TECHNISCHE UNIVERSITÄT MÜNCHEN
Fachgebiet für Bioinformatik

Computational approaches to enhance mass spectrometry-based proteomics

Nadin Neuhauser

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technische Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzende:		Univ.-Prof. Dr. I. Antes
Prüfer der Disseration:	1.	Univ.-Prof. Dr. D. Frischmann
	2.	Hon.-Prof. Dr. M. Mann Ludwig-Maximilians-Universität München

Die Disseration wurde am 07.03.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 25.04.2013 angenommen.

Table of Contents

Abbreviations **iii**

Summary **v**

Zusammenfassung **vii**

1 Introduction **1**

1.1 Historical Background 1
1.2 Mass spectrometry 3
1.3 Shotgun proteomics 8
1.4 Computational analysis 13
1.5 MaxQuant - Software environment 24

2 Results **29**

2.1 article 1 - Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment 29
2.2 article 2 - Expert System for Computer Assisted Annotation of MS/MS Spectra 44
2.3 article 3 - A Systematic Investigation into the Nature of Tryptic HCD Spectra 56
2.4 article 4 - High performance computational analysis of large-scale proteome datasets to assess incremental contribution to coverage of the human genome 71

3 Conclusions and outlook **97**

4 References **101**

Acknowledgments **113**

Curriculum Vitae **115**

Abbreviations

CID	collision induced dissociation
CPU	central processing unit
DNA	deoxyribonucleic acid
EBI	european Bioinformatics Institute
ESI	electrospray ionization
ETD	electron transfer dissociation
FDR	false discovery rate
FFT	fast fourier transform
FT	fourier transform
FWHM	full width at half maximum
GPFS	general parallel file system
HCD	higher energy collisional dissociation
HPC	high performance computing
HPLC	high pressure liquid chromatography
I/O	input and output
ICAT	isotope coded affinity tags
ICR	ion cyclotron resonance
IPI	international protein index
IT	ion trap
iTRAQ	isobaric tags for relative and absolute quantification
IUPAC	international union of pure and applied chemistry
LC-MS	liquid chromatography coupled to mass spectrometry
LIT	linear ion trap
LTQ	linear trap quadrupole
m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption ionization
mRNA	messenger ribonucleic acid
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NCBI	National Center for Biotechnology Information

PEP ..... posteriori error probability
PIF ..... precursor intensity fraction
ppm ..... parts per million
PSM ..... peptide spectrum match
PTM ..... post-translational modification
Q ..... quadrupole
RefSeq ..... Reference Sequence
RF ..... radio frequency
SIB ..... Swiss Institute of Bioinformatics
SILAC ..... stable isotope labeling by amino acids in cell culture
SSD ..... solid state disk
TIC ..... total ion current
TOF ..... time-of-flight
TQ ..... triple quadrupoles

Summary

Mass spectrometry based proteomics has now matured into a technology that enables routine identification and quantification of the expressed proteome and their post-translational modifications to an unprecedented depth. Due to the massive amount of data being generated in a relatively short time, data analysis poses next challenge in the field demanding parallel development of an efficient computational platform. The overall aim of the projects described in this thesis is to reduce the time taken for such computational analysis and to improve the interpretation of shotgun proteomics experiments using high-confidence analysis.

A central step in this endeavor was the development and implementation of our novel peptide search engine termed Andromeda (article 2.1). Integrated in the existing data analysis pipeline MaxQuant, Andromeda now enables high confidence identification of peptides using the raw data acquired on state of the art mass spectrometry instruments. Since the search engine is developed in-house it provides flexibility to tailor peptide identification to innovations not only in the protein quantification, but also in instrumentation and hardware. For instance, it allowed us to use the recent improvement in mass accuracy for better confidence in the identification of peptides. Additionally, we are now able to accommodate large databases and to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides. Using several large-scale datasets we proved that our search engine is at least comparable with the commercially available software such as the widely used Mascot program. In contrast to existing software Andromeda is freely available and runs also on a normal desktop computer.

The developments in the mass spectrometry instrumentation have led to generation of thousands of MS/MS spectra per hour and reliable interpretation of these spectra can be achieved only by automated analysis using sophisticated software. Apart from the regular fragment series explained by most peptide search engines, a MS/MS spectrum contains plenty of unassigned peaks resulting from various fragmentation types. To increase the confidence in the peptide identifications of Andromeda, we apply domain knowledge in the interpretation of peptide fragment spectra using a computer-based

'Expert System' (article 2.2). The goal of this approach was to find at least for all high abundant peaks a valid explanation. This bioinformatics application was designed for biologists to help them understand the complex mechanisms involved in peptide fragmentation by manual inspection of their data. This rule-based system represents a combination of theoretical understanding and a collection of heuristic problem-solving strategies that experience has shown to be effective. To estimate the risk of false annotations, we calculated a false discovery rate (FDR) for the used set of rules. With this expert system we were able for the first time to statistically verify, based on thousands of fragment spectra, the peptide fragments obtained by higher energy induced collisional dissociation (HCD) a new fragmentation method(article 2.3).

For large-scale projects in proteomics, such as clinical studies, the amount of time needed for analysis can be an essential bottleneck. This project was focused on reducing the processing time of MaxQuant by enhancements on the software and hardware side (article 2.4). Various sections in the MaxQuant pipeline underwent refactoring that otherwise performed poorly and are now executed in parallel. This parallelization has a dramatic effect when the hardware provides multiple central processing units (CPUs) such as on a computer cluster. Surprisingly, a hardware configuration which is optimized for high input and output (I/O) workload is equally efficient in faster data processing when compared to a computer cluster with high number of processors. This investigation uncovered important principles in computational analysis required towards our final aim of complete coverage of the human proteome.

Together, the projects presented in this thesis provide a substantial advancement in computational proteomics, which in turn will advance the proteomics workflow and accelerate biological and biomedical discoveries.

Zusammenfassung

Massenspektrometrie basierte Proteomforschung hat sich zu eine Technologie entwickelt, die Identifizierung und Quantifizierung des momentanen Proteomes und desen Modifikationenen, in noch nie da gewesener weise routinemäßig ermöglicht. In diesem Feld, stellt die innerhalb kurzer Zeit generierten und zunehmend größer werdende Datenmengen, eine neue Herausforderung für die Datenverarbeitung dar, welche die Entwicklung einer effizienten Computerplattform notwendig macht. Das große Ziel der Projekte die in dieser Arbeit beschrieben werden, ist die Zeit für die Datenauswertung zu reduzieren und die Interpretation von „Shotgun"-Experimenten in der Proteomforschung durch glaubwürdige Analyse zu optimieren.

Ein zentraler Schritt in diesen Bemühungen, war die Entwicklung und Implementierung unser neuartigen Peptidsuchmaschine Andromeda (Artikel 2.1). Eingebunden in unsere Analysepipeline MaxQuant erlaubt Andromeda die zuverlässige Identifizierung von Peptiden durch Rohdaten von modernen Massenspektrometern. Dadurch das die Suchmaschine von uns entwickelt wird, gibt es uns die Flexibilität die Peptideidentifikation abzustimmen, auf neue Innovationnen nicht nur im Bereich Proteinquantifizierung sondern auch auf Neuerungen in der Geräte und Computer Technologie. Zum Beispiel erlaubte es uns die neueste Verbesserung in der Massengenauigkeit für eine bessere Korrektheit in der Peptididentifizierung zu nutzen. Des Weiteren, sind wir nun auch in der Lage extrem große Datenbanken zu benutzen und komplexe Muster von post-translationalen Modifikationen zu identifizieren und zu bewerten, wie beispielsweise extrem phosphorylierte Peptide. Durch die Verwendung vieler großer Datensätze konnten wir beweisen, das unsere Suchmaschine zu mindestens vergleichbar ist mit kommerzielle erhältlicher Software, wie beispielsweise Mascot. Im Gegensatz zu bereits existierender Software ist Andromeda frei erhältlich und läuft auch auf gewöhnlichen Bürorechnern.

Die Geräteentwicklungen in der Massenspektrometie hat dazu geführt das tausende von MS/MS Spektren innerhalb einer Stunde generiert werden können und zuverlässige Interpretation dieser Spekten nur durch Automatisierung der Analyse durch fortgeschrittene Software erreicht werden kann. Ausgenommen von der regulären Frag-

mentierungsserie welche von den meisten Peptidsuchmaschinen erklärt werden kann, enthält ein MS/MS Spektrum eine Menge unerklärte Signale welche von verschiedenen Fragmentierungsarten kommen. Um die Zuverlässigkeit der Peptididentifizierung durch Andromeda zu stärken, wenden wir ein rechnerbasiertes Expertensystem für die Interpretation von Peptidfragmentspekten an (Artikel 2.2). Das Ziel dieses Ansatzes ist zu mindest für die prominenten Signale eine Erklärung zu finden. Die bioinformatische Anwendung wurde als Hilfestellung für Biologen entwickelt, um die komplexen Mechanismen die in der Peptidfragmentierung durch manuelle Inspektion Ihrer Daten zu zeigen. Das regelbasierte System repräsentiert eine Kombination von theorischen Erkenntnissen und einer Sammlung von heuristischen Lösungsansätzen, welche in der Praxis bestätigt wurden. Um die Risiken für falsche Erklärungen abzuschätzen zu können, haben wir eine Fehlerquote (FDR) für unsere Regelbasis berechnet. Mit dem Expertensystem waren wir durch die Automatisierung zum ersten mal in der Lage statistisch anhand von tausenden Fragementspekten zu überprüfen, welche Peptidfragmente in der neuen Fragmentierungsmethode „HCD“ zu finden sind (Artikel 2.3).

Für Projekte mit hohem Durchsatz in der Proteomforschung, wie beispielsweise klinische Studien, kann die Zeit für die Auswertung eine wesentliche Schwachstelle sein. Dieses Projekt hatte das Ziel, die Datenauswertung durch unserer MaxQuant Programm durch Verbesserungen auf Software- und Hardwareseite, zu beschleunigen(Artikel 2.3). Hier wir der Überarbeitungsprozess für Bereichen in der MaxQuant Programmablauf beschrieben, die vorher sehr langsam waren und nun gleichzeitig ausgeführt werden. Diese Parallelisierung hat eine bedeutenden Effekt, wenn die Hardware mehrere Prozessoren enthält, wie beispielsweise ein Computercluster. Überraschenderweise, ist Hardware welche für nur für die viele Schreib- und Leseauslastung optimierte wurde vergleichbar mit einem Computercluster der Gegensatz dazu viele Prozessoren hat. Diese Projekt eröffnete wichtige Erkenntnisse in der Datenauswertung welche notwendig waren um unserem Ziel der kompletten Vermessung des menschlichen Proteomes näher zu kommen.

Zusammengefasst, haben die Projekte die in dieser Doktorarbeit vorgestellt werden, wesentlichen zur Verbesserung der Datenverarbeitung in der Proteomforschung beigetragen, was sich als Vorteil in der Proteomforschungsablauf auswirkt und somit die Entdeckung neuer biologischer oder medizinischer Ansätze ermöglicht.

1 Introduction

1.1 Historical Background

A comprehensive understanding of complex biological systems requires the identification and functional characterization of its key components. In-depth analyses on a global systems wide scale have first been done in the field of genetics. The success of the Human Genome Project[1,2] has provided a blueprint for the gene-encoded proteins potentially active in all of the approximately 230 cell types that comprise the human body[3]. However, in the following years it quickly became clear that mapping static genomes is not sufficient to decipher the biology of the mammalian cell. A complete understanding of cellular function requires quantification of global messenger ribonucleic acid (mRNA) and protein levels as well as post-translational modifications and protein-protein interactions (chapter 1 from Hein *et al* [4]).

The first genome-wide method for expression analysis was the large-scale hybridization of mRNA to complementary sequences immobilized on chips. Despite their ubiquity and tremendous usefulness, microarrays have certain limitations. Besides the lack of reproducibility across platforms and laboratories[5], the nonquantitative nature in predicting the amount of change in the active mature protein, are problematic. Proteins are almost always the effectors of biological functions, but protein levels depend not only on the levels of the corresponding messages but also on a host of translational controls and regulated degradation[6,7]. These factors may be just as important as increased synthesis of mRNA and they cannot be measured directly by microarrays or the new deep sequencing technologies.

Proteomics is the systematic study of the many and diverse properties of proteins in a parallel manner with the aim of providing detailed descriptions of the structure, function and control of biological systems in health and disease. Advances in methods and technologies have catalyzed an expansion of the scope of biological studies from a single protein to proteome-wide measurements. The word proteomics, a chimera of the words, **prot**eins and geno**mics**, was invented by Professor Mark Wilkins in the early 1990s.

The technological basis of most current proteomics studies is biological mass spectrometry (MS), first catapulted to mainstream prominence with the development of the electrospray[8] and MALDI[9] ionization techniques[10]. This advance made biological molecules readily amenable to mass spectrometry and garnered the Chemistry Nobel Prize in 2002[11]. Fueled by substantial advances in instrumentation, sequence databases, specialized software, and innovative methodologies, the field has quickly transformed into a high-throughput analytical tool for the identification and quantification of hundreds to thousands of proteins per experiment.

Yet we still have limited knowledge regarding the majority of the approximately 20,500 protein-coding human genes discovered through the Human Genome Project[12]. As an extension of the genome project, the aim of the Human Proteome Project[3;13;14] is to map the entire human protein set, mainly using MS based proteomics. In 2008, the first complete proteome of the model organism yeast was published[15]. Most recently, high performance but robust MS instruments have further increased the power of MS-based proteomics and researchers succeeded in identifying 10,000 proteins from a human cancer cell line[16;17]. However, based on the UniProt database content, about 30% of these genes lack stringent experimental evidence at the protein level, and for many others there is very little information related to protein abundance levels, subcellular localization, and function[3]. High-throughput proteomics experiments produce large volumes of complex data that would address some of these shortfalls but this also requires sophisticated computational analyses. As such, the field of proteomics offers many challenges for data interpretation. Mass spectrometry based proteomics has undergone an immense development within the last decades[18]. Due to its high sensitivity and speed it now vastly outperforms traditional methods like Edman degradation or two-dimensional gel electrophoresis for sequencing proteins or analyzing complex protein mixtures. In terms of protein identification and quantification, MS-based proteomics has grown into a comprehensive technology, applicable even to complex protein mixtures of higher organisms.

In order to streamline the description of the methods to follow, I next introduce some important mass spectrometry concepts and terminology that will be used throughout this thesis.

1.2 Mass spectrometry

The basic principle of mass spectrometry is to generate ions from either peptides and proteins by any suitable method, to separate these ions by their mass-to-charge ratio (m/z), and to detect them qualitatively and quantitatively by their respective m/z and abundance.

Components of a mass spectrometer

A mass spectrometer consists of an *ion source*, a *mass analyzer* and a *detector* which operated under high vacuum conditions. To measure its molecular mass, a molecule must be ionized. This happens in the ion source of the mass spectrometer. However, proteins and peptides are polar, nonvolatile, and thermally unstable species that require an ionization technique that transfers an analyte into the gas phase without extensive degradation. Therefore, the *ion source* can be based either on electro-spray ionization (ESI)[19], which is appropriate for liquid samples; or on matrix assisted laser desorption ionization (MALDI)[20], which is appropriate for samples that have been mixed with a matrix and crystallized on a metallic plate. The discoveries of ESI and MALDI were the introduction of soft ionization methods that allow for proteins and peptides to be analyzed by MS.

Mass analyzers are an integral part of each instrument because they can store ions and separate them based on their mass-to-charge ratios. Ion trap (IT), Orbitrap, and ion cyclotron resonance (ICR) mass analyzers separate ions based on their m/z dependent resonance frequencies, quadrupoles (Q) use m/z stability in a radio frequency (RF) field, and time-of-flight (TOF) analyzers use flight time. Hybrid mass spectrometers have been built that combined more than one mass analyzer to answer specific needs during analysis[21].

The *detector* measures the value of an indicator quantity and thus provides data for calculating the abundances of each ion present. The detector records either the charge induced or the current produced when an ion passes by or hits a surface. Mass spectrometers do not measure mass directly, but rather the mass-to-charge ratio. Hence the measurements obtained are dependent on the charge state(s) of the molecule[22].

Each *mass analyzer* has defining properties, such as resolution, mass accuracy, scan rate, mass range, and sensitivity. The terms resolution and resolving power are derived from optical spectroscopy. Older publications around 1920-1940 always refer to R as resolving power. During the years there was some confusion about the terms and

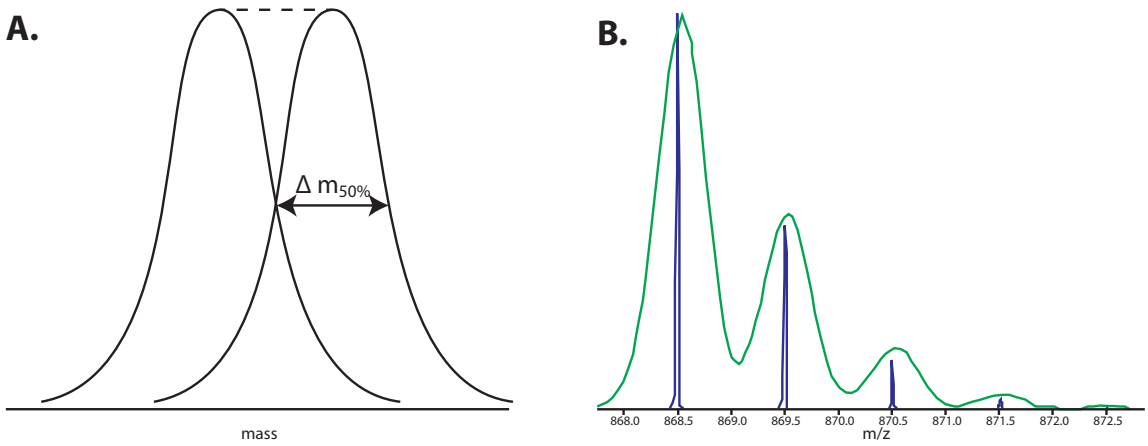


Figure 1: Mass resolving power and mass resolution. (A) Two equal-magnitude mass spectral peaks of equal width. The mass resolving power is the peak width at half-maximum peak height $\Delta m_{50\%}$ representing is the minimum mass difference to separate the two peaks. (B) The mass resolution is a performace parameter of a mass analyzer. The isotope pattern in green measured in low resolution compared to the high resolution in blue.

since 2006 they are defined by the international union of pure and applied chemistry (IUPAC). Mass resolving power $\Delta m = m_2 - m_1$ usually refers to the ability of separating two narrow mass spectral peaks (see figure 1A). Mass resolution $R = \frac{m}{\Delta m}$ is defined as the fraction of a designated mass m divided by the minimum peak width Δm necessary for separation at mass m. A specific m/z value and also the method like 10% valley or 50% valley or full width at half maximum (FWHM) generally specified. On current state of the art instruments, resolution is usually a large number (up to 2,000,000 for ICR FT-MS) and is useful for evaluating mass analyzer performance because it is a measure of quality over a wide range of m/z (see figure 1B). Very high mass resolving power enables ions of different elemental composition to be distinguished.

The mass accuracy refers to the deviation between the actual (calculated) and the experimentally determined mass of a compound and it is dependent on the resolution of the mass analyzer[23]. Mass accuracy is usually measured in parts per million (ppm), a dimensionless quantity, or in milli mass units. The ppm concept is important because experimental data often contains a linear systematic error that reaches an absolute maximum at higher masses[24]. Acquisition speed refers to the time frame of the experiment and ultimately is used to determine the number of spectra per unit time that can be generated. The mass range is the range of m/z values amenable to analysis by a given analyzer. As illustrated in figure 2 the mass precision of a peak from liquid chromatography coupled to mass spectrometry (LC-MS) is related to the number of scans. With increasing number of scans defining a eluting peak, the number of data points available for determination of the precise mass increases. This also has an direct influence on the

mass accuracy[25].

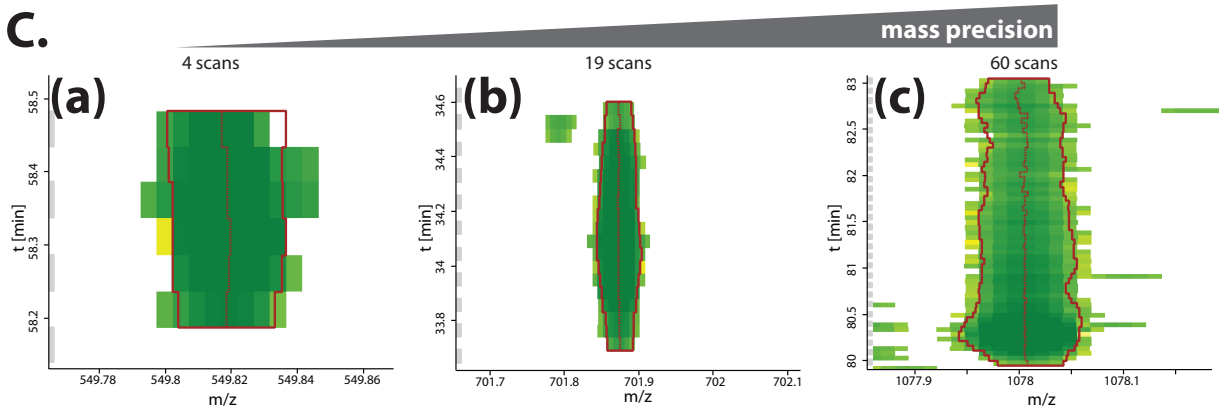


Figure 2: Mass precision Contour plots of whole elution profiles of the peak, in which the mass precision is dependent on the number of scans. The dotted line indicates the positions of the the individual peak centroids. With increasing number of scans the number of data points for determination of the precursor mass increases.

MS instruments in proteomics

The basic components of mass spectrometers are to some extent independent of each other, and as such it is possible to combine the different technological aspects to produce different types of mass spectrometers[22]. There are three broad categories of mass analyzers: the scanning and ion-beam mass spectrometers, such as TOF and quadrupoles, respectively; and the trapping mass spectrometers, such as IT, Orbitrap, and Fourier transform (FT)-ICR analyzers. The scanning mass analyzers like TOF are usually interfaced with MALDI to perform pulsed analysis, whereas the ion-beam and trapping instruments are frequently coupled to a continuous ESI ion source. The following instrument configurations are the most widely used solutions in the field of proteomics today[26]: ion traps such as the linear ion trap (LIT), triple quadrupoles (TQ), linear trap quadrupole (LTQ)-Orbitrap hybrid instrument, LTQ-FTICR, and the TQ-FTICR hybrid instruments Q-TOF and IT-TOF.

Ever since its introduction by Thermo Fisher Scientific in 2005, our laboratory has used instruments with the Orbitrap as the central element. The Orbitrap mass analyzer[27;28] features high resolution (routinely up to 150,000), high mass accuracy (from 2 to 5 ppm), a mass-to-charge range of 6000, and a dynamic range greater than 10,000[26]. When coupled to an LTQ ion trap, the hybrid instrument has the advantages of both high resolution and mass accuracy of the Orbitrap and the speed and the sensitivity of the LTQ. Similarly to ICR instruments, the Orbitrap use a fast Fourier transform

(FFT) algorithm to convert the time-domain signal at the detector into a mass-to-charge spectrum. The LTQ-Orbitrap gained rapid acceptance because it offered resolution and mass accuracy comparable to the LTQ-FTICR at a lower price tag and a lower maintenance cost, enabling for proteomic applications for a broad user base.

In proteomics tandem mass spectrometry (MS/MS or MS^2) is used to obtain sequence information by breaking the peptide at different bonds between the constituent amino acids. For example in collision induced dissociation (CID) and higher energy collision dissociation (HCD), the kinetic energy which is converted to internal energy by collision of protonated ions with a neutral gas (helium, nitrogen or argon), which results in bond breakage.

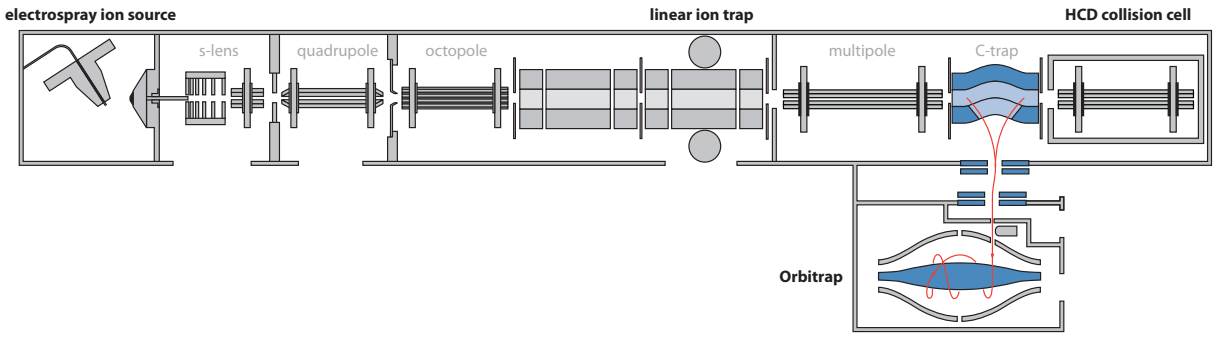


Figure 3: Schematic of the LTQ Orbitrap Velos MS instrument. This instrument is equipped with an ESI ion source and a linear ion trap that functions either only mass selection or also as the mass analyzer in case of low resolution CID MS/MS. The additional collision cell is attached to provide HCD fragmentation. Ions are transfered to the C-trap for accumulation[29] and injected as an ion package into the Orbitrap to obtain high resolution spectra. [Adapted from Olsen *et al* [30]]

The first hybrid instrument, which combined the LTQ ion trap with the Orbitrap was the LTQ Orbitrap[31]. In this instrument, the linear ion trap was used for CID fragmentation and analysis of fragment ions. The following instrument, the LTQ Velos[32] (see figure 3), had a separate collision cell. In this device the HCD fragmentation takes place and the fragment ions are then injected into the Orbitrap mass analyzer. Unlike the LTQ Orbitrap where only tandem spectra were acquired by low-resolution CID, spectra of the follower are obtained preferable in high-resolution HCD. To use the high resolution on both MS and MS^2 levels is called a 'high-high' strategy[33] and it results in higher peptide identification rates and a better confidence in these identification. An advantage of the 'low-high' strategy on LTQ Orbitrap instruments is that fragmentation events can in principle be performed in parallel with the high resolution measurements of the precursor ions in the Orbitrap analyzer.

Recently, two new Orbitrap instruments were introduced by Thermo Fisher Scientific: The Q Exactive is the first benchtop proteomics instrument[34]. In this instrument the linear ion trap is replaced by a quadrupole, in which this device is used only as a mass filter and not as mass analyzer. With this structural alteration the instrument has lost the capability for CID fragmentation. However, the acquisition time for HCD spectra decreased enormously and is now comparable in terms of speed to CID. Importantly, the instrument is much simpler and gains all the advantages in targeted acquisition that are usually associated with quadrupole instruments. The Orbitrap Elite is also a instrument of the new generation and here the size of the Orbitrap was reduced from 30 mm to 20 mm[35]. This compact Orbitrap has the advantage that the spectra can be acquired with twice the resolution than before. Both new instruments gain from a enhanced Fourier transformation algorithm, which by itself doubles resolution by including phase information[36].

1.3 Shotgun proteomics

The term 'shotgun proteomics' is the protein equivalent to shotgun sequencing in genomics in which the deoxyribonucleic acid (DNA) is sheared and sequenced, followed by alignment of small overlaps[37]. This 'bottom-up' peptide sequencing approach is the most popular and widely used technique when tackling high-complexity samples for large-scale analyses[22]. The first step after protein purification is to cleave proteins into peptides using a sequence-specific protease. A challenge in sample preparation is that not all proteins are soluble under the same conditions and many detergents interfere with the mass spectrometric analysis[38].

The mass spectrometer is most efficient at obtaining sequence information from peptides up to 20 residues long. This is one reason why peptides, and not intact proteins are more commonly measured. In addition, the sensitivity of the mass spectrometer for proteins is much lower than for peptides. One reason for this is that the combinatorial effect due to isoforms and modifications and the absence of sequence information make the identification almost unsolvable[38]. Nevertheless, in the 'top-down' approach[39–44] it is now possible to derive partial sequence information from intact proteins. Top down sequencing is used for identification purposes or the analysis of protein modifications in the context of the entire protein molecule. This technique reaches its limitation with increasing number of proteins in a sample, which make it not applicable for a large proteome.

Experimental setup

MS-based proteomics can deal with a wide variety of input materials, from prokaryote or eukaryote cells to entire tissues and body fluids. For this reason the object of investigation is nearly always a protein mixture[45–47]. These mixtures range in complexity from hundreds of proteins in affinity purifications (because of their inevitable background) to more than 10,000 different proteins in complete mammalian proteomes. The main technological goal of MS-based proteomics is the accurate characterization of as many proteins as possible in these mixtures[48]. A regular shotgun proteomics experiment proceeds in three steps, as illustrated in figure 4.

From proteins to peptides

A key step in shotgun proteomics is the digestion of proteins into peptides using a proteolytic enzyme (optionally, using multiple different enzymes).For this purpose trypsin,

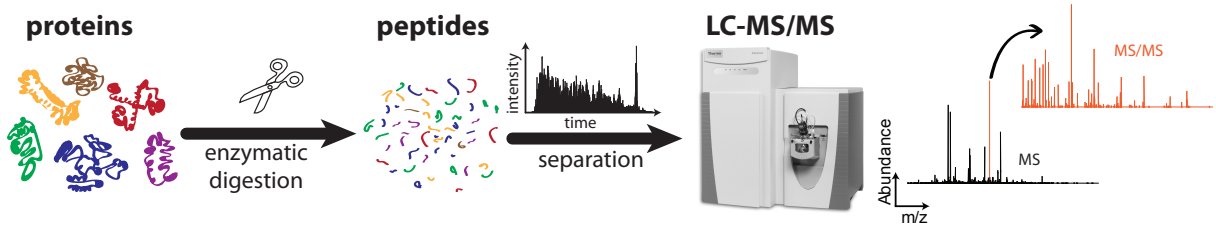


Figure 4: Overview of shotgun proteomics data production. The three steps - cleaving proteins into peptides, separation of peptides using liquid chromatography, and tandem mass spectrometric analysis - are described in the text. [Adapted from Nobel *et al* [49]]

which cleaves peptides on the carboy-terminal side of arginine and lysine residues, is most often used. Most of the resulting peptides are within the preferred mass range for sequencing and they should also have no or just a few internal trypsin sites (missed cleavages)[50]. The protein digestion step is often followed by a selective peptide enrichment or depletion strategy designed to capture peptides having certain specific properties of interest (e. g. phosphorylated peptides)[51].

Reduction of complexity

In our laboratory ESI is exclusively used as ionization method, which produces ions from a solution. To reduces the complexity at the ion source, the peptide mixture is separated by nano scale high-performance liquid chromatography (HPLC) column[38]. The HPLC uses a solvent gradient of increasing organic content to separate the peptide species based on a particular chemical property (e. g., their hydrophobicity)[49]. After separation, the eluting peptides are ionized by the ESI and proceed into the mass spectrometer. The single dimension of peptide separation that is provided by an HPLC column may not provide sufficient resolution if highly complex protein mixtures are analyzed. In this case, the probe can be divided at the protein- or at the peptide level into fractions, which produces less complex mixtures[38]. Furthermore, the analysis is thereby subdivided into several independent analysis runs, which increases confidence in database identification and it increases the dynamic range of the measurement (the difference between the most abundant and least abundant proteins can be identified in an experiment)[52].

Within the mass spectrometer

Peptides, as they elute from the reverse phase column at a particular time (retention time) are ionized and transfered into the gas phase. After ionization the peptides are

detectable for the mass spectrometer. For sequencing two rounds of mass spectrometry are performed (see figure 5).

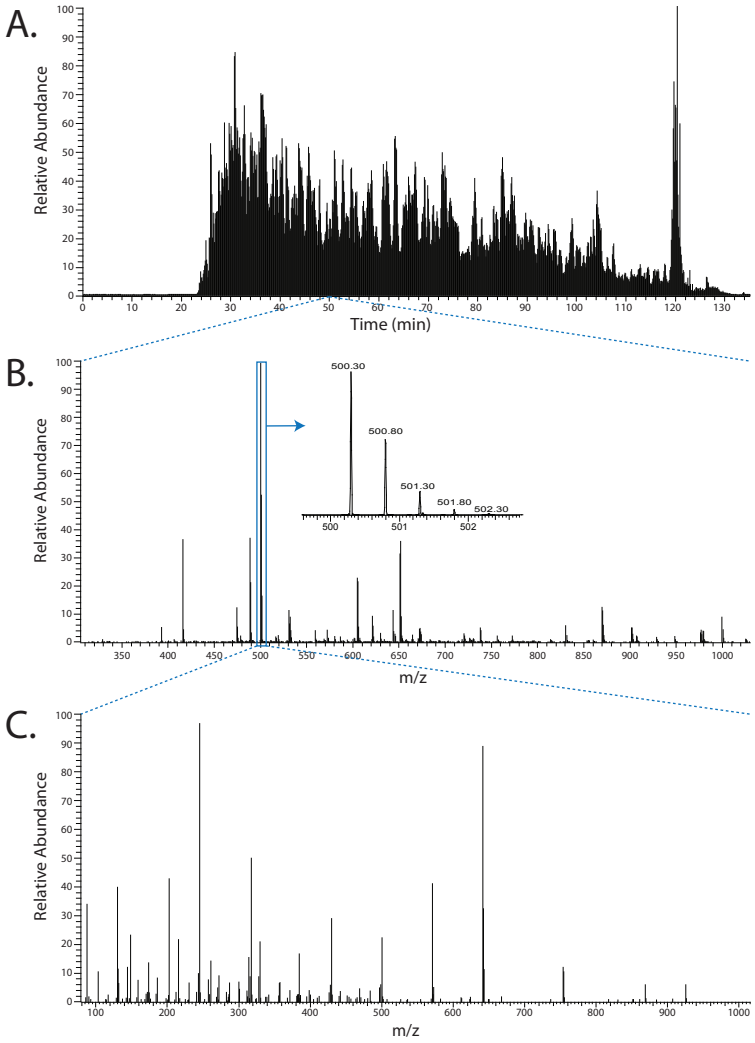


Figure 5: Tandem mass spectrometry (MS/MS). (A) Total ion current (TIC) - that is the sum of all ion intensities from all mass spectra being recorded during the LC-MS run - as a function of time. (B) Precursor-ion (MS^1) scan depicting peptide ions arising from peptides eluting and electro sprayed at a certain time point. The insert shows the isotopic distribution of a specific peptide ion. A clear separation of individual isotope masses is indicative of a high resolution instrument. (C) Fragment spectrum of the peptide ion of interest. Distinct mass increments between individual peaks allow for partial or complete deduction of the amino acid sequence [Adapted from Schwanhäuß er[53]].

In the first round of mass spectrometry, all peptide ion species eluting from the column are introduced into the instrument. The resulting full scan, also called MS^1 or survey scan, consists of mass-to-charge ratios and intensities of all peptide ions eluting at this time point. Based on the survey scan, the acquisition software of the mass spectrometer picks a preset (typically 5-20) number of peptides and proceeds to isolate each one of them. Selected peptide ions (precursor or parent ions) are broken down into smaller pieces (fragment ions) in the collision cell of the MS instrument. This is

called a data driven or data dependent topN method.

Fragmentation of peptides is usually achieved by collision with inert gas atoms, like in CID[54] or HCD[55], or chemical reactions with radicals in the gas phase as in electron transfer dissociation (ETD)[56]. Bond breakage mainly occurs through the lowest energy pathways - that is, cleavage of the amide bounds[57;58]. Note that fragments will only be detected if they carry at least one charge. Figure 6 explains how peptides fragment and how their fragment ions are designated. Briefly, the resulting ions are called *a*, *b* or *c*-ions when the charge is retained at the amino-terminal (N-terminal) fragment or *x*, *y* or *z*-ions when it is retained at the carboxy-terminal (C-terminal) fragment. This nomenclature was first proposed by Roepstorff and Fohlman[59], and subsequently modified by Johnson *et al* [60].

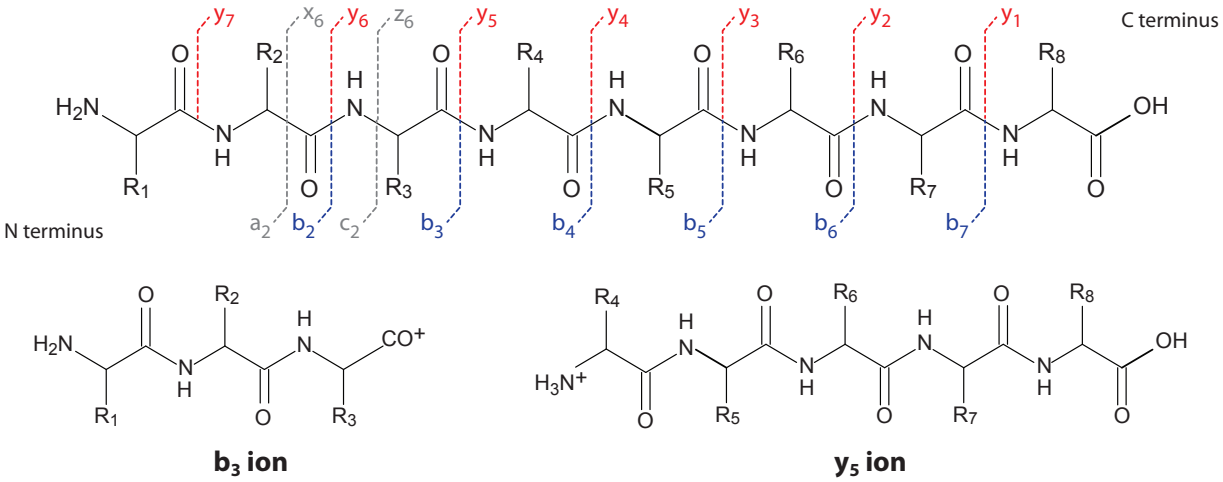


Figure 6: Peptide fragmentation. Ions are labeled consecutively from the amino terminus (N-terminus) a_m, b_m and c_m. If the ions contain the carboxyl terminus (C-terminus) they are named z_n, y_n and x_n. The subscript of the ions (n or m) indicates the number of residues in the fragment. The subscript of the R groups in the figures represents the side chain of the amino acid at that position. [Adapted from Steen *et al* [38]]

The appearance of the ion types dependents on the type of fragmentation. For instance in CID and HCD fragmentation, mostly b and y series are expected, and ETD produces predominantly c and z ions[50;61]. Fragmentation pattens are also strongly dependent on the chemical and physical properties of the amino acids and the primary structure of the peptide[62;63]. While the chemistry involved in peptide fragmentation is still not completely understood, the mobile proton model is currently the most widely accepted framework to describe the dissociation process[64;65]. Moreover, different fragmentation pathways of protonated peptides have been extensively investigated and modeled with respect to both kinetic and thermodynamic aspects[58]. The inspection of the generated fragment ions by HCD is part of my thesis and is further described in

article 2.3 on page 56. During the fragmentation process, each amino acid sequence is typically cleaved once, so cleavage of the population results in a variety of observed prefix and suffix sequences. The acquired tandem mass (MS/MS or MS^2) spectrum is a list of m/z values and intensities of all the fragment ions generated by fragmenting an isolated precursor ion. The fragmentation pattern encoded by the MS/MS spectrum allows identification of the amino acid sequence of the peptide that produced it[49].

Importantly, the mass accuracy and resolution of the MS analyzer have a significant effect on the information content of the spectrum, which is of great importance for the subsequent peptide identification step[25]. The accuracy with which an MS instrument can measure peptide ion m/z values ranges from as precise as several parts per million (ppm) in the case of high mass accuracy instruments such as the Orbitrap, to more than 500 ppm in case of low mass accuracy instruments. Similarly, the mass resolution of the instruments governs the ability to accurately determine the charge state of the peptide ion. The ability to isolate precursor ions for MS/MS sequencing within a narrow window around a particular m/z is dependent on the instrument. Even in high-resolution MS, the selection of the precursor ion for fragmentation is always performed with low resolution (typically a few Th) to ensure adequate sensitivity for MS/MS. In complex mixtures, this results in frequent co-fragmentation of co-eluting peptides with similar masses. These 'chimerical' MS/MS spectra[66] can be detrimental for identification of the peptide of interest, especially if the co-fragmented peptide is of comparable intensity. Co-fragmentation generally reduces the number of peptides identified in database searches and poses special problems for reporter fragment based quantification methods because both peptides contribute to the measured ratios[67].

Improvements in MS instrumentation have led to tremendous growth in the field of proteomics. For instance high-resolution is necessary to resolve overlapping isotopic distributions and identify the charge state. In addition, developments in accurate mass measurement dramatically improve identification confidence and limit search space leading to faster data processing. In general, if an instrument is correctly calibrated, high resolution can provide ppm mass accuracy[68].

