Gabriel, Erik, Helena, and others:* for being awesome!

I can't believe we're all doctors soon, *Ufuk* and *Marianne*. Thank you for keeping me sane from the biology people, for showing the way. *Ufuk:* we had some great trips in the name of science and good adventure! *Paolo & Linn:*

for the good times! *Peter:* for being scary and impressive, and for teaching MS/MS interpretation to the new kids.

There is a special place in my heart for all the wacko *Immunotech PhD-students and post-docs*. Thanks for taking all the dissertation parties so seriously and always igniting on all cylinders. This is fun. Keep dancing.

A special thanks is directed to the Prefects of Immunotechnology, *Carl* and *Sara*, whos silent lack of personalized objections has enabled me to efficiently navigate the duties of two departments, since the Malmström group moved to MedFak.

I am full of gratitude to the students I've been supervising. It is truelly amazing that there is a thirst for knowledge that can fuel so many questions! But since nothing will sharpen understanding like slighly off questions, you have challenged me to grow.

Two special pieces of thanks to *Stefan* at MediaTryck, and my uncle *Henrik* who designed the cover.

There would not have been a book to hold, without all the invaluable help and support of all friends, in particular during this hectic spring. Thank you in particular to *Astrid and Johan* for taking the kids, making us food, lending us stuff and building our home.

I too, believe you can fly, *Tommie*.

This book would of course not be here today unless my *mother* helped me store my Lego in HPLC-column boxes when I was just a kid. I also want to credit my *father*, who cruelly never got the chance to affect this thesis, even though he claimed I delayed his. You are both a part of me in all I do, I can hear you in my words, and see you in my thoughts.

Without you there would be no recovery, no shelter from the harsh rational realm. *My family*, I adore you in all ways, and praise you for forcing me to the here and now. *Kikki*, I love you, because of you.

References

(1) Lim, S. S.; Vos, T.; Flaxman, A. D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; AlMazroa, M. A.; Amann, M.; Anderson, H. R.; Andrews, K. G.; et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **2012**, *380* (9859), 2224–2260.

(2) Lozano, R.; Naghavi, M.; Foreman, K.; Lim, S.; Shibuya, K.; Aboyans, V.; Abraham, J.; Adair, T.; Aggarwal, R.; Ahn, S. Y.; et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **2012**, *380* (9859), 2095–2128.

(3) Ledford, H. Translational research: 4 ways to fix the clinical trial. *Nature* **2011**, *477* (7366), 526–528.

(4) Holden, J. P.; Lander, E. Report to the president on propelling innovation in drug discovery, development, and evaluation. September 2012.

(5) Roy, A. S. A. STIFLING NEW CURES: The True Cost of Lengthy Clinical Drug Trials. March 2012.

(6) Hatthew Herper. The Truly Staggering Cost Of Inventing New Drugs. *Forbes* **2012**.

(7) Aylin Sertkaya; Anna Birkenbach; Ayesha Berlind; John Eyraud. Examination of Clinical Trial Costs and Barriers for Drug Development. July 25, 2014.

(8) Downing, N. S.; Aminawung, J. A.; Shah, N. D.; Krumholz, H. M.; Ross, J. S. Clinical Trial Evidence Supporting FDA Approval of Novel Therapeutic Agents, 2005-2012. *JAMA* **2014**, *311* (4), 368.

(9) Polanski, M.; Anderson, N. L. A list of candidate cancer biomarkers for targeted proteomics. *Biomark. Insights* **2007**, *1*, 1–48.

(10) Chalasani, P.; Livingston, R. Differential Chemotherapeutic Sensitivity for Breast Tumors With “BRCAness”: A Review. *The Oncologist* **2013**, *18* (8), 909–916.

(11) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics MCP* **2002**, *1* (11), 845–867.

(12) Surinova, S.; Schiess, R.; Hüttenhain, R.; Cerciello, F.; Wollscheid, B.; Aebersold, R. On the development of plasma protein biomarkers. *J. Proteome Res.* **2011**, *10* (1), 5–16.

(13) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409* (6822), 860–921.

(14) Venter, J. C. The Sequence of the Human Genome. *Science* **2001**, *291* (5507), 1304–1351.

(15) Simon Tripp; Martin Grueber. Economic Impact of the Human Genome Project. Battelle Memorial Institute May 2011.

(16) Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) http://www.genome.gov/sequencingcosts/ (accessed Mar 31, 2016).

(17) Strohman, R. Epigenesis: The Missing Beat in Biotechnology? *Bio/Technology* **1994**, *12* (2), 156–164.

(18) Karr, J. R.; Sanghvi, J. C.; Macklin, D. N.; Gutschow, M. V.; Jacobs, J. M.; Bolival, B., Jr; Assad-Garcia, N.; Glass, J. I.; Covert, M. W. A whole-cell computational model predicts phenotype from genotype. *Cell* **2012**, *150* (2), 389–401.

(19) Modrek, B.; Resch, A.; Grasso, C.; Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **2001**, *29* (13), 2850–2859.

(20) Johnson, J. M.; Castle, J.; Garrett-Engele, P.; Kan, Z.; Loerch, P. M.; Armour, C. D.; Santos, R.; Schadt, E. E.; Stoughton, R.; Shoemaker, D. D. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **2003**, *302* (5653), 2141–2144.

(21) Lim, L. P.; Lau, N. C.; Garrett-Engele, P.; Grimson, A.; Schelter, J. M.; Castle, J.; Bartel, D. P.; Linsley, P. S.; Johnson, J. M. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **2005**, *433* (7027), 769–773.

(22) Selbach, M.; Schwanhäusser, B.; Thierfelder, N.; Fang, Z.; Khanin, R.; Rajewsky, N. Widespread changes in protein synthesis induced by microRNAs. *Nature* **2008**, *455* (7209), 58–63.

(23) Krishna, R. G.; Wold, F. Post-Translational Modifications of Proteins. In *Methods in Protein Sequence Analysis*; Imahori, K., Sakiyama, F., Eds.; Springer US: Boston, MA, 1993; pp 167–172.

(24) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4* (6), 1534–1536.

(25) Carter, R. J.; Parsons, J. L. Base excision repair: A pathway regulated by post-translational modifications. *Mol. Cell. Biol.* **2016**.

(26) Piacentini, M.; D'Eletto, M.; Farrace, M. G.; Rodolfo, C.; Del Nonno, F.; Ippolito, G.; Falasca, L. Characterization of distinct sub-cellular location of transglutaminase type II: changes in intracellular distribution in physiological and pathological states. *Cell Tissue Res.* **2014**, *358* (3), 793–805.

(27) Griffith, J. S. Self-replication and scrapie. *Nature* **1967**, *215* (5105), 1043–1044.

(28) Prusiner, S. B. Novel proteinaceous infectious particles cause scrapie. *Science* **1982**, *216* (4542), 136–144.

(29) Alberti, S.; Halfmann, R.; King, O.; Kapila, A.; Lindquist, S. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell* **2009**, *137* (1), 146–158.

(30) Anderson, L. Six decades searching for meaning in the proteome. *J. Proteomics* **2014**, *107*, 24–30.

(31) Malmström, J.; Beck, M.; Schmidt, A.; Lange, V.; Deutsch, E. W.; Aebersold, R. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **2009**, *460* (7256), 762–765.

(32) Chawade, A.; Lindlöf, A.; Olsson, B.; Olsson, O. Global Expression Profiling of Low Temperature Induced Genes in the Chilling Tolerant Japonica Rice Jumli Marshi. *PLoS ONE* **2013**, *8* (12), e81729.

(33) Unlü, M.; Morgan, M. E.; Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **1997**, *18* (11), 2071–2077.

(34) Arentz, G.; Weiland, F.; Oehler, M. K.; Hoffmann, P. State of the art of 2D DIGE. *Proteomics Clin. Appl.* **2015**, *9* (3-4), 277–288.

(35) Borrebaeck, C. A. . Antibodies in diagnostics – from immunoassays to protein chips. *Immunol. Today* **2000**, *21* (8), 379–382.

(36) Sykes, K. F.; Legutki, J. B.; Stafford, P. Immunosignaturing: a critical review. *Trends Biotechnol.* **2013**, *31* (1), 45–51.

(37) Domon, B.; Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **2010**, *28* (7), 710–721.

(38) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012**, *404* (4), 939–965.