1.4 Computational analysis

The data amounts is shotgun proteomics studies can often add up to hundreds of gigabytes. The ability to generate such large volumes of information has correspondingly increased the pressure on the downstream data processing algorithms and pipelines[69]. Indeed, the data processing was in the past a considerable bottleneck in proteomics experiments, lagging behind developments in the other areas of proteomics research.

Developments in last years have produced efficient algorithms for the different steps of informatics analysis[70], which can be subdivided into two major areas. The first part covers the evaluation of the raw mass spectrometry data up to a list of all identified proteins. In the second part, whole data sets have to be analyzed from a functional point of view, leading to biologically interpretable results. This thesis mainly deals with the first part.

Preparing data for search

The effectiveness of peptide identification algorithms is limited by the quality of the input spectra. The dominant ions in fragment spectra of peptides are often b-, y-ions and their derivatives resulting from the cleavage of the peptide bonds. However, MS/MS spectra typically contain many more peaks[71]. These result not only from isotope variants and multiply charged replicates of the peptide fragmentation products but also from unknown fragmentation pathways, chemical contaminations or from noise generated by the electronic detection system[72]. The presence of this background complicates spectrum interpretation. Consequently, an efficient preprocessing of MS/MS spectra can increase the sensitivity of peptide identification at reduced file sizes and run time without compromising its specificity.

Whereas MALDI mass spectra typically contain singly charged ions, ESI generate multiply charged fragment ions[8], which have the advantage of shifting heavy ions into lower m/z ranges where they are more easily detectable[71;72]. The problem is that the multiple charged ions can complicate the spectrum by causing replicates of otherwise identical ions at different charge states. In general, these multiply charged signals occur as isotope clusters. For the purpose of spectrum interpretation, peak replicates originating from different charge states have to be unified. Indeed, most of the common peptide search engines prefer simplified fragment spectra with singly charged ions without isotope patterns. One first step for preprocessing is therefore to detected multiply charged peak clusters, which are removed and converted into a singly charged mono-isotopic

peak that is added to the spectrum. This reliable data filtering is desirable without suppressing any signals or losing mass accuracy[72].

Another preprocessing step is to identify a subset of peaks in a given MS/MS spectrum that is suited to be submitted to the following analysis pipeline[73]. Since most peptide search engines have a optimum number of peaks per spectrum, different problem-specific techniques have been developed to filter the fragment spectra by intensity. For example the simplest MS/MS filter method (e.g. Hansen *et al* [74]), 'top X intensity', sorts all ions in a MS/MS scan by decreasing intensity and only keeps the first X ions. If there are less than X ions, all existing ions are selected. A more advanced approach is filter of 'top X intensity in a window of 100 Da', which selects the Top X most intensive peaks within 100 m/z intervals in its default setting[75]. Among all peaks within this window, only the top X most intense peaks are retained for further analysis. The static window 100 Da is selected in such a way that per interval the fragments of one amino acid residue is covered on average.

One additional point for a successful peptide search is to improve the quality of the information about the intact peptide. Here the focus is on enhancing the information retrieved from the MS^1 scans. Especially with high resolution data it generally is possible to identify the precursor charge state with near certainty[73]. Of main importance is the exact precursor mass and this can be for instance be improved by recalibration in the time and m/z-range[50;75].

Assemble peptide fragments

The interpretation of the preprocessed MS data and identification of the peptides is a central element in the computational pipeline[76]. The spectrum identification problem is difficult to solve primarily because of noise in the observed spectra. In general, the x-axis of the observed spectra is known with relatively high precision and accuracy. However, in any given spectrum, many expected fragment ions will fail to be observed, and the spectrum is also likely to contain a variety of additional, unexplained peaks. These unexplained peaks may result from unusual fragmentation events, in which small molecular groups are shed from the peptide during fragmentation, or from contaminating molecules (peptides or other small molecules) that are present in the mass spectrometer along with the target peptide species. This topic will be discussed in article 2.2 on page 44. In general, peptide identification is performed by correlating acquired experimental MS/MS spectra with theoretical spectra predicted for each peptide contained in a protein sequence data base or against spectra from a spectra library.

Alternatively, peptide sequences can be extracted directly from the spectra, i.e., without referring to a sequence database for help (*de novo* sequencing approach). There are also hybrid approaches, such as those based on the extraction of short sequence tags (3-5 residues long) followed by database searching[77].

De novo identification With complete fragmentation information, it is in principle possible to determine the peptide sequence from the spectrum (*de novo* sequencing). The most intuitive technique might would be to directly 'read off' the sequence information from the acquired MS/MS spectrum (see in figure 7). Such spectrum interpretation relies on the occurrence of ladders of fragments in the spectrum. These occur when a peak in a spectrum corresponding to the nth fragment ion is followed by subsequent peaks corresponding to fragment ions that contain the $n + 1$, $n + 2$, and in general $n + x$ residues. The distance between these consecutive peaks will then correspond to the mass of a single (perhaps modified) amino acid residue, allowing the sequence to be determined. Even with complete fragmentation, defining the peptide sequence is not a simple task. Due to the b-y ambiguity of backbone fragmentation peaks and the distracting presence of 'noise peaks', which may include doubly charged ions, internal fragments, and neutral losses, as well as true chemical noise.

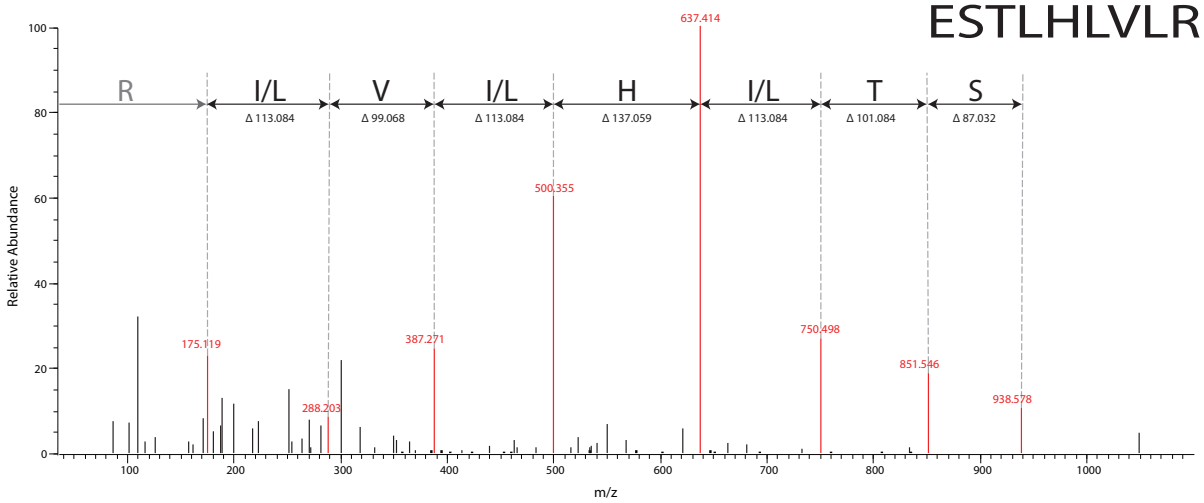


Figure 7: *De novo* peptide sequencing. The figure depicts, how the peptide sequence ESTLHLVLR can be derived by *de novo* from a fragmentation ladder. Starting at the m/z value of 175.119 which represents the mono isotopic mass of arginine (R) + the carboxy terminus (OH). The next residue can be inferred from the delta mass of 113.084 which matches to the mono isotopic mass of leucine (L) or isoleucine (I). Note that one can generally not distinguish between I and L in MS[60]. The ladder of fragment ions can be follow up to the serine (S). Finally, the N-terminal amino acid can be calculated by subtracting intact peptide mass from the last fragment mass.

De novo sequencing is an important approach in the case of organisms with unsequenced or only partially sequenced genomes[78]. In those cases, tools such as MS-BLAST and similar approaches or extensions[79–82] can assist with the downstream analysis of the *de novo* derived peptide sequences to infer the identities of the sample proteins. In addition, the database search compared with *de novo* approach is in the first instance not able to handle wide spread deviations from the database sequence such as mutations or unexpected modification. There are a number of *de novo* sequencing programs including PEAKS[83] and PepNovo[84]. Fragmentation is rarely complete, so even the best *de novo* sequencing programs are less sensitive than database search programs. In comparison to database search, a *de novo* algorithm typically find a correct partial sequence with 6 or more residues, in 60 to 90 percent of the spectra if the input data is of good quality[85]. *De novo* sequencing is not widely used for large scale data analysis because it is computationally intensive and requires high quality MS/MS spectra, and even then does not always guarantee that the peptide can be found reliably.

Database search approaches Sequence database searching remains the dominant method for assigning peptide sequences to MS/MS spectra. In cases where the spectrum contains no recognizable fragment series at all to define some parts of the sequence, database searching is the only option. The basic concept of such search programs it to take as input the experimental MS/MS spectrum and compare it against theoretical fragmentation spectra generated for peptides from the searched protein database (see figure 8). Importantly, the comparison is performed not against all possible peptide sequences, but against a much smaller set of candidate peptides deriving from the database. The candidate peptide list is generated by the program using in silico database digestion and application of several criteria. The most important of these criteria are the parent ion mass tolerance, enzyme digestion constraint (e.g. allowing tryptic peptides only), and which if any post-translational or chemical modifications are allowed. Additional search parameters include the type of fragment ions expected in the spectrum (e.g. y and b ions in HCD), and the fragment ion mass tolerance. The output from the program is a list of peptide spectrum matches (PSM), ranked according to the search score.

The search score essentially measures the degree of similarity between the experimental MS/MS spectrum and the theoretical spectrum. Three basic approaches[87] have been used to determine a match between a spectrum and sequence: autocorrelation between calculated and theoretical spectrum (first used in SEQUEST[88]), extraction of a short sequence with flanking mass values (first implemented by PeptideSearch[77]) and

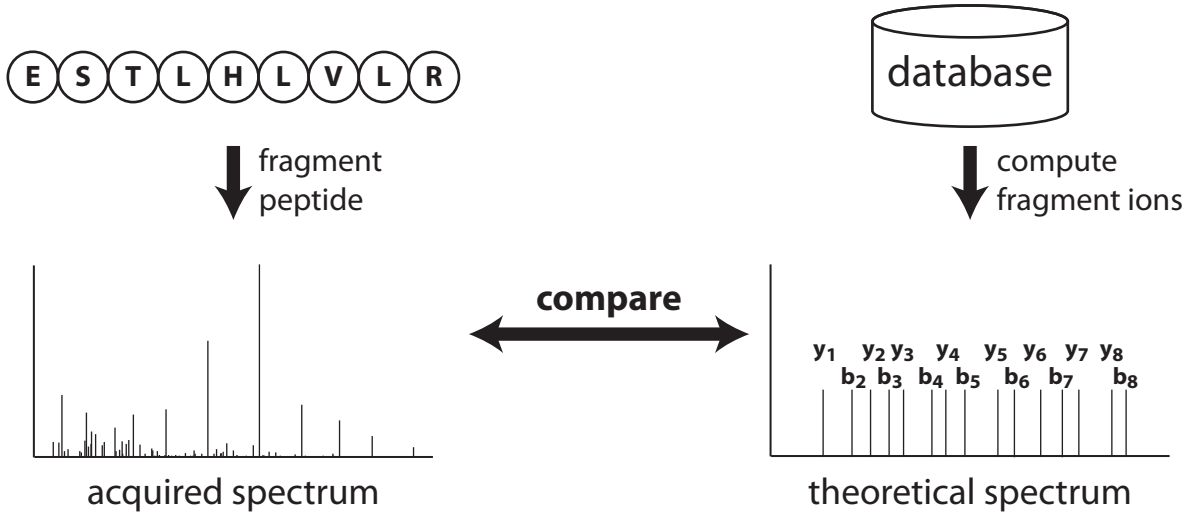


Figure 8: Peptide identification by database search. An acquired MS/MS spectrum is correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the theoretical and actual spectra. [Adapted from Nesvizhskii *et al* [86]]

probability-based matching between calculated and theoretical spectra (pioneered in Mascot[89]) matching. Note that in almost all search engines only the m/z values and not the relative intensity of the fragment ions are compared[90].

The basis of the probability-based this approach, is to calculate the chance that the observed match between the experimental data set and each sequence database entry is a random event. Andromeda, a novel implementation of this approach, is exclusively used in our and many other laboratories. The basis of this approach, is to calculate the probability that the observed match between the experimental data set and each sequence database entry is a chance event. In Andromeda, the most intense ion peaks are extracted within windows of 100 Th, matched with expected fragment masses and the chances of random matches are calculated by means of a binomial distribution function (for more details see article 2.1). The match with the lowest probability to be random is reported as the best match. The score of the match is reported as the negative logarithm of this probability. Whether the best match is also statistically significant depends on the size of the database.

One of the many challenges of large-scale proteomics experiments is to find the correctly identified peptides while maintaining control over false positive identifications[91]. Current methods can never definitively prove that a result is true[92], but an appropriate choice of algorithm can provide a measure of the statistical risk that a result is false, i.e., the statistical significance[93]. The determination of a correct identification is often done

as a subjective assessment based on manual inspection of parameters like the number of MS/MS ions explained by the proposed peptide sequence, biochemistry rules (e.g., the *proline rule*[94]), delta mass values between the measured peptide and the proposed peptide sequence and predicted versus measured peptide retention[91;95]. This determination becomes less subjective with high mass accuracy measurements, but for most investigators, nominal mass accuracy from ion trap mass spectrometers has been typical in the past[91]. In the beginning days of proteomics, the lack of control for false positive identifications led to many reports with incorrectly identified proteins, lack of statistical estimates of false positive error rates in the report and missing details on the search parameters used for identified peptides.

The false discovery rate (FDR), rigorously defined as the proportion of significant results that are expected to be false discoveries in a claimed set of findings, is now routinely used[91]. The dominant method for calculating the FDR is the 'target-decoy' approach[96]. This strategy is based on appending reversed, randomized or shuffled sequences to the original (target) database before performing the search. Then these artificial (decoy) sequences are used to evaluate the portion of false positive among all positive identifications. A simple and powerful way to create a decoy database is to simply reverse each protein in the original database and perform an *in silico* digestion. With some small adoptions, the reverse transformation preserves amino acid frequencies, protein and (approximate) peptide length distributions as well as approximate mass distributions of theoretical peptides[96]. The decoy technique works at both the peptide and protein levels, so that one can send the decoys through a succession of tools (for example, database search, significance analysis, and protein assembly) in order to measure the false discovery rate of the complete pipeline[97]. A FDR cutoff can then be set to limit the maximum number of accepted false-positive matches. Typical cutoff values range between 1 and 5 %, which means that a small portion of any identified peptides will be incorrect. If a large database is searched, these will typically be proteins with a single peptide hit, also called 'one-hit wonders'.

Protein sequence databases Generally the database used for identification should be as inclusive as possible, to allow finding the proteins that have been measured. However, searching large databases also reduces the sensitivity of peptide identification by introducing more false identifications (the likelihood of obtaining a high scoring random match increases with increasing database size)[76;93]. Because of this, the choice of the protein database plays an important role in MS data identification. The most commonly used protein sequence databases[86] for searching MS/MS spectra include

1. Entrez Protein database from the US National Center for Biotechnology Information (NCBI)
2. Reference Sequence (RefSeq) database from NCBI
3. UniProt, consisting of SWISS-PROT and its supplement, TrEMBL
4. International Protein Index (IPI) database, maintained by the European Bioinformatics Institute

Databases vary in terms of their completeness, degree of redundancy and quality of sequence annotation. Entrez Protein is the most complete database; however, it contains many redundant sequences (partial mRNAs, sequencing errors and so forth), and the entries are not as accurately annotated as those in SWISS-PROT or RefSeq[86].The International Protein Index (IPI)[98] was founded in 2001 to cover the gaps in gene predictions between different databases and to provide non-redundant complete sets of sequences for search with MS data. Due to the enormous effort for manual curation and problems with the identifier stability, IPI was closed in 2011. The standard in use today are the complete proteome databases of UniProtKB/SWISS-PROT[99–101] which was introduced in 2006, and also contains manually annotated representation of all protein coding genes. However, this protein knowledge base consists of two sections: First, SWISS-PROT, which is manually annotated and strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Second, TrEMBL, a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT and therefore not reviewed. These two databases are developed by the SWISS-PROT groups at the Swiss Institute of Bioinformatics (SIB) and at the European Bioinformatics Institute (EBI).

The choice of the sequence database for MS analysis depends on the goal of the experiment. For many organisms, multiple sequence databases are available. In most cases, using a better annotated database such as UniProt of RefSeq should be sufficient. When the identification of sequence polymorphisms is particularly important, one may attempt to perform searches against a larger database such as Entrez Protein[76]. Search algorithm which use huge sequence databases with much redundancy are also computer intensive and such searches should be done using only high quality MS/MS spectra.

Modified proteins During peptide search the size of the database is also affected by the number of allowed modifications. It is possible that some amino acids are modified (post-translational or chemical modifications), resulting in mass shifts. Such changes in

mass need to be taken into account to correctly compute theoretical MS/MS spectra.

The simplest cases are fixed modifications, e.g. carbamidomethyl cysteine (+57.0214 Da). All cysteine residues in protein are assumed to be alkylated and the nominal amino acid mass is replaced by a shifted mass in all computations. A fixed modification assumes that every instance of that residue has been modified, so there is no computational overhead to the search, and the score will not be adversely affected.

By contrast, variable modifications are not present systematically. A variable modification is defined by the mass difference to the unmodified amino acid and its localization is in most cases restricted to one amino acid. The database is extended by the modified versions of the peptides containing the specified amino acid during the search. The level of complexity becomes even more dramatic if one considers that the number of combinations of possible modification states for a protein increases more than exponential. For example a protein with two modification can be in four states and a protein with ten modification sites can have 1024 states[102]. This dramatic effect on the search space markedly increase the search time and the FDR. In practice it is therefore not feasible to allow many diverse variable modifications. Instead the number of modification sites is restricted to an appropriate number when searching MS data against a database. The settings are also dependent on the sample preparation, for instance oxidation of methionine residues (+15.9949 Da) or acetylation of the peptide N-terminus (+42.0105 Da) are expected to occur during by the normal workflow. In cases were a specific post-translational modification (PTM) e.g. phosphorylation is the key element of a study, the modification is enriched by biochemical methods during sample preparation and has to be specified in the peptide search[51].

From peptides to proteins

Since most biologist are interested in the proteins present in their samples, the next step in the bottom-up approach is to map the identified peptides to their proteins of origin[103]. However, this 'protein inference problem' is far from trivial[104]. The ultimate goal of inferring protein identities based upon peptide assignments remains a challenge, even when statistical models are employed for validating those assignments. Most of the problems associated with protein identification are caused by so-called degenerate peptides shared by multiple proteins[105]; see figure 9 for an example. Such cases often occur in eukaryotic databases, which contain homologous and redundant entries, and make it difficult to infer the particular corresponding protein(s) present in the original sample[106]. When two or more sequences in the database are identified on

the basis of the same peptides, it is impossible to know with certainty which protein is present in the sample[104]. Protein assembly tools typically rank proteins by confidence, treating distinct peptides as independent evidence of the presence of a protein, but discounting duplicate identifications. These tools output 'protein groups', rather than single proteins, in the case that the peptide identifications do not distinguish between homologous proteins[97].

The pioneering software called Protein Prophet[105] initially groups all assigned peptides according to their corresponding proteins in the database. Once grouping is complete, the assigned peptides corresponding to an individual protein, and their probabilities, must be combined to compute a single protein confidence measure that is effective at distinguishing the correct from incorrect protein identifications. A particular challenge in that regard is the detection of correct protein identifications with only a single corresponding assigned peptide in the data set, since the majority of incorrect protein identifications also have only one corresponding peptide.

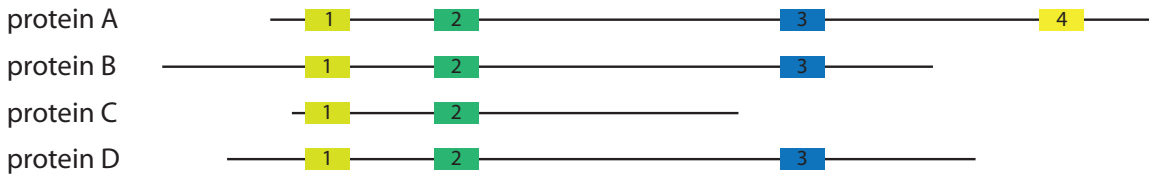


Figure 9: Issues in protein identification. Four proteins (A, B, C, D) are identified by four distinct peptides (box). Although B and D are different, it is impossible to ascertain which molecule is present, as both have been identified by the same (shared) peptides. A variation is shown in C. Protein A shares three peptides with B and D, and two with C, but also has a unique fourth peptide. From this information it can concluded that D is in the sample.

Protein quantification

Protein identification is only the first step and quantification is necessary for most biological studies to estimate the protein concentration in a sample. Especially system-wide or systems biology studies require the capability to quantify proteins of the cell from large-scale proteomics experiments. The overall goal of such measurements is to obtain a snapshot of concentrations of proteins associated with different states, such as healthy or diseased. The various methods in MS-based proteomics (see figure 10) to estimate the protein concentration can be divided in two groups, relative and absolute quantitative proteomics.

The most popular approaches for relative quantification are based on labeling proteins or peptides in at least on of the compared samples with compounds enriched

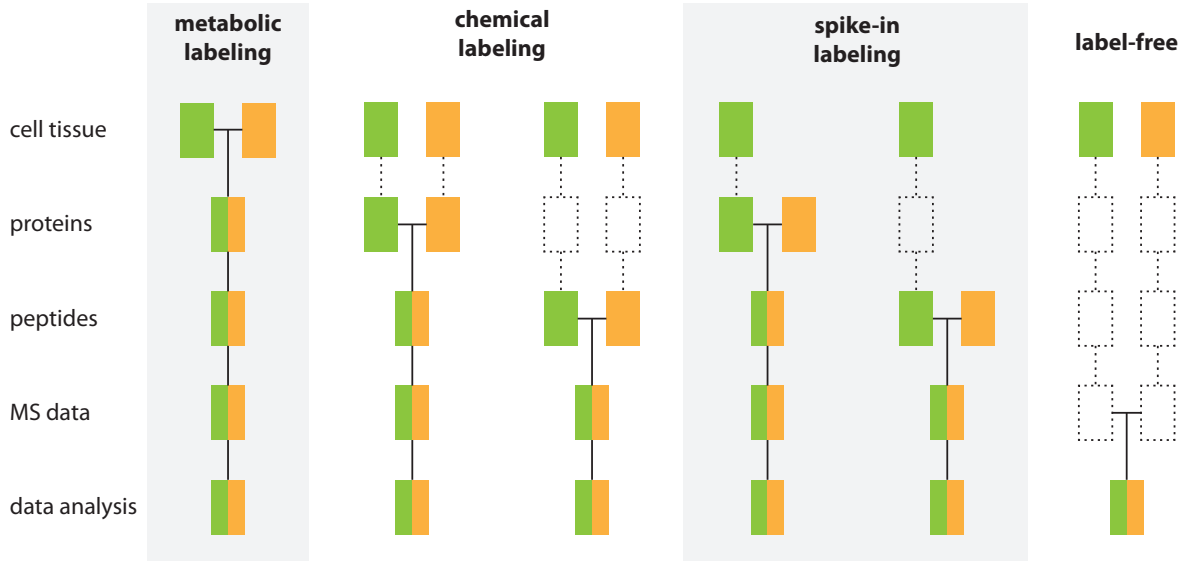


Figure 10: Collection of common quantitative workflows in MS-based proteomics. Green boxes and orange boxes represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate the points at which experimental variation and thus quantification errors can occur. [Adapted from Bantscheff *et al* [107]]

in stable heavy isotopes of hydrogen, carbon, nitrogen or oxygen[108]. These labeling techniques exploit the fact that labeled molecules behave almost identically during chromatographic separation, ionization and in the mass analyzers. In metabolic labeling, the label is introduced to the whole cell or organism *in vivo*, through the growth medium. In contrast, in chemical labeling the label is added to proteins or tryptic peptides through chemical derivatization or enzymatic modification *in vitro*, after sample collection.

A popular metabolic labeling method is stable isotope labeling by amino acids in cell culture (SILAC)[109]. In SILAC, essential amino acids such as arginine and lysine are provided in 'light' and 'heavy' forms to the two cell populations and are incorporated into each protein after a few cell doublings, leading to a well-defined mass difference. 'Chemical labeling' makes use of externally introduced isotopic or isobaric reagents. Examples for the first category include dimethyl labeling[110] and isotope-coded affinity tags (ICAT)[111]. Isobaric mass tagging, illustrated by isobaric tag for relative and absolute quantification(iTRAQ)[112], differs from the methods described above in that labeled peptides have exactly the same mass and are thus indistinguishable in the survey spectra. In this case, the different mass tags separate only upon fragmentation and quantitation relies on the intensity ratios of so-called reporter ions in the fragment spectra.

Although protein relative quantification using labeling strategies has been success-

fully used in many studies, these techniques also have several limitations, for instance in the number of samples that can be directly compared, side products generated during labeling, costs of reagents and so on. Consequently, there is much interest in methods that do not require isotope labels and that rely on direct comparison of peptide signals across different experiments. These so-called 'label-free' methods offer a simpler sample preparation and direct comparison of multiple samples. In its most primitive form, the number of peptide fragmentation events is taken as an estimate of the amount of protein. This spectral counting technique is used to provide a semi-quantitative measure of protein abundance[97] but has been found to often result in imprecise or irreproducible data[113]. In contrast to spectral counting, common label-free methods are based on the comparison of normalized intensities from two separate runs. Intensity-based label-free quantification is more powerful, but requires careful and reproducible sample preparation techniques along with sophisticated software. This approach does not require continually re-identifying peptides in every sample under study because it decouples profiling from identification and subsequently links the profiling and identification data sets in silico via accurate m/z and retention time. The most common readouts are extracted ion chromatograms (XIC) of the parent ion.

Relative quantification by its nature cannot provide information about absolute protein abundance. Especially in a medical context, knowing absolute amounts of disease-specific biomarkers can provide diagnostic information of high relevance[114]. Also, modeling approaches require absolute molecule numbers to quantitatively describe dynamic systems. Absolute measurements of protein concentrations can be achieved with 'spiked synthetic peptides', as in AQUA[115] or by artificial proteins derived from detected peptides, as in QconCAT[116], and 'Absolute SILAC'[114].

1.5 MaxQuant - Software environment

As described before, the data analysis typically involves several steps and is not confined solely to peptide identification by a peptide search engine. There are only a few software environments that provide data analysis in a single environment performing all or most of the steps from acquired raw data to final protein lists. Examples are the Trans-Proteomic Pipeline[117], OpenMS Proteomics Pipeline[118] or Skyline[119]. These efforts were usually not directed at high-resolution data of the type readily attainable today and they did not approach the quality of a skilled human expert. Our own laboratory has developed the MaxQuant computational proteomics environment, which is freely available to academic and commercial users and which has been widely adopted by the community. MaxQuant[75] is a set of algorithms that efficiently and robustly extracts information from raw MS data and allows very high peptide identification rates as well as high-accuracy protein quantification for several thousand proteins in complex proteomes. MaxQuant takes advantage of high resolution data such as those obtained by Orbitrap instruments and employs algorithms that determine the mass precision and accuracy of peptides individually. This leads to greatly enhanced peptide mass accuracy that can be used as a filter in database searching[25].

The analysis pipeline (see figure 11) consists on five main tasks: feature detection, recalibration, peptide identification, protein group assembly and writing tables.

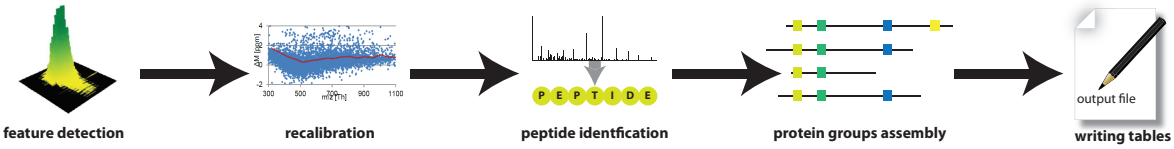


Figure 11: MaxQuant Tasks of the MaxQuant analysis pipeline grouped in five mayor parts.

The first MaxQuant version, published in 2008, focused on quantification by stable amino acid isotope labeling (SILAC)[120;121]. Later on quantification options were extended with the implementation of quantitation using chemical labeling (such as iTRAQ) and a sophisticated label-free algorithm. The MaxQuant environment originally used the popular Mascot peptide search engine to match tandem mass spectra to possible peptide sequences. This was later replaced by Andromeda, our in-house developed search engine. For more information see article 2.1 on page 29. The MaxQuant software was initially developed for instruments of Thermo Fisher Scientific, and in contrast to other computational tools, MaxQuant is using the pure raw file from the instrument as input. One advantage of this is that no information is lost due to conversion as often happens when using a general open file format (such as mzXML).

Feature detection The first element in the pipeline is the detection of the peptide features. In contrast to other software, MaxQuant makes use of all three dimensions (3D), meaning m/z range, abundance and retention time, rather than single MS^1 scans, to take maximum advantage of high resolution and mass accuracy (see figure 12A). For the boundary construction of all peaks a gaussian peak shape is fitted into the m/z-retention time plane (see figure 12B). This is followed by the determination of the isotope patterns using undirected graphs with the detected peaks as vertices. The result are thousands of features which represents the eluted peptides and are defined as the 3D peak shape of the isotope pattern (see figure 12C).

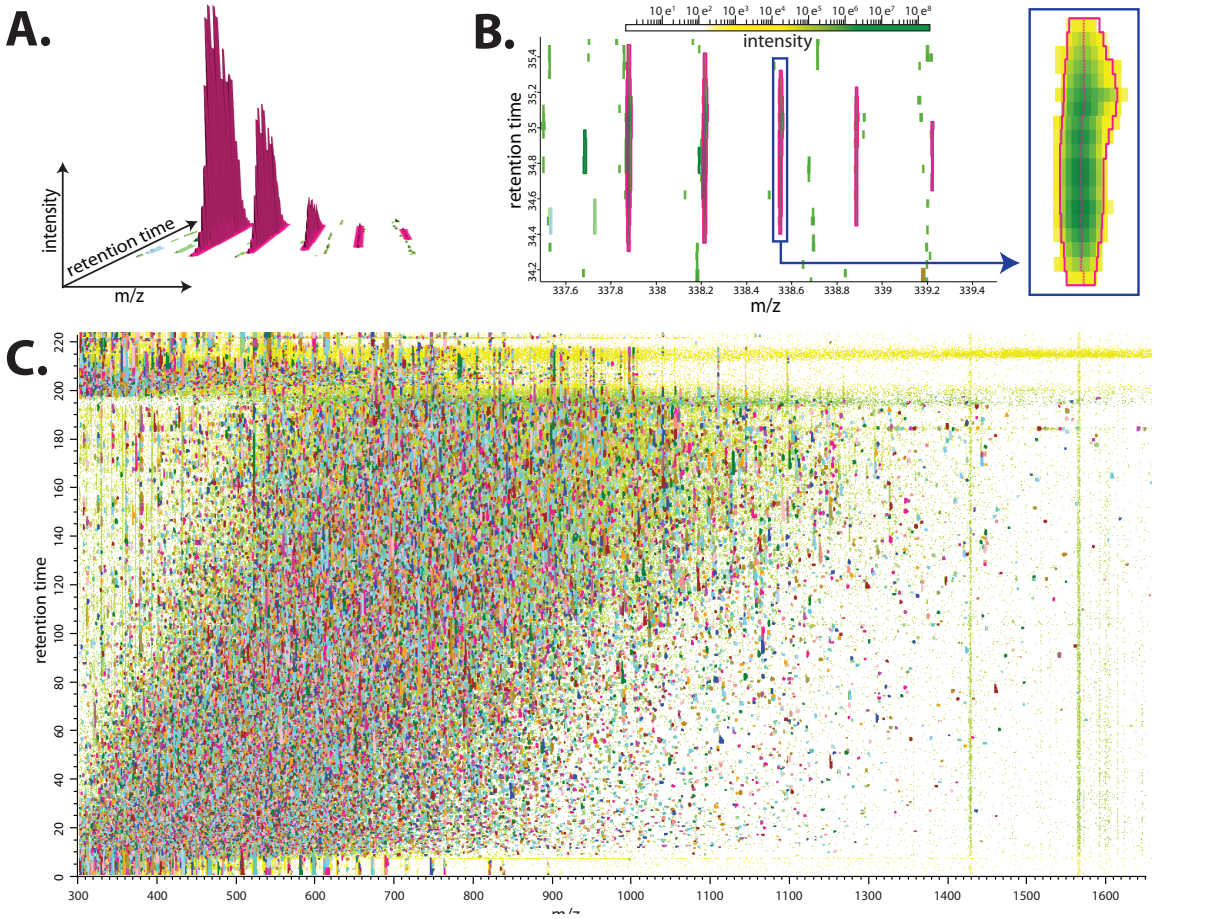


Figure 12: Feature detection. (A) The 3D representation of the isotope pattern of a feature. (B) Signals eluting from the column are drawn with color-coded intensity, decreasing from green over yellow to white. After peak detection a clear isotope pattern is represented by the pink shape. To the right, a peak is shown in detail, in which the centroids are displayed as dotted line. (C) Visualization of all MS runs in one experiment where retention time in minutes is depicted along the Y axis and m/z along the X axis. An optional layer enables the visualization of all detected features using varying colors.

In case of heavy isotopic labeling the next step is to detect label pairs (light-heavy) or even multiplets (for example light-medium-heavy). For this tasks the algorithm looks

for shifted isotope patterns, according to the mass shift expected for a maximum of two labeled amino acids. The requirements of a potential labeled peptide multiplet are equal charge states and sufficient intensity correlation over elution time. Additionally in each of these cases the multiplets need to have the same atomic composition which is ascertained by convolution of the measured isotope patterns with the theoretical patterns of the difference atoms. For normalization reasons the resulting isotope pattern are scaled to each other, which then yields the fold-change between labeled peptides. Additionally, the median logarithm of the entire set of all normalized peptide ratios is shifted to zero. This normalization is done for each LC-MS run separately and is necessary to correct for unequal protein loading. It relies on the reasonable assumption that the majority of proteins show no differential regulation. Already at this stage, the quantification step for peptides with isotopic label can be performed, without knowing their identity.

Using the full information of all MS scans belonging to a 3D peaks has several advantages compared to the information from a single MS^1 scan. For example, the peptide mass is precisely determined from all full scans of the elution profile from the peptide feature using their averaged intensity-weights. The complete 3D peak cluster also improves the accuracy of quantification. A priori knowledge of the label state allows searching with fixed modification for labeled peptides. In this way, the information about label state and the precise peptide mass enhance the later peptide database search.

The last task in feature detection is the preparation of the peak list for the peptide database search. For this purpose the tandem spectra are preprocessed, which involves converting peaks from profile to centroid mode, transformation of isotope patterns to one peak including the transfer to single charge state and finally the filtering of the top x (x is dependent on the resolution) most abundant peaks per 100 Th window.

Mass recalibration Small changes in the instrument changes its behavior over time and for instance temperature drift can lead to noticeable changes in mass accuracy. This results in a systematic mass errors[122]. To compensate for drifts in instrument calibration, a compound of known mass is often employed[29]. This 'lock mass' provides an internal mass standard in every spectrum. The source of this compound can come from a separate ion source or the compound can be mixed with the analyte. One problem with this approach is that the internal standard can interfere with the analyte, yield low abundant signals that are difficult to pick up, especially in the presence of high abundant samples. The electrospray process itself enables the usage of chemicals from

the laboratory air as internal standard. In the LTQ-Orbitrap family special components present in the laboratory air are isolated, mixed with the sample ions and measured. This technique is used to automatically adjust the mass scale in real time[29]. However, the entire procedure of adding the lock mass and recalibration can have a negative effect on cycle time.

These practical problems are sidestepped in MaxQuant through the use of a so called 'software lock mass'[123]. Here the complexity of typical peptide mixtures in proteomics is employed to eliminate the requirement for a physical lock mass. Since we integrated our own search engine Andromeda, an initial peptide search can be performed at larger mass tolerances, resulting in a large number of identified peptides. These data points are used to apply a nonparametric calibration curve determined from the difference in observed versus calculated peptide mass. The software lock mass corrects mass errors both on the retention time scale and on the m/z axis. In this way sub-ppm mass accuracy can be obtained even without lock mass injection during the data acquisition, with none of the experimental cost of a physical lock mass[123].

Peptide identification The precursor mass obtained by the recalibrated survey spectra together with the MS/MS spectra are then subjected to a database search employing the integrated search engine Andromeda. False discovery rates (FDR) are estimated by searching against a concatenated target-decoy database as explained above[96;124]. This database contains all true protein sequences, concatenated with reversed versions of these sequences. To avoid spurious correlations because half of the reversed tryptic peptides have the same mass as the forward sequence, we also swap every arginine and lysine with the preceding amino acid in the reversed sequences. This approach still retains the local amino acid relations - leading to the same length and mass distribution of peptides. After the database search, the list of top fifteen peptide candidates is sorted according to their peptide score or p-score and filtered for consistency with a priori information, retaining the best scoring one. By default a 1% FDR is applied, which means the peptide list is cut at 1% reverse hits. For this purpose, peptides are ranked according to their individual peptide posteriori error probability (PEP)[125], which is dependent on the identification score and on the peptide length. The PEP, also known as the 'local FDR', represents the probability that a given PSM is incorrect, given the distribution of all the PSMs with same amino acis length in the experiment[92].

Due to the complexity of peptide mixtures and the relatively low resolution of precursor isolation, two peptides are frequently 'co-fragmented'. These 'chimerical' MS/MS spectra[66] can be detrimental for identification of the peptide of interest, especially if the

co-fragmented peptide is of comparable intensity. Co-fragmentation generally reduces the number of peptides identified in database searches and poses special problems for reporter fragment based quantification methods because both peptides contribute to the measured ratios.

The novel second peptide identification algorithm (see article 2.1) turns this problem into an advantage by adding a additional peptide search. Signals coming from the already identified peptide are removed. The remaining fragment peaks are submitted to a new database search with the precursor mass from the peptide that was not intentionally targeted for MS/MS. Note that due to statistical reasons, a separate FDR is applied for peptides coming from the second peptide search.

Protein group assembly The FDR controlled peptide identifications are used to assemble protein groups. For this, peptides are distinguished into unique peptides (present in only one group), group unique peptides and non-unique peptides (present in more than one group). Non-unique peptides are assigned to the protein group with the most peptides for quantification (razor peptides). Thereby overestimation of protein identifications is prevented. The list of proteins is sorted according to the protein PEP and by default cut at 1% false positives. The protein PEP is calculated as the product of the individual peptide posterior error probabilities. The protein ratios are calculated as the median of the SILAC peptide ratios.

Write tables In the last section of MaxQuant the results of the previous steps are combined and provided as several text files. This output tables can be used for downstream analysis using different statistical packages such as R[126] or Perseus[127]. The Viewer software, part of the MaxQuant environment, enables further inspection of the MaxQuant identifications. In this software, the cross references are used to connect several of the output tables. Also the visualization of the LC-MS data such as contour plot (see figure12C) or annotated MS/MS spectra, is made possible by this program.

2 Results

New technological developments in the field of proteomics generally have the goal to improve on the one hand the number of identified proteins to maximize the coverage of the proteomic sample, and on the other hand prove the correctness of the result. Furthermore the time needed for this analysis is also an important consideration. The following articles made contribution to all three areas.

2.1 article 1 - Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment

Jürgen Cox, **Nadin Neuhauser**, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen and Matthias Mann
Journal of proteome research 2011, 10, 1794-1805

The peptide search is the central element in computational shotgun proteomics. A widely used program for this, is the Mascot search engine[89]. This program is implemented as client-server application and is commercial available, which means that the exact algorithms used are not publicly known. As an alternative to this 'black box' situation, we decided to develop our own search engine Andromeda that would be free of these restrictions. In contrast to Mascot, it can handle data with arbitrarily high fragment mass accuracy and is therefore able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides.

Andromeda can be run independently as a standalone version. In this case, the user is required to perform the downstream statistical processing, so as to rigorous control the protein false discovery rate. More usually, Andromeda run as an integral part of the MaxQuant platform. The MaxQuant pipeline performs several search cycles (initial search, main search and second peptide search), which enables analysis of large data sets in one workflow on a desktop computer.

In Andromeda, we chose a probability based approach for peptide identification, using binomial distribution probability and defining a so-called p-score. This score originated from the determination of the localization probability in modified peptides[128] and was already used from the beginning of MaxQuant for ranking the peptide candidates. We perform a rigorous comparison of the above mentioned Mascot software with our new search engine on several large-scale data sets. Indeed, the paper also demonstrates the ability of Andromeda to accurately handle many modifications of the same peptide.