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

(40) Karas, M.; Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **1988**, *60* (20), 2299–2301.

(41) de Hoffmann, E. Mass Spectrometry. In *Kirk-Othmer Encyclopedia of Chemical Technology*; John Wiley & Sons, Inc., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005.

(42) Paul, W.; Steinwedel, H. Apparatus for separating charged particles of different specific charges. US2939952, 1960.

(43) Miller, P. E.; Denton, M. B. The quadrupole mass filter: Basic operating concepts. *J. Chem. Educ.* **1986**, *63* (7), 617.

(44) Paul, W.; Steinwedel, H. A New Mass Spectrometer Without a Magnetic Field. *Zeitschrift fuer Naturforschung* **1953**.

(45) Paul, W.; Steinwedel, H. Separation and indication of ions with different specific charges. DE 944900 19560628, 1956.

(46) Stephens, W. E. A pulsed mass spectrometer with time dispersion. *Phys. Rev.* **1946**, *69* (11-1).

(47) Hipple, J. A.; Thomas, H. A. A Time-of-Flight Mass Spectrometer with Varying Field. *Phys. Rev.* **1949**, *75* (10), 1616–1616.

(48) Comisarow, M. B. Resolution-enhanced Fourier transform ion cyclotron resonance spectroscopy. *J. Chem. Phys.* **1975**, *62* (1), 293.

(49) Makarov, null. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **2000**, *72* (6), 1156–1162.

(50) Hardman, M.; Makarov, A. A. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal. Chem.* **2003**, *75* (7), 1699–1705.

(51) Hunt, D. F.; Buko, A. M.; Ballard, J. M.; Shabanowitz, J.; Giordani, A. B. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biol. Mass Spectrom.* **1981**, *8* (9), 397–408.

(52) Hunt, D. F.; Bone, W. M.; Shabanowitz, J.; Rhodes, J.; Ballard, J. M. Sequence analysis of oligopeptides by secondary ion/collision activated dissociation mass spectrometry. *Anal. Chem.* **1981**, *53* (11), 1704–1706.

(53) Hunt, D. F.; 3rd, J. R. Y.; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci.* **1986**, *83* (17), 6233–6237.

(54) Röst, H.; Malmström, L.; Aebersold, R. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol. Cell. Proteomics MCP* **2012**, *11* (8), 540–549.

(55) Roepstorff, P.; Fohlman, J. Proposal for a common nomeclature for sequence ions in mass-spectra of peptides. *Biol. Mass Spectrom.* **1984**, *11* (11), 601–601.

(56) Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biol. Mass Spectrom.* **1988**, *16* (1-12), 99–111.

(57) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399–1406.

(58) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.

(59) Gucinski, A. C.; Dodds, E. D.; Li, W.; Wysocki, V. H. Understanding and Exploiting Peptide Fragment Ion Intensities Using Experimental and Informatic Approaches. In *Proteome Bioinformatics*; Hubbard, S. J., Jones, A. R., Eds.; Humana Press: Totowa, NJ, 2010; Vol. 604, pp 73–94.

(60) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908–3922.

(61) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77* (19), 6364–6373.

(62) Jones, J. L.; Dongre, A. R.; Somogyi, A.; Wysocki, V. H. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *J. Am. Chem. Soc.* **1994**, *116* (18), 8368–8369.

(63) Barsnes, H.; Eidhammer, I.; Martens, L. A global analysis of peptide fragmentation variability. *Proteomics* **2011**, *11* (6), 1181–1188.

(64) Toprak, U. H.; Gillet, L. C.; Maiolica, A.; Navarro, P.; Leitner, A.; Aebersold, R. Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol. Cell. Proteomics* **2014**.

(65) Cooks, R. G. Special feature: Historical. Collision-induced dissociation: Readings and commentary. *J. Mass Spectrom.* **1995**, *30* (9), 1215–1221.

(66) Mitchell Wells, J.; McLuckey, S. A. Collision-Induced Dissociation (CID) of Peptides and Proteins. In *Methods in Enzymology*; Elsevier, 2005; Vol. 402, pp 148–185.

(67) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709–712.

(68) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **1998**, *120* (13), 3265–3266.

(69) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci.* **2004**, *101* (26), 9528–9533.