A key advantage of Andromeda is its extensibility. For example, proteomics with high accuracy MS and MS/MS data (high-high mode), is becoming increasingly common. Andromeda, in contrast to Mascot, allows arbitrarily accurate MS/MS requirements specified in ppm. Similarly, Mascot precludes identification of SILAC pairs if the same amino acid can bear a fixed and a variable modification. Apart from describing the score we have also made the actual code used in Andromeda available with this publication. Both the standalone and also the integrated MaxQuant version of Andromeda, is freely available.

One of my specific contributions was 'AndromedaConfig' a user interface to specify allowed modifications, enzymes and databases. The user can add novel modifications by their elemental composition and amino acid sites. Compared to the common used UNIMOD interface, the user can also specify for each modification, the neutral losses of each individual amino acid separately as well as so-called diagnostic ions. It is also possible to enter modifications that are interpreted as labels by MaxQuant, such as SILAC states. In the configuration program new cleavage rules of known or new enzymes can be defined as well. Since the rules have to be specified by regular expressions, which can be error prone, the program has a verification utility assuring that the correct cleavage rules are used. For the extraction of the identifier of a FASTA file, the Andromeda software has to know the parsing pattern. The user can achieve this by adding new protein databases via the AndromedaConfig interface. Similar to the enzyme specification, regular expression are employed in the configuring process of new databases. The user can select from a number of pre-configured parse rules or enter a new regular expression. Again, this can be tested on the current FASTA file within the user interface. All changes made are stored in XML files (modifications.xml, enzymes.xml and databases.xml), which are located in the configuration folder of the Andromeda or MaxQuant software. When using Andromeda in the MaxQuant pipeline, the definition of modifications, enzymes and databases, has to be completed before opening the graphical user interface.

In addition I developed a web server for searching individual spectra using Andromeda. The scoring results of the 15 best peptide candidates can be inspected by the annotated spectrum for the highest scoring and all other candidate peptide sequences. In this standalone web server, the submission is limited to five peak lists to avoid overload.

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment

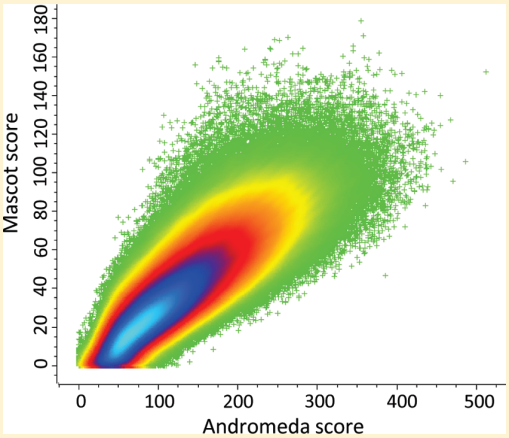
Jürgen Cox,[*,†] Nadin Neuhauser,[†] Annette Michalski,[†] Richard A. Scheltema,[†] Jesper V. Olsen,[‡] and Matthias Mann[*,†,‡]

[†]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[‡]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark

Supporting Information

ABSTRACT: A key step in mass spectrometry (MS)-based proteomics is the identification of peptides in sequence databases by their fragmentation spectra. Here we describe Andromeda, a novel peptide search engine using a probabilistic scoring model. On proteome data, Andromeda performs as well as Mascot, a widely used commercial search engine, as judged by sensitivity and specificity analysis based on target decoy searches. Furthermore, it can handle data with arbitrarily high fragment mass accuracy, is able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides, and accommodates extremely large databases. The algorithms of Andromeda are provided. Andromeda can function independently or as an integrated search engine of the widely used MaxQuant computational proteomics platform and both are freely available at www.maxquant.org. The combination enables analysis of large data sets in a simple analysis workflow on a desktop computer. For searching individual spectra Andromeda is also accessible via a web server. We demonstrate the flexibility of the system by implementing the capability to identify cofragmented peptides, significantly improving the total number of identified peptides.



KEYWORDS: tandem MS, search engine, spectrum scoring, post-translational modifications, mass accuracy, collision induced dissociation, higher-energy collisional dissociation, Orbitrap

INTRODUCTION

Mass spectrometry (MS)-based proteomics is becoming a commonly used technology in a wide variety of biological disciplines.[1−6] In a "shotgun" format, very complex peptide mixtures are produced by enzymatic digestion of protein mixtures, which are analyzed by liquid chromatography followed by tandem mass spectrometry.[7,8] Per LC−MS/MS run, thousands of MS and MS/MS scans are acquired, often producing gigabytes of high resolution data per day and per mass spectrometer. Computational proteomics has become a key research area, dealing with the challenges of how to most efficiently extract protein identification and quantification results from the raw data. Both the proteomics community and the bioinformatics community have dealt with many areas of this novel field, and there is already a large literature outlining and reviewing the general tasks involved,[9−17] particular computational aspects of the field[18−22] and integrated data analysis pipelines.[23−30]

In this context, our group has developed the MaxQuant environment, a computational proteomics workflow that addresses the above tasks with a focus on high accuracy and quantitative data. It includes peak detection in the raw data, quantification, scoring of peptides and reporting of protein groups.[31] MaxQuant takes advantage of high resolution data such as those obtained by the linear ion trap−Orbitrap instruments and employs algorithms that determine the mass precision and accuracy of peptides individually. This leads to greatly enhanced peptide mass accuracy that can be used as a filter in database searching.[32] MaxQuant was also specifically designed to achieve the highest possible quantitative accuracy in conjunction with stable isotope labeling with amino acids in cell culture (SILAC).[33,34] Using high resolution data combined with individualized mass accuracies and robust peptide and protein scoring results in high peptide identification rates of typically 50% and even higher on SILAC peptide pairs.[31] This was an important foundation for the quantification of the first complete model proteome, that of budding yeast.[35]

Received: October 23, 2010
Published: January 21, 2011

Journal of Proteome Research ARTICLE

The MaxQuant environment originally used the Mascot peptide search engine[36] to match tandem mass spectra to possible peptide sequences. Mascot together with SEQUEST[37] are commonly used search tools in proteomics today. However, there are many others including Protein prospector,[38] ProbID,[39] X!Tandem;[40] OMSSA,[41] ProSight[42] and Inspect[43] (see Nesvizhskii et al. for a review[14]). Mascot takes a probability based approach to match sequences from a database to tandem mass spectra.[36] Because it is a commercial program the exact algorithms it employs are neither known nor available for modification. Furthermore, Mascot is implemented in a client-server configuration, which imposes practical restrictions for some applications such as real-time searches. We therefore set out to develop a new search engine that would be free of these restrictions. We aimed at performance at least on par with Mascot, which has become a "gold standard" in proteomic analysis, and robustness for scaling up to extremely large and complex data sets. In combination with MaxQuant, the new search engine would then enable analysis of complex data sets on desktop machines by any proteomics researcher or biologist wishing to employ proteomics.

Database searching with fragment mass spectra typically follows one of three approaches:[44,45] (i) deriving a partial or full peptide sequence with associated mass information (first implemented by PeptideSearch[46] and graph theory based *de novo* methods[47]), (ii) autocorrelation between the experimental and a calculated spectrum (first used in SEQUEST) or (iii) calculating a probability that the observed number of matches between the calculated and measured fragment masses could have occurred by chance (pioneered in Mascot). We chose the probability based approach based on the binominal distribution probability and started from a score that we had originally developed for analyzing MS^3 data for which no search software was available at the time.[48] This score has already been used for ranking the peptides in MaxQuant searches from the beginning and it also determines the localization probability of modifications in peptides.[48]

In this paper, we describe the architecture of the Andromeda search engine and its scoring function. We perform a rigorous comparison against the Mascot search engine on several large-scale data sets. The ability of Andromeda to accurately handle many modifications of the same peptide is demonstrated. Due to the complexity of peptide mixtures in shotgun proteomics and the relatively low resolution of precursor isolation, two peptides are frequently 'cofragmented' and there are algorithms that try to identify them from mixture spectra.[49−52] We demonstrate the flexibility of the Andromeda search engine by implementing a novel second peptide identification algorithm.

MATERIALS AND METHODS

Benchmark Data Sets

Raw data from 84 LC−MS runs was taken from Luber et al,[53] a label-free proteome study of mouse dendritic cells to a depth of 5780 proteins. Cell subpopulations were obtained by FACS sorting, proteins were separated by 1D SDS-PAGE and digested with trypsin. Peptides from the gel pieces were analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific).

As a phosphoproteomics benchmark data set we took the raw data from 117 LC−MS runs produced in a phosphatase knockdown analysis.[54] Drosophila Schneider SL2 cells were differentially SILAC labeled as pairs with Lys-8/Arg-10 and Lys-0/Arg-0. Proteins were separated by 1D SDS-PAGE and digested with trypsin or in solution digested without gel separation. Peptides were subjected to TiO_2 chromatography and strong cation exchange chromatography and analyzed on a nanoflow HPLC system connected to a hybrid LTQ-Orbitrap (Thermo Fisher Scientific). For the analysis, we used only those MS/MS spectra that were acquired on a recognized SILAC pair. Modifications due to labeling with Lys-8 and Arg-10 can then be taken as fixed.

The benefits of second peptide analysis were investigated using data that was acquired on an LTQ-Orbitrap Velos. Briefly, HeLa cell lysate was in solution digested with trypsin, the peptide mixture was separated on a nanoflow HPLC system and analyzed using a data-dependent "top 10" method. Several runs were acquired with varying isolation windows. The precursor ions were isolated in selection windows of 1, 2, 4, 8, 16, and 32 Th followed by HCD fragmentation and high resolution data acquisition of the MS/MS spectra in the Orbitrap.

Data Preparation

MaxQuant, version 1.1.1.25, generated peak lists from the MS/MS spectra for the database searches. For the low-resolution MS/MS spectra recorded in "centroid" mode the 6 most abundant peaks per 100 Th mass intervals are kept for searching. High-resolution profile MS/MS data is deconvoluted (deisotoping and transfer of all fragment ions to single charge state) before extraction of the ten most abundant peaks per 100 Th. All statistical filters in MaxQuant like peptide and protein false discovery rates and mass deviation filters were disabled in order to score all submitted MS/MS spectra. Peptide masses were recalibrated by MaxQuant prior to both Andromeda and Mascot searches. For the Mascot search (using Mascot server version 2.2.04), peak lists written out by MaxQuant were converted to mgf format, the standard Matrix Science data format. Oxidation of methionine and N-terminal protein acetylation were used as variable modifications for all searches. A mass tolerance of 6 ppm was used for the peptide mass. To make Mascot and Andromeda searches comparable, we did not use the individual peptide mass tolerances in MaxQuant. A tolerance of 0.5 Th was used for matching fragment peaks produced by CID. The HCD fragment ion data used in the co-fragmentation study were searched with a 20 ppm window in Andromeda. A maximum of two missed cleavages were allowed in all searches. The "instrument" parameter was set to "ESI-TRAP" in the Mascot search. Mascot and Andromeda scores were matched to each other based on raw file name and scan number.

The search was performed against a concatenated target-decoy database with modified reversing of protein sequences as described previously.[31] Mouse and human data was searched against the respective IPI databases,[55] version 3.68, while the drosophila data was searched against protein sequences from flybase[56] version 5.24.

Search Engine Configuration

In Andromeda, the user specifies allowed peptide and protein modifications, enzymes used for protein cleavages and the protein sequence databases to be searched in the program AndromedaConfig.exe. Modifications are specified by their elemental composition. Neutral losses and diagnostic ions can be specified separately for each type of amino acid with the modification in question. Modifications that are interpreted as labels by MaxQuant can be defined here, such as SILAC labels. Searches with semispecific enzymes are supported as well, where

Journal of Proteome Research ARTICLE

Table 1. Most Important Regular Expressions Defining How Protein Identifiers Are Extracted from the Headers of Fasta File Entries

regular expression	description
>(.*)	Everything after ">"
>([^])	Up to first space
>IPI:([^| .]*)	IPI accession
>(gi|[0−9]*)	NCBI accession
>([^\t]*)	Up to first tab character
>.*|(.*)|	Uniprot identifier

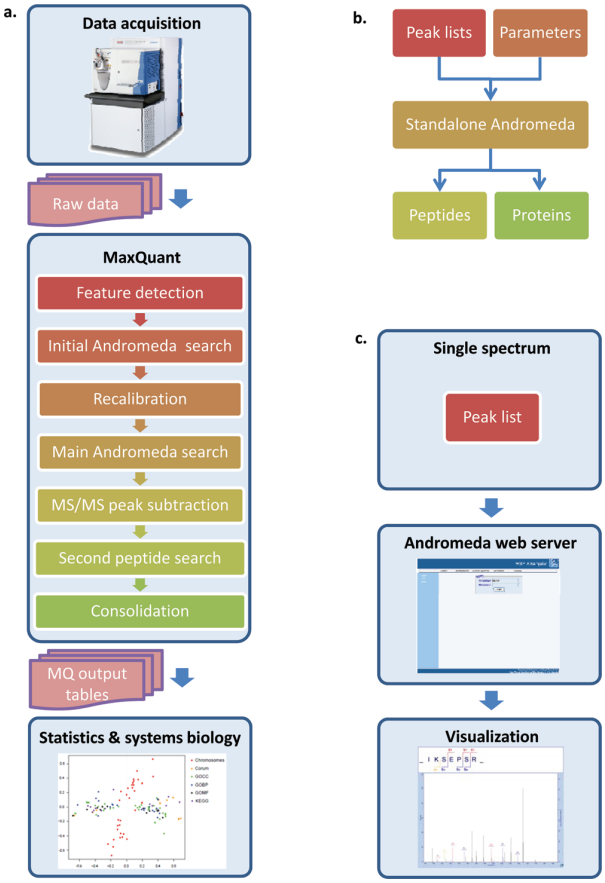


Figure 1. Three Andromeda configurations: (a) integrated in MaxQuant, (b) standalone search engine, and (c) web server.

only one peptide terminus needs to be a cleavage site according to the given protease digestion rule while the other terminus can be an arbitrary position in the protein. An unspecific search is also supported where both of the peptide termini can be arbitrary positions in a protein. Parse rules for regular expressions as defined in the Microsoft .NET framework (msdn.microsoft.com/en-us/library/az24scfc.aspx) are used to define how a protein identifier is extracted from the header line of a FASTA database file entry. Some of the most important regular expressions can be found in Table 1.

Input and Output Formats

Input files for peak lists and parameter values as well as output files for peptide identifications and a tentative protein list are all human-readable text files. Parameter files have the ending ".apar" and contain a list of key-value pairs where each pair is separated by a "=" sign. Expressions used for modifications, labels, enzymes and databases must have been defined previously in the AndromedaConfig.exe program. Peak list files have the extension ".apl" and can consist of arbitrarily many spectra, one following the other, each spectrum entry being enclosed by "peaklist start" and "peaklist end" lines. Some key-value pairs with peaklist-specific parameters are followed by two columns of numbers containing the m/z and intensity values. The peptide result files (".res") contain up to 15 candidate peptide matches for each peak list. For each candidate the peptide sequence, modification state, score, mass, mass deviation and all corresponding protein IDs are given.

Software Availability

MaxQuant with Andromeda as the integrated search engine can be downloaded from www.maxquant.org. A standalone version of Andromeda is available at www.andromeda-search.org. The source code is provided as Supporting Information 1. Both applications require Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. The Andromeda web server can be accessed at www.biochem.mpg.de/mann/tools/ for a limited number of submissions of MS/MS spectra. Andromeda has been written in the programming language C#, using the Microsoft .NET framework version 3.5.

RESULTS

Andromeda is a search engine based on a probability calculation for the scoring of peptide−spectrum matches. A version of it is fully integrated into the MaxQuant quantitative proteomics platform. Hence, all the data processing from the acquired raw data to the list of quantified peptides and proteins can be performed in a single end-to-end workflow (Figure 1a). In addition to the regular search Andromeda can be used in different contexts: for example in MaxQuant it is used for determining the mass-dependent recalibration function based on a preliminary database search, and for the identification of one or more cofragmented peptides (see below). We also provide a standalone version of Andromeda that produces scored peptide candidates, given a collection of MS/MS peak lists and a parameter file (Figure 1b). In this option, many of the statistical processing algorithms that are part of MaxQuant are not applied to the data and the reported list of identified proteins is only tentative without rigorous control of protein false discovery rate (FDR). The output consists of a raw list of scored peptide candidates per spectrum together with the protein list. Furthermore, there is a web server version of Andromeda for the submission of a limited set of spectra (Figure 1c), www.biochem.mpg.de/mann/tools/. In addition to the scoring results of the 15 best peptide candidates, the annotated spectrum can be inspected for the highest scoring and all other candidate peptide sequences. Despite these alternative uses, we anticipate that Andromeda will most commonly be employed as the search engine for MaxQuant.

Indexing Peptides and Proteins

To efficiently score an MS/MS spectrum it is important to be able to quickly retrieve all candidate peptides that have a suitable calculated precursor mass within a given tolerance. First we generate a list of all peptides obtained by the specified digestion

Journal of Proteome Research ARTICLE

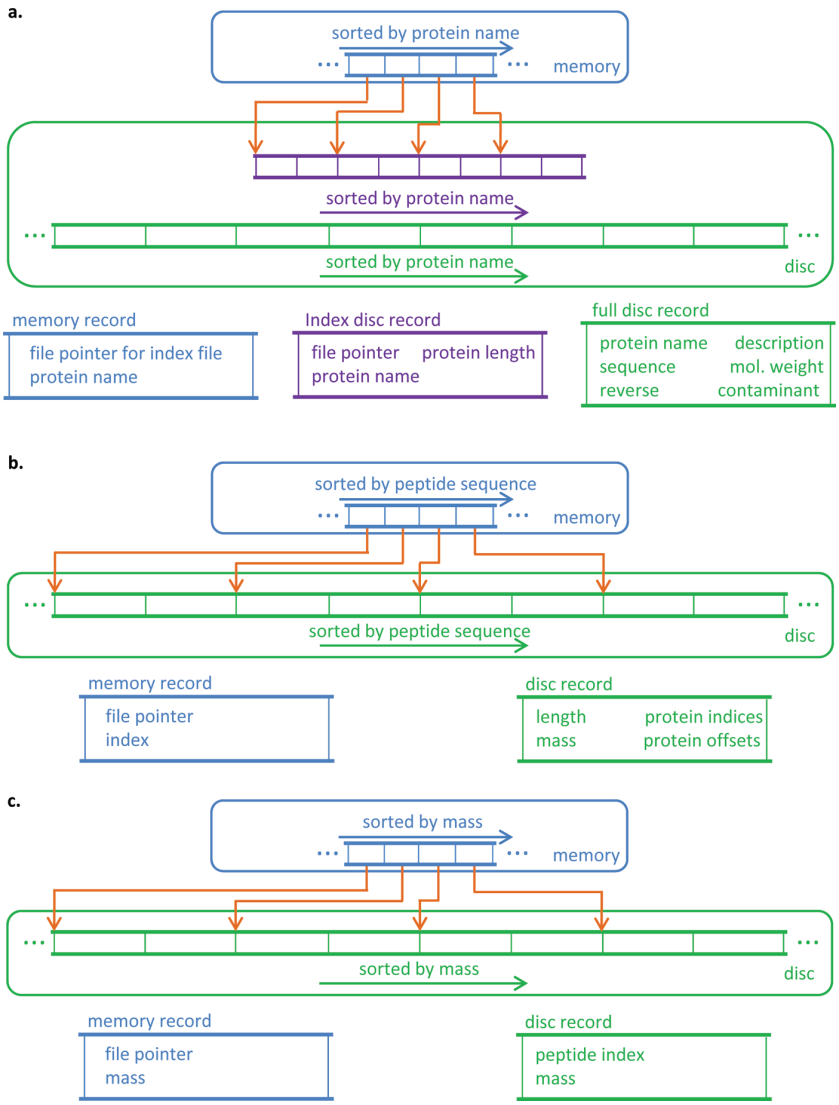


Figure 2. Memory and disk structure. (a) Protein list has a two-layer index structure. One small index is kept in memory whose entries point to blocks of multiple entries in the secondary index that is kept on disk. Each entry of the disk index points to the position of the protein entry in the file containing the complete information for each protein including the amino acid sequence. The protein lists are sorted alphabetically by the protein names. (b) Peptide index that resides in memory points to equally sized blocks of peptide entries, which are kept on disk. (c) Similar structure for the list of all combinations of peptide sequence and variable modifications. Index and disk entries are sorted by the peptide mass to allow for quick retrieval of all peptide candidates within a given mass interval.

rule from the protein sequences considering all possible combinations of preset variable modifications. At this stage we are only interested in the peptide masses, therefore only the number but not the positions of the modifications are important. The list of all of these peptides is sorted by mass for quick search access, which only grows slowly with increasing size (proportional to the log of the number of peptides for a binary search). The number of peptides with specific modifications can become very large, either when searching in an extended protein sequence database or by specifying many variable modifications. One common setting is to search the human IPI database including reverse sequences and common contaminants digested with trypsin and allowing for up to two missed cleavages. The number of modifications to consider can also grow rapidly. For example, in a phosphoproteomic experiment with triple SILAC labeling of lysine and arginine, one may simultaneously deal with phosphorylation of serine, threonine and tyrosine, Lys4, Lys8, Arg6, Arg10 and oxidation of methionine as variable modifications. (This is the case for those MS/MS spectra where the SILAC state could not be determined prior to the database search; otherwise the modification state of Arg and Lys are set by MaxQuant.) For the human IPI database and including the reversed sequences, this corresponds to a list of 174 618 protein sequences resulting in 7 837 653 peptide sequences and 76 937 183 modification-specific peptides (without taking modification positioning into account). These numbers can become even larger, for example in cases where one wants to search against a six-frame translation of the whole genome.

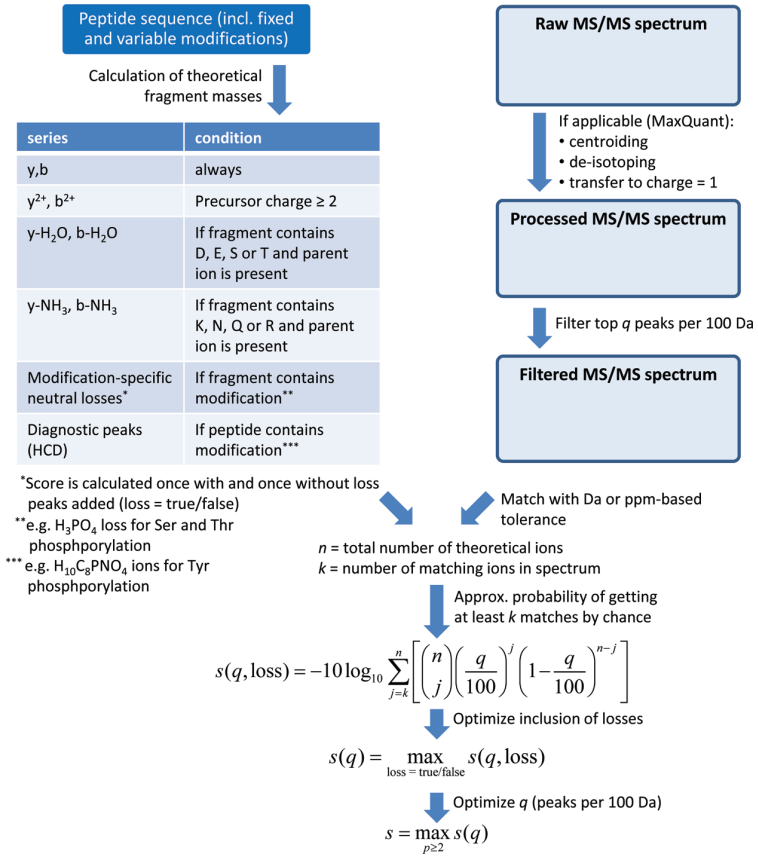


Figure 3. Schematic of the peptide scoring algorithm. The upper left branch shows the calculation of the theoretical fragment ion masses while the right branch indicates the processing of the experimental MS/MS spectra. In particular, all ion types that are used for the scoring can be found in the table on the left. The final score involves an optimization of the number of highest intensity peaks that are taken into account per 100 Da m/z interval and over the inclusion of modification-specific neutral losses.

We therefore wished to be able to handle protein sequence information without limitation on the sizes of calculated protein and peptide lists. Our goal was to work within the memory limits of 32-bit operating systems, which is around 1.6 GB from within the Microsoft .NET framework. The data structures for the search engine have to have an even smaller memory footprint since other data might be required to be in memory at the same time. Obviously the full modification-specific peptide list is too large to keep in memory and it has to reside on the hard disk (or solid state disk for improved performance). This is also true for the peptide and protein lists because unlimited scalability is desired. Only an index for each of the files is kept in memory, which contains positions of the records relative to the beginning of the file. These memory indices can already exceed the memory limitations for very large numbers of peptides. Therefore the index points to beginnings of blocks of elements in the file with a suitably chosen block size such that the lengths of the indices in memory never exceed a fixed size. In Figure 2, the structure of these lists and the relationships between memory and files residing on the hard disk are shown for proteins, peptides and modification-specific peptides. The records always contain indices to the respective items in the hierarchy above, assuring easy navigation from a candidate peptide to all the proteins that it occurs in. The modification-specific peptide list is the one that is directly accessed in database searches. It is sorted by mass, which allows quick retrieval of peptides within the given mass window. Protein and peptide list are instead sorted alphabetically by protein name and peptide sequence, respectively.

Scoring Model

The probabilistic score employed in Andromeda is derived from the p-score that was introduced for the identification of MS^3 spectra.[48] Given a peptide sequence together with a configuration of fixed and variable modifications for that peptide, first the theoretical fragment ions are calculated (Figure 3). For CID and HCD the list of theoretical fragment ion masses always contains the singly charged b- and y-ions. If the precursor charge is greater than one, the doubly charged b- and y-ions are added. In case of low resolution ion trap MS/MS spectra the charge state of fragments usually cannot be determined. The calculated doubly charged m/z values are then added explicitly if it is desired to match more highly charged fragments. For high-resolution MS/MS the charge state can be assigned to a fragment if more than one isotopic peak is detected. For these cases we remove peaks of fragments with charge higher than 1 from the spectrum and reintroduce them into the spectrum as singly charged fragment ions. If there are several charge states for a fragment their intensities are added, taking account of the fact that signal is proportional to charge in the Orbitrap analyzer. We noticed that even for high-resolution MS/MS data, where charge state

detection is possible in general, it is beneficial to consider doubly charged b- and y-ions as well. This is because for lower mass fragments sometimes only the monoisotopic peak is detectable precluding charge state determination and hence also the transformation to charge state one. For example assuming that the elemental composition of fragments follows the averagine model[57] the ratio between the ^{13}C and monoisotopic peak intensities for a fragment of 400 Da is 4.6:1. For less abundant fragments this can obviously lead to nondetection of the ^{13}C peak while the monoisotopic peak is above the noise level.

Calculated peaks corresponding to water and ammonia losses are only offered for matching as singly charged ions in those cases where the main b- and y-ion fragment is present and contains the amine-, amide- or hydroxyl-containing amino acid side-chains that tend to lead to the respective side chain loss. Modification-specific losses are configurable in the program AndromedaConfig, which is included in the MaxQuant distribution. The above-mentioned modification-specific neutral losses, as well as ions that are diagnostic for the presence of a particular modification of an amino acid type can be freely configured there. For example, the loss of phosphate from a phosphorylated serine or threonine is much more likely than from a tyrosine, which instead produces a highly specific immonium ion at mass 216.0426 (see, e.g., Steen et al.[44]). If Andromeda is used within MaxQuant, the report for each modification site includes presence or absence of a diagnostic peak in the MS/MS spectrum. The score is calculated once including configurable neutral losses and once excluding them and the maximum of the two scores is chosen. (Note that all scoring procedures are carried out identically for sequences from the reverse database, so they do not introduce a bias.)

The first step in the actual calculation of the score is to count the number of matches k between the n theoretical fragment masses and the peaks in the spectrum. The higher k is compared to n, the lower the chance that this happened by chance.[48] Because there are many signals in MS/MS spectra, including many low intense noise signals, the number of peaks in a defined mass interval—here 100 Th, which is the typical distance between consecutive members of fragment series (average mass of amino acids)—are limited to a maximum number. The parameter q is defined as the number of allowed peaks in the mass interval and it is needed to calculate the probability of a single random match. If the difference between calculated and measured masses is less than a predefined value, a match is counted. This can be done with an absolute mass tolerance window specified in Th or a relative mass window specified in ppm. While the former is appropriate for ion trap spectra, the latter is more suitable for high-resolution FT-ICR or Orbitrap spectra.

The Andromeda score is calculated as -10 times the logarithm of the probability of matching at least k out of the n theoretical masses by chance as shown in Figure 3. This is slightly different from Olsen et al.,[48] where the probability of matching exactly k out of n theoretical masses is determined. The formula used here is more similar to a definition of a p value for the null hypothesis that there is no similarity between the theoretical mass list and list of the spectrum masses. In particular, the score has the desirable property to vanish for $k = 0$. The calculation of the probability is only approximate since the probability for a single random match is taken to be $q/100$, which is exact if there was only one possible match per nominal mass. For high resolution MS/MS data the true random match probability is considerably less than this and the true score would be higher but more complicated to calculate. However, this simplification is conservative as it decreases the calculated score and is justified by the excellent performance of the search algorithm on high-accuracy MS/MS data.

The intensities of the peaks in the MS/MS spectra are indirectly taken into account by calculating the score for all values for q (number of peaks per 100 Th) up to the specified maximum. The best of these scores for varying q is selected. Therefore two spectrum-sequence comparisons with the same values for n and k can result in different scores depending on the intensities of the matched peaks. Generally, the score is higher if the matches are among the more intense peaks because the optimal value of q will be lower (see formula in Figure 3). However, we have found it crucial that this intensity weighting is not done on the overall intensity scale over the whole spectrum, but that it is restricted to local mass regions (e.g., the 100 Th mass range intervals.). This compensates for underlying global peak density distributions which typically favor small fragment masses.

The inclusion of additional information like peptide length, number of modifications or of missed cleavages can aid the specificity of peptide assignments to spectra. Ideally this is done in a data-dependent manner in which different weights for different classes of peptides can be derived from the data by machine learning in a Bayesian framework. We wished to include such a weighting of peptide classes into the score while retaining a basic search engine score that is deterministic and only depends on the spectrum being scored rather than the ensemble of all other spectra. To capture the dependence of the score on peptide mass and on the number of modifications we introduced a fixed additive component to the Andromeda score, which depends on the number of modifications and is a linear function of the mass. The specific values are determined in a manner that adjusts the distributions of reverse hits from a target-decoy search so that they become equal. The net effect of this procedure is to minimize the FDR for a given cutoff value, because it does not depend on peptide mass and modification state any longer. We used a large data set of MS/MS spectra and incorporated the specific weights into the scoring function. A data-dependent Bayesian scoring can still be applied to the output of the Andromeda search engine. For instance, MaxQuant additionally performs a peptide length dependent Bayesian analysis in a data dependent manner.[31]

Comparison to the Mascot Search Engine

Mascot[36] is a widely used standard for database searching and most other search engines have been compared to Mascot. Therefore we investigated how Andromeda compares to Mascot in terms of scoring of peptide-spectrum matches. As the exact details of the Mascot scoring system are not known, we compared the performance of Andromeda vs Mascot empirically on very large sets of proteomic data.

In Figure 4a, we plot the Mascot score against the Andromeda score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement as described in Materials and Methods. For each MS/MS spectrum the highest scoring peptide is taken which is not necessarily the same for the Mascot and the Andromeda scoring. In Figure 4b, the fraction of cases for which the top-scoring Andromeda and Mascot peptide sequences coincide is displayed as a histogram depending on the Andromeda score. As can be seen, above an Andromeda score of 100 the top-scoring peptides coincide in almost all cases.

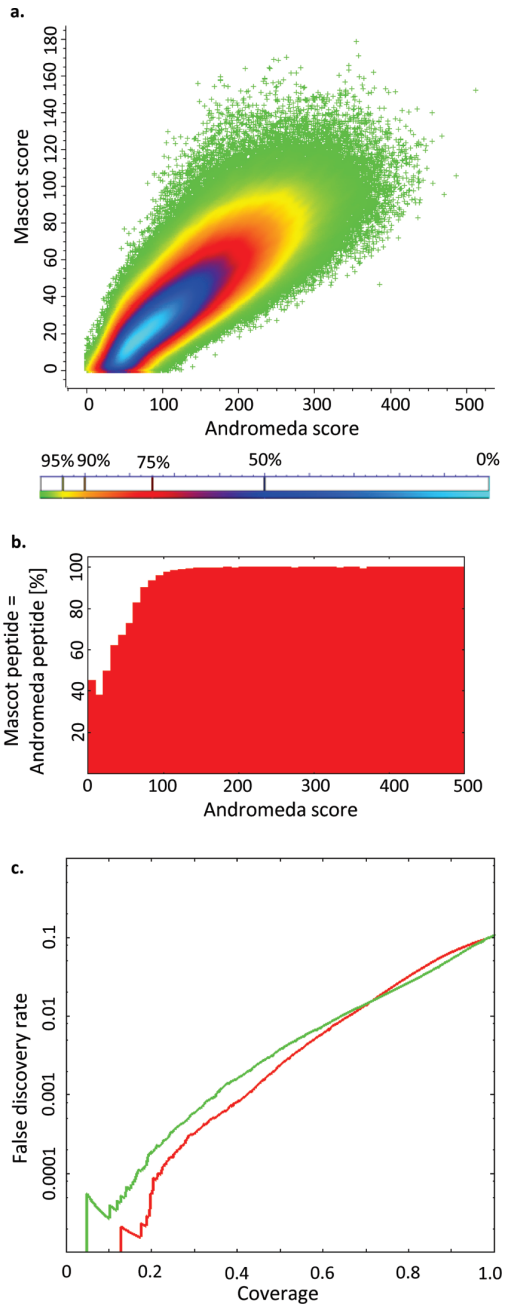


Figure 4. (a) Andromeda vs Mascot score for a data set of 732 287 MS/MS spectra derived from a label-free mouse proteome measurement.[53] The score for the top-scoring peptide for each MS/MS spectrum is shown which is not necessarily the same peptide sequence for the Mascot and the Andromeda identification. The color code indicates the percentage of points that are included a region of a specific color. (b) Histogram of the percentage of cases in which the top-scoring Andromeda and Mascot peptide sequences are equal as a function of Andromeda score. For the comparison leucine and isoleucine were treated as the same amino acid. (c) False discovery rate as a function of coverage for the same data set calculated based on the reverse hits from the target-decoy search.

Of the recorded MS/MS spectra, 89.1% correspond to unmodified peptides and most of the identified modified peptides have an oxidized methionine. The point density is indicated by the color code in Figure 4a which encodes the percentage of points that are included a region of a specific color. For example, the yellow line in Figure 4a encloses 95% of all data points. This visualization allows the visual detection of outliers (like a two-dimensional data plot), while at the same time retaining information about the density of points that would normally only be visible in a 3D data plot. It is immediately apparent from the figure that the scores correlate well overall. There are no distinct populations of peptides that are only identified by one of the search engines. A linear regression results in the equation $M = 0.311 * A - 32.231$, where M is the Mascot score and A the Andromeda score, with an R^2 value of 0.708. This indicates that Andromeda scores are generally about 3-fold larger than Mascot scores. However, this does not indicate a 3-fold larger confidence. The statistical power is better determined by calculating coverage and false discovery rates as a function of score threshold as is done below. A rough conversion between Andromeda and Mascot scores can be performed by a division by three or application of the regression line. Note that there are only very few and dispersed outliers on either side; of the order of tens of spectra out of the total of more than 700 000. Furthermore, there are virtually no high-scoring outliers near either axis, indicating an absence of spectra that were ranked highly with one method but scored close to zero with the other. This demonstrates that no populations of peptides would be lost entirely by employing one score or the other.

Next we compare the performance of the Andromeda and Mascot search engines as a function of False Discovery Rates estimated as the number of hits from the reverse database divided by the number of forward hits at any given minimum score. The sensitivity of the database search is defined as the number of accepted forward hits relative to the total number of forward hits at the same score. Mascot and Andromeda have very similar characteristics over the whole range of FDRs, in particular including the often used 1% FDR rate (Figure 4b). This shows that the two scores are very close in discriminatory power.

Scoring of Phosphopeptides

Figure 5a shows the same type of plot as in Figure 4a but for a data set that is enriched for phosphopeptides. Of the recorded 586 883 MS/MS spectra in Figure 5a, 27.4% have one or more phosphorylations. Outliers are visible in the region of high Andromeda and low Mascot score and most of them correspond to peptides with three to five phosphorylation events. Figure 5b displays the MS/MS spectrum of a peptide with five phosphorylation sites that has a Mascot score of 5.2 and an Andromeda score of 199.3. The y-series coverage is almost complete with most fragments occurring with a neutral loss of a phosphate molecule. An FDR coverage curve for the phosphopeptide data set is depicted in Figure 5c. The performances of Mascot and Andromeda are similar over the entire range with an advantage for Andromeda in the high specificity region. At the typical operation point of 1% FDR results are very close. We speculate that the better scoring in the region of higher specificity may be due to a better matching of spectra of phosphopeptides in Andromeda due to more comprehensive combinatorics of positioning of phospho-groups on the available serine, threonine and tyrosine sites in the peptide sequences, including a more complete offering of neutral losses. During the Andromeda search we offer up to 1000 positionings of variable modifications within any given peptide which is exhaustive for most situations.

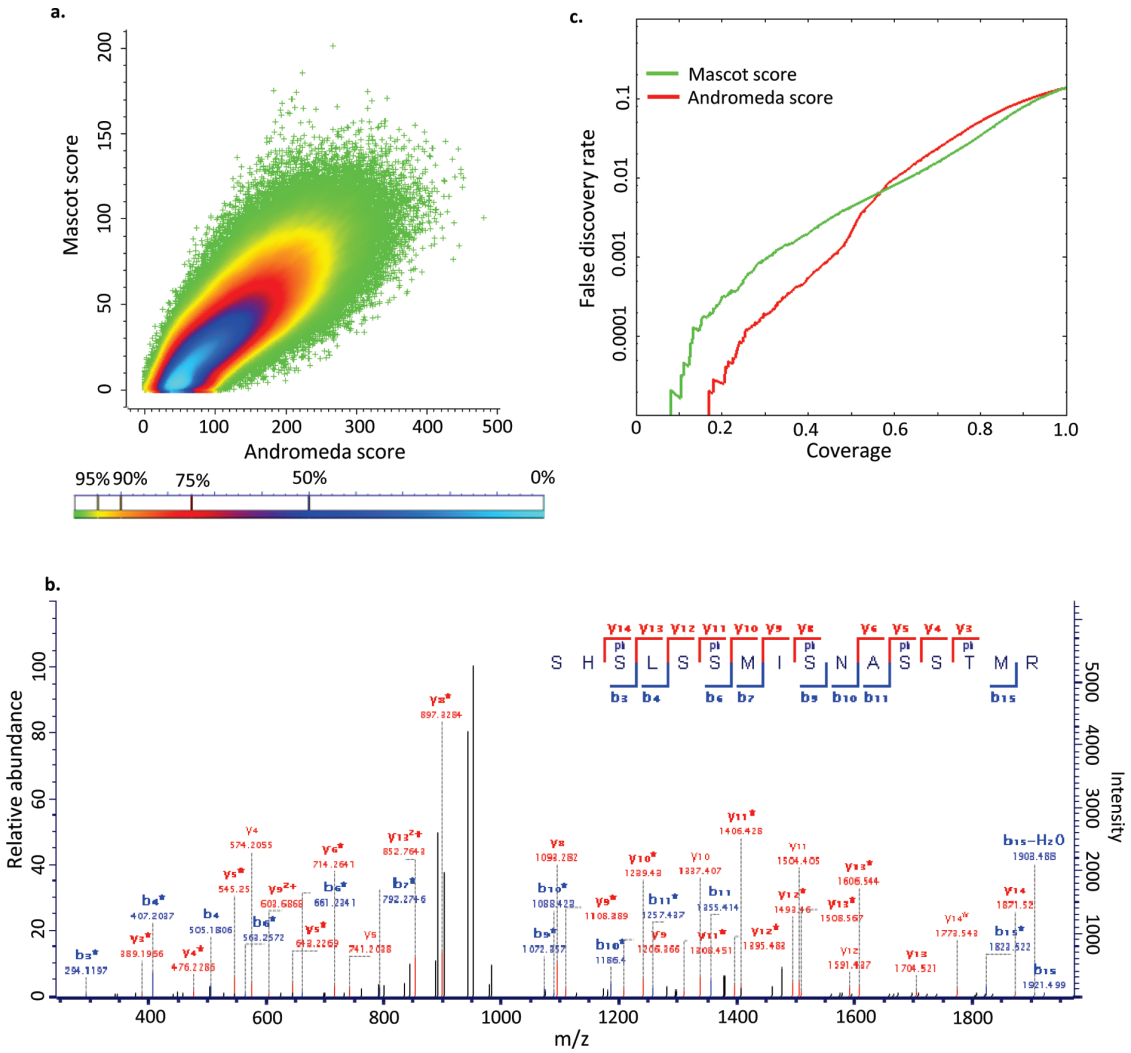


Figure 5. (a) Andromeda vs Mascot score for 586,883 MS/MS spectra from the phospho-proteome data by Hilger et al.[54] (b) Annotated MS/MS spectrum of the peptide SHpSLSpSMIpSNApSSpTMR. Mascot and Andromeda produce the same top-scoring peptide sequence with a Mascot score of 5.2 and an Andromeda score of 199.3. (c) False discovery rate as a function of coverage for the same data set calculated in the same way as in Figure 4c.