(70) Tabb, D. L.; Saraf, A.; Yates, J. R. GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Anal. Chem.* **2003**, *75* (23), 6415–6421.

(71) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem.* **2005**, *77* (18), 5800–5813.

(72) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. *Mol. Cell. Proteomics* **2007**, *6* (11), 1942–1951.

(73) Li, W.; Ji, L.; Goya, J.; Tan, G.; Wysocki, V. H. SQID: An Intensity-Incorporated Protein Identification Algorithm for Tandem Mass Spectrometry. *J. Proteome Res.* **2011**, *10* (4), 1593–1602.

(74) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–682.

(75) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999**, *17* (10), 994–999.

(76) Andrews, G. L.; Simons, B. L.; Young, J. B.; Hawkridge, A. M.; Muddiman, D. C. Performance Characteristics of a New Hybrid Quadrupole Time-of-Flight Tandem Mass Spectrometer (TripleTOF 5600). *Anal. Chem.* **2011**, *83* (13), 5442–5446.

(77) Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; et al. Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Anal. Chem.* **2013**, *85* (24), 11710–11714.

(78) Kuhn, E.; Wu, J.; Karl, J.; Liao, H.; Zolg, W.; Guild, B. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and13C-labeled peptide standards. *PROTEOMICS* **2004**, *4* (4), 1175–1186.

(79) Anderson, L.; Hunter, C. L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics MCP* **2006**, *5* (4), 573–588.

(80) Wolf-Yadlin, A.; Hautaniemi, S.; Lauffenburger, D. A.; White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci.* **2007**, *104* (14), 5860–5865.

(81) Gallien, S.; Duriez, E.; Demeure, K.; Domon, B. Selectivity of LC-MS/MS analysis: Implication for proteomics experiments. *J. Proteomics* **2012**.

(82) Purvine, S.; Eppel, J.-T.; Yi, E. C.; Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **2003**, *3* (6), 847–850.

(83) Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **2004**, *1* (1), 39–45.

(84) Plumb, R. S.; Johnson, K. A.; Rainville, P.; Smith, B. W.; Wilson, I. D.; Castro-Perez, J. M.; Nicholson, J. K. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom. RCM* **2006**, *20* (13), 1989–1994.

(85) Panchaud, A.; Scherl, A.; Shaffer, S. A.; von Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **2009**, *81* (15), 6481–6488.

(86) Weisbrod, C. R.; Eng, J. K.; Hoopmann, M. R.; Baker, T.; Bruce, J. E. Accurate Peptide Fragment Mass Analysis: Multiplexed Peptide Identification and Quantification. *J. Proteome Res.* **2012**, *11* (3), 1621–1632.

(87) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics MCP* **2012**, *11* (6), O111.016717.

(88) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L.-Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics MCP* **2015**, *14* (5), 1400–1410.

(89) Villars, R. L.; Olofson, C. W.; Eastwood, M. Big data: What it is and why you should care. IDC 2011.

(90) Hashem, I. A. T.; Yaqoob, I.; Anuar, N. B.; Mokhtar, S.; Gani, A.; Ullah Khan, S. The rise of “big data” on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115.

(91) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **2014**, *1*, 140031.

(92) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.

(93) Malmström, E.; Kilsgård, O.; Hauri, S.; Smeds, E.; Herwald, H.; Malmström, L.; Malmström, J. Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat. Commun.* **2016**, *7*, 10261.

(94) Michalski, A.; Cox, J.; Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793.

(95) Johnson, J. M.; Edwards, S.; Shoemaker, D.; Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **2005**, *21* (2), 93–102.

(96) Stamm, S.; Ben-Ari, S.; Rafalska, I.; Tang, Y.; Zhang, Z.; Toiber, D.; Thanaraj, T. A.; Soreq, H. Function of alternative splicing. *Gene* **2005**, *344*, 1–20.

(97) Engelson, V.; Fritzson, D.; Fritzson, P. Lossless Compression of High-volume Numerical Data from Simulations. *Data Compression Conf.* **2000**, 574–586.

(98) Ratanaworabhan, P.; Ke, J.; Burtscher, M. Fast lossless compression of scientific floating-point data. *Data Compression Conf. 2006 DCC 2006 Proc.* **2006**, *1*, 133–142.

(99) King, R.; Bonfiglio, R.; Fernandez-Metzler, C.; Miller-Stein, C.; Olah, T. Mechanistic investigation of ionization suppression in electrospray ionization. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (11), 942–950.

(100) Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clin. Chem.* **2003**, *49* (7), 1041–1044.

(101) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics MCP* **2013**.

(102) Ghaemmaghami, S.; Huh, W.-K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. Global analysis of protein expression in yeast. *Nature* **2003**, *425* (6959), 737–741.

(103) Wang, C. P.; Isenhour, T. L. Time-warping algorithm applied to chromatographic peak matching gas chromatography/Fourier transform infrared/mass spectrometry. *Anal. Chem.* **1987**, *59* (4), 649–654.

(104) Prince, J. T.; Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal. Chem.* **2006**, *78* (17), 6140–6152.

(105) Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008**, *9* (1), 375.

(106) McCalley, D. V. Rationalization of Retention and Overloading Behavior of Basic Compounds in Reversed-Phase HPLC Using Low Ionic Strength Buffers Suitable for Mass Spectrometric Detection. *Anal. Chem.* **2003**, *75* (14), 3404–3410.

(107) Cech, N. B.; Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **2001**, *20* (6), 362–387.

(108) Mei, H.; Hsieh, Y.; Nardo, C.; Xu, X.; Wang, S.; Ng, K.; Korfmacher, W. A. Investigation of matrix effects in bioanalytical high-performance liquid chromatography/tandem mass spectrometric assays: application to drug discovery. *Rapid Commun. Mass Spectrom. RCM* **2003**, *17* (1), 97–103.

(109) Mallet, C. R.; Lu, Z.; Mazzeo, J. R. A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts. *Rapid Commun. Mass Spectrom.* **2004**, *18* (1), 49–58.

(110) Tang, K.; Page, J. S.; Smith, R. D. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (10), 1416–1423.

(111) Furey, A.; Moriarty, M.; Bane, V.; Kinsella, B.; Lehane, M. Ion suppression; a critical review on causes, evaluation, prevention and applications. *Talanta* **2013**, *115*, 104–122.

(112) Yost, R. A.; Enke, C. G. Triple Quadrupole Mass Spectrometry. *Anal. Chem.* **1979**, *51* (12), 1251A – 1264A.

(113) Zubarev, R. A.; Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **2013**, *85* (11), 5288–5296.

(114) Maclean, B.; Tomazela, D. M.; Abbatiello, S. E.; Zhang, S.; Whiteaker, J. R.; Paulovich, A. G.; Carr, S. A.; Maccoss, M. J. Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry. *Anal. Chem.* **2010**, *82* (24), 10116–10124.

(115) Yadav, M.; Patel, D.; Singhal, P.; Prasad, R.; Goswami, S.; Shrivastav, P. S.; Pande, U. C. Effect of collision-activated dissociation gas and collision energy on the fragmentation of dipyridamole and its rapid and sensitive liquid chromatography/electrospray ionization tandem mass spectrometric determination in human plasma. *Rapid Commun. Mass Spectrom. RCM* **2008**, *22* (4), 511–518.

(116) Prakash, A.; Tomazela, D. M.; Frewen, B.; Maclean, B.; Merrihew, G.; Peterman, S.; Maccoss, M. J. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *J. Proteome Res.* **2009**, *8* (6), 2733–2739.

(117) Picotti, P.; Rinner, O.; Stallmach, R.; Dautel, F.; Farrah, T.; Domon, B.; Wenschuh, H.; Aebersold, R. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* **2010**, *7* (1), 43–46.

(118) Hüttenhain, R.; Soste, M.; Selevsek, N.; Röst, H.; Sethi, A.; Carapito, C.; Farrah, T.; Deutsch, E. W.; Kusebauch, U.; Moritz, R. L.; et al. Reproducible

quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci. Transl. Med.* **2012**, *4* (142), 142ra94.

(119) Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Innovation: Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (7), 577–583.

(120) Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinforma. Oxf. Engl.* **2006**, *22* (14), e481–e488.

(121) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125–131.

(122) Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.* **2009**, *27* (2), 190–198.

(123) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteomics MCP* **2011**.