In MaxQuant, the top-scoring peptide is furthermore rescored with essentially exhaustive positioning of modifications. We merely restrict the combinatorics to 100 000 possibilities to exclude the rare instances where single peptides cause long calculation times due to "combinatorial explosion". In Supplementary Figure 1 (Supporting Information), the same data as in Figure 5a is shown six times—each time highlighting another population of top-scoring peptides with a fixed number of phosphorylations. Peptides with higher phosphorylations tend to have many data points in the high Andromeda score but low-to-moderate Mascot score region further indicating that Andromeda performs better on highly phosphorylated peptides.

Identification of Second Peptides

Even in high-resolution MS, the selection of the precursor ion for fragmentation is always performed with low resolution (typically a few Th) to ensure adequate sensitivity for MS/MS. In complex mixtures, this results in frequent cofragmentation of coeluting peptides with similar masses. These 'chimerical' MS/MS spectra[52] can be detrimental for identification of the peptide of interest, especially if the cofragmented peptide is of comparable intensity. Co-fragmentation generally reduces the number of peptides identified in database searches and poses special problems for reporter fragment based quantification methods because both peptides contribute to the measured ratios.

However, this situation can be turned to an advantage if both peptides can be identified. In particular, this presents the opportunity to identify peptides that have not been targeted for MS/MS and to obtain two or more peptide identifications from a single MS/MS spectrum. Although this problem has been addressed before,[49−52] to our knowledge it has not been adopted in mainstream search engines yet. Here we describe a second peptide identification algorithm that we have integrated into the Andromeda/MaxQuant workflow.

To illustrate the principles of our algorithm, Figure 6a shows an LC−MS map, where 3D peaks are indicated as lines marking the peak boundaries. The blue isotope pattern has been selected for fragmentation at the position of the cross on the

dx.doi.org/10.1021/pr101065j |J. Proteome Res. 2011, 10, 1794–1805

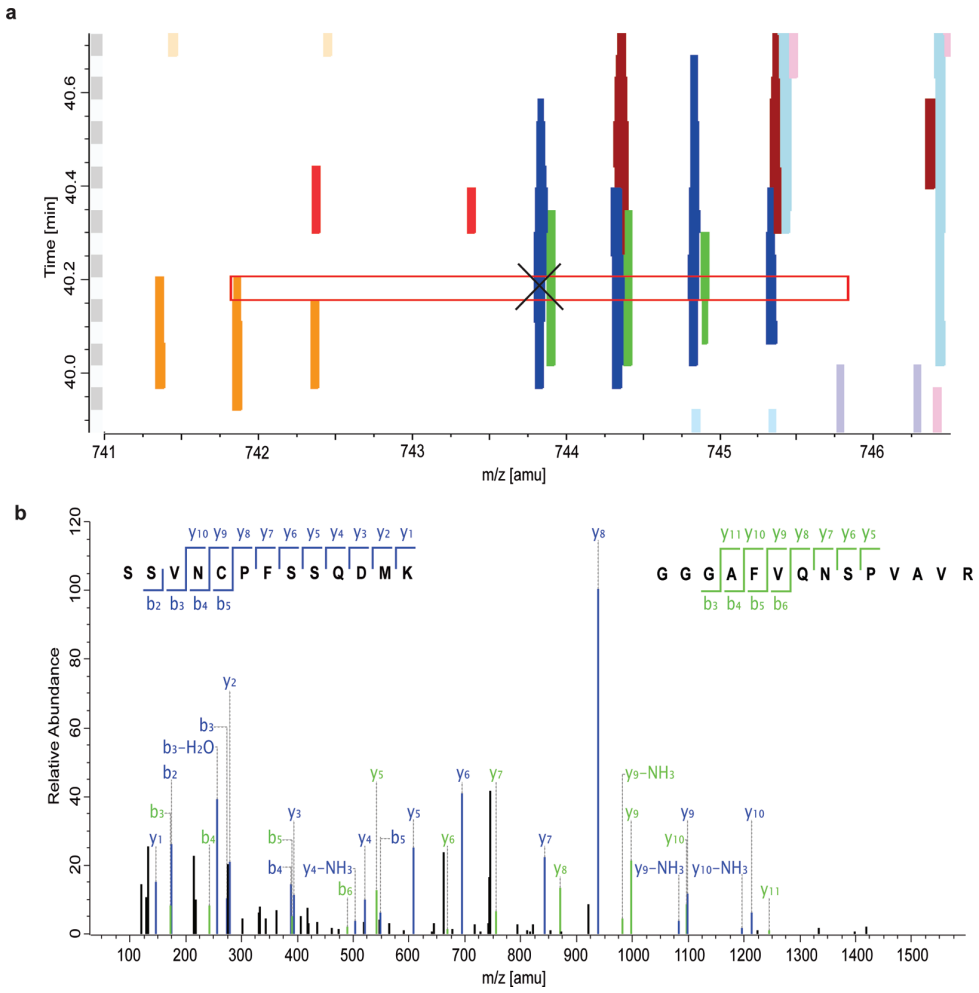


Figure 6. Second peptide identification. (a) LC−MS map of the sequenced (blue) and cofragmented (green) peptide described in the main text. The blue peptide has been selected for fragmentation at the position of the cross. The red rectangle indicates the isolation window. (b) MS/MS spectrum leading to the identification of both peptides. Fragments of the two peptides are indicated in blue and green, respectively. The blue peptide is identified in the conventional database search while the green peptide has been identified as "second peptide".

monoisotopic peak of that peptide. The red rectangle indicates the region from which ions have been isolated for fragmentation. Clearly the peptide corresponding to the green isotope pattern that has not been selected for sequencing intersects with the isolation rectangle. Therefore its fragments should be present in the MS/MS spectrum as well. The actual fragment spectrum is shown in Figure 6b where the fragments originating from the targeted and identified peptide (blue isotope pattern) are indicated in blue. This process is repeated for the entire LC−MS/MS run. For every 3D MS isotope pattern that has not been selected for sequencing the algorithm checks whether it intersects with the isolation window of any MS/MS spectrum. If this is the case then the fragments in this MS/MS spectrum that have already been assigned to a peptide sequence during the main Andromeda search are subtracted. The remaining fragments are submitted to a new database search with the precursor mass from the peptide that was not targeted for MS/MS. The collection of these "subtracted" peak lists is submitted to Andromeda in the same way as in a conventional search. However, the results of this second peptide search are further processed with their own peptide length based posterior error probability and precursor mass filtering. Since these spectra are on average of lower quality than the original MS/MS spectra we have found it to be crucial that they have their own data-dependent statistical model for peptide identification. The resulting peptides are then accepted up to a 1% FDR and integrated into the usual protein identification and quantification workflow.

The HCD data set used for testing (see Materials and Methods) was acquired with a total isolation width of 4 Th for every MS/MS spectrum. The identification rate for the set of second peptide spectra is much lower compared to the normal MS/MS identification rate of 50%. Nevertheless, since the number of the second peptide spectra is quite high compared to normal MS/MS spectra considering cofragmentation still leads to a considerable increase in peptide identifications. In our example, the number of identified peptide features increased by 10.7% by the inclusion of second peptide identifications. The gain in the number of identified peptides depends on the isolation width for the acquisition of MS/MS spectra. For instance, at an isolation width of 2 Th we observe that the increase

Journal of Proteome Research ARTICLE

in identified peptides through second peptide identifications is only 5.7%. The relative gain is larger at increased isolation width because the average number of additional peptides within the window increases. However, the chance to identify the main peptide decreases due to the mixing of the spectrum with fragments from other peptides. The dependence of the number of peptide identifications for conventional and second peptides is shown in Supplementary Figure 2 (Supporting Information).

DISCUSSION AND OUTLOOK

Here we have described Andromeda, a novel search engine for matching MS/MS spectra to peptide sequences in a database. Andromeda can either be used in a stand-alone mode or—more typically—as part of the MaxQuant environment. Apart from an optimal scoring model our intention was to develop a very robust architecture with unlimited scalability. We have demonstrated this on large scale data sets with hundreds of thousands of spectra. Andromeda has been "stress tested" in ongoing studies and has been the default search engine in our laboratory for some time. A practical advantage of the MaxQuant/Andromeda combination is that it runs locally on the user's computer. This eliminates client-server set up and communication issues. The computational proteomics pipeline starting from raw data files to reported protein groups and their quantitative ratios now appears unified to the user. Despite the local search architecture, processing speeds are generally not different from the previous MaxQuant/Mascot environment in which Mascot was run on an external server. Furthermore, we have added a separate module called Perseus (www.maxquant.org), which performs bioinformatic analysis of the output of the MaxQuant/Andromeda workflow. Perseus is already available and in use[58] and completes the pipeline for computational proteomics analysis but will be described in a future publication (Cox et al., in preparation).

The scoring function at the heart of Andromeda is built on a simple binominal distribution probability formula (Figure 3), which we have previously used in scoring MS^3 spectra and localizing PTMs.[59] Andromeda divides the MS/MS spectrum into mass ranges of 100 Th. In each of these ranges the number of experimental peaks offered for matching is dynamically tested in an intensity prioritized manner.

False discovery rates for the same initial probability score can still depend on the number of modifications and on the mass of the peptide. This is accounted for in Andromeda by an additive component to the score. Comparison to Mascot on very large data sets reveals very few outliers—in particular almost no peptides are exclusively identified by one of the two search engines. Furthermore, the coverage of identified peptides at any given FDR is likewise similar, including at the generally used operating point of 1% expected false positives. We did notice improved identification of heavily modified peptides in Andromeda compared to Mascot, which we attribute to the more exhaustive combinatorial analysis of placing PTMs on all possible amino acids. As the Mascot search engine has become one of the standards in proteomics, equivalent performance fulfills the goal that we had set for the development of Andromeda and likely implies favorable comparison to other search engines as well. Apart from describing the score we have also made the actual code used in Andromeda available for inspection with this publication (Supporting Information 1).

A key advantage of Andromeda is its extensibility. For example, proteomics with high accuracy MS and MS/MS data (high−high mode[60]), is becoming increasingly common. Andromeda, in contrast to Mascot, allows arbitrarily accurate MS/MS requirements specified in ppm. Similarly, Mascot precludes identification of SILAC pairs if the same amino acid can bear a fixed and a variable modification. This causes a substantial loss of quantification information, for example in the analysis of lysine acetylated peptides[61] because all MS/MS spectra of lysine-acetylated peptides that were sequenced on the heavy SILAC partner will not be identified by Mascot. All these quantitative ratios are retrieved in the MaxQuant/Andromeda workflow.

More generally, additional scoring modes can be added to Andromeda. We demonstrated this by implementing a second peptide identification algorithm into the MaxQuant/Andromeda workflow. For each isotope cluster that is detected in the LC−MS data but that was not targeted for fragmentation the algorithms checks if the precursor isotope pattern intersects the selection window of any MS/MS event. If so, fragment ions belonging to the identified peptide are subtracted and the search is repeated with the cofragmented peptide in a statistically rigorous way. As demonstrated here, this leads to an appreciable increase in peptide and protein identifications in complex mixtures. As another example, special algorithms are necessary for peptide identification in data independent MS/MS where the whole mass range is fragmented.[62,63] Using the MaxQuant/Andromeda infrastructure our group recently developed an implementation of this principle on the Exactive instrument, which consists only of an Orbitrap analyzer with HCD capability.[64]

In conclusion, we have developed, described and tested a robust and scalable search engine that in combination with MaxQuant represents a powerful and unified analysis pipeline for quantitative proteomics, which is freely available to the community.

ASSOCIATED CONTENT

Supporting Information

Supplemental figures and materials. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*J.C. e-mail cox@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2088 or M.M. e-mail mmann@biochem.mpg.de, fax +49 (89) 8578 3209, phone +49 (89) 8578 2557.

ACKNOWLEDGMENT

We thank other members of our groups in Martinsried and Copenhagen for fruitful discussion and help. We also thank early adopters of Andromeda in other laboratories for helpful feedback. This work was partially supported by the European Union seventh Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.

(2) Yates, J. R., 3rd; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6* (9), 702–14.

(3) Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science* **2006**, *312* (5771), 212–7.

(4) Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007** *130* (3), 395–8.

(5) Vermeulen, M.; Selbach, M. Quantitative proteomics: a tool to assess cell differentiation. *Curr. Opin. Cell Biol.* **2009**, *21* (6), 761–6.

(6) Choudhary, C.; Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell. Biol.* **2010**, *11* (6), 427–39.

(7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–82.

(8) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–7.

(9) Nesvizhskii, A. I.; Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today* **2004**, *9* (4), 173–81.

(10) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (4), 419–34.

(11) Chalkley, R. J.; Hansen, K. C.; Baldwin, M. A. Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Methods Enzymol.* **2005**, *402*, 289–312.

(12) Colinge, J.; Bennett, K. L. Introduction to computational proteomics. *PLoS Comput. Biol.* **2007**, *3* (7), e114.

(13) Matthiesen, R. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* **2007**, *7* (16), 2815–32.

(14) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787–97.

(15) Deutsch, E. W.; Lam, H.; Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **2008**, *33* (1), 18–25.

(16) 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.

(17) Matthiesen, R.; Jensen, O. N. Analysis of mass spectrometry data in proteomics. *Methods Mol. Biol.* **2008**, *453*, 105–22.

(18) Bandeira, N.; Clauser, K. R.; Pevzner, P. A. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell. Proteomics* **2007**, *6* (7), 1123–34.

(19) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. Clustering millions of tandem mass spectra. *J. Proteome Res.* **2008**, *7* (1), 113–22.

(20) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **2008**, *7* (1), 286–92.

(21) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 47–50.

(22) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245–53.

(23) Rauch, A.; Bellew, M.; Eng, J.; Fitzgibbon, M.; Holzman, T.; Hussey, P.; Igra, M.; Maclean, B.; Lin, C. W.; Detter, A.; Fang, R.; Faca, V.; Gafken, P.; Zhang, H.; Whiteaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M. W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006**, *5* (1), 112–21.

(24) Rinner, O.; Mueller, L. N.; Hubalek, M.; Muller, M.; Gstaiger, M.; Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.* **2007**, *25* (3), 345–52.

(25) Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R., 3rd A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **2008**, *5* (4), 319–22.

(26) Brusniak, M. Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinform.* **2008**, *9*, 542.

(27) May, D.; Law, W.; Fitzgibbon, M.; Fang, Q.; McIntosh, M. Software platform for rapidly creating computational tools for mass spectrometry-based proteomics. *J. Proteome Res.* **2009**, *8* (6), 3212–7.

(28) Deutsch, E. W.; Shteynberg, D.; Lam, H.; Sun, Z.; Eng, J. K.; Carapito, C.; von Haller, P. D.; Tasman, N.; Mendoza, L.; Farrah, T.; Aebersold, R. Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. *Proteomics* **2010**, *10* (6), 1190–5.

(29) Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S. E.; Rigbolt, K. T.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J.; Blagoev, B.; Andersen, J. S.; Mann, M. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* **2010**, *9* (1), 393–403.

(30) Kumar, C.; Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **2009**, *583* (11), 1703–12.

(31) 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), 1367–72.

(32) Cox, J.; Mann, M. Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1477–85.

(33) Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7* (12), 952–8.

(34) 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* **2002**, *1* (5), 376–86.

(35) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251–4.

(36) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–67.

(37) 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*, 976–89.

(38) Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999**, *71* (14), 2871–82.

(39) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2* (10), 1406–12.

(40) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–7.

(41) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958–64.

(42) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W701–6.

(43) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally

modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–39.

(44) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5* (9), 699–711.

(45) Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(46) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–9.

(47) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6* (3−4), 327–42.

(48) Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (37), 13417–22.

(49) Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**, *5* (16), 4096–106.

(50) Bern, M.; Finney, G.; Hoopmann, M. R.; Merrihew, G.; Toth, M. J.; MacCoss, M. J. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **2010**, *82* (3), 833–41.

(51) Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **2010**, *9* (7), 1476–85.

(52) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–60.

(53) Luber, C. A.; Cox, J.; Lauterbach, H.; Fancke, B.; Selbach, M.; Tschopp, J.; Akira, S.; Wiegand, M.; Hochrein, H.; O'Keeffe, M.; Mann, M. Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **2010**, *32* (2), 279–89.

(54) Hilger, M.; Bonaldi, T.; Gnad, F.; Mann, M. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **2009**, *8* (8), 1908–20.

(55) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **2004**, *4* (7), 1985–8.

(56) Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R.; Zhang, H. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* **2009**, *37* (Database issue), D555–9.

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

(58) Geiger, T.; Cox, J.; Mann, M. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* **2010**, *6* (9), e1001090.

(59) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, In Vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635–48.

(60) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (47), 18132–8.

(61) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **2009**, *325* (5942), 834–40.

(62) Geromanos, S. J.; Vissers, J. P.; Silva, J. C.; Dorschel, C. A.; Li, G. Z.; Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **2009**, *9* (6), 1683–95.

(63) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77* (7), 2187–200.

(64) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **2010**, *9* (10), 2252–61.

2.2 article 2 - Expert System for Computer Assisted Annotation of MS/MS Spectra

Nadin Neuhauser, Annette Michalski, Jürgen Cox, and Matthias Mann
Mol Cell Proteomics 2012, 11, 1500-1509

As described before, the main method to identify a peptide by a MS/MS spectrum is by comparing the observed with a theoretical spectrum. Probability based search engines - like Mascot or Andromeda - use mainly the regular fragment series such as b- and y-ions for this purpose. However, peptide fragmentation is far from trivial and some peaks can not be described by common backbone breakage, regardless of the identification score. If some high abundant peaks are not explained by conventional fragmentation rules, the user may doubt the identification. In principle the remaining ions can be explained by peptide fragmentation rules described in the specialized literature, but this would require substantial expertise and effort. For this reason we wanted to give novices in the area of peptide fragmentation a tool to find a explanations for - at least - all high abundant peaks. This tool is a so-called 'Expert System' and it supports beginners and advanced users alike with specialized knowledge from the mass spectrometric fragmentation part of their experiments. For the latter, it can also be used to focus on unusual fragmentation events, perhaps leading to the discovery of new fragmentation rules. This would be especially valuable in case of modified peptides, which often have complicated fragmentation spectra. Using a rule base technique has the advantage that the extension of the knowledge is also possible for researchers without a background in computer science. It also enables exclusion of particular rules from the existing knowledge base.

The knowledge base of the Expert System consists of fragmentation rules for tryptic HCD spectra that are known from the literature or by intensive testing of hypotheses (for more details see next article). Differently from a human expert, in our approach we can estimate the risk of false annotation by calculating a FDR for the rule set using well controlled high-throughput projects with thousands of spectra interpretations. For this purpose, I devised a novel approach in which the FDR is calculated by counting the false annotations of inserted independent peaks into a spectrum. A measurement about the completeness of the annotation is the intensity coverage, which represents the percentage of peak intensities explained by the Expert System. I showed that the Expert System can significantly increase the intensity coverage from 58% with regular annotation to 98%, while the chance for a false positive annotation is below 5%.

So far, the annotation by the Expert System is not part of the search engine score nor does it have any influence on the peptide identification. One reason for this is that increasing the number of possible fragment types would lower the score for statistical reasons. This is because the likelihood of observing all fragment types is not the same. However, Andromeda and the Expert System were developed at the same time and this meant that knowledge gained through the Expert System helped to improve Andromeda. For example, Andromeda uses backbone fragments that have lost water (H_2O) or ammonia (NH_3) when the peptide sequence contains specific basic amino acid residues. This is common knowledge, but it became part of Andromeda, only after statistical relevance was shown by investigations with the Expert System.

The Viewer software - which was partly developed by me - is part of the MaxQuant framework and allows the user to manually inspect the quality of all identified MS/MS spectra using the Expert System. To full fill the publication criteria of some journals[129], these spectra can be efficiently exported to bitmap images (png, jpg) or also vector graphics (pdf). Additionally a web service exists where a user can submit an individual spectrum - together with the peptide sequence - to inspect the annotation by the Expert System independent from the MaxQuant environment. I have developed the knowledge base in cooperation with my co-author Annette Michalski. My responsibility were all computational tasks including the implementation of knowledge base and the Expert System.

Technological Innovation and Resources

Expert System for Computer-assisted Annotation of MS/MS Spectra*[S]

Nadin Neuhauser‡¶, Annette Michalski‡¶, Jürgen Cox‡, and Matthias Mann‡§

An important step in mass spectrometry (MS)-based proteomics is the identification of peptides by their fragment spectra. Regardless of the identification score achieved, almost all tandem-MS (MS/MS) spectra contain remaining peaks that are not assigned by the search engine. These peaks may be explainable by human experts but the scale of modern proteomics experiments makes this impractical. In computer science, Expert Systems are a mature technology to implement a list of rules generated by interviews with practitioners. We here develop such an Expert System, making use of literature knowledge as well as a large body of high mass accuracy and pure fragmentation spectra. Interestingly, we find that even with high mass accuracy data, rule sets can quickly become too complex, leading to over-annotation. Therefore we establish a rigorous false discovery rate, calculated by random insertion of peaks from a large collection of other MS/MS spectra, and use it to develop an optimized knowledge base. This rule set correctly annotates almost all peaks of medium or high abundance. For high resolution HCD data, median intensity coverage of fragment peaks in MS/MS spectra increases from 58% by search engine annotation alone to 86%. The resulting annotation performance surpasses a human expert, especially on complex spectra such as those of larger phosphorylated peptides. Our system is also applicable to high resolution collision-induced dissociation data. It is available both as a part of MaxQuant and via a webserver that only requires an MS/MS spectrum and the corresponding peptides sequence, and which outputs publication quality, annotated MS/MS spectra (www.biochem.mpg.de/mann/tools/). It provides expert knowledge to beginners in the field of MS-based proteomics and helps advanced users to focus on unusual and possibly novel types of fragment ions. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M112.020271, 1500–1509, 2012.***

In MS-based proteomics, peptides are matched to peptide sequences in databases using search engines (1–3). Statistical criteria are established for accepted *versus* rejected peptide spectra matches based on the search engine score, and usually a 99% certainty is required for reported peptides. The search engines typically only take sequence specific backbone fragmentation into account (*i.e.* a, b, and y ions) and some of their neutral losses. However, tandem mass spectra—especially of larger peptides—can be quite complex and contain a number of medium or even high abundance peptide fragments that are not annotated by the search engine result. This can result in uncertainty for the user—especially if only relatively few peaks are annotated—because it may reflect an incorrect identification. However, the most common cause of unlabeled peaks is that another peptide was present in the precursor selection window and was cofragmented. This has variously been termed "chimeric spectra" (4–6), or the problem of low precursor ion fraction (PIF)[1] (7). Such spectra may still be identifiable with high confidence. The Andromeda search engine in MaxQuant, for instance, attempts to identify a second peptide in such cases (8, 9). However, even "pure" spectra (those with a high PIF) often still contain many unassigned peaks. These can be caused by different fragment types, such as internal ions, single or combined neutral losses as well as immonium and other ion types in the low mass region. A mass spectrometric expert can assign many or all of these peaks, based on expert knowledge of fragmentation and manual calculation of fragment masses, resulting in a higher degree of confidence for the identification. However, there are more and more practitioners of proteomics without in depth training or experience in annotating MS/MS spectra and such annotation would in any case be prohibitive for hundreds of thousands of spectra. Furthermore, even human experts may wrongly annotate a given peak—especially with low mass accuracy tandem mass spectra—or fail to consider every possibility that could have resulted in this fragment mass.

Given the desirability of annotating fragment peaks to the highest degree possible, we turned to "Expert Systems," a well-established technology in computer science. Expert Systems achieved prominence in the 1970s and 1980s and were meant to solve complex problems by reasoning about knowl-

From the ‡Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

Received May 5, 2012, and in revised form, July 19, 2012
⌘ *Author's Choice*—Final version full access.
Published, MCP Papers in Press, August 10, 2012, DOI 10.1074/mcp.M112.020271

[1] The abbreviations used are: PIF, Precursor Intensity Fraction; FDR, False Discovery Rate; MS/MS, Tandem mass spectrometry; HCD, Higher Energy Collisional Dissociation; PEP, Posterior Error Probability; PDF, Portable Document Format; IM, immonium ion; SC, side chain fragment ion; Th, Thomson.

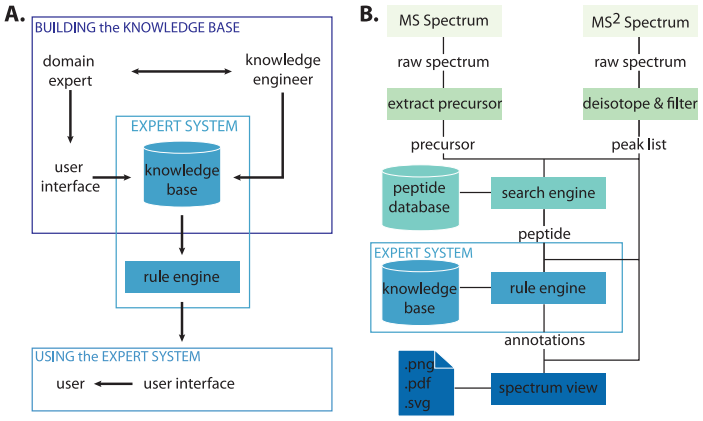


FIG. 1. **Basic concept of the Expert System.** *A*, An Expert System is constructed by interviewing an expert in the domain (here peptide fragmentation and the accumulated literature) and devising a set of rules with associated priority and dependence on each other. The knowledge base contains the rules whereas the rule engine is generic and applies the rules to the data. *B*, Data are automatically processed following the steps depicted.

edge (10, 11). Interestingly, one of the first examples was developed by Nobel Prize winner Joshua Lederberg more than 40 years ago, and dealt with the interpretation of mass spectrometric data. The program's name was Heuristic DENTRAL (12), and it was capable of interpreting the mass spectra of aliphatic ethers and their fragments. The hypotheses produced by the program described molecular structures that are plausible explanations of the data. To infer these explanations from the data, the program incorporated a theory of chemical stability that provided limiting constraints as well as heuristic rules.

In general, the aim of an Expert System is to encode knowledge extracted from professionals in the field in question. This then powers a rule-based system that can be applied broadly and in an automated manner. A rule-based Expert System represents the information obtained from human specialists in the form of IF-THEN rules. These are used to perform operations on input data to reach appropriate conclusion. A generic Expert System is essentially a computer program that provides a framework for performing a large number of inferences in a predictable way, using forward or backward chains, backtracking, and other mechanisms (13). Therefore, in contrast to statistics based learning, the "expert program" does not know what it knows through the raw volume of facts in the computer's memory. Instead, like a human expert, it relies on a reasoning-like process of applying an empirically derived set of rules to the data.

Here we implemented an Expert System for the interpretation for high mass accuracy tandem mass spectrometry data of peptides. It was developed in an iterative manner together with human experts on peptide fragmentation, using the published literature on fragmentation pathways as well as large data sets of higher-energy collisional dissociation (HCD) (14) and collision-induced dissociation (CID) based peptide identifications. Our goal was to achieve an annotation performance similar or better than experienced mass spectrometrists (15), thus making comprehensively annotated peptide spectra available in large scale proteomics.

EXPERIMENTAL PROCEDURES

The benchmark data set is from Michalski *et al*.[2] Briefly, *E. coli*, yeast and HeLa proteomes were separated on 1D gel electrophoresis and in gel digested (16). Resulting peptides were analyzed by liquid chromatography (LC) MS/MS on a linear ion trap - Orbitrap instrument (LTQ Velos (17) or ELITE (18), Thermo Fisher Scientific). Peptides were fragmented by HCD (14) or by CID, but in either case fragments were transferred to the Orbitrap analyzer to obtain high resolution tandem mass spectra (7500 at *m/z* 400). We scanned tandem mass spectra already from *m/z* 80 to capture immonium ions as completely as possible. Data analysis was performed by MaxQuant using the Andromeda search engine (8, 9). Maximum initial mass deviation for precursor peaks was 6 ppm and maximum deviation for fragment ions for both the search engine and for the Expert System was 20 ppm. MaxQuant preprocessed the spectra to be annotated by the Expert System in the same way as it does for the Andromeda search engine: Peaks were filtered to the 10 most abundant ones in a sliding 100 *m/z* window, de-isotoped and shifted to charge one where possible. From this data, sequence-spectra pairs were selected that had a certainty of identification of 99.99% PIF values (7) larger than 95% and that were sequence unique (more than 16,000 peptides).

The Expert System was written in the programming language C#, using the Microsoft .NET framework version 3.5 and the Workflow. Activities library, which contains a rule engine to implement an Expert System (Microsoft Corporation, Redmond, WA).

MaxQuant contains the Expert System as an integrated option in its Viewer—the component that allows visualization of raw and annotated MS data. MaxQuant can freely be downloaded from www.maxquant.org. It requires Microsoft .NET 3.5, which is either already installed with Microsoft Windows or can be installed as a free Windows update. In our group we have implemented the Expert System both on a Windows cluster and in a desktop version. Additionally, we provide an Expert System web server, which can be accessed at

[2] Michalski, A., Neuhauser, N., Cox, J., and Mann, M., unpublished data.

www.biochem.mpg.de/mann/tools/. Although MaxQuant allows the Expert System annotation of arbitrary numbers of MS/MS spectra, the webserver is currently limited to the submission of one MS/MS spectrum at a time. After upload of a list of peaks with *m/z* value and their intensities—together with the corresponding peptide sequence—the spectrum with all annotations is displayed. This can then be exported in different graphical formats.

RESULTS AND DISCUSSION

Construction of the Expert System—Human experts perform a generic set of tasks when solving problems such as the interpretation of an MS/MS spectrum. These rules have to be codified in the Expert System, mainly in the form of a series of IF-THEN rules. Fig. 1 shows the major steps involved in building and using the Expert System. It is important to acquire all relevant rules to interpret MS/MS spectra as comprehensively as possible. However, to avoid over-annotation leading to false positives (see below), the number of rules and their interactions should not become too large. This balance was struck by evaluating the performance of different set of rules on large data sets in conjunction with human experts.

Rules were encoded in a table-like structure, where they could be activated, deactivated or modified. To create the knowledge base, the extent of interactions of the rules also had to be determined—for instance, which combination of neutral losses to allow. After iterative construction of the knowledge base, the rule engine then applied the encoded knowledge to MS/MS spectra and displayed the result to the user (Fig. 1*A*). The processing steps that are performed on the raw MS and MS/MS spectra are shown in Fig. 1*B* (see also EXPERIMENTAL PROCEDURES). Note that the workflow is entirely automated and that user interaction is possible but not required. Arbitrary numbers of annotated spectra of peptides of interest can be produced as interactive screen images or high resolution, printable PDF files. The Expert System is very fast, and 16,000 spectra can be annotated in less than four hours on a desktop system.

The IF-THEN constraints of our Expert System can be divided into four major parts (Fig. 2). At first the Expert System calculates any specific backbone fragments (a, b, and y-ion series), the charged precursor ion, the immonium ions as well as side chain fragments in the low-mass region and places them into a queue. In the second part of the workflow every element in this queue is filtered with respect to the actual MS/MS spectrum. Even if there is a peak corresponding to a calculated item in the queue, it may still be filtered out (symbolized by missing annotations after the filter in Fig. 2). For instance, a b_1 ion is only allowed in very restricted circumstances.

In the third step, neutral losses and internal fragments for the filtered values are calculated and added to the queue. They are then subjected to the same filtering rules as in step 2. Step 3 is iterative, as several subsequent neutral losses may be allowed.

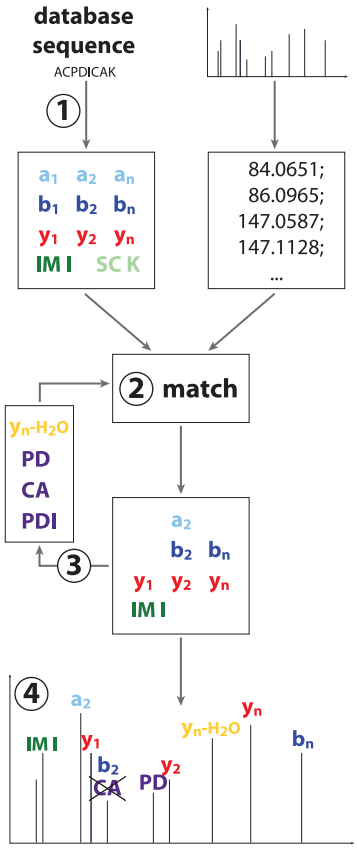


FIG. 2. **Work flow of the Expert System.** ① From the database sequence of the peptide identified by the search engine, a list of possible fragment ions is created. ② Peaks from the measured spectrum are compared with the possible fragments and preliminarily annotated if they pass the rules of the Expert System. ③ Neutral losses and internal fragments are generated from the candidate, annotated peaks and exposed to the Expert System rules. ④ Potential conflicts are resolved via the priority of the annotations and peaks are labeled. Note that possible internal fragment 'CA' is crossed out because the b_2 ion has the higher priority.

In the fourth and last step each potential annotation is given a priority. If there is more than one possible annotation, the one with the highest priority is chosen (*i.e.* the one that triggered the rules with higher priority). However, in this case the Expert System provides a pop-up (or "tool-tip") containing the other possibility when hovering the mouse over the peak. (This can still happen if the FDR is properly controlled and is then typically caused by two different chemical designations for the same ion; or by different ions with the same chemical composition, such as small internal fragments with different sequence but the same amino acids).

Determining a False Discovery Rate for Peak Annotation—Use of a very high threshold for peptide identification (99.99%) ensured that virtually none of the peptides in our collection should be misidentified. However, when building

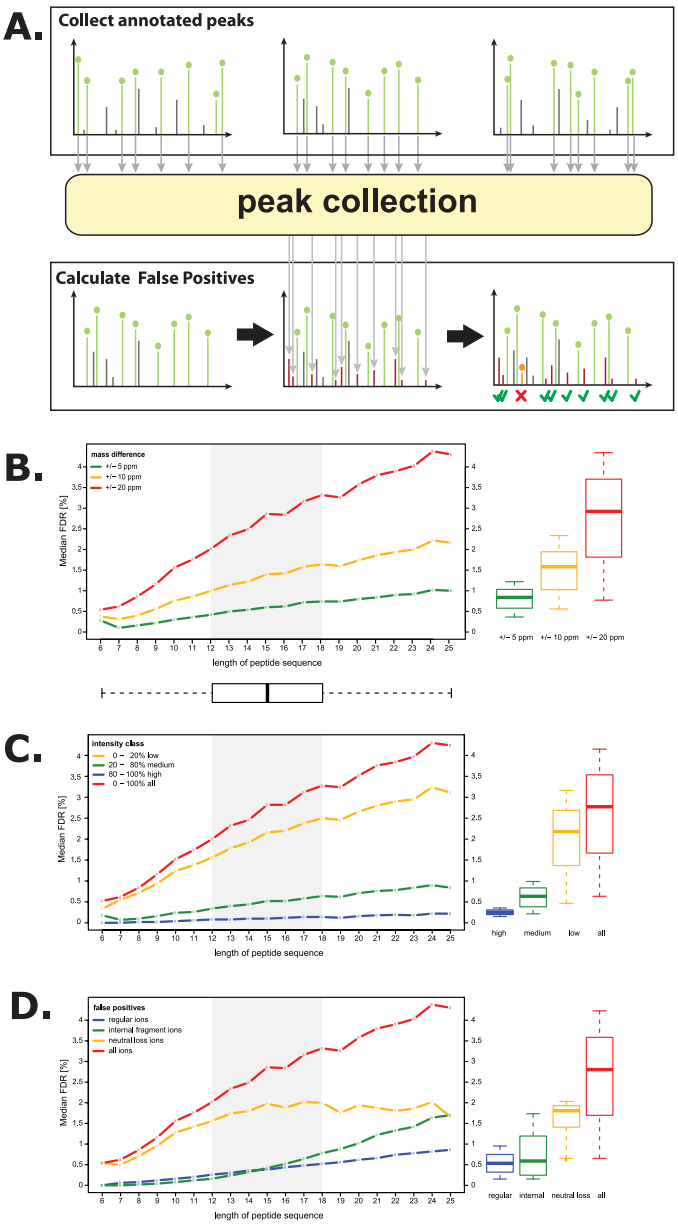


FIG. 3. **Calculation of false discovery rate for peak annotations.** *A*, The upper panels represent a large number of identified MS/MS spectra from which annotated peaks are drawn to form a large peak collection of possible fragment masses. From each identified spectrum in the data set, 10 random fragments are inserted and the number of annotations by the Expert System is counted. This process is repeated 500 times for each peptide. *B*, Median FDR as determined in *A* as a function of peptide length distinguished by the mass difference of fragment ion and theoretical mass. The FDR for peak annotation rises with peptide length and is strongly dependent on the mass difference. Box plot at the bottom shows that 50% of the peptides were between 12 and 18 amino acids long. The box plots on the right summarize the range of FDR values regardless of peptide length. *C*, Graph of the median FDR as a function of peptide length but separated by intensity classes of the false annotated fragment peaks. Most false positives come from the low abundant peaks (*blue*) rather than the medium (*green*) or high abundance fragment peaks (*yellow*). *D*, Same plot as above but differentiated by the fragment ion type of the false positives. Getting lower number of false positives from regular fragment annotations (*blue*), compared with internal fragment (*green*) and neutral loss annotations (*yellow*).

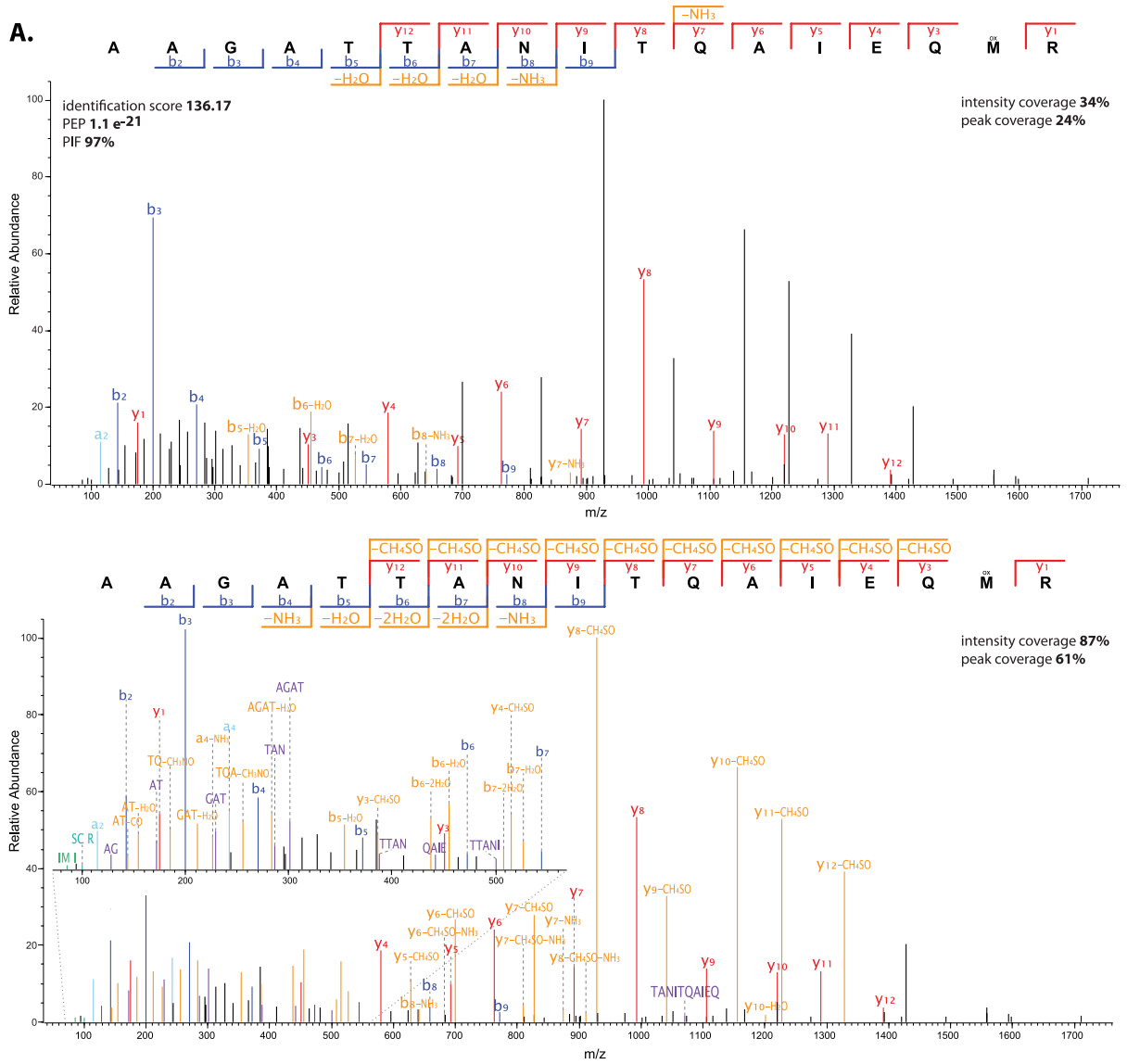


FIG. 4. **Example spectra before and after Expert System annotation.** *A*, Based on the search engine result, 34% of the fragments by peak intensities and 24% by peak number are explained, whereas the Expert System almost completely annotates the spectrum (for further explanation see main text). Posterior Error Probability (PEP) a statistical expectation value for peptide identification in Andromeda. Apart from the large fraction of a-, b-, and y-ions (*pale blue/dark blue/red*) and ions with neutral losses (*orange*), one can find internal fragment ions (*purple*) and in the low mass region one immonium ion of Isoleucine (*green*) and a side chain loss from arginine (*turquoise*). *B*, Expert System annotation of a phosphorylated peptide. Apart from the internal ions, several phosphorylation-related fragment ions were found. The asterisk (*) denotes loss of H3O4P with a delta mass of 97.9768 from the phosphorylated fragment ion.

the Expert System, we noticed that it was still possible to over-interpret the MS/MS spectra. This was initially surprising to us because our large scale data set had good signal to noise and peaks was only candidates for annotation when their calculated mass was less than 20 ppm from the observed mass. The over-interpretation became apparent through conflicting annotations for the same peak, and was typically caused by a combination of rules, such as several neutral losses from major sequence specific backbone or internal ions. Because conflicting or wrong annotations would undermine the entire rational for the Expert System, we devised a scheme to stringently control the false discovery rate for peak annotation.