(124) Qeli, E.; Omasits, U.; Goetze, S.; Stekhoven, D. J.; Frey, J. E.; Basler, K.; Wollscheid, B.; Brunner, E.; Ahrens, C. H. Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *J. Proteomics* **2014**, *108*, 269–283.

(125) Searle, B. C.; Egertson, J. D.; Bollinger, J. G.; Stergachis, A. B.; MacCoss, M. J. Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. *Mol. Cell. Proteomics MCP* **2015**, *14* (9), 2331–2340.

(126) Picotti, P.; Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **2012**, *9* (6), 555–566.

(127) Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **2015**, *10* (3), 426–441.

(128) Malmström, L.; Malmström, J.; Selevsek, N.; Rosenberger, G.; Aebersold, R. Automated workflow for large-scale selected reaction monitoring experiments. *J. Proteome Res.* **2012**, *11* (3), 1644–1653.

(129) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **2012**, *12* (8), 1111–1121.

(130) Panse, C.; Trachsel, C.; Grossmann, J.; Schlapbach, R. specL--an R/Bioconductor package to prepare peptide spectrum matches for use in targeted proteomics. *Bioinformatics* **2015**, *31* (13), 2228–2231.

(131) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **1993**, *195* (1), 58–64.

(132) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (11), 5011–5015.

(133) Mann, M.; Højrup, P.; Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **1993**, *22* (6), 338–345.

(134) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol. CB* **1993**, *3* (6), 327–332.

(135) Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **1993**, *214* (2), 397–408.

(136) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(137) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7* (1), 40–44.

(138) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (4), 320–332.

(139) Jarvis, R. M.; Goodacre, R. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinforma. Oxf. Engl.* **2005**, *21* (7), 860–868.

(140) Listgarten, J.; Emili, A. Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (4), 419–434.

(141) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* **2007**, *79* (15), 5620–5632.

(142) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. OpenMS – An open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008**, *9* (1), 163.

(143) Jaitly, N.; Mayampurath, A.; Littlefield, K.; Adkins, J. N.; Anderson, G. A.; Smith, R. D. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics* **2009**, *10*, 87.

(144) Gorshkov, V.; Verano-Braga, T.; Kjeldsen, F. SuperQuant: A Data Processing Approach to Increase Quantitative Proteome Coverage. *Anal. Chem.* **2015**, *87* (12), 6319–6327.

(145) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–1372.

(146) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627–1639.

(147) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (4), 229–233.

(148) Reiter, L.; Rinner, O.; Picotti, P.; Hüttenhain, R.; Beck, M.; Brusniak, M.-Y.; Hengartner, M. O.; Aebersold, R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **2011**, *8*, 430–435.

(149) Wenger, C. D.; Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J. Proteome Res.* **2013**, *12* (3), 1377–1386.

(150) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

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

(152) Häkkinen, J.; Vincic, G.; Månsson, O.; Wårell, K.; Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.* **2009**, *8* (6), 3037–3043.

(153) Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **2009**, *9* (5), 1220–1229.

(154) Nahnsen, S.; Bertsch, A.; Rahnenführer, J.; Nordheim, A.; Kohlbacher, O. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *J. Proteome Res.* **2011**, *10* (8), 3332–3343.

(155) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics MCP* **2013**, *12* (9), 2383–2393.

(156) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(157) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **2003**, *100* (16), 9440–9445.

(158) Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2002**, *64* (3), 479–498.

(159) Hendrickson, E. L.; Xia, Q.; Wang, T.; Leigh, J. A.; Hackett, M. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *The Analyst* **2006**, *131* (12), 1335.

(160) Mueller, L. N.; Brusniak, M.-Y.; Mani, D. R.; Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **2008**, *7* (1), 51–61.

(161) Fenselau, C. A review of quantitative methods for proteomic studies. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2007**, *855* (1), 14–20.

(162) Li, Z.; Adams, R. M.; Chourey, K.; Hurst, G. B.; Hettich, R. L.; Pan, C. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *J. Proteome Res.* **2012**, *11* (3), 1582–1590.

(163) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895–1904.

(164) Ono, M.; Shitashige, M.; Honda, K.; Isobe, T.; Kuwabara, H.; Matsuzuki, H.; Hirohashi, S.; Yamada, T. Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics MCP* **2006**, *5* (7), 1338–1347.