The false discovery rate (FDR) is meant to represent the percent probability that a fragment peak is annotated by

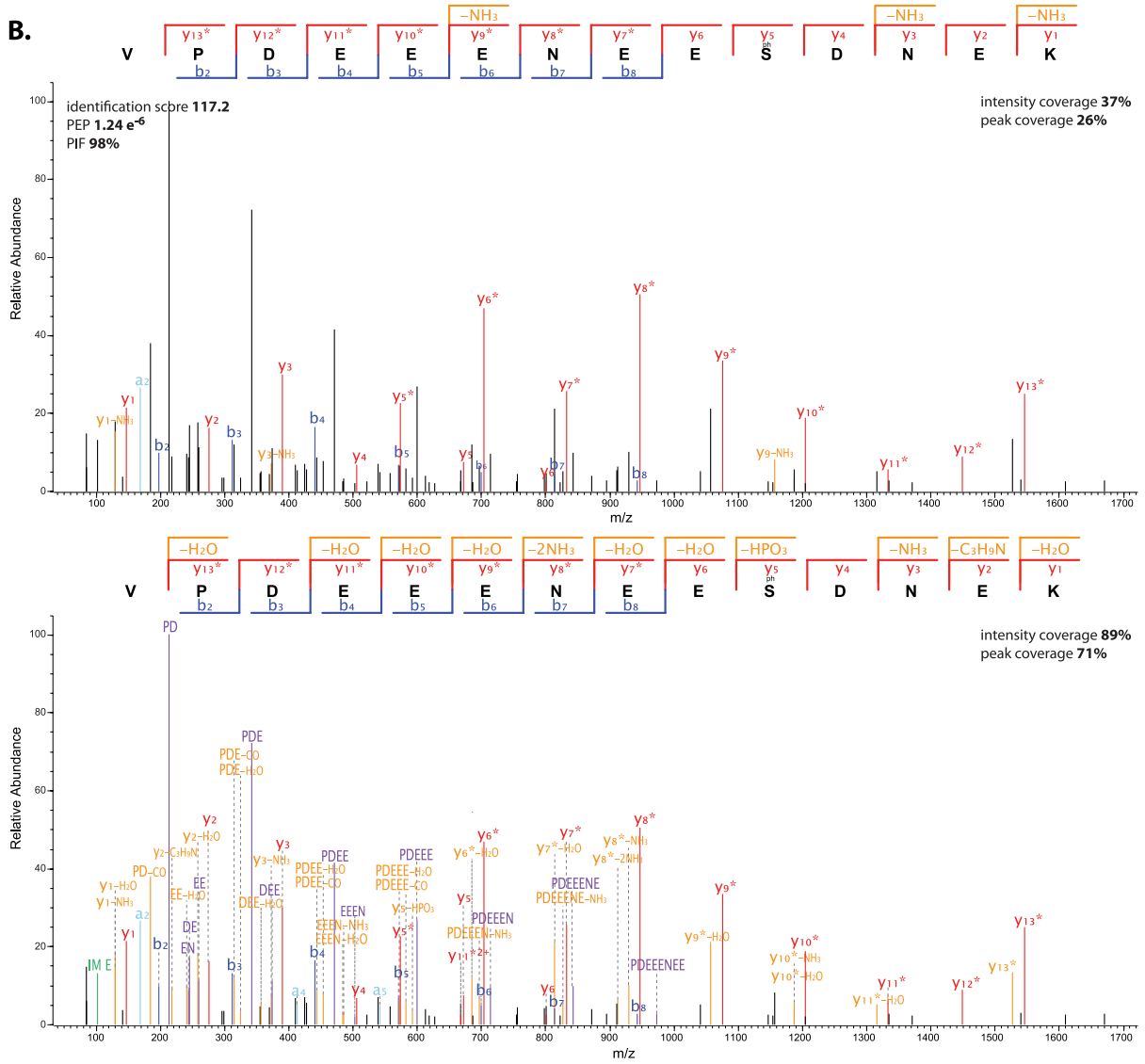


FIG. 4—*continued*

chance because its mass fits one of the Expert System rules for the peptide sequence. To calculate a proper FDR, we therefore needed to provide a set of background peaks that would represent false positives when they are labeled by the Expert System. Producing realistic background peaks turned out to be far from trivial because they need to have possible masses that can in principle be generated from peptide sequences and they need to be independent of the sequence of the peptide in question. The principle of our solution to this problem is shown in Fig. 3*A*. From the large data set underlying this study, we collect the m/z values of all annotated peaks, except those coming from immonium or side chain ions. They were stored in a large peak collection of several million entries, together with the respective peptide sequences and the relative intensity of the peak. For each spectrum in which we wanted to determine the FDR, we then inserted a random set of 10 peaks from the collection, where after we checked if the sequence of the selected peaks was independent from the sequence of the current spectrum. If one of the inserted peaks overlapped with an existing peak, it was discarded. By definition these 10 peaks represent possible peptide fragments and, because they are chosen randomly from millions of other peaks, they collectively represent a good approximation to a true background set. This would

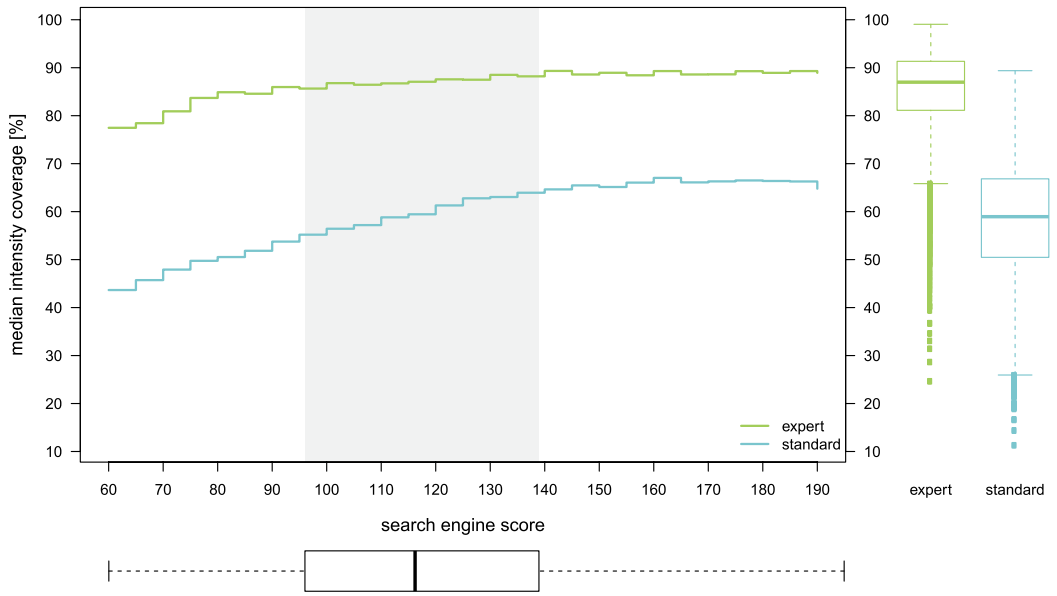


FIG. 5. **Expert System performance on a large data set.** Median sequence coverage by summed fragment ion intensity is plotted as a function of identification score. Statistics is based on more than 16,000 spectra. For every identification score, the Expert System adds a large proportion of explainable peaks. Box plot below the graph indicates that 50% of peptides in the set have an Andromeda score between 98 and 140. Box plots on the right indicate the range of values for the intensity coverage for standard and Expert System annotation.

not be the case for permutation of the sequence of the precursor in question, for instance, because many of the fragment peaks in permutated sequences are identical. Whenever the Expert System annotated one of these peaks, it was counted as a false positive. To find the number of repeats necessary to obtain a stable FDR for this procedure, we chose a set of spectra and simulated a thousand times on each one. We found that the FDR was constant after 500 iterations. For the final FDR calculation, for each spectrum we added a different set of 10 random peaks from the collection and repeated this 500 times. This was then applied to each of the more than 16,000 pure (high PIF) spectra in the large scale data set.

Beyond providing a solid FDR estimate for each rule set, this procedure also allowed us to identify the rules or rule combinations that were responsible for miss-annotation, *i.e.* the rules that falsely annotated the inserted peaks. These mostly turned out to be chains of subsequent neutral losses. In conjunction with detailed evaluation of the frequency of ion types, we iteratively designed an optimal rule set (supplemental Table S1). For instance, neutral losses from a particular amino acid were allowed if they occurred in more than five percent of the fragment sequences that contained that amino acid. Likewise, of a set of about 42 possible neutral side chain losses, only six were sufficiently important to retain them in the Expert System. The Figs. 3*B*–3*D* show the results of the median FDR as a function of the peptide length based on this final rule set. The overall FDR—indicated in red—is the same in all plots and shows a clear growing trend in the number of false positives with the length of the peptides. For small peptides of 12 amino acids or less, the FDR was less than 2.1% and all peptides in the range investigated had a peak annotation FDR of less than 5%. With these settings, the annotations are correct in more than 97% of the cases for the vast majority of MS/MS spectra. The Expert System could of course be pruned to provide a lower FDR by narrowing the mass tolerance window; however, this would come at the expense of discarding correct annotations. To explore the influence of mass accuracy on potential false positive annotations, we repeated these calculations with required mass deviations no larger than 5 ppm or no larger than 10 ppm. As can be seen in Fig. 3*B*, this further reduced possible errors to less than 1%, or less than 0.3%, respectively. This highlights the value of high mass accuracy in unambiguously identifying fragment mass identity.

Furthermore, peaks with a low signal to noise are more likely to be miss-annotated than more intense peaks. In Fig. 3*C* we sorted the peak intensity of the false positives into three intensity classes (Fig. 3*C*). The median FDR of peaks with high or medium abundance are only 0.1 or 0.5%. For low abundance peaks it is higher but still with a median of no more than 2.1%.

Next we separately investigated the FDR as a function of peptide length for the different fragment ion types. As can be seen in Fig. 3*C*, regular ions and internal fragments contribute very little to overall false annotation (0.4 and 0.5%), whereas neutral loss ions are wrongly annotated in 1.8% of the case or even more.

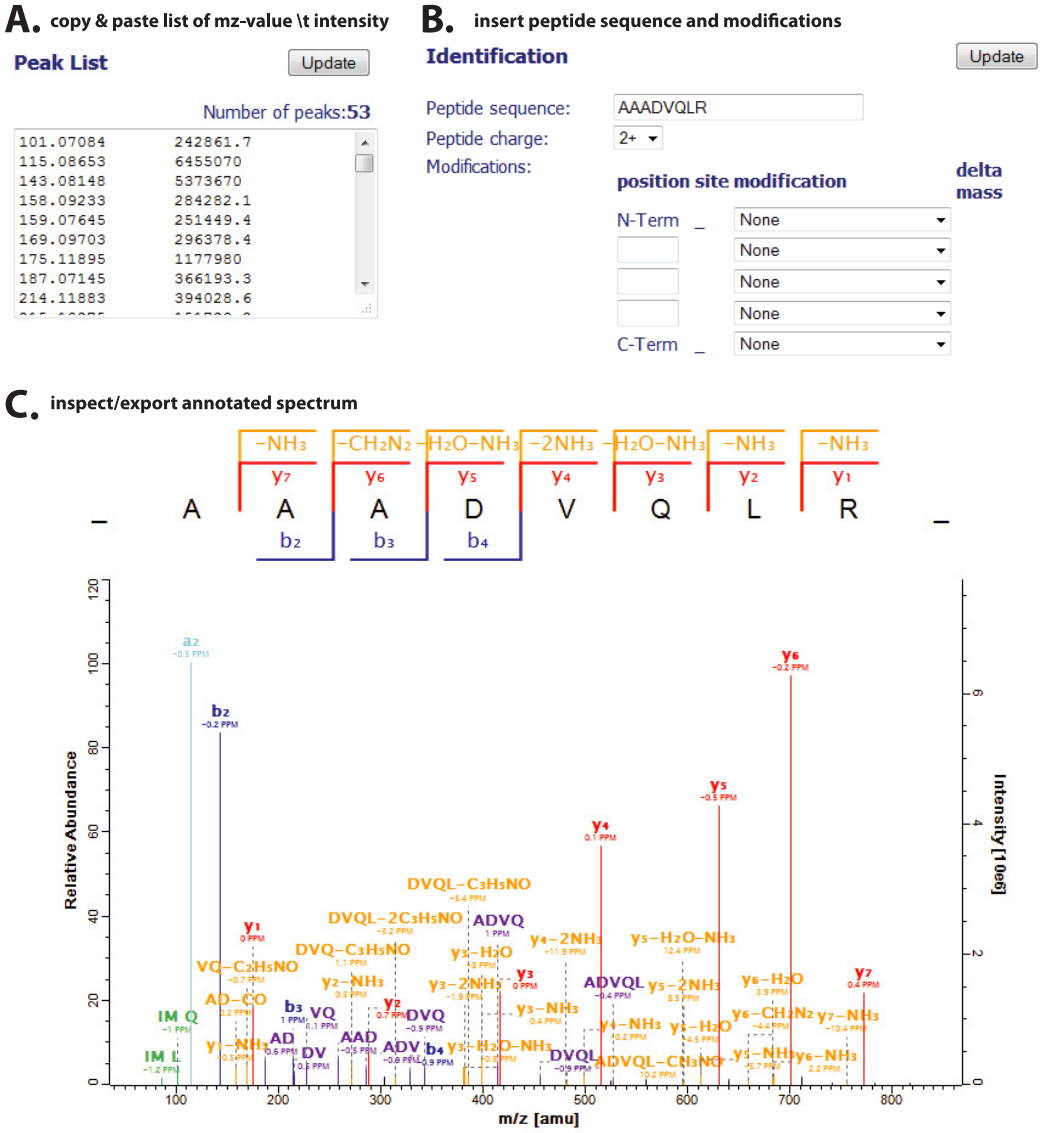


FIG. 6. **Web interface for the Expert System.** *A*, Text field to paste the spectrum in text format (*m/z* value; intensity in arbitrary units). *B*, Form to enter the peptide sequence, modifications and their positions. *C*, Detected backbone fragments and their neutral losses are indicated in the peptide logo. Scalable spectrum annotated by the Expert System. Note that neutral loss peaks are very small compared with the major backbone fragments. The spectrum can be downloaded with the desired resolution and in the desired graphical format.

Performance of the Expert System—Fig. 4 shows an illustrative example of an HCD fragmented peptide before and after Expert System evaluation. The peptide was identified with an Andromeda score of 136 and posterior error probability (PEP) of 1.1E-21 (the corresponding Mascot score was 83). The spectrum features an uninterrupted b-ion series from b_2 to b_9 and an uninterrupted y-ion series from y_1 to y_{12}, together covering the entire peptide sequence. Despite this unambiguous identification, the peaks used by the search engine to identify the peptide only accounted for 35% of the summed intensity of the peaks in the fragmentation spectrum. Coverage by number of explained peaks was even lower at 24% (allowing up to 10 peaks per 100 Th in the measured spectrum see EXPERIMENTAL PROCEDURES). There is a series of high abundance, high *m/z* fragments as well as a large number of low abundance peaks in the low and medium *m/z* range that are unexplained by the search engine. After annotation by the Expert System, this situation changes entirely. The high m/z series is revealed to be a prominent loss of CH_4SO from oxidized methionine. The low

mass ions are neutral losses, internal fragments and combinations between them and they were unambiguously and correctly assigned. Altogether, the Expert System accounted for almost all prominent ions and explained a total of 88% of the ion current. Manual annotation of this spectrum would have been possible but would have been very time consuming.

Interpretation of phosphorylated peptides, especially large ones, is more difficult than that of unmodified peptides. Furthermore, accurate placement of the phosphorylation site can be challenging. We used literature knowledge (19, 20) and the results of a large-scale investigation into the fragmentation of phosphorylated peptides to derive suitable fragmentation rules for the Expert System. This led to an additional six rules, which were easily integrated, illustrating the extensibility of the Expert System. Fig. 4*B* depicts an example annotation of the relatively complex fragmentation spectra typical of phosphorylated peptides. The large ion series from the low mass range to about mass 1000 is caused by an extensive and uninterrupted internal ion series starting from the proline in the second position of the peptide sequence. As these internal fragments contain several glutamines, they lead to additional water and ammonia losses. However, there are also newly annotated fragments resulting from neutral losses in addition to loss of the phosphorylation site. Moreover, the neutral loss of HPO_3 is annotated.

Large-scale Evaluation of the Performance of the Expert System—We used the population of 16,000 spectra with high PIF—identified with a false discovery rate of 0.01% by the search engine—and annotated them automatically using the Expert System. For each spectrum we calculated the intensity coverage obtained by the fragments used by the search engine and the fragments explained by the Expert System. Higher scoring fragmentation spectra would be expected to have a larger fraction of their ion current annotatable than lower scoring peptides. Fig. 5*A* shows a plot of the median of these values for all search engine scores. A total of 95% of these Andromeda scores are within a range of 96 to 138. Here the median intensity coverage by standard annotation varies from 55% at 96 to 64% at 138. The Expert System, in contrast, annotated between 86 and 89% of the total ion current in the fragment spectra of the same peptides. This represents an average increase of 28%. There was only a small percentage of peptides that were lower scoring than 96 and for these the increased annotation percentage of the Expert System was even larger (34%). Interestingly, even in very high scoring HCD fragment spectra there are still many peaks not directly annotated by the search engine. For these, the average increase of annotated ion current because of the Expert System was still 23%.

The rule set of the Expert System was derived from HCD data. However, HCD and CID appear to produce similar ion types, although with different abundances. We therefore tested if the derived rule set was also applicable to high resolution CID data. This was indeed the case, and a total of 85% of the ion current in high resolution CID spectra explained by the Expert System, although in CID spectra a higher percentage (79%) of the peaks are already accounted for by standard ion types. Therefore we conclude that the Expert System can be used equally well for high resolution HCD and CID data although the benefits for CID are not as large as they are for HCD.

Webserver for Expert System Annotation of Spectra—The Expert System is now part of the Viewer component of MaxQuant, which is freely available at www.maxquant.org. In this environment, the Expert System can annotate arbitrarily large data sets of identified peptides and visualize and export them in different graphical formats such as PDF. Additionally, we established a webserver to make the Expert System available to any proteomics scientist, regardless of the computational workflow that he or she is using. The webserver is located at http://www.biochem.mpg.de/mann/tools/and its graphical interface is shown in Fig. 6. The user needs to supply a mass spectrum in the form of an m/z and peak intensity list as well as the sequence of the identified peptide (Figs. 6*A*, 6*B*). Common modifications and their position in the sequence can also be specified. The webserver then provides an annotation of the spectrum within the stated mass tolerance as shown in Fig. 6*C*. The graph is scalable to enable detailed study of complex fragmentation spectra. Mass deviations in ppm (calculated mass – measured mass) can also be depicted. This annotated spectrum can be downloaded in a number of graphical formats for use in publications.

CONCLUSION AND OUTLOOK

Here we have made use of Expert Systems—a well-known technology in computer science—to automatically but accurately interpret the fragmentation spectra of identified peptides. We have shown that the Expert System performs very well on high mass accuracy data, annotating the large majority of medium to high abundance peaks. For HCD spectra it explains on average 28% more of the peak intensities than the search engine results alone. We derived a rigorous false positive rate, ensuing that less than 5% of peaks can be missannotated—this rate is even lower for spectra with at least median scores and fragment ion intensities of at least moderate abundance. The rule set was derived by iterative interpretation of large HCD data set but we show that the Expert System is equally applicable to high resolution CID spectra.

We envision different uses for the Expert System: For beginners in MS-based proteomics, it enables efficient training in the interpretation of MS/MS spectra without requiring much input from a specialist. For advanced users, it allows focusing on unusual and potentially novel types of fragments. One caveat is that the Expert System currently cannot explain fragment peaks that belong to cofragmented precursors; a very common occurrence that we deliberately avoided here by selecting only pure MS/MS spectra. This limitation can be

addressed if both precursors are identified and communicated to the Expert System. Such a feature might be particularly useful for instruments that allow deliberate multiplexing of precursors, which leads to complex MS/MS spectra (21).

The Expert System has been in routine use in our laboratory for a number of months. During this time we have found that it provides helpful confirmation of the identification of the peptide and the identity of the previously unlabeled fragment ions. This is particularly welcome in the case of complicated spectra of important peptides, such as the ones regulated in the biological function in question. Compared with a human expert, the principal advantages of the Expert System are its speed, its ability to check for all supplied rules in a consistent manner as well as its rigorously controlled false positive rate. Obviously, the Expert System is limited to the knowledge supplied whereas an experienced mass spectrometrist can go beyond these rules and discover the origin of novel fragmentation mechanisms.

As we have shown here, Expert Systems can readily be applied to problems in computational proteomics. Given their relative ease of implementation, they may become useful in other areas in MS-based proteomics, too.

Acknowledgments—We thank Forest White for critical comments on this manuscript.

* This work was supported by funding from the European Union 7th Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645).

S This article contains supplemental Table S1.

¶ These authors contributed equally.

§ To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711
2. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–797
3. Granholm, V., and Käll, L. (2011) Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **11,** 1086–1093
4. Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., and Old, W. M. (2010) Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **9,** 4152–4160
5. Zhang, N., Li, X. J., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005) ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **5,** 4096–4106
6. Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J., and MacCoss, M. J. (2010) Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82,** 833–841
7. Michalski, A., Cox, J., and Mann, M. (2011) 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.* **10,** 1785–1793
8. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
9. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
10. Giarratano, J. C., and Riley, G. (2005) *Expert systems: principles and programming.* PWS Pub. Co., Boston
11. Liao, S. H. (2005) Expert system methodologies and applications - a decade review from 1995 to 2004. *Expert Syst. Appl.* **28,** 93–103
12. Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J. (1969) Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. *J. Am. Chem. Soc.* **91,** 7440–7445
13. Russell, S. J., Norvig, P., and Davis, E. (2010) *Artificial intelligence: a modern approach*. Prentice Hall, Upper Saddle River, NJ
14. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
15. Bin, M., and Johnson, R. (2012) De novo sequencing and homology searching. *Mol. Cell. Proteomics* **11,** O111.014902
16. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1,** 2856–2860
17. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8,** 2759–2769
18. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11,** 10.1074/mcp.O111.013698
19. Boersema, P. J., Mohammed, S., and Heck, A. J. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **44,** 861–878
20. Kelstrup, C. D., Hekmat, O., Francavilla, C., and Olsen, J. V. (2011) Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *J. Proteome Res.* **10,** 2937–2948
21. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** 10.1074/mcp.M111.011015

2.3 article 3 - A Systematic Investigation into the Nature of Tryptic HCD Spectra

Annette Michalski, **Nadin Neuhauser**, Jürgen Cox, and Matthias Mann
Journal of proteome research 2012, 11, 5479-5491

For many years, collision induced dissociation (CID) has been the workhorse of tandem mass spectrometry[97], breaking the analyte into characteristic fragments in order to provide structural information. Here we take a closer look at the recently developed higher energy collision induced dissociation (HCD) fragmentation technique. In contrast to CID in ion trap instruments, HCD is almost always acquired in high resolution and with high mass accuracy, which can even enable determination of elemental composition especially in the low mass region. To characterize the fragmentation pattern obtained by HCD, a significant number of spectra had to be considered.

Here we investigate a large-scale analysis to find out what is the difference between CID and HCD and which fragment ions can be found in HCD spectra. In order to automate this process, this investigation accrued in context of the development of the Expert System (article 2). Using this computational support it was then possible to apply novel and already known fragmentation rules to determine several thousands of high quality peptide spectrum matches. For this I was implementing a program that automates the spectrum interpretation and provides the results in tables which were used for further statistical analysis.

One observation during the development process was that co-fragmenting peptides occur quite often. A measurement of the pureness in MS/MS spectra is the precursor intensity fraction (PIF) reporting the percentage of the picked precursor in the selection window. The fact that a multiple peptide species can be a source for interfering peaks was already known, but the observed magnitude of this effect caused us to perform further investigations. Michalski *et al* [130] asked how often peptides are close enough in mass and retention time to elute in the same selection window, and which influence this has on peptide identification. As a result these observations also prompted us to develop the 'second peptide search' in Andromeda. For further investigations, the datasets in this study where filtered to contain only pure representations of fragment spectra (PIF >95%).

Most of the peptide fragments found in this investigation are already known from the literature especially from extensive research on CID fragmentation[57;131]. But we also

evaluated fragmentation mechanism known from other fragmentation techniques and inferred new hypotheses. For instance, a list of possible neutral losses were generated by taking chemical aspects into consideration. For practical and statistical reasons, the list was reduced to contain only a few significant candidates.

In contrast to CID, which has the low mass cutoff, HCD spectra contains more signals in the lower mass range. This has an positive effect for reporter ion based quantification like iTRAQ or TMT. Additionally, immonium ions and side chain losses can be found in the lower mass area. Due to its higher energy, internal fragments are frequently observed, which is not common in CID spectra obtained in ion trap.

While the original aim of this study was to find new fragmentation types, we concluded that CID and HCD fragmentation are in fact quite comparable in their fragmentation behavior.

Journal of proteome research

Article

pubs.acs.org/jpr

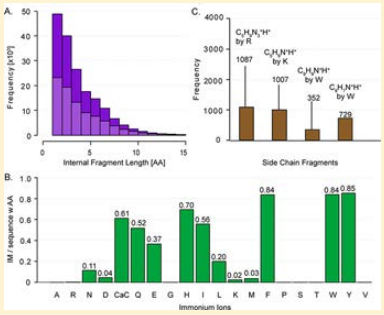
A Systematic Investigation into the Nature of Tryptic HCD Spectra

Annette Michalski, Nadin Neuhauser, Jürgen Cox, and Matthias Mann*

Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany

Supporting Information

ABSTRACT: Modern mass spectrometry-based proteomics can produce millions of peptide fragmentation spectra, which are automatically identified in databases using sequence-specific *b*- or *y*-ions. Proteomics projects have mainly been performed with low resolution collision-induced dissociation (CID) in ion traps and beam-type fragmentation on triple quadrupole and QTOF instruments. Recently, the latter has also become available with Orbitrap instrumentation as higher energy collisional dissociation (HCD), routinely providing full mass range fragmentation with high mass accuracy. To systematically study the nature of HCD spectra, we made use of a large scale data set of tryptic peptides identified with an FDR of 0.0001, from which we extract a subset of more than 16 000 that have little or no contribution from cofragmented precursors. We employed a newly developed computer-assisted "Expert System", which distills our experience and literature knowledge about fragmentation pathways. It aims to automatically annotate the peaks in high mass accuracy fragment spectra while strictly controlling the false discovery rate. Using this Expert System we determined that sequence specific regular ions covering the entire sequence were present for almost all peptides with up to 10 amino acids (median 100%). Peptides up to 20 amino acid length contained sufficient fragmentation to cover 80% of the sequence. Internal fragments are common in HCD spectra but not in high resolution CID spectra (10% vs 1%). The low mass region contains abundant immonium ions (6% of fragment ion intensity), the characteristic a_2, b_2 ion pair (72% of spectra), side chain fragments and reporter ions for peptide modifications such as tyrosine phosphorylation. *B*- and *y*-ions account for only 20% of fragment ions by number but 53% by ion intensity. Overall, 84% of the fragment ion intensity was unambiguously explainable. Thus high mass accuracy HCD and CID data are near comprehensively and automatically interpretable.



KEYWORDS: *tandem mass spectrometry, fragmentation mechanisms, shotgun proteomics, ion types, CID, HCD, Expert System, spectrum annotation*

INTRODUCTION

Rapid technological development of mass spectrometric instrumentation in conjunction with advanced bioinformatics analysis capabilities now allow relatively streamlined and in depth analysis of proteomic samples.[1−3] Modern proteomics projects routinely generate millions of fragmentation spectra, making entirely automated software tools a necessity. These include search engines that match MS/MS spectra to the most probable peptide sequence in a database, typically relying on sequence-specific backbone fragments, referred to as "regular ions" in this article, as well as associated neutral losses.[4] However, there are many other fragment ions in tandem mass spectra, and it has been argued that detailed interpretation of at least the more abundant peaks should be a requirement for confident peptide assignment.[5] Likewise, detailed understanding of the fragmentation process and discovery of potential new fragment types requires knowledge of the identity of the majority of fragmentation peaks.

While there are many different ways to fragment peptides, in proteomics collision-induced fragmentation has by far been the most frequently used technique (for a recent tutorial of peptide fragmentation and spectrum interpretation, see ref 6). While there are differences in how the ions are activated, the general ion types are the same and are summarized in Figure 1. The backbone fragments are designated as *a*, *b*, *c* for N-terminal and *x*, *y*, *z* for C-terminal types depending on the cleavage position on the peptide backbone.[7−10] A full series of either *b*- or *y*-type ions in principle allows reading out the entire amino acid sequence from a fragment ion spectrum. In collision-induced fragmentation techniques, cleavage of the peptide bond is preferred, but labile post-translational modifications such as phosphorylation or glycosylation also partially or (rarely) completely detach. While the chemistry involved in peptide fragmentation is still not completely understood, the mobile proton model is currently the most widely accepted framework to describe the dissociation process.[11,12] Moreover, different fragmentation pathways of protonated peptides have been extensively investigated and modeled with respect to both kinetic and thermodynamic aspects.[13]

In addition to the standard backbone ions, tandem mass spectra can contain many additional fragment ions.[14] Numerous studies of peptide dissociation behavior have been carried out to investigate the abundance and structure of ion types such as internal ions, immonium ions or neutral losses from these (Figure 1).[15,16] Some programs such as Protein

Received: July 30, 2012
Published: September 23, 2012

dx.doi.org/10.1021/pr3007045 | J. Proteome Res. 2012, 11, 5479−5491

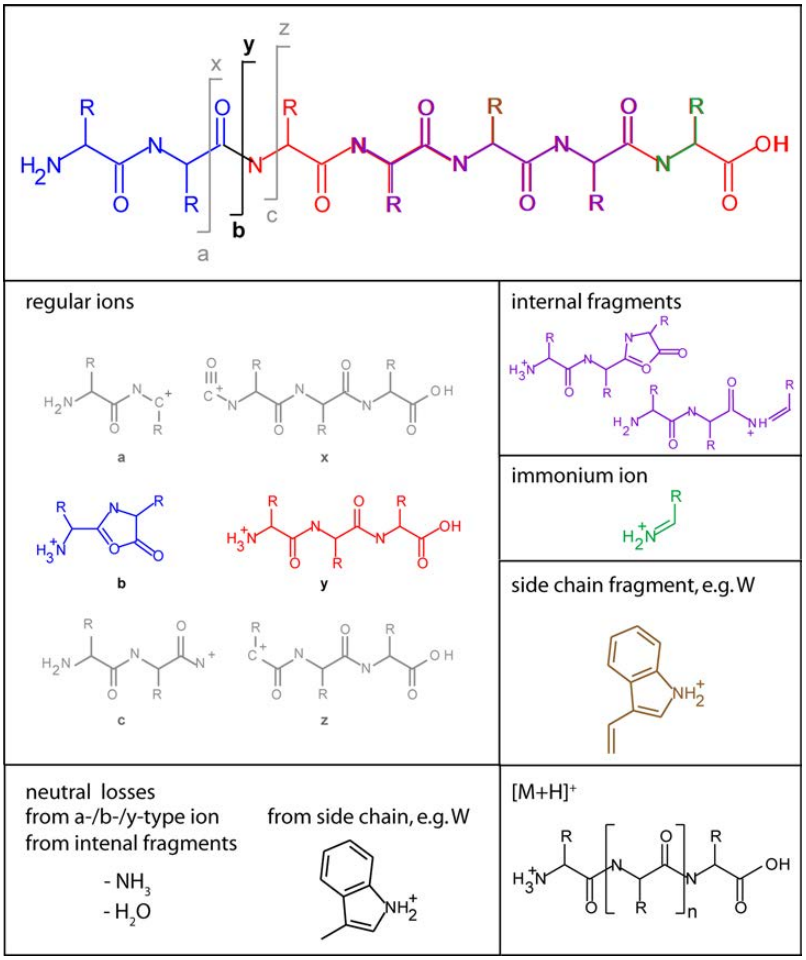


Figure 1. Cleavage sites of the peptide backbone giving rise to N-terminal *a*-, *b*- or *c*-type ions and the corresponding C-terminal *x*-, *y*- or *z*-type ions, respectively. The most prominent cleavage in CID and HCD fragmentation happens at the peptide bond. The boxes below represent the most frequent ion types of collision induced fragmentation processes; the color code provides their origin in the peptide sequence.

Prospector provide comprehensive lists of produced ion types for different fragmentation mechanisms and instrument types and even consider the latter for scoring of tandem mass spectra.[17] Furthermore, special types of ions have been characterized, for instance, b_1 ions of N-terminally acetylated ions,[18] c_1 ions in case glutamine is the second amino acid from the N-terminus,[19,20] specific side chain losses such as from oxidized methionine[21] and many more. Finally, novel fragmentation processes continue to be discussed controversially, such as the extent of scrambling of *b*-ions due to their formation of a cyclic peptide structures followed by random cleavage, which could interfere with determination of the correct amino acid sequence from the data.[22−25]

Furthermore, the observed types of fragment ions in a tandem mass spectrum depend on the instrument type. Triple quadrupole and quadrupole time-of-flight (TOF) fragmentation are beam-type dissociation processes,[26] where primary fragments retain kinetic energy and are therefore more likely to fragment again in the multiple collision conditions typical of these instruments. In 3D or 2D ion traps the excitation and activation step is only applied to the selected precursor mass. Any primary fragmentation product is off-resonance with the applied radio frequency and therefore usually remains intact. When collision-induced dissociation is performed in ion traps (often primarily associated with CID fragmentation), the low mass fragments are typically not retained, leading to a low mass cutoff in the tandem mass spectra.[27]

Higher energy collisional dissociation (HCD), first described in 2007, made beam type fragmentation available on the Orbitrap analyzer platforms.[28] Recently, HCD fragments have also been analyzed at low resolution in an ion trap[29,30] but are in general always detected in the Orbitrap analyzer at high resolution and mass accuracy. Since the introduction of the LTQ Orbitrap Velos mass spectrometer, which features improved sensitivity and HCD capability compared to its predecessors, routine acquisition of tandem mass spectra in the Orbitrap analyzer has become feasible.[31] This approach is termed "high−high" strategy because both the full scans (MS) and the fragment ion scans (MS/MS) have high resolution and high mass accuracy in comparison to previous strategies with

dx.doi.org/10.1021/pr3007045 | *J. Proteome Res.* 2012, 11, 5479−5491

acquisition of CID scans (MS/MS) in the ion trap ("high–low").[32] Note that high–high strategies have been the default in quadrupole-TOF instruments for many years; however, this did not necessarily imply high mass accuracy in the MS/MS mode, primarily due to issues with ion statistics. Because of the dedicated collision cell, HCD fragment ion spectra cover nearly the entire mass range and are therefore particularly suitable for observing the low mass region, which contains an a_2/b_2 pair, immonium ions, fragments resulting from the amino acid side chains as well the reporter ions[18] used for quantification in the TMT or iTRAQ methods.[33−35] Importantly, high mass accuracy of fragment ions helps to unambiguously annotate the fragment ion peaks. Especially in the low mass region, an accurate mass measurement may even uniquely determine the elemental composition of the fragment.

In contrast to ion trap CID data, high resolution HCD has been relatively little studied. Although HCD ion types are expected to recapitulate fragmentation rules known from older CID type instruments, those have not been tested on large-scale and high accuracy data. Here, we wished to take advantage of the excellent signal-to-noise, dynamic range and mass accuracy of HCD spectra on the Orbitrap analyzer to systematically investigate features of HCD spectra. This was facilitated by a rule-based "Expert System", which was developed in an iterative manner with this study and is described elsewhere.[36] This Expert System synthesizes well-established knowledge about peptide fragmentation pathways mechanisms. It is capable of annotating large-scale MS/MS data sets based on the rules chosen by the researcher. We apply the Expert System for a comprehensive statistical investigation into the nature of HCD tandem mass spectra of tryptic peptides.

■ EXPERIMENTAL PROCEDURES

Sample Preparation

Total cell extracts of *E. coli*, yeast and HeLa cells were separated by 1D-SDS PAGE (4–12% Novex mini-gel, Invitrogen) in three separate lanes. Colloidal Coomassie (Invitrogen) was used for staining of the proteins before each lane was cut into 8 or 10 slices. All gel slices were subjected to reduction of the proteins with 10 mM DTT in 50 mM ammonium bicarbonate and subsequently alkylated with 55 mM IAA in 50 mM ammonium bicarbonate. In-gel digestion with 12.5 ng/μL trypsin (Promega) in 50 mM ammonium bicarbonate was carried out at 37 °C for 12 h followed by extraction of the tryptic peptides with 3% TFA in 30% ACN.[37] Peptides were loaded on C_{18} StageTips[38] before eluting them with 80% ACN in 0.5% acetic acid prior to analysis.

HeLa cell lysate was digested according to the filter-aided sample preparation (FASP) method.[39] Briefly, the lysate was solubilized in SDS-containing buffer and loaded onto Microcon YM-30 devices (Millipore, Billerica, MA, USA) to remove SDS and exchange it by urea. The protein mixture was alkylated with 50 mM iodoacetamide before urea was replaced with 20 mM ammonium bicarbonate. The proteins were digested overnight at 37 °C with trypsin (Promega) (1 μg of trypsin/100 μg of protein). Peptides were collected from the filter after centrifugation. For enrichment of phosphorylated peptides, the mixture was acidified with trifluoroacetic acid to pH 2.7 and ACN was added to a final concentration of 30%. Incubation with TiO_2 beads[40] (MZ Analysentechnik, Germany) prepared in 30 mg/mL solution of dihydrobenzoic acid (Sigma) was carried out for 30 min, before the beads were washed with 30% ACN and 3% TFA (twice) followed by two washes with 75% ACN and 0.3% TFA. The phosphopeptides were eluted with buffer containing 15% ammonium hydroxide and 40% ACN. Finally, the eluted phosphopeptides were loaded on C_{18} StageTips before they were eluted with 60% ACN in 0.5% acetic acid prior to analysis.

LC–MS/MS Analysis

For the analysis of proteome samples, the peptide mixture was separated on a C18-reversed phase column (15 cm, 75 μm ID, packed in-house with ReproSil-Pur C_{18}-AQ 3 μm resin, Dr. Maisch GmbH). An Easy-nLC (Thermo Scientific, Odense) with IntelliFlow system was used for sample loading and operated at a constant flow rate of 250 nL/min during the 110 min linear gradient of 8–60% buffer B (80% ACN and 0.5% acetic acid). A nanoelectrospray ion source (Thermo Scientific, Odense) was used for online coupling to the LTQ Orbitrap Velos mass spectrometer.[31] Mass spectra were measured in positive ion mode applying a data-dependent "top 10" method for the acquisition of a survey scan followed by MS/MS spectra of the 10 most abundant precursors. High resolution data was acquired in the Orbitrap analyzer with a resolution of 30 000 (m/z 400) for MS and 7500 (m/z 400) for MS/MS scans. For peptide fragmentation higher energy collisional dissociation (HCD) was used applying a normalized collision energy of 40 eV. The minimal signal threshold required was set to 5000. The target value in the Orbitrap analysis was 1×10^6 for the MS scans and 5×10^4 for the MS/MS scans with 2 Th isolation window and the first mass was set to 80 Th for HCD spectra. Fragmented precursors were dynamically excluded from targeting for 90 s. High resolution CID data was acquired on an Orbitrap Elite (Thermo Scientific) the same parameters; however, the resolution for MS scans was 120 000 (m/z 400) and for MS/MS scans 15 000 (m/z 400); the normalized collision energy was set to 35 eV.