(165) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics MCP* **2002**, *1* (5), 376–386.

(166) America, A. H. P.; Cordewener, J. H. G. Comparative LC-MS: a landscape of peaks and valleys. *Proteomics* **2008**, *8* (4), 731–749.

(167) Nanni, P.; Levander, F.; Roda, G.; Caponi, A.; James, P.; Roda, A. A label-free nano-liquid chromatography–mass spectrometry approach for quantitative serum peptidomics in Crohn’s disease patients. *J. Chromatogr. B* **2009**, *877* (27), 3127–3136.

(168) Waldemarson, S.; Krogh, M.; Alaiya, A.; Kirik, U.; Schedvins, K.; Auer, G.; Hansson, K. M.; Ossola, R.; Aebersold, R.; Lee, H.; et al. Protein expression changes in ovarian cancer during the transition from benign to malignant. *J. Proteome Res.* **2012**, *11* (5), 2876–2889.

(169) Smith, R.; Ventura, D.; Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinform.* **2015**, *16* (1), 104–117.

(170) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinforma. Oxf. Engl.* **2010**, *26* (7), 966–968.

(171) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32* (3), 219–223.

(172) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **2015**, *12* (3), 258–264, 7 p following 264.

(173) Luethy, R.; Kessner, D. E.; Katz, J. E.; MacLean, B.; Grothe, R.; Kani, K.; Faça, V.; Pitteri, S.; Hanash, S.; Agus, D. B.; et al. Precursor-Ion Mass Re-Estimation Improves Peptide Identification on Hybrid Instruments. *J. Proteome Res.* **2008**, *7* (9), 4031–4039.

(174) Niu, M.; Mao, X.; Ying, W.; Qin, W.; Zhang, Y.; Qian, X. Determination of monoisotopic masses of chimera spectra from high-resolution mass spectrometric data by use of isotopic peak intensity ratio modeling. *Rapid Commun. Mass Spectrom. RCM* **2012**, *26* (16), 1875–1886.

(175) Zhang, B.; Pirmoradian, M.; Chernobrovkin, A.; Zubarev, R. A. DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. *Mol. Cell. Proteomics MCP* **2014**, *13* (11), 3211–3223.

(176) Shteynberg, D.; Mendoza, L.; Hoopmann, M. R.; Sun, Z.; Schmidt, F.; Deutsch, E. W.; Moritz, R. L. reSpect: Software for Identification of High and Low Abundance Ion Species in Chimeric Tandem Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2015**.

(177) Smith, R.; Mathis, A. D.; Ventura, D.; Prince, J. T. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics* **2014**, *15 Suppl 7*, S9.

(178) Smith, R.; Ventura, D.; Prince, J. T. Novel algorithms and the benefits of comparative validation. *Bioinformatics* **2013**, *29* (12), 1583–1585.

(179) Krueger, C. W. Software reuse. *ACM Comput. Surv.* **1992**, *24* (2), 131–183.

(180) Lim, W. C. Effects of reuse on quality, productivity, and economics. *IEEE Softw.* **1994**, *11* (5), 23–30.

(181) Shaw, M. Architectural issues in software reuse: it's not just the functionality, it's the packaging. *ACM SIGSOFT Softw. Eng. Notes* **1995**, *20* (SI), 3–6.

(182) Martin, R. C. *Agile software development: principles, patterns, and practices*; Alan Apt series; Prentice Hall: Upper Saddle River, N.J, 2003.

(183) Knuth, D. E. The errors of tex. *Softw. Pract. Exp.* **1989**, *19* (7), 607–685.

(184) Holzinger, A. Usability engineering methods for software developers. *Commun. ACM* **2005**, *48* (1), 71–74.

(185) Brown, P. J. Error messages: the neglected area of the man/machine interface. *Commun. ACM* **1983**, *26* (4), 246–249.

(186) Knuth, D. E. Structured Programming with go to Statements. *ACM Comput. Surv.* **1974**, *6* (4), 261–301.

(187) Haller, P.; Odersky, M. Scala Actors: Unifying thread-based and event-based programming. *Theor. Comput. Sci.* **2009**, *410* (2-3), 202–220.

(188) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics MCP* **2005**, *4* (10), 1419–1440.

(189) Duncan, M. W.; Aebersold, R.; Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **2010**, *28* (7), 659–664.