For the phosphoproteome data, the enriched peptide mixtures were separated on a C_{18}-reversed phase column (20 cm, 75 μm ID, packed in-house with ReproSil-Pur C_{18}-AQ 1.8 μm resin, Dr. Maisch GmbH) applying a 90 min linear gradient of 5–30% buffer B (80% ACN and 0.1% formic acid) and analyzed on the Orbitrap Elite instrument[41] that was online-coupled to an Easy-nLC 1000 (Thermo Scientific, Odense). The MS data was acquired with resolution of 120 000 (m/z 400) and target value of 1×10^6 and MS/MS (HCD fragmentation) with resolution of 15 000 (m/z 400) and target value of 5×10^4 in a data-dependent "top 15" method with a dynamic exclusion of 30 s. The signal threshold was set to 5000 for an isolation window of 2 Th and the first mass of HCD spectra to 80 Th. The collision energy was set to 35 eV.

Data Analysis

All spectra were processed with MaxQuant[42] version 1.2.5.2 using the Andromeda search engine[43] to search the MS/MS spectra with trypsin specificity against the IPI human database (version 3.68, 87 061 entries) combined with 262 common contaminants. We allow for up to 2 missed cleavages and N-terminal acetylation and methionine oxidation were selected as variable, carbamidomethylation of cysteine was selected as fixed modification. For MS spectra an initial mass accuracy of 7 ppm was allowed, and the MS/MS tolerance was set to 20 ppm for fragment detection in the Orbitrap analyzer for high resolution CID and HCD. A sliding mass window was applied to filter the MS/MS spectra for the 10 most abundant peaks in 100 Th. For identification, the peptide FDR was set to 0.0001. (The protein

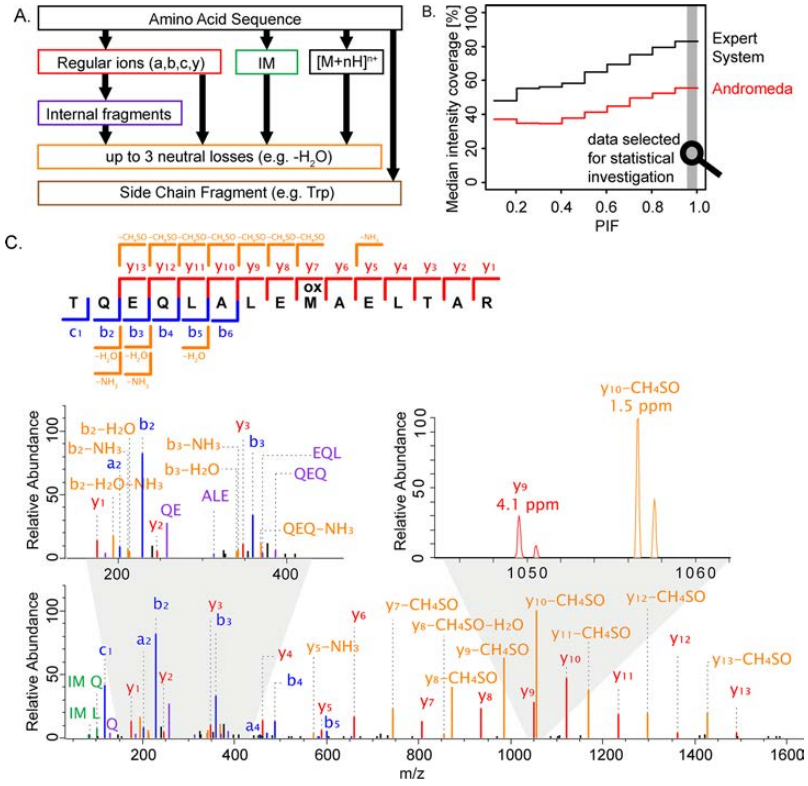


Figure 2. Peak annotation by the Expert System. (A) Ranking of the six major ion types: intact precursor mass $[M+nH]^{n+}$, regular ions, immonium ions (IM), internal fragments, neutral losses and side chain fragments that are considered for peak annotation by the Expert System. (B) Average intensity coverage of the total intensity of >100 000 MS/MS spectra by standard search engine annotation (Andromeda, red line) and by the Expert System (black line) vs the precursor intensity fraction "PIF" provides a measure for the purity of precursor isolation. The high quality data set (16 000 spectra) that was selected for statistical investigation is highlighted in gray. (C) Typical MS/MS spectrum with PIF 0.99 annotated by the Expert System reaching an intensity coverage of 87%. A zoom window displays the high mass accuracy of two fragment ion peaks.

FDR remained at the standard setting of 0.01, but protein identifications were not directly used in this paper.) The shortest peptide length was set to 6 amino acids, and the Max Quant feature to treat the isobaric amino acids leucine and isoleucine as indistinguishable for improved statics was disabled. This setting ensures that either amino acid matches the fragmentation spectrum as HCD in our setup cannot distinguish them; however, side chain losses can then be assigned correctly because the isoleucine/leucine ambiguity is absent after database search. MaxQuant and Andromeda data processing provides access to the peptide sequences that were identified from the MS/MS spectra. Detailed annotation of the MS/MS spectra was then carried out using the Expert System.[36] Results were further analyzed within the R scripting and statistical environment.[44] Raw mass spectrometric data are available at Tranche (www.proteomecommons.org) using the following hash code:

pI2oaLaSi7gPxUWNbesdXCgR17sWvMY6qVkHL+MtWA0Q5sqn/UxZVSjk3KpFTfrmDYpf3y/Iv6WfaAi6-HaILdZL0YocAAAAAAAAT7Q==

RESULTS AND DISCUSSION

Generation of a High Quality Data Set

To produce a diverse set of fragmentation spectra of tryptic peptides, we separated proteomes of *E. coli*, yeast and HeLa cells by one-dimensional gel electrophoresis, excised eight slices and in-gel digested them (Experimental Procedures). This generated a total of 24 complex peptide mixtures, which were analyzed using a "high–high" strategy on a linear ion trap–Orbitrap instrument (LTQ Orbitrap Velos) using HCD as the fragmentation method. For a smaller number of fractions, we also employed CID fragmentation followed by high resolution detection of fragments in the Orbitrap analyzer (Experimental Procedures).

We wished to work with an extremely high quality set of fragmentation spectra in order to enable us to unambiguously attribute the observed fragments to the precursors. Therefore, we set the false discovery rate (FDR) for peptide identification by MaxQuant using the Andromeda search engine[42,43] to 0.0001 rather than the customary 0.01. From our data set, we obtained more than 100 000 MS/MS spectra that were identified with this very stringent criterion. We and others have recently introduced the notion of the precursor intensity fraction (PIF),[45] chimeric or mixture MS/MS spectra,[46,47] which refers to the fact that precursor ions are frequently

dx.doi.org/10.1021/pr3007045 | J. Proteome Res. 2012, 11, 5479–5491

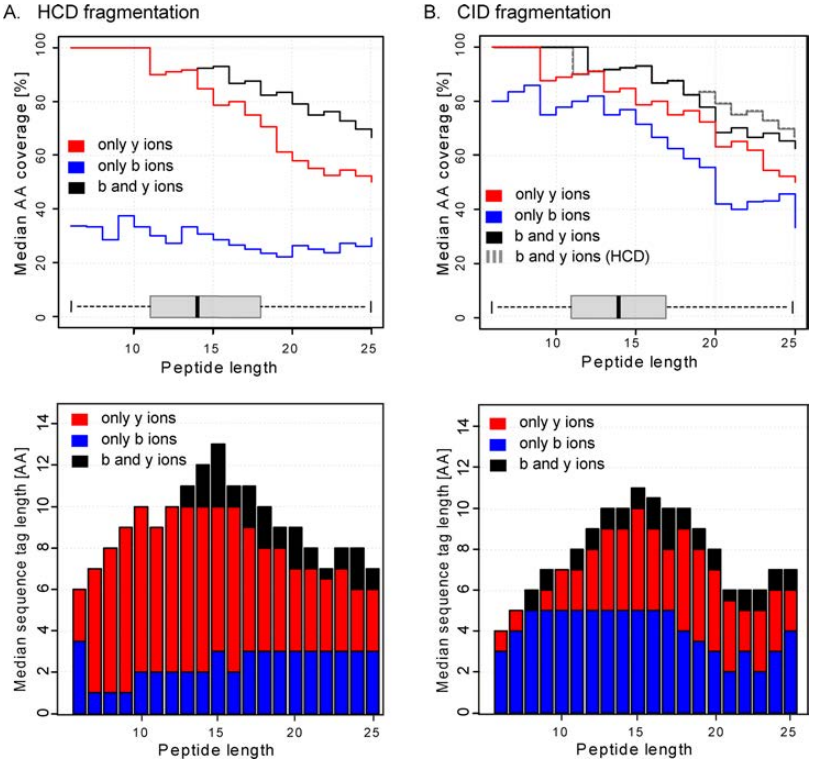


Figure 3. Sequence information content in HCD and CID spectra (A) Median coverage of amino acids by *y*-type ions (red), *b*-type ions (blue) and both together (black) in the upper panel. The boxplot displays the distribution of the peptide length within the data set (>16 000 spectra). The lower panel shows the median length of the longest sequence tag based on *y*-type ions (red), *b*-type ions (blue) and both together (black). (B) Same as (A) for a data set of 3290 high resolution CID spectra. The dashed gray line in the upper panel repeats the median amino acid coverage in HCD from panel (A) for comparison.

cofragmented unintentionally in the analysis of complex peptide mixtures. For our purposes we needed to minimize the occurrence of coeluting precursor ions in the isolation window so that they could not "contaminate" the MS/MS spectra with unassignable peaks. This was achieved by only retaining spectra with a PIF greater than 0.95. If there was more than one spectrum for a particular sequence, the one with the highest PIF was kept. Furthermore, we required the peptide length to be smaller than 26 amino acids and the charge state to be 2+, 3+ or 4+. These filters reduced the number of MS/MS spectra to about 16 000, which were nearly free of any contaminating peaks and which represented a broad sampling of typical tryptic peptides.

Computer-Assisted Annotation by the Expert System

We recently developed a computer Expert System,[36] which is now integrated into the Viewer component of the MaxQuant software environment. Briefly, the Expert System features a knowledgebase that was supplied with peptide fragmentation mechanisms described in the literature (see Introduction) and with knowledge gained from manual evaluation of small and large-scale HCD data sets. These facts are implemented in a rule-engine that assigns annotations to the peaks in the MS/MS spectra. In order to avoid incorrect assignments, the Expert System follows strict dependencies among its rules. We derived a rigorous FDR for peak annotation, which made it possible to derive a minimal yet relative comprehensive set of rules.[36]

Some MS/MS peaks can have an elemental composition that corresponds to more than one ion type, and we have developed a strict ranking of the possible annotations to address this particular issue (Figure 2A). On the basis of the identified peptide sequence, regular ions that result from cleavage of peptide bonds (*b*- and *y*-type ions), *a*-type ions that derive from the corresponding *b*-type ion by losing CO and *c*-type ions that occur in specific cases,[20] are assigned the highest priority for annotation. The chemical structures of regular ions and immonium ions are different, and as a consequence, there is no possible overlap between them. Therefore the order of assignment is of no consequence, and they are treated with the same priority. The second step covers annotations of neutral losses and internal fragment ions; these types derive from regular backbone ions. Importantly, neutral losses are specific to N- or C-termini of fragments or to a single or several amino acids. These are required to be contained in the peptide sequence to allow an annotation. Internal fragment ions originate from regular ions that have undergone a second cleavage of the peptide backbone. The side chains of the amino acids tryptophan (W), arginine (R) and lysine (K) are prone to produce specific fragment ions that can carry a proton because of the heteroatom in their chemical structure. Their mass is sufficiently large (>100 Da) that they are recorded in HCD fragment ion spectra as side chain fragment ions. They are assigned a low priority because they are independent of any

other ion type. Finally, incomplete fragmentation results in protonated precursor ions remaining in the MS/MS spectra, which are annotated as $[M+nH]^{n+}$.

The Expert System greatly improves on the number and intensity coverage of assigned peaks in the fragmentation spectra calculated by adding the signal for the 10 largest peaks per sliding 100 Th window. Standard annotation by the Andromeda search engine results in an intensity coverage of up to 58% for pure spectra (PIF > 0.95; highlighted in gray in Figure 2B). Including the additional ion types that are covered by the Expert System increased the intensity coverage to 84%. With the Expert System in hand, we next annotated all of the about 100 000 high scoring fragment spectra in the initial set. This showed that even for impure MS/MS spectra (PIF less than 0.5), the intensity coverage of assigned peaks in MS/MS spectra was still above 50%. A typical MS/MS spectrum with a high PIF precursor that was comprehensively annotated by the Expert System is displayed in Figure 2C. Virtually all major peaks are correctly annotated and fragment intensity coverage reaches 87%. The figure also illustrates the mass accuracy typically achieved in our experiment. Even though the lock mass feature during data acquisition was enabled,[48] data analyzed with Max Quant is routinely independently recalibrated.[49]

Sequence Related Information Content of HCD Spectra

The most important information imbedded in tandem mass spectra relates directly to the amino acid sequence of the peptide. Cleavage of all peptide bonds, resulting in *b*- and *y*-type ion series, would in principle allow read out of the peptide sequence from the MS/MS spectrum in two directions starting from the N- or the C-terminus, respectively. Moreover, combining the *b*- and *y*-ion series highlights complementary *b*- and *y*-type ions pairs that together match the mass of the unfragmented peptide. Complementary pairs provide strong constraints for correct peptide identification and can be used in scoring algorithms even of multiplexed spectra.[50]

In our large collection of HCD data, we found nearly universal evidence for such pairs. Typical spectra have the prominent a_2/b_2 pair (observed in 72% of the peptide sequences) followed by at least a few more *b*-ions. *Y*-ion series were very abundant in our spectra, especially in the middle mass range (450−800 Da). For peptides that were not too long (<20 amino acids), the low mass *b*-ion series almost always had a corresponding, complementary *y*-ion series of high intensity. These trends are well-known from triple quadrupole and quadrupole time-of-flight spectra.

We next evaluated all 16 000 HCD spectra in the collection (Figure 3A). Remarkably, for peptides up to 12 amino acids the *y*-ion series alone provided for at least half of the sequences complete sequence coverage (median 100%), indicating that complete sequencing of such peptides even in routinely acquired large-scale data sets is in principle possible. This includes the order of the two first amino acids, which is normally inaccessible because of the missing y_{n-1} and b_1 ions (see below). With increasing peptide length, the amino acid coverage slowly drops to a median of 50% at a peptide length of 25 amino acids, which was the upper limit in our collection (Figure 3A). The *b*-ion series, in contrast, remains at a constant level, providing about 30% amino acid coverage independent of the peptide length. Taking both ion series together yields median amino acid coverage of 80% percent even for a peptide length of 20 AA.

Besides the percentage of the sequence that is covered by backbone fragmentation, another important parameter is the number of amino acids that can be read out from the MS/MS spectrum as an uninterrupted part of the sequence, i.e., the maximum sequence tags length.[51] A sequence tag of six amino acids is generally unique in the human genome even without added peptide mass information.[6,52] In addition to peptide identification, such stretches are useful for partial de novo sequencing or homology searching. The lower panel in Figure 3A depicts the median sequence tag lengths based on the two different ion series of the identified sequence. Peptides up to 10 amino acids contain a complete *y*-ion based sequence tag, but above this length, the $y_n - 1$ ion is often of too low intensity to be recorded. Even small peptides contain short sequence tags of three amino acids, which are sufficient for peptide identification. When combined with the *y*-ion series, the *b*-ion series helps to increase the sequence tag length for peptides larger than 14 amino acids. The largest median sequence tag length is about 12 amino acids, and it starts to drop from a peptide length of 16 amino acids.

We next compared the sequence related information content of HCD with that of high resolution CID spectra both acquired in the Orbitrap analyzer. A prominent difference is the much larger contribution of the *b*-ion series in CID spectra (Figure 3B). This is due to the higher stability of *b*-ions in ion trap fragmentation processes. Although lower than the *y*-ion series, the *b*-ion series continued to provide a median of more than 50% sequence coverage up to a peptide length of 19 amino acids. Nevertheless, the combined contribution from *y*-ions and *b*-ions was slightly higher for HCD than for CID, which partly reflects the more extensive fragmentation in beam type instruments and the fact that ion series in CID spectra are limited by the low mass cutoff that is inherent to ion trap fragmentation. As a consequence, maximum sequence tag length was likewise higher in HCD spectra.

We have previously investigated maximum sequence tag lengths in low resolution CID spectra. In more than 85% of the identified spectra sequence tags of at least three amino acids and only in half of the spectra sequence tags of six or more amino acids were detected.[52] Despite the potential for overcounting due to the lower mass accuracy, these sequence tags were substantially shorter than tags from either high resolution HCD or high resolution CID.

Neutral Loss Fragments in HCD

During collision-induced dissociation processes, peptides can follow numerous fragmentation pathways and consequently give rise to various ion types beyond those produced by the typical peptide backbone cleavage. A large class of such ions are those involving neutral losses from different fragment species. These occur from nearly all ion types, however, the chemical structures of the diverse ion types as well as the amino acid side chains allow specific neutral losses (Figure 2A). In some cases, these can result either from the peptide terminus or from one of the side chains of the amino acids, and localization of the origin is not straightforward. However, such losses can still be unambiguously assigned to the fragment ion. We carried out a systematic study considering 45 possible chemical compositions that could formally occur as neutral losses from amino acid residues. We then used our large scale data set to determine the primary neutral losses for all of the fragments in the collection that contained the amino acid in question. The median absolute mass accuracy of all neutral losses is 2.7 ppm with 97.5% of the

peaks within 5 ppm, therefore they can unambiguously be connected to their precursor fragments. Only first neutral losses which happened in at least 5% of the cases were considered and encoded in the Expert System.[36] Table 1 summarizes the observed frequencies of the primary neutral losses that occur in different combinations in more than 270 000 fragments. While *b*-type ions frequently lose a water molecule, the chemical structure of *y*-type ions allows both water and ammonia losses. These are by far the most frequent neutral losses. Furthermore, acidic amino acids as well as serine and threonine are likely to lose water. However, it was possible in about 48% of the cases to assign the neutral loss to either a specific amino acid or the C-terminus of the fragment, because there was only one possible origin for the water loss. At least 33% of the spectra from sequences that contain glutamic acid, serine or threonine exhibit water losses from those amino acids. This is the case in only 29% of spectra where the water loss can be confidently assigned to aspartic acid. The rate of ammonia losses is comparable to water losses and this also holds true for confidently assignable losses from glutamine (29%), asparagine (30%) and arginine (21%). Further frequently observed neutral losses that are specific to certain amino acids include CH_3NO from glutamine (20%) and from asparagines (29%) or C_2H_4O from threonine (26%). While other neutral losses may exist, our large data set suggests that they are unlikely to occur at substantial frequencies in HCD spectra.

Table 1. Neutral Losses Considered by the Expert System and Fraction of Spectra That Contain the Loss from the Corresponding Amino Acid[a]

NH_3	45% (N-term)	30% (N)	29% (Q)	21% (R)	
H_2O	48% (C-term)	37% (S)	44% (T)	21% (D)	33% (E)
CO	84% (internal)[b]				
CO_2	5% (D)				
CH_2N_2	8% (R)				
CH_3NO	29% (N)	20% (Q)			
CH_4O	5% (S)				
CH_4SO	89% (Mox)				
C_2H_4	5% (I)				
C_2H_5NO	9% (N)	6% (Q)			
C_2H_4O	26% (T)				
$C_2H_4O_2$	6% (D)	6% (E)			
C_3H_6	6% (L)				
$C_3H_9N_3$	6% (R)				
C_3H_6SO	6% (Mox)				
C_3H_8SO	12% (Mox)				
C_4H_8	5% (L)				
C_8H_7N	6% (W)				
C_9H_9N	12% (W)				

[a]Only examples allowing unambiguous localization of the origin of the neutral loss were considered. [b]This ion is formally equivalent to an *a*-type internal fragment.

Internal Fragments

Internal fragments in the MS/MS spectra are characteristic of beam-type fragmentation because these result from ions undergoing a second cleavage resulting in a C-terminal carboxyl-group and an N-terminal oxazolone structure.[13,53] In our large-scale data set, the length of internal fragments varied between two and more than 10 amino acids, depending on peptide length. The majority of internal fragments, however, are shorter than five amino acids. Proline is most often the first amino acid of an internal fragment since N-terminal cleavage is very pronounced at this amino acid; this is called the proline effect.[54] However, we found that on the basis of peak presence, rather than peak intensity, proline initiated internal sequences were more than four times as common as those of the median of other amino acids (Supporting Information, Figure S1A). For cleavage at the C-terminal amino acid of an internal fragment there is a slight preference for aspartic acid, glutamic acid, glutamine, tryptophan and histidine (Supporting Information, Figure S1B). Proline is the least common amino acid at the C-terminus of internal ions.

Low Mass Region

HCD fragmentation takes place in a dedicated collision cell and is not subject to the low mass cutoff of ion trap CID spectra, therefore in principle allowing observation of the entire mass range. In practice, HCD spectra are normally acquired from m/z 100, but for a more extensive investigation of the low mass region we acquired data in our study from m/z 80, which was the lowest practical m/z without reducing the scan speed of the instrument. Therefore our data set does not contain immonium ions with an m/z lower than 80 Th.

Figure 4B displays the frequency of immonium ions in the MS/MS spectra. The most prominent immonium ions originate from phenylalanine (F), tryptophan (W) and tyrosine (Y) and can be observed in at least 84% of all peptide sequences containing the respective amino acid. This is due to their chemical structure containing both a heteroatom and an aromatic system that are prone to stabilize a positive charge and for the same reason, the immonium ion of histidine (H) is often present (70%). Carbamidomethylated cysteine (caC), glutamine (Q) and glutamic acid (E) immonium ions (61, 52, and 37%, respectively) can also be found relatively abundantly in the spectra. Aspartic acid (D) and asparagine (N) produce a significantly lower rate of immonium ions. Interestingly, immonium ions of isoleucine (I) and leucine (L) are detected in the MS/MS spectra with different frequencies. Immonium ions of glycine (G), alanine (A), serine (S), proline (P), valine (V) and threonine (T) are not observed in our, data as their m/z is below 80 Th. Arginine (R) and lysine (K) represent special cases due to their position at the N-termini of tryptic peptides. A very frequently observed ion is the immonium ion of lysine with an ammonia loss (IM K − NH_3). In fact, this ammonia loss often occurs even without immonium ion, and this was therefore implemented as an exception to the strict requirement for a detected precursor fragment in the Expert System. Immonium ions can be used to support the peptide sequence assignment. In special cases, such as phosphotyrosine (pY), immonium ions can be used as reporter ions to verify the existence and the nature a phosphorylation site (see below).[55,56]

Another fragment ion type in the low mass region are fragment ions that result from cleavage of amino acid side chains in which the molecular structure can stabilize a proton. This is the case for some of the amino acids that contain a nitrogen atom, such as arginine, lysine and tryptophan. The chemical compositions of the side chain fragments and their frequency of occurrence are displayed in Figure 5C. Note that these side chain fragments are different from the *v*-, *w*- and *d*-type ions from high energy CID dissociation carried out on TOF/TOF instruments.[57,58] In addition to the general ion types, certain amino acid side chains follow different

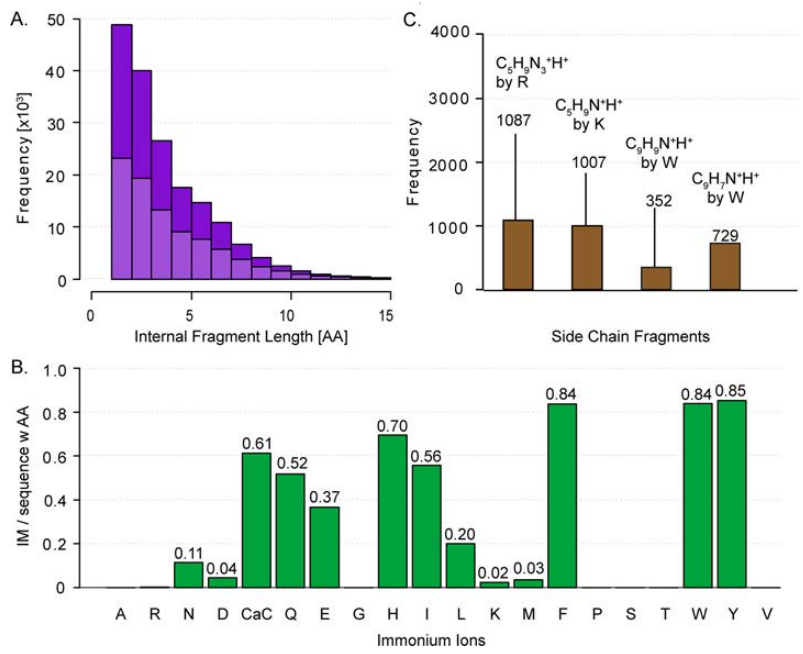


Figure 4. Statistics on the low mass region fragment ions from 16 000 MS/MS spectra. (A) Histogram of the length of all internal fragment ions in purple; the fraction of internal fragment ions starting with proline is highlighted by light color. (B) Percentage of immonium ion (IM) occurrence if the amino acid corresponding amino acid was at least once contained in the peptide sequence. Immonium ions of Alanine, Glycine, Proline, Serine and Threonine were not considered, because their m/z value is lower than 80 Th. (C) Bar plot displaying the five most abundant side chain fragment ions that are automatically assigned by the Expert System with their total number of occurrences within the data set and their chemical structures.

fragmentation pathways resulting in unusual ion types. Lehmann and co-workers observed c_1 ions resulting from the N-terminal amino acid of the peptide, if the second amino acid is glutamine (Q).[19,20,59] Along these lines, we investigated asparagine and carbamidomethylated cysteine and found the same behavior for these two candidates. Furthermore, b_1 ions are usually not observed because of their chemical instability. However, we did observe b_1 ions from acetylation of methionine, serine or alanine at the protein N-terminus. This is thought to be due to stabilization of this fragment by the acetyl group.[18,60] Besides the qualitative information contained in the variety of ion types of natural peptides, the low mass region in HCD fragmentation also gives access to the reporter ions for the TMT[61] and iTRAQ[33] quantification methods.[34,35] The reporter ions of TMT and iTRAQ are at m/z 126.1277, 127.1248, 128.1344, 129.1315, 130.1411, 131.1382 and m/z 114.1112, 115.1146, 116.1116, 117.1150, respectively. Investigation of our large-scale and high accuracy data set revealed no interfering ions of the same m/z. Therefore problems in quantification by these methods are confined to cofragmentation of other labeled peptides rather than other ion types that have the same mass as these reporter ions.

Global Composition of Tryptic HCD Spectra

The different ion types covered by the Expert System, such as regular ions, neutral losses, internal fragments, immonium ions, side chain fragments and the intact peptide mass $[M+nH]^{n+}$ by their nature occur in MS/MS spectra with different frequencies (Figure 5A). However, for high confidence of peptide identification it is predominantly the highly abundant MS/MS peaks that are of interest. Figure 5B displays the contribution of each of the ion types to the overall intensity coverage: Regular ions (a, b, c and y) account for 54% of total MS/MS spectra intensity and peaks that result from neutral losses for a further 15%. Immonium ions can originate from several amino acids, and these signals are added as singly charged peaks at defined masses in the low mass region. Together, their mean contribution to the total intensity coverage is 6%. Unlike immonium ions, internal fragments are spread over the low to middle mass range of the MS/MS spectrum because they can be generated by any two cleavages of the peptide backbone, and hence they are not as obvious in tandem mass spectra. As described above, in HCD internal fragment ions are frequently observed. However, their abundance is lower than that of immonium ions or y-ions, and together they contribute 10% to the total fragment intensity. The protonated unfragmented peptide precursor only has an average intensity coverage less than 1% in our data set. Side chain fragments account for only 0.1% of total peaks and an intensity coverage of less than 0.1% and are therefore not displayed in the pie chart. The fraction of unannotated peaks accounts for 44% on the basis of the 10 largest peaks per hundred Th but only for 15% with regard to total intensity coverage. This provides evidence that remaining peaks are mainly of low abundance. Note that those, beyond potentially being noise peaks, could also result from combinations of multiple neutral losses without precursor fragments or similar, which were not allowed by the Expert System to maintain a strict false positive rate. Furthermore, cofragmentation of other precursors still occurs in our data set to some degree. Together, our data suggests that nearly all fragment peaks in HCD are

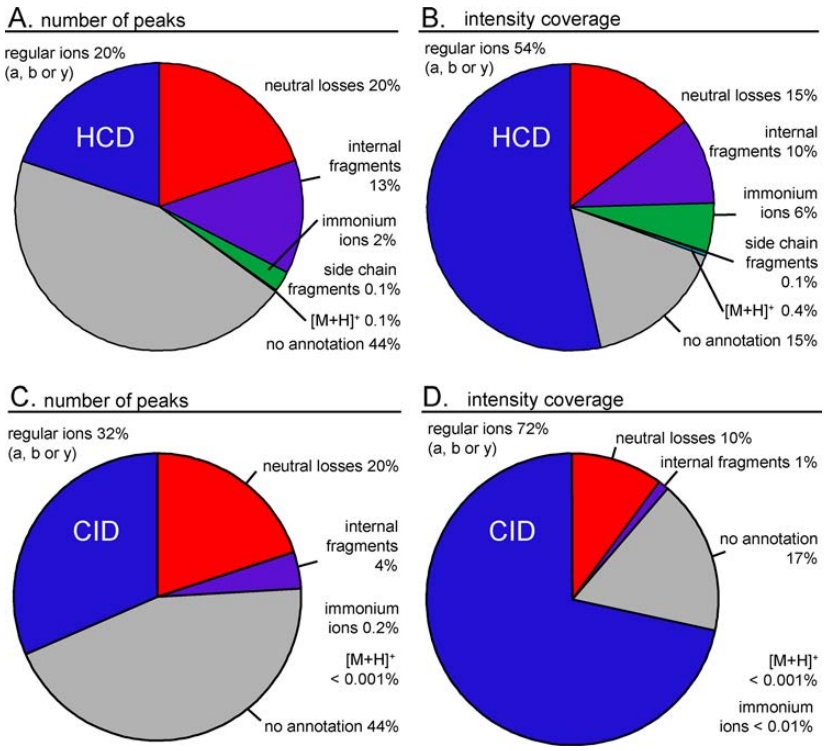


Figure 5. Intensity distribution of different ion types. (A) Average proportions of the six major ion types in HCD spectra by peak count based on a sliding mass window filtering for the 10 most abundant peaks per 100 Da; >16 000 tandem mass spectra. (B) Same as (A) but referring to the intensity coverage of the MS/MS spectrum. (C and D) Same as (A) and (B) for >3200 high resolution CID tandem mass spectra for comparison to the HCD ion type distribution.

explainable on the basis of current understanding of fragmentation pathways.

We next repeated the same analysis as above for high resolution CID spectra, which resulted in quite similar findings for the number of peaks. As expected, the number of immonium ions and internal fragments was drastically reduced since ion trap fragmentation is not capable of retaining the low mass region of the tandem mass spectra and their formation requires double cleavage. Together with the higher preponderance of high mass *b*-ions, this has the effect of increasing the fraction of regular ions to 32% as compared to the 20% of HCD fragmentation. On the basis of intensity coverage, this effect is less pronounced (72% for CID compared to 54% for HCD). Interestingly, using the Expert System the fraction of unannotated peaks by intensity is very similar between CID (17%) and HCD (15%).

Characteristics of Phosphorylated Peptides

Protein phosphorylation is among the most important and best studied post-translational modifications and is almost always located at serine, threonine or tyrosine in mammalian cells. Because of its chemical nature, the phosphogroup easily detaches from serine and threonine during collision induced fragmentation processes resulting in very characteristic and abundant neutral loss peaks such as HPO_3 and H_3PO_4. Furthermore, as already mentioned above, phosphotyrosine leads to a unique and characteristic immonium ion with m/z 216.0426.

We investigated large scale phosphorylation data with the Expert System, incorporating rules for the above-mentioned phosphospecific fragment ions. We found that the occurrence of both neutral losses from phosphorylated serine is about four times as high (65% for HPO_3 and 49% for H_3PO_4) as from threonine (18 and 12%, respectively). Table 2 summarizes the frequencies of these neutral losses. Their absolute number reveals an average of three H_3PO_4 losses and two HPO_3 losses per spectrum.

Table 2. Fraction of 1157 Spectra of Modified Sequences (Phospho STY) Containing Neutral Losses, Reporter Ions from Phosphorylated Serine (S) and Threonine (T) or the Characteristic *X*-Ion at Least Once[a]

	$-HPO_3$	$-H_3PO_4$	pS	pT	x_n (S,T)
S (1094)	65% (713)	49% (540)	29		279
T (585)	18% (103)	12% (68)		3	

[a]The first column lists the total number of sequences that contain the amino acid S or T at least once.

Finally, we investigated the frequency of x_n ions pinpointing the localization of a serine or threonine phosphor site in the peptide sequence very recently described by Kelstrup et al.[62] Our data set consisting of 1157 spectra of phosphorylated peptide sequences contains this characteristic x_n ion in 279 of the fragmentation patterns (24%).

dx.doi.org/10.1021/pr3007045 | *J. Proteome Res.* 2012, 11, 5479–5491

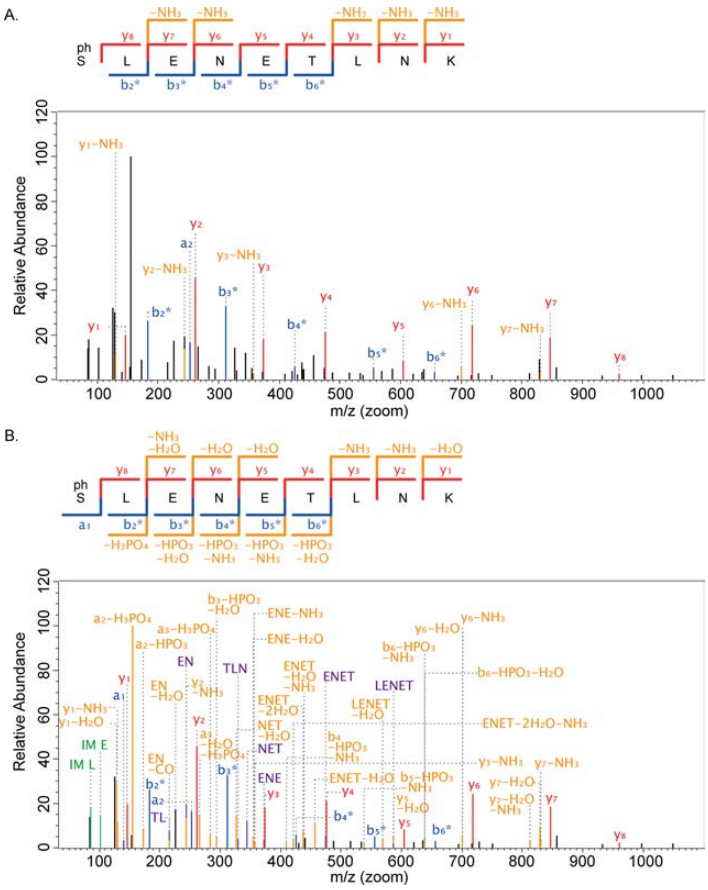


Figure 6. Annotated spectrum of phosphorylated peptide fragmented with HCD. (A) The phosphorylated peptide phSLENETLNK was identified and annotated by the Andromeda search engine assigning regular ions and single neutral losses. (B) The Expert System was modified for phosphorylated peptides to enable comprehensive annotation: Several additional neutral losses, internal fragments and immonium ions increase the intensity coverage to 82%.

■ CONCLUSION AND OUTLOOK

In 2007, beam-type fragmentation was introduced on Orbitrap instrumentation. This HCD mode of fragmentation has become especially popular since some limitations of ion source brightness and ion extraction from the collision cell were removed.[31] In our group, for instance, both proteome and PTM-based investigations are routinely done with HCD rather than low or high resolution CID. This was one reason why it was important to investigate the ion types that are produced by HCD. However, even though the general dissociation mechanisms operative in CID have been studied for decades,[63,64] large data sets with very high quality thresholds have previously not been studied. This was made possible here by very stringent filtering of peptide fragment spectra on the basis of identification score as well as near absence of cofragmenting peaks. Most importantly, we developed and made use of an Expert System, which annotated peptide peaks with high comprehensiveness but low false positive rates.

Our investigation of HCD yielded a broad and quantitative overview of the ion types produced. It turns out that HCD spectra are somewhat more complex than CID spectra but that the peaks are assignable to the same degree. The low mass region is particularly straightforward to interpret given the very high resolution of the Orbitrap analyzer in this region, coupled to the high mass accuracy, which generally allows determination of the chemical composition of these fragments. The information content of HCD spectra is mostly related to very extensive series of *y*-ions, supplemented by relatively short series of low mass *b*-ions. This is in contrast to ion trap CID spectra, in which the high mass *b*-ions are also very prominent. Nevertheless, the coverage of peptide sequence overall and in particular with continuous ion series is somewhat higher in HCD than it is in CID. Remarkably, for tryptic peptides up to 15 amino acids, the fragment contents is almost complete, meaning that there is sufficient information in principle for de novo sequencing or at least very long sequence tags.

Our quantification of the overall contribution of different ion types to the entire MS/MS spectrum revealed that only a relatively small proportion remains unassigned by the rules that we have implemented into the Expert System. This proportion would further shrink if noise and remaining cofragmentation was further reduced and if the rules of the Expert System were relaxed. This means that the ion types produced in HCD and by extension by CID are already very well understood. New

fragmentation pathways of standard peptides could of course be discovered in the future, but it is unlikely that such ions would contribute very much to the overall ion intensity. For modified peptides, our Expert System and quantification of fragmentation frequencies could help to discover potential new fragment types. In this connection, we have already demonstrated straightforward extension of our approach to phosphorylated peptides. In conclusion, we have here reported the most extensive investigation into HCD of peptides and hope that the results will be useful for both small and large scale investigation of the proteome.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1: Barplot displaying the number of internal fragments of an amino acid as (A) first (N-terminal) (B) last (C-terminal) amino acid of an internal fragment divided by the number of occurences of the amino acid in the dataset. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mmann@biochem.mpg.de. Fax: +49 89 8578 2219.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank our colleagues at the Max Planck Institute of Biochemistry for help and fruitful discussions, in particular Kirti Sharma for the preparation of the phosphoproteome samples. The research leading to these results has received funding from the European Commission's Seventh Framework Programme (Grant Agreement HEALTH-F4-2008-201648/PROSPECTS).

■ ABBREVIATIONS

CID, collision-induced dissociation; ETD, electron transfer dissociation; FDR, false discovery rate; FT, Fourier transform; HCD, higher energy collisional dissociation; HPLC, high performance liquid chromatography; ICR, ion cyclotron resonance; IM, immonium ion; IPI, international protein index; LTQ, linear trap quadrupole; MS/MS, tandem mass spectrometry; PIF, precursor intensity fraction; Q TOF, quadrupole time-of-flight instrument; TOF, time of flight

■ REFERENCES

(1) Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R. Generating and navigating proteome maps using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **2010**, *11* (11), 789−801.
(2) Mallick, P.; Kuster, B. Proteomics: a pragmatic perspective. *Nat. Biotechnol.* **2010**, *28* (7), 695−709.
(3) Cox, J.; Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **2011**, *80*, 273−99.
(4) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787−97.
(5) White, F. M. The potential cost of high-throughput proteomics. *Sci. Signaling* **2011**, *4* (160), pe8.
(6) Ma, B.; Johnson, R. De novo sequencing and homology searching. *Mol. Cell. Proteomics* **2012**, *11* (2), O111 014902.
(7) Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biomed. Environ. Mass Spectrom.* **1988**, *16* (1−12), 99−111.
(8) Oomens, J.; Young, S.; Molesworth, S.; van Stipdonk, M. Spectroscopic evidence for an oxazolone structure of the b(2) fragment ion from protonated tri-alanine. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (2), 334−9.
(9) Bythell, B. J.; Somogyi, A.; Paizs, B. What is the structure of b(2) ions generated from doubly protonated tryptic peptides? *J. Am. Soc. Mass Spectrom.* **2009**, *20* (4), 618−24.
(10) Perkins, B. R.; Chamot-Rooke, J.; Yoon, S. H.; Gucinski, A. C.; Somogyi, A.; Wysocki, V. H. Evidence of diketopiperazine and oxazolone structures for HA b(2)(+) Ion. *J. Am. Chem. Soc.* **2009**, *131* (48), 17528−9.
(11) 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), 1399−406.
(12) Boyd, R.; Somogyi, A. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (8), 1275−8.
(13) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508−48.
(14) Medzihradszky, K. F. Peptide sequence analysis. *Methods Enzymol.* **2005**, *402*, 209−44.
(15) Papayannopoulos, I. A. The interpretation of collision-induced dissociation tandem mass-spectra of peptides. *Mass Spectrom. Rev.* **1995**, *14* (1), 49−73.
(16) Falick, A. M.; Hines, W. M.; Medzihradszky, K. F.; Baldwin, M. A.; Gibson, B. W. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass-spectrometry. *J. Am. Soc. Mass Spectrom.* **1993**, *4* (11), 882−93.
(17) Chalkley, R. J.; Baker, P. R.; Huang, L.; Hansen, K. C.; Allen, N. P.; Rexach, M.; Burlingame, A. L. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell. Proteomics* **2005**, *4* (8), 1194−204.
(18) Hung, C. W.; Schlosser, A.; Wei, J. H.; Lehmann, W. D. Collision-induced reporter fragmentations for identification of covalently modified peptides. *Anal. Bioanal. Chem.* **2007**, *389* (4), 1003−16.
(19) Winter, D.; Lehmann, W. D. Sequencing of the thirteen structurally isomeric quartets of N-terminal dipeptide motifs in peptides by collision-induced dissociation. *Proteomics* **2009**, *9* (8), 2076−84.
(20) Lee, Y. J.; Lee, Y. M. Formation of c1 fragment ions in collision-induced dissociation of glutamine-containing peptide ions: a tip for de novo sequencing. *Rapid Commun. Mass Spectrom.* **2004**, *18* (18), 2069−76.
(21) Reid, G. E.; Roberts, K. D.; Kapp, E. A.; Simpson, R. J. Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *J. Proteome Res.* **2004**, *3* (4), 751−9.
(22) Harrison, A. G.; Young, A. B.; Bleiholder, C.; Suhai, S.; Paizs, B. Scrambling of sequence information in collision-induced dissociation of peptides. *J. Am. Chem. Soc.* **2006**, *128* (32), 10364−5.
(23) Bleiholder, C.; Osburn, S.; Williams, T. D.; Suhai, S.; Van Stipdonk, M.; Harrison, A. G.; Paizs, B. Sequence-scrambling fragmentation pathways of protonated peptides. *J. Am. Chem. Soc.* **2008**, *130* (52), 17774−89.
(24) Goloborodko, A. A.; Gorshkov, M. V.; Good, D. M.; Zubarev, R. A. Sequence scrambling in shotgun proteomics is negligible. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (7), 1121−4.
(25) Yu, L.; Tan, Y.; Tsai, Y.; Goodlett, D. R.; Polfer, N. C. On the relevance of peptide sequence permutations in shotgun proteomics studies. *J. Proteome Res.* **2011**, *10* (5), 2409−16.
(26) Xia, Y.; Liang, X. R.; McLuckey, S. A. Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions. *Anal. Chem.* **2006**, *78* (4), 1218−27.

(27) Schwartz, J. C.; Senko, M. W.; Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (6), 659−69.
(28) 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), 709−12.
(29) McAlister, G. C.; Phanstiel, D. H.; Brumbaugh, J.; Westphall, M. S.; Coon, J. J. Higher-energy collision-activated dissociation without a dedicated collision cell. *Mol. Cell. Proteomics* **2011**, *10* (5), O111 009456.
(30) Horner, J. A.; Remes, P.; Biringer, R.; Huhmer, A.; Specht, A. Achieving increased coverage in global proteomics survey experiments using higher-energy collisional dissociation (HCD) on a linear ion trap mass spectrometer. *Application Note 538*; Thermo Fisher Scientific: San Jose, CA, 2011.
(31) Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **2009**, *8* (12), 2759−69.
(32) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (47), 18132−8.
(33) Ross, P. L.; Huang, Y. L. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3* (12), 1154−69.
(34) Bantscheff, M.; Boesche, M.; Eberhard, D.; Matthieson, T.; Sweetman, G.; Kuster, B. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2008**, *7* (9), 1702−13.
(35) Pichler, P.; Kocher, T.; Holzmann, J.; Mohring, T.; Ammerer, G.; Mechtler, K. Improved precision of iTRAQ and TMT quantification by an axial extraction field in an Orbitrap HCD cell. *Anal. Chem.* **2011**, *83* (4), 1469−74.
(36) Neuhauser, N.; Michalski, A.; Cox, J.; Mann, M. Expert System for computer assisted annotation of MS/MS spectra. *Mol. Cell. Proteomics* **2012**.
(37) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1* (6), 2856−60.
(38) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75* (3), 663−70.
(39) Wisniewski, J. R.; Zougman, A.; Mann, M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* **2009**, *8* (12), 5674−8.
(40) Pinkse, M. W.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal. Chem.* **2004**, *76* (14), 3935−43.
(41) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **2012**, *11* (3), O111 013698.
(42) 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), 1367−72.
(43) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10* (4), 1794−805.
(44) Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graphical Stat.* **1996**, *5* (3), 16.
(45) 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), 1785−93.
(46) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152−60.
(47) Wang, J.; Bourne, P. E.; Bandeira, N. Peptide identification by database search of mixture tandem mass spectra. *Mol. Cell. Proteomics* **2011**, *10*, 12.
(48) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4* (12), 2010−21.
(49) Cox, J.; Michalski, A.; Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (8), 1373−80.
(50) Ledvina, A. R.; Savitski, M. M.; Zubarev, A. R.; Good, D. M.; Coon, J. J.; Zubarev, R. A. Increased throughput of proteomics analysis by multiplexing high-resolution tandem mass spectra. *Anal. Chem.* **2011**, *83* (20), 7651−6.
(51) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390−9.
(52) Cox, J.; Hubner, N. C.; Mann, M. How much peptide sequence information is contained in ion trap tandem mass spectra? *J. Am. Soc. Mass Spectrom.* **2008**, *19* (12), 1813−20.
(53) Ballard, K. D.; Gaskell, S. J. Sequential mass-spectrometry applied to the study of the formation of internal fragment ions of protonated peptides. *Int. J. Mass Spectrom.* **1991**, *111*, 173−189.
(54) Harrison, A. G.; Young, A. B. Fragmentation reactions of deprotonated peptides containing proline. The proline effect. *J. Mass Spectrom.* **2005**, *40* (9), 1173−86.
(55) Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal. Chem.* **2001**, *73* (7), 1440−8.
(56) Boersema, P. J.; Mohammed, S.; Heck, A. J. Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **2009**, *44* (6), 861−78.
(57) Johnson, R. S.; Martin, S. A.; Biemann, K. Collision-induced fragmentation of (M+H)+ ions of peptides—side-chain specific sequence ions. *Int. J. Mass Spectrom.* **1988**, *86*, 137−54.
(58) Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **2000**, *72* (3), 552−8.
(59) Winter, D.; Seidler, J.; Hahn, B.; Lehmann, W. D. Structural and mechanistic information on c(1) ion formation in collision-induced fragmentation of peptides. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (10), 1814−20.
(60) Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. Why are B ions stable species in peptide spectra? *J. Am. Soc. Mass Spectrom.* **1995**, *6* (12), 1165−74.
(61) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895−904.
(62) Kelstrup, C. D.; Hekmat, O.; Francavilla, C.; Olsen, J. V. Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *J. Proteome Res.* **2011**, *10* (7), 2937−48.
(63) Khatun, J.; Ramkissoon, K.; Giddings, M. C. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem.* **2007**, *79* (8), 3032−40.
(64) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., 3rd Statistical characterization of ion trap tandem mass

spectra from doubly charged tryptic peptides. *Anal. Chem.* **2003**, *75* (5), 1155−63.

2.4 article 4 - High performance computational analysis of large-scale proteome datasets to assess incremental contribution to coverage of the human genome

Nadin Neuhauser, Nagarjuna Nagaraj, Peter McHardy, Sara Zanivan, Richard Scheltema, Jürgen Cox, and Matthias Mann
submitted on 28/02/2013 to Journal of proteome research
accepted on 23/04/2013 by Journal of proteome research

Since the data amount produced in proteomics laboratories is continuously increasing, the aim of this project was to figure out how to accelerate our computational analysis platform MaxQuant. For this purpose there are two major focus areas: On the one hand the algorithms in MaxQuant can be improved for instance by efficient parallelization of processes in the pipeline. On the other hand we tested different hardware setups beginning with a standard desktop computer, proceeding to an I/O optimized high end computer up to a large-scale computer cluster to evaluate the best cost and time efficient solution.

The goal of the software optimization was to adapt MaxQuant so that the software can run efficiently on different hardware configurations. For this reason I first identified the major bottlenecks of a standard proteome analysis. For instance the peptide search was parallelized using the data decomposition technique, where the MS/MS spectra are efficiently distributed and analyzed in parallel.

Concurrently to the parallelization, I adjusted MaxQuant to run on the computer cluster of our institute, which consists of 44 nodes each with 8 virtual cores. Hereof two of the nodes are taken only for job submission and the remaining 42 are allocated as working nodes. Since, our analysis pipeline is restricted to the Windows operating system the Windows Server HPC 2008 R2 is installed on all nodes. For running MaxQuant on the computer cluster, I implemented an interface which handles the interactions between the pipeline and the job handler of the cluster. The input, output and error streams are redirected to the MaxQuant interface as well as to text files, which became quite important during the developing and evaluating phase. Since the distribution of the computational tasks is at the node not at the core level (-due to technical reasons), they are packed in a way, that each node will handle eight parallel processes.

After efficient parallelization of the software, we compared different hardware settings. A standard desktop computer with only four or eight central processing units

(CPUs) has only a small fraction of the computational power of the cluster with 336 CPUs. This difference is clearly visible in the performance analysis - primarily for larger datasets. We also included in our evaluation an I/O optimized computer where the data is stored on a solid state disk (SSD). In comparison to the cluster, this configuration was performing similarly, which is mainly explained by the high I/O demand of MaxQuant. We conclude that in the future the I/O optimized hardware will be used.

On the computer cluster we then applied the new version of MaxQuant to determine the coverage of protein groups mapped to the human protein coding genes. For this, I collected six data sets from cell lines, cancer tissue and a body fluid of current high quality measurements which resulted in more than 1000 LC-MS raw files. With the conventional setup on a standard desktop computer, analysis of this large data set would have taken many weeks, but the cluster finished this task in less than six days. As an result we have identified more than 13,000 protein groups - corresponding to 12,000 coding genes, which is around 60% of the entire genome. Given the fact that 30% of the gene coding regions have no stringent evidence at the protein level[3], this is already a substatial coverage from just af few experiments in a single laboratory. The reasons why coverage was not higher still, are not clear at present. Perhaps a combination of more specialized and distinct proteome sources and additional advances in proteomic technology are required to approach complete coverage of the human genome.

High performance computational analysis of large-scale proteome datasets to assess incremental contribution to coverage of human genome

Nadin Neuhauser[1], Nagarjuna Nagaraj[1], Peter McHardy[2], Sara Zanivan[2], Richard Scheltema[1], Jürgen Cox[1] and Matthias Mann[1],*

[1]From the Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[2]From the Vascular Proteomics Lab, Beatson Institute for Cancer Research, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK

Keywords: tandem mass spectrometry, shotgun proteomics, performance, analysis pipeline

*Corresponding author:
Matthias Mann
Department of Proteomics and Signal Transduction
Max-Planck Institute of Biochemistry
Am Klopferspitz 18
D-82152 Martinsried
Germany
Email: mmann@biochem.mpg.de
Phone: +49 (89) 8578 - 0
Fax: +49 (89) 8578 - 37 77

Abbreviations: MS, mass spectrometry; MS/MS, tandem mass spectrometry; PC, personal computer; HPC, high performance computing; HDD, hard disk drive; SSD, solid state disk; GPFS, general parallel file system; FDR, false discovery rate; ATA, advanced technology attachment; SATA, serial ATA; SCSI, small computer system interface; SAS, serial attached SCSI; Th, Thomson; PSM, peptide spectrum match; CPU, central processing unit; I/O input/output

Abstract

Computational analysis of shotgun proteomics data can now be performed in a completely automated and statistically rigorous way, as exemplified by the freely available MaxQuant environment. The sophisticated algorithms involved and the sheer amount of data translate into very high computational demands. Here we describe parallelization and memory optimization of the MaxQuant software with the aim of executing it on a large computer cluster. We analyze and mitigate bottlenecks in overall performance and find that the most time consuming algorithms are those detecting peptide features in the MS1 data as well as the fragment spectrum search. These tasks scale with the number of raw files and can readily be distributed over many CPUs as long as memory access is properly managed. Here we compared the performance of a parallelized version of MaxQuant running on a standard desktop, an I/O performance optimized desktop computer ('game computer'), and a cluster environment. The modified gaming computer and the cluster vastly outperformed a standard desktop computer when analyzing more than 1000 raw files. We apply our high performance platform to investigate incremental coverage of the human proteome by high resolution MS data originating from in-depth cell line and cancer tissue proteome measurements.

Introduction

The technology of mass spectrometry (MS)-based proteomics has been improving at a very fast rate during the last two decades[1-4]. Advances in instrumentation have reduced acquisition time and increased resolution and sensitivity, which in combination with the high resolving-power of current mass analyzers in both MS and MS/MS mode have led to very large data sets. For instance, in our laboratory we routinely acquire 400,000 to 650,000 data-dependent MS/MS spectra for every quadrupole - Orbitrap instrument[5] per day, requiring approximately 14 GB of storage space. The growing file size and number of raw files dramatically increases the burden on the computational tools used to analyze these data[6]. This analysis is becoming so computationally intensive that it can preclude processing on a standalone personal computer (PC) or make it so slow that it prevents researchers from trying different scenarios or hypotheses[7].

Once the basic algorithms in the computational proteomics pipeline have been thoroughly optimized, overall performance improvements rely on better computational hardware. In addition to faster processors, developments in computer science have increasingly focused on the use of multiple processors in parallel. Parallelism as such has been employed for many years, mainly in high-performance computing (HPC). With parallel computing, computational problems are divided into smaller ones and solved concurrently, distributing the computational effort in many cases over multiple CPUs (central processing unit). This can be among multiple cores within a single processor, a multiprocessor system or a network of computers - a so-called computing cluster. However, this new hardware requires parallelized algorithms, to benefit from the increased hardware capacities. This is by no means trivial and most existing applications cannot exploit multi-core systems yet. Typically only some parts of computational problems can be completely parallelized and overall performance is frequently limited by access to shared resources or communication between tasks. Despite these obstacles, the number of applications that use parallelization is gradually increasing. As an example from bioinformatics on next-generation sequencing data, an algorithm for sequence alignment has recently been re-implemented using the principle of parallelization to use the power of a multi-core environment[8]. The parallel algorithm had an analysis time 20 times faster than the non-parallelized (serial) version.

In computational proteomics, data analysis typically involves several steps and is not confined solely to peptide identification by a peptide search engine such as Mascot[9]. There are few software solutions that aim to provide data analysis from acquired raw data to final protein lists in a single environment. Examples are the Trans-Proteomic

Pipeline[10], OpenMS Proteomics Pipeline[11] or Skyline[12]. Our own laboratory has developed the MaxQuant computational proteomics framework, which is freely available to academic and commercial users and which has been widely adopted by the research community[13, 14]. MaxQuant enables processing of raw MS data files, incorporates its own probabilistic search engine called Andromeda[15] and has recently been supplemented by the extensible Perseus environment for statistical and functional analysis[16]. To increase the performance of our analysis pipeline we here adapt MaxQuant to exploit non-shared memory parallel computing, so it can be run on a high-performance computer cluster. Due to the fact that many of the component tasks are independent of each other, the MaxQuant pipeline could be substantially parallelized. This led to dramatically increased performance, which we demonstrate here by analyzing the coverage of the human genome by large-scale data sets from high resolution shotgun proteomics. We also compare performance of the cluster to intermediate hardware solutions such as high performance personal computers, which would be economically accessible to all groups using state of the art proteomics.

Experimental Methods

Implementation of MaxQuant - Originally MaxQuant was developed to run on desktop computers with one or multiple cores, which can support a semi-parallelized instance of the software. The cluster instead has a large number of nodes, consisting of multiple cores (see below). To keep MaxQuant independent from the hardware setup during parallelization, the original implementation was refactored. This step left the core algorithms for desktop and cluster versions identical, only differing in the way that the single tasks in the analysis pipeline are called from the exchangeable framework. For the desktop version the MaxQuant software itself is in charge of executing the code at the correct time in the pipeline. For the cluster this control needs to be relinquished to a job manager, requiring a new interface that uses the Job Manager provided by Windows HPC 2008 R2. MaxQuant automatically generates a job instance spanning several tasks and passes the instance to the job manager, which then distributes the tasks over all available nodes. The job manager is aware of all resources and takes care of the task queue which can originate from different users. For this reason the graphical interface of MaxQuant can be closed after submitting the job, in contrast to the desktop version.

Next we set out to adapt MaxQuant to efficiently use the power of the high perfor-

mance cluster. The principle units of parallelization are the raw files from the project to be analyzed; and the basic structure is to allocate each raw file to a physical or virtual core. In the desktop version, we had used multiple processes, which enabled semi-parallelization because the number of cores is limited on a standard PC. The challenge was therefore to correctly distribute the different tasks over several nodes.

The code in MaxQuant is structured in so-called 'task groups' which are co-dependent and have to run one after the other. As an example, detecting the features in MS1 scans is a task group. Each of these task groups consists of several instances, where the number of instances is dependent on the number of raw files. As these instances can run in parallel, we distribute them over the available nodes according to how many cores are available on each node. The code executed on each of the nodes is the same as on the desktop version. Implementing this basic parallelization initially led to low usage of the computing power of the nodes, because only few of the necessary tasks truly ran concurrently.

To enable more efficient parallelization we first identified the bottlenecks in performance. In this process the poorly performing sections were iteratively identified that could safely be executed in parallel. For instance, protein group assembly consumed a disproportional amount of time (see Results and Discussion). Within this task group we identified functionality that can run in parallel and split this task group into three new task groups: 'Prepare protein assembly', 'Assembling protein groups' and 'Finish protein assembly'. Of these tasks, 'Assembling protein groups' can be broken up into many small parts that can be executed in parallel (i.e. each protein group can be processed independently); whereas the other tasks cannot be performed in parallel as they consist of a single task. With this improvement we obtained an enormous speedup in this part of the pipeline. This process was performed on the most time consuming task groups in the pipeline. Compared to previous versions consisting of 22 task groups (Version 1.2.0.0) under default conditions we now have 38 task groups of which 20 groups are parallelized. This division has the advantage that so called fallback positions are created enabling partial processing where the researcher can for example reprocess a part of the pipeline with different settings.

All of the parallelization improvements made for the cluster version also benefit the normal PC version when many CPU cores are available. As a last step we identified the major bottleneck for these types of machine, which turned out to be the input and output (I/O) access to hard drives. To mitigate this bottleneck we optimized the hardware as described below.

Hardware setup - For performance benchmarking we used three different hardware setups (see Table 1). As a representative normal desktop PC we used an Intel Core i7-2600 processor with 3.4 GHz, 16 GB of RAM and 460 GB space on a conventional hard disk drive (HDD) with a serial advanced technology attachment (SATA) connection (purchased from Dell Computers). Since such a computer is meant to still be available for normal office work, we use only 4 of the 8 virtual cores for processing the data sets. Additionally we chose a high performance desktop computer, custom-built for advanced video gaming, which is employed in our department for highly demanding computations. This type of computer has the similar processor, but is equipped with 1 TB of solid state disks (SSD) configured in RAID 0 connected via a PCI-Express RAID controller with a battery backup unit and full cache enabled. A RAID configuration is providing a potential factor of two in read access speed as the data is duplicated on both drives. The I/O optimized machine also uses faster memory (DDR3 1866MHz Quad Channel). This computer was purchased from Eclipse Computing, Ayrshire, UK and costs two to three times the amount of a desktop computer designed for typical computational tasks (for current configuration employed in our department see www.maxquant.org). We store our data on the solid state drive, which has a dramatic effect on the crucial I/O performance bottlenecks. These configurations were compared to our Windows cluster equipped with 44 nodes, two of which are exclusively used to submit MaxQuant jobs. Each node consists on an Intel Xeon E5540 processor with 2.53 GHz and 24 GB of RAM. For the global data storage we found it advantageous to install a high performance general parallel file system (GPFS) with 10 TB of storage space on a HDD using the SAS protocol.

Table 1: Key parameters of the three hardware platforms

	standard desktop PC	**high-end gaming PC**	**computer cluster**
computing capacity	using 4 virtual cores	using 8 cores	using 336 virtual cores
computing power	3.4 GHz, 16 GB of RAM	3.4 GHz, 32 GB of RAM	2.53 GHz, 24 GB of RAM
I/O performance	HDD, SATA	SSD, PCI Express RAID	HDD, SAS + GPFS client

A 64-bit version of Windows 7 is installed on the standard desktop computer and on the I/O optimized high end computer, whereas the cluster was run with a 64-bit version of Windows HPC 2008 R2. The installation of the freely available Thermo MS FileReader and .NET Framework 4.5 is necessary for all three platforms. For more

information see www.maxquant.org/requirements.htm

We have also begun testing two rack mounted configurations with 64 logical processors, where multiple cores share the same memory. Both machines have 128 GB of memory (16x9 GB Dual Rank RDIMM) for 4 CPUs with 1600 MHz. The major difference for these two setups is that one is using 4 AMD processors (Opteron 6276 with 2.3 GHz) and the other is designed with 4 Intel processors (Intel Xeon E5-4640 with 2.4GHz). Additional a RAID Controller PERC H700 or PERC H710p with 1GB NV cache is installed, respectively. The storage space is basically the same 6x 900 GB HDD using SAS protocol. The operating system on both solutions is the Microsoft Windows Server 2012 Standard 64-bit.

Datasets for human proteome analysis - To obtain a large dataset for evaluating the performance of the different hardware setups, we combined raw files from different published experiments from our group. In total we used data from in-depth proteomics studies of 30 different cell lines17, 18 resulting in 763 raw files. To cover more of the human proteome we also included raw files from two tissue19, 20 and one body fluid projects21. For the estimation of the measured human proteome we employed a total of 1004 raw files from the studies listed in Table 2. For an estimation of the runtime behavior of our application we tested five data sets with varying raw file numbers (6, 18, 198, 343 and 763 raw files) on the three different hardware setups.

Table 2: Key parameters of the three hardware platforms

	name	**enzyme**	**type**	**instrument**	**raw files**	**scans**
1	11 cell lines[17]	trypsin	cell lines	LTQ Orbitrap XL	198	7,779,031
2	8 cell lines	trypsin	cell lines	Q Exactive	145	17,477,801
3	breast cancer[18]	trypsin	cell lines	LTQ Orbitrap XL	420	7,448,447
4	colon cancer - I[19]	trypsin	tissue	LTQ Orbitrap XL	135	4,461,151
5	colon cancer - II[20]	typsin lys-C	tissue	Q Exactive	24	2,000,864
6	urinary proteome [21]	lys-C	body fluid	LTQ Orbitrap XL	82	1,495,293
					1004	**40,662,587**

Data analysis - All data were processed with MaxQuant13 version 1.3.7.4 using Andromeda15 to search the MS/MS spectra with trypsin or LysC specificity against the

complete human dataset of the UniProt database22 (release January 2013, 87,638 entries) combined with 262 commonly detected contaminants. We allow for up to two missed cleavages and N-terminal acetylation and methionine oxidation were selected as variable, carbamidomethylation of cysteine was selected as fixed modification. For MS spectra an initial mass accuracy of 4.5 ppm was allowed and the MS/MS tolerance was set to 20 ppm. A sliding mass window was applied to filter the MS/MS spectra for the 10 most abundant peaks in 100 Th. For identification, the FDR at the peptide spectrum matches (PSM) and protein level was set to 0.01.

Availability - The desktop version of MaxQuant as well as the special cluster version are freely available at http://www.maxquant.org/downloads.htm

Results and Discussion

Many of the tasks in computational proteomics place very challenging demands on the computational hardware. These demands can be thought of as a combination of three different factors: (i) processing power of the computer chips or cores employed (ii) the number of these cores and (iii) the speed of read and write operations. Importantly, an improvement in any one can fail to improve the analysis time when other factors still act as a bottleneck to the whole system. For example, extremely high processing speed may be practically unimportant if reading of raw data or distribution to the relevant cores is slow.

The processing power of single cores improves over the years. For setting up a computational pipeline one typically selects the fastest and most cost effective version of mainstream and mass produced products, such as Intel or AMD chips. The trend in high-performance computing has been to group multiple processors together, both in single chips and by connecting large numbers of chips (computing clusters). In principle the computational capacity is multiplied by the number of chips, however, this requires efficient parallelization of the software (discussed below). Furthermore, data for processing needs to be available to the cores and intermediate results need to be written out sufficiently fast so as not to slow down overall performance. This may require equipping the cores with large individual memory stores and advanced overall memory management.

Given efficient hardware for computationally intensive tasks, the software needs to be structured to take optimal advantage of the resources. In general, one tries to di-

vide the computational workflow for one proteomic analysis (a 'job') into largely self-contained units of 'tasks' that run independently as separate processes. With this strategy, tasks normally do not communicate with each other. Designing a parallel workflow therefore involves 'decomposition', which entails breaking down a complex system into smaller pieces, to find tasks that can run concurrently in parallel applications. There are two major decomposition methods in parallel programming, functional and data decomposition23. Functional decomposition requires a restructuring of the algorithms into independent units, which can be very challenging. Data decomposition is used more often, because it only requires a solid understanding of the data and how the algorithms process it. In the context of computational proteomics data decomposition can take the form of processing each of the raw files on a different core.

Once a significant fraction of the proteomic analysis pipeline is separated into independent tasks executed on different cores, it is crucial to minimize communication between the cores. Likewise, the tasks of the different cores must be balanced, so that ideally no single core does more work than the others. Furthermore, when working with large amounts of data on a distributed computing system, the speed and latency of the network can be a bottleneck. (This largely makes cloud computing solutions impractical in current computational proteomics.)

Implementing MaxQuant on a cluster

When MaxQuant was released in 2008, it was designed to be executed on conventional desktop PCs. The requirements to run MaxQuant efficiently were to have sufficient processing power and space on a local disk (see Figure 1A). Already in the original release, the program was semi-parallelized using multiple processes. The user had to enter the maximum number of threads to be used, depending on available virtual cores and other uses of the computer14 (when the number of threads selected is the same or higher than the number of available computing cores, the computer will become unresponsive). In the new release, MaxQuant was extended to use a computer cluster, where the processes are distributed over several computational nodes. A typical computer cluster contains many nodes, ideally with the same configuration and a global file system that is accessible from each of the nodes and where the data is stored (see Figure 1B).

As shown in Figure 2A the computational pipeline, which appears as a single and unified whole to the user, can conceptually be broken down into consecutive 'task groups' were some can be parallelized and others not. Limiting factors for the performance are (i) their demand on I/O speed, (ii) the CPU load of the particular compu-

tations and (iii) the degree to which the task groups can be parallelized.

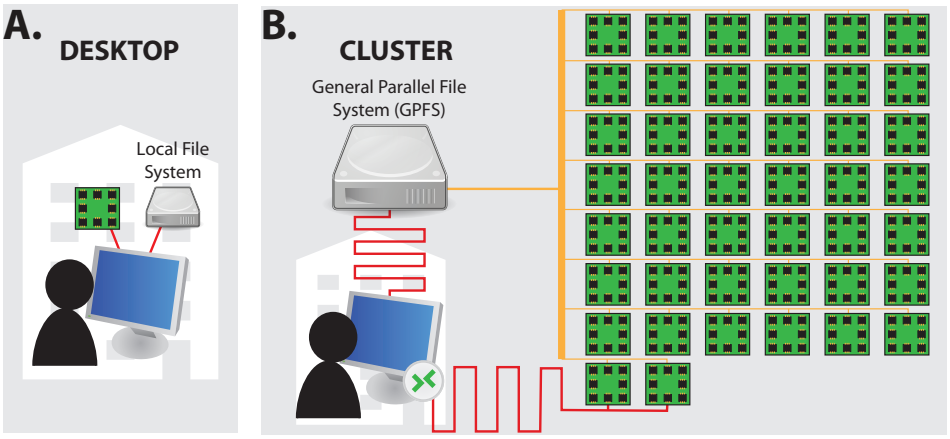


Figure 1: Distinct hardware setups. (A) In the field of proteomics desktop computers with a single (multicore) processor and data located on a local file system are generally used. (B) In contrast a computer cluster has multiple nodes, where one node represents one desktop computer. Additionally a global file system is required where the raw and meta data are stored and is accessible from all nodes. Usually, the cluster has a remote location, such as in a large computing center. In the figure we compare a quad core desktop computer with a cluster consisting of 42 nodes and 336 cores, both running the Windows operating system.

The computational task groups of MaxQuant can be conceptualized as 14 fundamental groups, which we briefly summarize below (see Figure 2B). In the 'initialize' phase the raw files are verified for intactness and readability, additionally an index file for each raw file is created containing scan meta-data that is accessed many times. The next step - 'feature detection' - extracts the peptide features present at the full scan (MS1) level, typically peptide precursor masses. The detection of the 3D peaks (m/z over the retention time) and of label pairs (typically SILAC-pairs) is also performed here. In the following section we perform an initial search using the Andromeda search engine15 to get a first list of identified peptides, which can be used in later steps. To correct for any systematic mass errors occurring during data acquisition the initial list of identifications is used to calculate mass calibration curves over time and m/z24. The search is now repeated with the updated m/z values resulting in the final list of peptide identifications, where for each fragmentation spectrum the up to 10 best scoring peptide sequences are retained. After peptide identification, the peptide spectrum matches (PSM) whose measured mass differences exceed the individually calculated mass tolerance are removed and the peptide identification with the next best score falling within the mass tolerance is retained13. The peptide false discovery rate (FDR) is calculated on this pre-filtered peptide list and peptide identifications below a specified threshold are discarded. Since different peptide species resulting in very similar m/z values can elute

A. tasks in MaxQuant

		I/O demand	computation	paralellization
1	Configuring	-	-	-
2	Testing files	-	+	++
3	Finish Testing files	-	-	-
4	Feature detection	++	++	++
5	Combining apl files for first search	+	-	-
6	Preparing searches	+	-	-
7	MS/MS first search	+	++	++
8	Read search results for Recalibration	+	-	-
9	Mass recalibration	+	++	++
10	MS/MS preapration for main search	+	++	++
11	Combining apl files for main search	+	-	-
12	MS/MS main search	+	++	++
13	Preparing combined folder	-	-	-
14	Calculating masses	+	++	++
15	Correcting errors	+++	-	-
16	Reading search engine results	+	++	++
17	Finish reading search results	-	++	++
18	Filter identifications (MS/MS)	-	++	++
19	Applying FDR	++	-	-
20	Assembling second peptide MS/MS	++	++	++
21	Combining apl files for second peptide search	+	-	-
22	Second peptide search	+	++	++
23	Reading search engine results second peptide	+	++	++
24	Finish reading search results second peptide	-	++	++
25	Filtering identifications second peptide	-	++	++
26	Applying FDR second peptide	++	-	-
27	Reporter quantification	-	++	++
28	Retention time alignment	++	+	+
29	Matching between runs	++	+	+
30	Prepare protein assembly	-	-	-
31	Assembling protein groups	+	++	++
32	Finish protein assembly	++	+	+
33	Updating identifications	-	++	++
34	Label-free normalization	++	+	+
35	Label-free quantification	+	+	+
36	Label-free collect	+	+	+
37	iBAQ	-	-	-
38	Estimating complexity	-	++	++
39	Prepare writing tables	-	-	-
40	Writing tables	++	++	++
41	Finish writing tables	+	+	+

B. tasks in computational proteomics

1	initialize
2	feature detection
3	initial search
4	mass recalibration
5	main search
6	mass pre-calculations
7	apply FDR
8	second peptide search
9	apply FDR (SP)
10	RT align & match runs
11	assemble proteins
12	label-free
13	write table
14	other

C. performance on a desktop PC

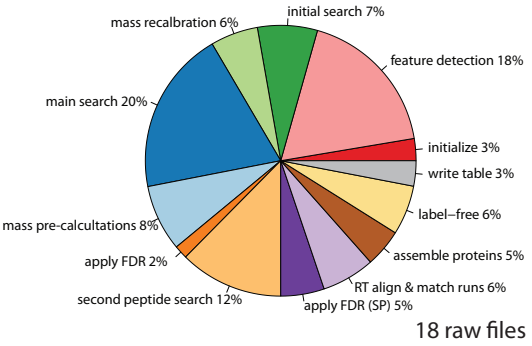


Figure 2: Time spent on tasks and groups of tasks in the MaxQuant pipeline. (A) Detailed list of task groups that are performed on the raw data in the course of a complete MaxQuant analysis. For each of them the demands or suitability to I/O, computational power and parallelization are indicated. (B) Task groups from A grouped into larger procedures. (C) Proportional times consumed by the procedures from B. during a typical analysis of 18 raw files of approximately 20 GB per dataset (total of 133,110 high resolution MS scans and 514,912 high resolution MS/MS scans).

at the same time, multiple precursors can occur in the same selection window, leading to fragmentation of several peptides in a single MS/MS spectrum. In cases where we identified the intended precursor, a third Andromeda search attempts to also identify the co-eluting and co-fragmented peptide. These additional peptides require a separate FDR correction15. For replicates or comparison of different runs, a sophisticated tree-based retention time alignment is performed. This alignment is used to transfer peptide identifications to raw files were a particular feature is observed but not identified ('match between runs' option in MaxQuant), increasing the number of identified peptides per raw file and reducing missing values for quantification[17]. The next step is to assemble the identified peptides to proteins. For this purpose we group proteins that are identified by the same peptides in a user-configurable manner. Depending on the user settings, label-free quantification is performed, correcting for systematic differences in quantification between the raw-files. The last step is to write the results to output tables, which can then be used for downstream analysis or loaded into the 'Viewer' part of MaxQuant for detailed visual inspection of the data. Notably, the output of the calculations themselves can reach gigabytes.

The time spent on each of these task groups is strongly dependent on the number and the size of the raw files. Figure 2C illustrates the percentages of the total time spent on each task group for a typical project using a standard desktop computer (see Experimental Methods). The data set contains measurements of a fractionated cell line in triplicate, giving rise to 18 raw MS files. Although most computation time is required for the peptide identification by the Andromeda search engine (initial search, main search and second peptide search), this takes less than half of the total (39%). The next largest item is the feature detection in these large data files (18%). Tasks like mass recalibration (6%), applying the peptide FDR (7%), match-between runs including retention time alignment (6%), label-free quantification (6%) and protein group assembly (5%) are also time-consuming. If the number of cores is limited, many of these computational times grow directly with the size of the dataset, quickly becoming impractical.

We next illustrate the benefits of parallelization on the main peptide search procedure. After preprocessing of the MS/MS spectra in MaxQuant the resulting fragment spectra are sorted by their precursor peptide mass, which has important advantages later on in the search. This task is not trivial, since the work should be distributed in a way that all processes finish at the same time, to avoid slowing down the overall pipeline with a single peak list file taking a disproportional amount of time to finish. For this reason, we make the number of spectra in each peak list file dependent on the peptide mass range to counterbalance the increasing combinatorial calculation time for

peptides with higher mass. The first step in the Andromeda search is the creation of the peptide sequence database with its associated peptide masses. The in silico digestion of the proteins and the creation of database search indices is parallelized using multi-threading of only one processor. Since we have split all spectra into independent peak list files at this point, the following peptide search can be executed in separate, parallel processes. This data decomposition is done in a similar manner for the initial and second peptide search. We compared serial and parallel execution by running either a single or 18 CPUs on the cluster. For the initial, main and second peptide search of a dataset of 547,900 MS/MS spectra that are distributed into 18 peak list files we decreased the run time almost 7-fold (1.1 h for the parallel and 7.9 h for the serial search, respectively).

Similarly to the peptide search tasks, we particularly concentrated our parallelization efforts on feature detection, mass recalibration, FDR application, protein group assembly and writing tables, constituting the major remaining bottlenecks.

Performance of desktop vs. cluster for data sets of variable size

In modern proteomics, large numbers of files are often analyzed together. These could for instance be generated during in-depth analysis of a proteome with many fractionation steps across several conditions. Furthermore, all files associated with a given project spanning many months or even years are best analyzed together in MaxQuant to guarantee overall comparability of results and to avoid inflation of the FDR[25]. Ideally, the computational proteomics infrastructure should not pose a limitation to such analyses. Here we investigate the gains of our optimization efforts using a computer cluster consisting of 42 nodes with 8 virtual cores each, resulting in the potential for 336 parallel operations. We compare this setup to a conventional desktop PC with a comparable processor configuration, in which 4 parallel cores are dedicated to MaxQuant. In Figure 3A the advantages of parallelization in terms of analysis time are visualized by horizontal bars, consisting of individual tasks that represent the processing time for each task group. If the task group can be parallelized, the bar is rotated vertically since a group of files is now analyzed as fast as a single file. In cases with only few raw files rotating these task groups vertically does not shorten the entire processing time appreciably. However, for larger number of files, the savings become dramatic. To test this on a specific example, we analyzed a small data set with 6 raw files and a large data set with 763 raw files on both the desktop and the computer cluster (Figure 3B). For the small data set, the saving in computation time were 41% (2.8 h vs. 1.7 h). However, for the very large data set, processing time on the cluster was 5 day whereas the desktop

calculation took almost 20 days (Suppl. Table 1).

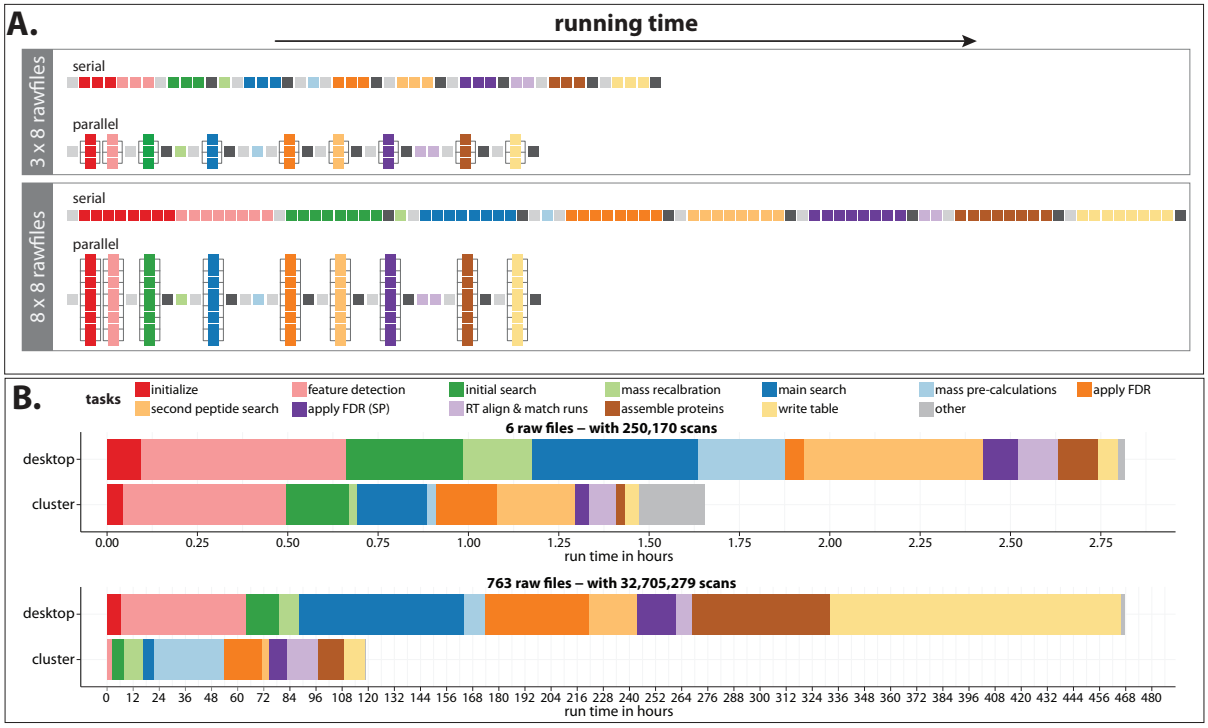


Figure 3: Comparison of the improvements by parallelization of different task groups. (A) Conceptual visualization of the effects of parallelization of some but not all task groups on total computing time. When only a few raw files need to be analyzed the gain is minimal, but with extension of the dataset the time savings become dramatic. (B) Total run times of two different datasets (6 vs. 763 raw files) with color coding of the different task groups.