(190) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480* (7376), 254–258.

(191) Mann, M.; Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994**, *66* (24), 4390–4399.

(192) Bahr, U.; Pfenninger, A.; Karas, M.; Stahl, B. High-Sensitivity Analysis of Neutral Underivatized Oligosaccharides by Nanoelectrospray Mass Spectrometry. *Anal. Chem.* **1997**, *69* (22), 4530–4535.

(193) Kelly, R. T.; Page, J. S.; Tang, K.; Smith, R. D. Array of Chemically Etched Fused-Silica Emitters for Improving the Sensitivity and Quantitation of Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **2007**, *79* (11), 4192–4198.

(194) Stergachis, A. B.; MacLean, B.; Lee, K.; Stamatoyannopoulos, J. A.; MacCoss, M. J. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat. Methods* **2011**, *8* (12), 1041–1043.

(195) Muntel, J.; Boswell, S. A.; Tang, S.; Ahmed, S.; Wapinski, I.; Foley, G.; Steen, H.; Springer, M. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). *Mol. Cell. Proteomics MCP* **2015**, *14* (2), 430–440.

(196) Schmidt, A.; Karas, M.; Dülcks, T. Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI? *J. Am. Soc. Mass Spectrom.* **2003**, *14* (5), 492–500.

(197) Iribarne, J. V. On the evaporation of small ions from charged droplets. *J. Chem. Phys.* **1976**, *64* (6), 2287.

(198) Taflin, D. C.; Ward, T. L.; Davis, E. J. Electrified droplet fission and the Rayleigh limit. *Langmuir* **1989**, *5* (2), 376–384.

(199) Kebarle, P.; Tang, L. From ions in solution to ions in the gas phase - the mechanism of electrospray mass spectrometry. *Anal. Chem.* **1993**, *65* (22), 972A – 986A.

(200) Li, K.-Y.; Tu, H.; Ray, A. K. Charge limits on droplets during evaporation. *Langmuir ACS J. Surf. Colloids* **2005**, *21* (9), 3786–3794.

(201) Page, J. S.; Kelly, R. T.; Tang, K.; Smith, R. D. Ionization and transmission efficiency in an electrospray ionization—mass spectrometry interface. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (9), 1582–1590.

(202) Enke, C. G. A predictive model for matrix and analyte effects in electrospray ionization of singly-charged ionic analytes. *Anal. Chem.* **1997**, *69* (23), 4885–4893.

(203) Cech, N. B.; Enke, C. G. Effect of Affinity for Droplet Surfaces on the Fraction of Analyte Molecules Charged during Electrospray Droplet Fission. *Anal. Chem.* **2001**, *73* (19), 4632–4639.

(204) Cech, N. B.; Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **2000**, *72* (13), 2717–2723.

(205) Marginean, I.; Page, J. S.; Tolmachev, A. V.; Tang, K.; Smith, R. D. Achieving 50% ionization efficiency in subambient pressure ionization with nanoelectrospray. *Anal. Chem.* **2010**, *82* (22), 9344–9349.

(206) Canterbury, J. D.; Merrihew, G. E.; MacCoss, M. J.; Goodlett, D. R.; Shaffer, S. A. Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (12), 2048–2059.

(207) Park, S.-G.; Anderson, G. A.; Navare, A. T.; Bruce, J. E. Parallel Spectral Acquisition with an Ion Cyclotron Resonance Cell Array. *Anal. Chem.* **2016**, *88* (2), 1162–1168.

(208) Antberg, L.; Cifani, P.; Levander, F.; James, P. Pathway-centric analysis of the DNA damage response to chemotherapeutic agents in two breast cell lines. *EuPA Open Proteomics* **2015**, *8*, 128–136.

(209) Stella, R.; Biancotto, G.; Arrigoni, G.; Barrucci, F.; Angeletti, R.; James, P. Proteomics for the detection of indirect markers of steroids treatment in bovine muscle. *PROTEOMICS* **2015**, *15* (13), 2332–2341.

(210) Maher, S.; Jjunju, F. P. M.; Taylor, S. Colloquium: 100 years of mass spectrometry: Perspectives and future trends. *Rev. Mod. Phys.* **2015**, *87* (1), 113–135.