Next, we systematically investigated the advantages of the computer cluster over the desktop computer for increasing number of raw files. Because different raw files can contain very different amounts of data, we scaled the x-axis in Figure 4A in scans instead of raw files (one raw file contains generally between 3,500 and 21,000 scans, depending on the instrument, gradient length and the chosen topN method). Recapitulating the results described above, a clear trend emerged, in which the saved computing time was negligible for small datasets and increased drastically at very large data sizes. We also plotted processing times for the different task groups separately (Figure 4B). This revealed that feature detection benefited most from parallelization, followed by the main peptide search. However, tasks like write out of the large output tables also profit extensively from parallelization (in this case because MaxQuant needs to access all the raw files in this task group, which is much faster in parallel mode).

Performance of an I/O optimized desktop computer

Given the time expenditure of the MaxQuant task groups on the large data sets, it ap-

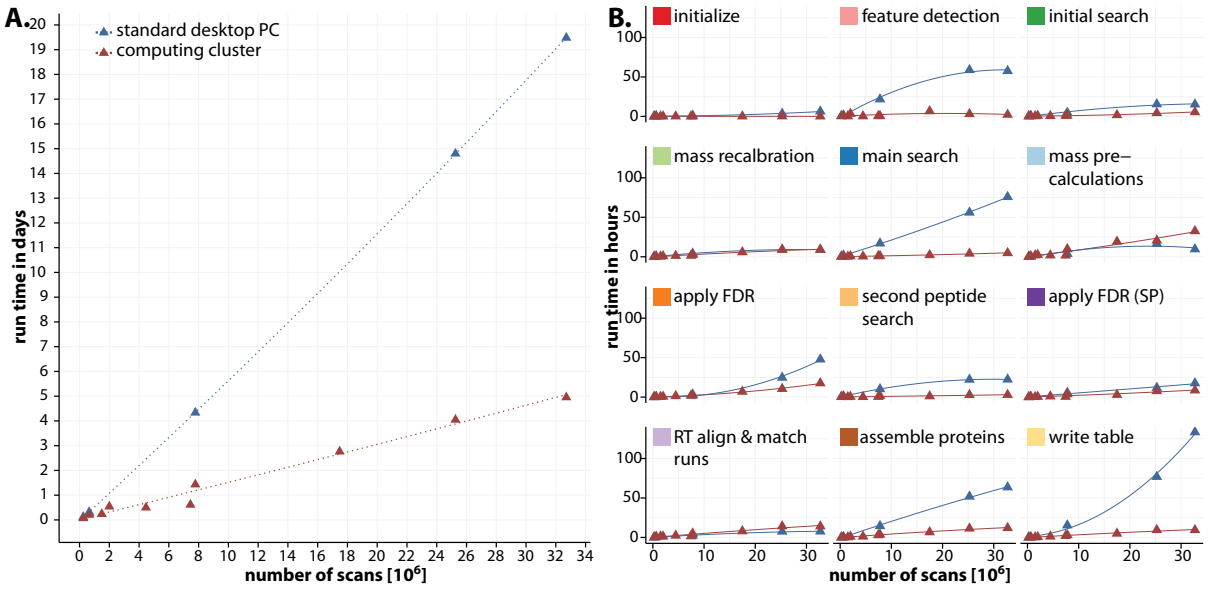


Figure 4: (A) Computational time as a function of data to be processed. The x-axis is in units of added MS scans in the individual raw files. (B) Same panel A, but for each task group separately.

peared that memory constraints during access to raw and intermediate data might play just as large a role as total processing power (Figure 4). We tested this notion using a custom-built computer that was optimized for applications such as high end gaming, equipped with 1 TB of solid state disks configured in RAID mode (see Experimental Methods). On this computer we used all 8 cores and processed small, medium and very large data sets. As can be seen in Figure 5, the processing time per raw file was very similar to that of the cluster, even for the very large data set. As the cluster has 336 cores and the I/O optimized high end desktop computer only 8, we conclude that the benefits of parallelization accrue mainly from better I/O access, whereas computing power is less of a limiting resource under these circumstances. In terms of expenditure, this makes high-end computational resources readily accessible to a large number of research groups lacking access to cluster computing facilities.

Incremental coverage of the human genome by large-scale data sets

Although the human genome has been sequenced more than ten years ago, it is still not clear how many different gene products it specifies. Estimates for the number of protein coding genes have been shrinking over the last ten years, from initial values of over 40,000 to a recent one that finds 20,225 open reading frames with at least some associated experimental or bioinformatics evidence[26]. Definitive proof of protein coding potential would be provided by solid data obtained by MS based proteomics. Accordingly, one of the goals of the chromosome centric Human Proteome Project is to map the entire human protein set to the set of protein-coding genes[27]. Currently, unam-

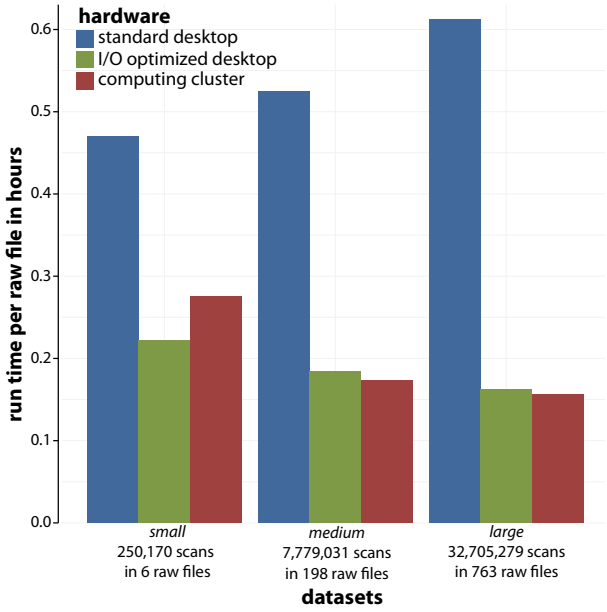


Figure 5: Comparison of total analysis time for a small, medium and very large data set using a desktop, I/O optimized, high end desktop and a computer cluster.

biguous protein level information is missing for up to 30% of human genes28. Here we employ a large collection of high resolution mass spectrometric data to determine the increase of coverage of the human genome as more and more experiments spanning multiple human sample types for different conditions are combined. Due to the large number of raw files involved, this task could not be carried out with a standard desktop configuration but rather required computational advances as described above.

We have previously found that in-depth proteomic sequencing of human cancer cell lines allows unambiguous identification of about 10,000 different protein groups using currently available technology[17, 29] and other groups have reported similar results[30, 31]. We therefore collected raw files from our laboratory from three deep cell line proteome projects together covering 30 cell lines[17, 18] (Experimental Methods). Furthermore, we added data from two recent studies of colon cancer tissues[19, 20] as well as a representative of a body fluid proteome21. Together these data comprised a collection of 1004 raw files, analyzed together on the cluster in a run time of only 5.5 days. At a 1% percent FDR at both peptide and protein levels, MaxQuant found a total of 13,242 protein groups in the UniProt database. We identify 255,432 different tryptic or LysC peptides, whereas 61,583 peptide sequences are unique within all proteins in the fasta file. A protein group contains on average 14.9 unique peptides, whereas the median is 9 (Figure 6A, Suppl. Table 2b). Just 1.9 or 3.6% of the proteome was identified with only one or only two peptides, respectively (Figure 6A). Sequence coverage was on average 42% (41.2% median). To our knowledge, this is the largest collection of

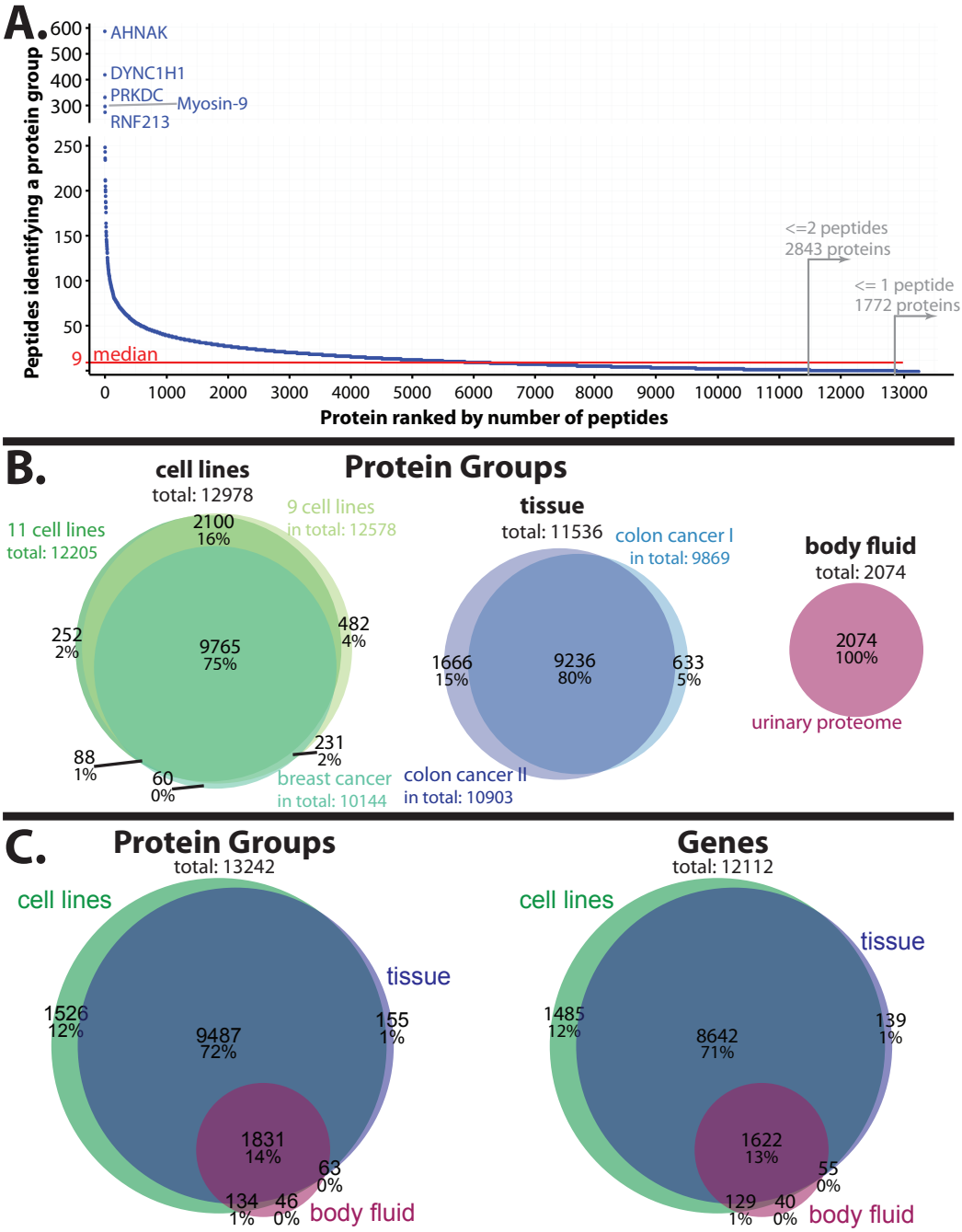


Figure 6: (A) Number of peptides identifying each protein group. Proteins are ranked by highest to lowest number of peptides. AHNAK, or desmoyokin an abundant and exceptionally large protein (700 kDa), is identified the largest number of unique peptides. Interestingly the relatively uncharacterized E3 ligase RNF213 is also among the top 5 proteins. (B) Number of proteins identified in the three cell line projects, two colon cancer projects and in the body fluid proteome. (C) Total number of protein groups identified from the cell line, colon cancer and body fluid proteomes shown individually in A (left). Same Venn diagram but showing different protein coding genes instead of protein groups (right).

unique peptides reported so far. For the expressed proteome that was identified here, close to half of the primary structure was therefore verified on average in this data set. Matching all separately identified protein groups to the human genome yielded 12,112 genes, which is almost 60%, assuming a total number of 20,225.

The data analyzed above originated from three main sources - three different cell line projects, two colon cancer tissue investigations and a study of the variability of the urinary proteome (Table 2). The three cell line studies together identified more than 13,000 different protein groups (Figure 6B). The depth of coverage in each project depended on the technology used (long columns and Q Exactive, vs. shorter columns and LTQ Orbitrap Velos mass spectrometers), but the main finding is that there is a very large overlap among the cell line proteomes. Notably, despite a very large number of raw files (420), the breast cancer cell line study added only 3% unique proteins to the other two cell line projects. This reflects the advance in shotgun proteomics technology and illustrates a general finding that accumulating large number of measurements by itself does not necessarily lead to larger identified proteomes.

The large overlap in cell line proteomes agrees with previous findings that found remarkably similarity in the identity - if not the abundance - of the expressed proteins[17, 32]. Naturally, the two cancer tissue proteomes have large overlap but interestingly the number of proteins identified in this single in vivo source was 11,536 - not much smaller than the total number from the different cell lines. The body fluid proteome identified about 2000 protein groups, partially reflecting the higher dynamic range of this proteome and the absence of fractionation.

Next we compared the cell line projects, colon cancer study and the body fluid study (Figure 6C). Again we found a large overlap, and intriguingly the in-depth colon cancer proteome only added 1% to the total identified proteome. This may reflect the fact that nearly all these proteomes are of cancer origin, but it also highlights the fact that the addition of tissue, per se, does not necessarily add many unique protein identifications. Likewise, the urinary proteome only added very few new proteins, indicating that body fluids may also not necessarily add substantially to overall coverage. When considering protein groups mapped to genes, we observe slightly smaller overall numbers, but the proportional contribution of the individual proteome sources remains largely unchanged (Figure 6C).

To study saturation properties of proteome coverage in large data sets in more detail, we investigated how quickly proteome coverage was reached as a function of the number of experiments used as input. Since this depends on the order in which the projects are added, we simulated the additional coverage from the analysis of the 1000 raw files.

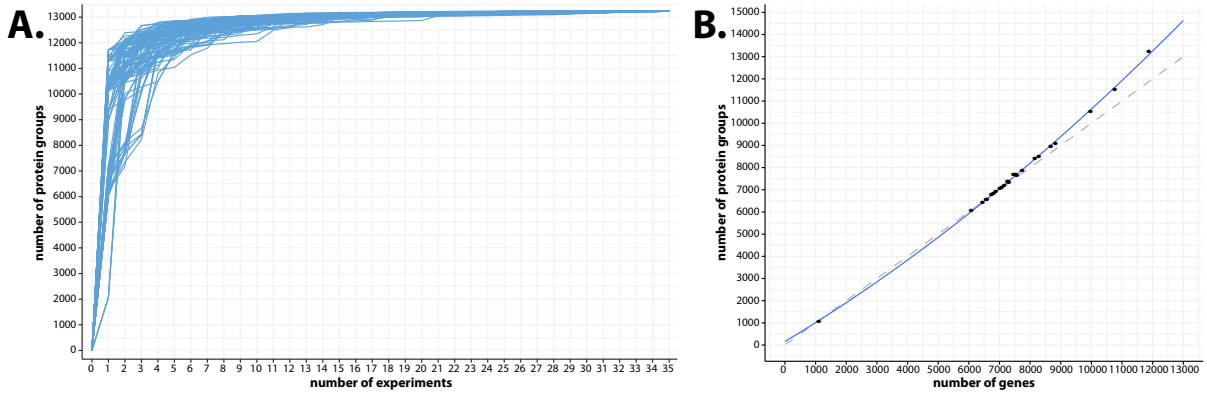


Figure 7: (A) Saturation curves of number of proteins identified when incrementally adding experiments. The traces represent 100 different simulations. (B) Number of identified and distinguishable protein groups as a function of the number of identified protein coding genes (see text for details).

Using 100 different combinations yielded the saturation curves shown in Figure 7A. In some of the simulations, the final proteome coverage was essentially reached with less than 5% of the total experiments. For instance, measurements with the Q Exactive and long columns reached already 95% of the total protein identifications with only 91 raw files. In each of the 100 simulations, nearly the final number of identified proteins was obtained after a third to half of the experiments had been added. This analysis again underscores that reaching a given depth depends more on the technology used than on the cumulative number of analyses.

Finally, we investigated the relationship between identified protein groups and genes. As can be seen in Figure 7B, at low numbers of identified genes the ratio between protein groups and genes is about one to one. Starting from 7000 genes, a larger number of isoforms is added as a function of additional genes. This is because the increasing depth of coverage necessary to identify many genes also concomitantly leads to increasing sequence coverage.

Conclusion and Outlook

Here we have analyzed the different task groups comprising the computational pipeline in the MaxQuant environment. This revealed specific bottlenecks, which were removed as far as possible. The resulting MaxQuant version is highly parallelized and memory optimized. For small sets of MS raw files it performs very fast on both the desktop or on a large cluster. For instance, in our laboratory we often analyze proteomes with six fractions, each measured in 4 h gradients, which in triplicate experiments results in 18 large raw files. These are processed in 7.6 h on a standard desktop (using 4 virtual cores) and 4.9 h on the cluster (Suppl. Table 1). For very large data sets, however, the cluster massively outperforms the desktop computer, to the extent that some analyses are only practical on the cluster.

We had expected that spread the workload on multiple processors will be the best solution and the processing time will be reduced by the number of processing units. But in the course of improving the computational speed of MaxQuant, we also tested an I/O optimized high end desktop PC. Surprisingly, this configuration performed essentially as well as the large cluster, at a small fraction of the costs and with much less administration overhead. Therefore, our recommendation at this point is to invest in this or similar configuration for laboratories or facilities with medium to large data production. Close to 1000 raw files can still be efficiently processed in the standard workflow in a matter of a few days.

What do these findings imply for potential bottlenecks in the computational analysis of deep proteome data? As we have shown here, current data sets can easily be handled on relatively inexpensive hardware. For the future, both the power of computational hardware and the size of the data acquired in proteomic investigations will increase. For instance, the number of MS and MS/MS scans used in standard acquisitions could increase several fold over the next few years, just as it has over the last several years. Countering this additional computational load, current desktop chips with 12 virtual cores already exist, rather than the 8 cores employed here, and chips with 16 cores are to be released shortly. Similarly, after initial submission of this manuscript, we have installed two rack mounted solutions with 64 logical processors, from Intel and AMD, respectively (see Experimental Methods). Both systems are essentially as easily administered as PCs, but combine improvements due to fast and local memory with increased number of computation units, and are still quite economical. In our initial tests, they performed equally well to the I/O optimized PCs on small numbers of files but were able to handle larger files sets without slow down.

Based on these trends we expect that the computational demands of the standard workflow for in depth shotgun proteomics can be comfortably handled for the foreseeable future. However, specialized tasks, such as searches in six frame translations of large genomes, and other extremely computing intensive tasks may benefit from large clusters.

We applied the improvements in software and hardware to investigate the incremental contribution to coverage of the human genome from large-scale data sets generated in our laboratory. This revealed that there is a large overlap in the identity of proteins in different cell line proteomes as well as an in-depth measured human tissue proteome, consistent with earlier findings. Together, the analysis of more than 1000 raw files identified more than 13,000 different protein groups, mapping to more than 12,000 of the roughly human 20,000 protein coding genes. Interestingly, this depth could be reached with a small subset of the raw MS data, namely the ones using the latest technology. In contrast, hundreds of raw files obtained with a workflow from just a few years ago made essentially no contribution to total identifications. The implications for current efforts to map the entire proteome would be to focus on technology development for in-depth measurements rather than predominantly on accumulation of large numbers of data sets.

Acknowledgments: We thank our colleagues at the Max Planck Institute of Biochemistry for help and fruitful discussions, in particular excellent hardware and configuration assistance by Mario Oroshi. We thank Martin Hoffmann and Bernhard Busch of our institute's computing center for help with the compute cluster. The research leading to these results has received funding from the European Commission's 7th Framework Programme (grant agreement HEALTH-F4-2008-201648 / PROSPECTS).

References

1. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. Nature 2003, 422, (6928), 198-207.

2. Mallick, P.; Kuster, B., Proteomics: a pragmatic perspective. Nat Biotechnol 2010, 28, (7), 695-709.

3. Cox, J.; Mann, M., Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem 2011, 80, 273-99.

4. Altelaar, A. F.; Munoz, J.; Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 2012, 14, (1), 35-48.

5. Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Na-

garaj, N.; Cox, J.; Mann, M.; Horning, S., Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 2011, 10, (9), M111 011015.
6. MacCoss, M. J., Computational analysis of shotgun proteomics data. Curr Opin Chem Biol 2005, 9, (1), 88-94.
7. 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.
8. Galvez, S.; Diaz, D.; Hernandez, P.; Esteban, F. J.; Caballero, J. A.; Dorado, G., Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment. Bioinformatics 2010, 26, (5), 683-6.
9. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20, (18), 3551-67.
10. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 2005, 1, 2005 0017.
11. Kohlbacher, O.; Reinert, K.; Gropl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M., TOPP–the OpenMS proteomics pipeline. Bioinformatics 2007, 23, (2), e191-7.
12. 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. Bioinformatics 2010, 26, (7), 966-8.
13. 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), 1367-72.
14. Cox, J.; Matic, I.; Hilger, M.; Nagaraj, N.; Selbach, M.; Olsen, J. V.; Mann, M., A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nat Protoc 2009, 4, (5), 698-705.
15. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 2011, 10, (4), 1794-805.
16. Cox, J.; Mann, M., 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics 2012, 13 Suppl 16, S12.
17. Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most

proteins. Mol Cell Proteomics 2012, 11, (3), M111 014050.
18. Geiger, T.; Madden, S. F.; Gallagher, W. M.; Cox, J.; Mann, M., Proteomic portrait of human breast cancer progression identifies novel prognostic markers. Cancer Res 2012, 72, (9), 2428-39.
19. Wisniewski, J. R.; Ostasiewicz, P.; Dus, K.; Zielinska, D. F.; Gnad, F.; Mann, M., Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol Syst Biol 2012, 8, 611.
20. Wisniewski, J. R.; Dus, K.; Mann, M., Proteomic workflow for analysis of archival formalin fixed and paraffin embedded clinical samples to a depth of 10,000 proteins. Proteomics Clin Appl 2012.
21. Nagaraj, N.; Mann, M., Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. J Proteome Res 2011, 10, (2), 637-45.
22. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 2012, 40, (Database issue), D71-5.
23. Mohammed, Y.; Shahand, S.; Korkhov, V.; Luyf, A. C. M.; van Schaik, B. D. C.; Caan, M. W. A.; van Kampen, A. H. C.; Palmblad, M.; Olabarriaga, S. D. In Data Decomposition in Biomedical e-Science Applications, e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on, 5-8 Dec. 2011, 2011; 2011; pp 158-165.
24. Cox, J.; Michalski, A.; Mann, M., Software lock mass by two-dimensional minimization of peptide mass errors. J Am Soc Mass Spectrom 2011, 22, (8), 1373-80.
25. Schaab, C.; Geiger, T.; Stoehr, G.; Cox, J.; Mann, M., Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Mol Cell Proteomics 2012, 11, (3), M111 014068.
26. Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M. F.; Kellis, M.; Lindblad-Toh, K.; Lander, E. S., Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 2007, 104, (49), 19428-33.
27. Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol 2012, 30, (3), 221-3.
28. Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Hu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S., The human proteome project: Current state and future direction. Mol Cell Proteomics 2011.

29. Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M., Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011, 7, 548.
30. Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R., The quantitative proteome of a human cell line. Mol Syst Biol 2011, 7, 549.
31. Munoz, J.; Low, T. Y.; Kok, Y. J.; Chin, A.; Frese, C. K.; Ding, V.; Choo, A.; Heck, A. J., The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol 2011, 7, 550.
32. Lundberg, E.; Fagerberg, L.; Klevebring, D.; Matic, I.; Geiger, T.; Cox, J.; Algenas, C.; Lundeberg, J.; Mann, M.; Uhlen, M., Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol 2010, 6, 450.

3 Conclusions and outlook

Analytical Mass spectrometry-based proteomics, specifically the shotgun approach, has now reached a high level of maturity with respect to sample processing, data acquisition and data analysis[86]. Nevertheless, proteomics still lags behind other large-scale approaches in biology and further technological advances are urgently needed. The main goals are to increase throughput and spectra quality so that spatiotemporal dimensions, population parameters and the complexity of protein modifications can be considered on a quantitative scale[102]. Traditionally, most proteomics analysis has been carried out using relatively inexpensive ion trap instruments, which offer fairly low precision and accuracy of mass determination. Higher resolution instruments, which achieve precision better than 10 ppm, were previously expensive and rare. Indeed with the introduction of new instruments such as the Orbitrap analyzer or improved quadrupole time-of-flight mass spectrometers, high resolution instruments have become much more commonplace[49]. However, the resulting data sets were very complex and their analysis requires several steps from raw data processing, to database search, statistical evaluation of the search result, quantitative algorithms and statistical analysis of quantitative data[102].

In this thesis, I have alleviated some of these bottlenecks through development of algorithms and robust and reliable software to analyze high-quality MS data. I contributed to the novel peptide search engine Andromeda, which produces results at least as good as the commonly used commercial Mascot software[67]. Despite the excellent performance of Andromeda in large-scale studies, generally half of the acquired MS/MS spectra remain 'unassigned' (i.e. without high confidence peptide identification)[50]. This can have several reasons: constrained data base search parameters (e.g. search for tryptic peptides only), the combinatorial problem caused by post-translational modifications, spectra containing fragmentation ions originating from multiple peptides, single amino acid polymorphisms, and splicing isoforms and the complexity of redundant peptides on peptide and protein identification[102;132]. With the Andromeda search engine MaxQuant is already able to identify co-fragmenting peptides when the triggered precursor peptide was identified in a previous step. This second

peptide search is routinely used in almost all experiment in our laboratory and contributes around 10% additional identifications. For further increasing the identification rate several approaches are under development. Most prominently the transfer of identifications from on dataset (where the peptide was more abundant) to the dataset under investigation ('Match-between-runs' feature in MaxQuant). The error tolerant (or 'blind') database search is looking for peptides which occur in a modified version of an already identified unmodified peptide[133;134]. The incompleteness of the searched protein sequence databases are likely to be at least partly solved by next generation sequencing of transcriptomes[102;135]. In the meantime the computational pipeline could be modified so that in addition to the database search, *de novo* sequencing is integrated for unassigned MS/MS spectra. This multi-step approach would enable the detection of novel peptides such as peptides with amino acid exchanges or those originating from splice variants[132].

In this thesis I also introduced and implemented a computer-based Expert System, which is used for automated annotation of high-resolution MS/MS spectra[136]. I used a knowledge base of peptides fragmentation rules, which was applied to complete the annotation of thousands of spectra. We figured out that the fragmentation products of HCD are comparable to CID and most of the fragmentation types were already known by literature[137]. The rule set used for annotating the spectra was developed in close contact with a human domain expert and it was stringently controlled by an FDR approach. Currently, the output of the Expert System is used as an add-on to the MaxQuant pipeline for manual inspection of the acquired MS/MS spectra and their assigned peptide identifications. For instance, in cases where two or more peptides are intentionally fragmented together (multiplexing), in-depth classification of which fragment peak belongs to which peptide will help in the identification process[138]. By removing peaks that belong to already identified peptides, the complexity of the multiplexed spectrum will be reduced and even low abundant fragment peaks of the co-fragmented unidentified peptides can now be used for the identification.

Recent efforts in MS-based proteomics led to the archiving of data sets as large as several million fragmentation spectra[102]. For such large data amounts, the efficiency of the computational analysis is an important practical consideration[132]. In this thesis, I have made great efforts to adapt MaxQuant to run efficiently on the currently available hardware platforms[139]. For this purpose, the bottlenecks in the pipeline were identified and alleviated - mainly by parallelization. For larger datasets, this parallelization has a dramatic effect when more than the normal four CPUs are used for data analysis. We tested several hardware configurations and conclude that machines with high I/O

performance have the best effect in reducing the analysis time. We used the optimized MaxQuant version to measure the current state of the proteome coverage of the human protein-coding genes and detected around 60% of the gene products. While this is already an impressive coverage, given the fact that only a relatively small number of projects contributed to it, new developments will be needed to identify proteins for the entire human genome. On the technological side further instrumental advances are likely needed, such as higher sensitivity for the detection of low abundant proteins, improvements in the scan rate of the instrument side so that also low abundant peaks are picked. Even more promisingly, the instrument can be made to pick peptides for sequencing with the help of a software based 'intelligent agent'[140]. In this connection an interesting consideration is that the scan speeds of the mass spectrometers increase, the difference between targeted and discovery proteomics will become more and more blurred[49]. It would be also a good idea to use rigorous statistical tests to select samples that will contribute to additional coverage of the genome, in a way that is exemplified our previous investigation (article 4) . Further on the computational side, enhancing the proportion of peptides identified remains an important goal for computation proteomics and in depth comparison of the de novo sequenced proteome with the genome will likely reveal many novel biological phenomena.

4 References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, and et al. Baldwin, J. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, and et al. Sutton, G. G. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[3] P. Legrain, R. Aebersold, A. Archakov, A. Bairoch, K. Bala, and et al. Beretta, L. The human proteome project: Current state and future direction. *Mol Cell Proteomics*, 2011.

[4] M. Walhout, M. Vidal, and J. Dekker. *Handbook of Systems Biology: Concepts and Insights*. Elsevier Science, 2012.

[5] R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, and et al. Boysen, C. Evaluation of dna microarray results with quantitative gene expression platforms. *Nat Biotechnol*, 24:1115–1122, 2006.

[6] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19:1720–1730, 1999.

[7] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25:117–124, 2007.

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

[9] F. Hillenkamp and M. Karas. Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol*, 193:280–295, 1990.

[10] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[11] J. Cox and M. Mann. Is proteomics the new genomics? *Cell*, 130:395–398, 2007.

[12] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, and et al. Lin, M. F. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A*, 104:19428–19433, 2007.

[13] A gene-centric human proteome project: Hupo–the human proteome organization. *Mol Cell Proteomics*, 9:427–429, 2010.

[14] The call of the human proteome. *Nat Methods*, 7:661–661, 2010.

[15] L. M. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, and et al. Frohlich, F. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455:1251–1254, 2008.

[16] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, and et al. Kelso, J. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol*, 7:548, 2011.

[17] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, and et al. Szymborska, A. The quantitative proteome of a human cell line. *Mol Syst Biol*, 7:549, 2011.

[18] S. D. Patterson and R. H. Aebersold. Proteomics: the first decade and beyond. *Nat Genet*, 33 Suppl:311–323, 2003.

[19] M. Mann and M. Wilm. Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci*, 20:219–224, 1995.

[20] F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem*, 63:1193A–1203A, 1991.

[21] G. L. Glish and D. J. Burinsky. Hybrid mass spectrometers for tandem mass spectrometry. *J Am Soc Mass Spectrom*, 19:161–172, 2008.

[22] J. Colinge and K. L. Bennett. Introduction to computational proteomics. *PLoS Comput Biol*, 3:e114, 2007.

[23] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol Cell Proteomics*, 6:377–381, 2007.

[24] R. Matthiesen. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics*, 7:2815–2832, 2007.

[25] J. Cox and M. Mann. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an orbitrap. *J Am Soc Mass Spectrom*, 20:1477–1485, 2009.

[26] J. R. Yates, C. I. Ruse, and A. Nakorchevsky. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng*, 11:49–79, 2009.

[27] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks. The orbitrap: a new mass spectrometer. *J Mass Spectrom*, 40:430–443, 2005.

[28] A. Makarov, E. Denisov, O. Lange, and S. Horning. Dynamic range of mass accuracy in ltq orbitrap hybrid mass spectrometer. *J Am Soc Mass Spectrom*, 17:977–982, 2006.

[29] J. V. Olsen, L. M. de Godoy, G. Li, B. Macek, P. Mortensen, and et al. Pesch, R. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a c-trap. *Mol Cell Proteomics*, 4:2010–2021, 2005.

[30] J. V. Olsen, J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, and et al. Denisov, E. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics*, 8:2759–2769, 2009.

[31] A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, and et al. Strupat, K. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem*, 78:2113–2120, 2006.

[32] T. P. Second, J. D. Blethrow, J. C. Schwartz, G. E. Merrihew, M. J. MacCoss, and et al. Swaney, D. L. Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. *Anal Chem*, 81:7757–7765, 2009.

[33] M. Mann and N. L. Kelleher. Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci U S A*, 105:18132–18138, 2008.

[34] A. Michalski, E. Damoc, J. P. Hauschild, O. Lange, A. Wieghaus, and et al. Makarov, A. Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol Cell Proteomics*, 10:M111 011015, 2011.

[35] A. Michalski, E. Damoc, O. Lange, E. Denisov, D. Nolting, and et al. Muller, M. Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down lc ms/ms and versatile peptide fragmentation modes. *Mol Cell Proteomics*, 11:O111 013698, 2012.

[36] O. Lange, A. Makarov, E. Denisov, and W. Balschun. Accelerating spectral acquisition rate of orbitrap mass spectrometry. *Proc. 58th Conf. Amer. Soc. Mass Spectrom*, 2010.

[37] J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller. Shotgun sequencing of the human genome. *Science*, 280:1540–1542, 1998.

[38] H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5:699–711, 2004.

[39] E. Mortz, P. B. O'Connor, P. Roepstorff, N. L. Kelleher, T. D. Wood, and et al. McLafferty, F. W. Sequence tag identification of intact proteins by matching

tanden mass spectral data against sequence data bases. *Proc Natl Acad Sci U S A*, 93:8264–8267, 1996.

[40] D. M. Horn, R. A. Zubarev, and F. W. McLafferty. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc Natl Acad Sci U S A*, 97:10313–10317, 2000.

[41] S. K. Sze, Y. Ge, H. Oh, and F. W. McLafferty. Top-down mass spectrometry of a 29-kda protein for characterization of any posttranslational modification to within one residue. *Proc Natl Acad Sci U S A*, 99:1774–1779, 2002.

[42] G. K. Taylor, Y. B. Kim, A. J. Forbes, F. Meng, R. McCarthy, and N. L. Kelleher. Web and database software for identification of intact proteins using "top down" mass spectrometry. *Anal Chem*, 75:4081–4086, 2003.

[43] F. W. McLafferty, K. Breuker, M. Jin, X. Han, G. Infusini, and et al. Jiang, H. Top-down ms, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS J*, 274:6256–6268, 2007.

[44] J. F. Kellie, J. C. Tran, J. E. Lee, D. R. Ahlf, H. M. Thomas, and et al. Ntai, I. The emerging process of top down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol Biosyst*, 6:1532–1539, 2010.

[45] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, and et al. Morris, D. R. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, 17:676–682, 1999.

[46] J. Peng and S. P. Gygi. Proteomics: the move to mixtures. *J Mass Spectrom*, 36:1083–1091, 2001.

[47] M. P. Washburn, D. Wolters, and 3rd Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19:242–247, 2001.

[48] J. Cox and M. Mann. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem*, 80:273–299, 2011.

[49] W. S. Noble and M. J. MacCoss. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol*, 8:e1002296, 2012.

[50] J. Cox, N. C. Hubner, and M. Mann. How much peptide sequence information is contained in ion trap tandem mass spectra? *J Am Soc Mass Spectrom*, 19:1813–1820, 2008.

[51] J. Fila and D. Honys. Enrichment techniques employed in phosphoproteomics. *Amino Acids*, 43:1025–1047, 2012.

[52] M. Schirle, M. A. Heurtier, and B. Kuster. Profiling core proteomes of human cell lines by one-dimensional page and liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 2:1297–1305, 2003.

[53] B. Schwanhaeussler. *Global analysis of cellular protein dynamics by pulse-labeling and quantitative mass spectrometry*. PhD thesis, Humboldt University Berlin, 2010.

[54] D. F. Hunt, 3rd Yates, J. R., J. Shabanowitz, S. Winston, and C. R. Hauer. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A*, 83:6233–6237, 1986.

[55] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nat Methods*, 4:709–712, 2007.

[56] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*, 101:9528–9533, 2004.

[57] I. A. Papayannopoulos. The interpretation of collision-induced dissociation tandem mass-spectra of peptides. *Mass Spectrom Rev*, 14:49–73, 1995.

[58] B. Paizs and S. Suhai. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev*, 24:508–548, 2005.

[59] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, 11:601, 1984.

[60] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal Chem*, 59:2621–2625, 1987.

[61] G. C. McAlister, W. T. Berggren, J. Griep-Raming, S. Horning, A. Makarov, and et al. Phanstiel, D. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *J Proteome Res*, 7:3127–3136, 2008.

[62] D. L. Tabb, L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin, and 3rd Yates, J. R. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*, 75:1155–1163, 2003.

[63] F. Schutz, E. A. Kapp, R. J. Simpson, and T. P. Speed. Deriving statistical models for predicting peptide tandem ms product ion intensities. *Biochem Soc Trans*, 31:1479–1483, 2003.

[64] V. H. Wysocki, G. Tsaprailis, L. L. Smith, and L. A. Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom*, 35:1399–1406, 2000.

[65] R. Boyd and A. Somogyi. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J Am Soc Mass Spectrom*, 21:1275–1278, 2010.

[66] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N. G. Ahn, and W. M. Old. Quantifying the impact of chimera ms/ms spectra on peptide identification in large-scale proteomics studies. *J Proteome Res*, 9:4152–4160, 2010.

[67] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *J Proteome Res*, 10:1794–1805, 2011.

[68] A. G. Marshall and C. L. Hendrickson. High-resolution mass spectrometers. *Annu Rev Anal Chem (Palo Alto Calif)*, 1:579–599, 2008.

[69] L. Martens. Bioinformatics challenges in mass spectrometry-driven proteomics. *Methods Mol Biol*, 753:359–371, 2011.

[70] O. Vitek. Getting started in computational mass spectrometry-based proteomics. *PLoS Comput Biol*, 5:e1000366, 2009.

[71] N. Mujezinovic, G. Raidl, J. R. Hutchins, J. M. Peters, K. Mechtler, and F. Eisenhaber. Cleaning of raw peptide ms/ms spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics*, 6:5117–5131, 2006.

[72] M. Wehofsky and R. Hoffmann. Automated deconvolution and deisotoping of electrospray mass spectra. *J Mass Spectrom*, 37:223–229, 2002.

[73] B. Y. Renard, M. Kirchner, F. Monigatti, A. R. Ivanov, J. Rappsilber, and et al. Winter, D. When less can yield more - computational preprocessing of ms/ms spectra for peptide identification. *Proteomics*, 9:4978–4984, 2009.

[74] K. C. Hansen, G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin, and A. L. Burlingame. Mass spectrometric analysis of protein mixtures at low levels using cleavable 13c-isotope-coded affinity tag and multidimensional chromatography. *Mol Cell Proteomics*, 2:299–314, 2003.

[75] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26:1367–1372, 2008.

[76] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*, 73:2092–2123, 2010.

[77] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, 66:4390–4399, 1994.

[78] J. Seidler, N. Zinn, M. E. Boehm, and W. D. Lehmann. De novo sequencing of peptides by ms/ms. *Proteomics*, 10:634–649, 2010.

[79] A. J. Liska and A. Shevchenko. Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 3:19–28, 2003.

[80] J. Grossmann, B. Fischer, K. Baerenfaller, J. Owiti, J. M. Buhmann, and et al. Gruissem, W. A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments. *Proteomics*, 7:4245–4254, 2007.

[81] M. Junqueira, V. Spirin, T. S. Balbuena, H. Thomas, I. Adzhubei, and et al. Sunyaev, S. Protein identification pipeline for the homology-driven proteomics. *J Proteomics*, 71:346–356, 2008.

[82] D. Tessier, P. Yclon, I. Jacquemin, C. Larre, and H. Rogniaux. Ovnip: an open source application facilitating the interpretation, the validation and the edition of proteomics data generated by ms analyses and de novo sequencing. *Proteomics*, 10:1794–1801, 2010.

[83] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, and et al. Doherty-Kirby, A. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 17:2337–2342, 2003.

[84] A. Frank and P. Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem*, 77:964–973, 2005.

[85] M. Bern and D. Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comput Biol*, 13:364–378, 2006.

[86] A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, 4:787–797, 2007.

[87] R. G. Sadygov, D. Cociorva, and 3rd Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods*, 1:195–202, 2004.

[88] 3rd Yates, J. R., J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, 67:1426–1436, 1995.

[89] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[90] S. J. Barton and J. C. Whittaker. Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom Rev*, 28:177–187, 2009.

[91] R. E. Higgs, M. D. Knierman, A. B. Freeman, L. M. Gelbert, S. T. Patil, and J. E. Hale. Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. *J Proteome Res*, 6:1758–1767, 2007.

[92] V. Granholm and L. Kall. Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics*, 11:1086–1093, 2011.

[93] J. Eriksson and D. Fenyo. The statistical significance of protein identification results as a function of the number of protein sequences searched. *J Proteome Res*, 3:979–982, 2004.

[94] S. J. Barton, S. Richardson, D. N. Perkins, I. Bellahn, T. N. Bryant, and J. C. Whittaker. Using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem ms. *Anal Chem*, 79:5601–5607, 2007.

[95] Y. Wang, J. Zhang, X. Gu, and X. M. Zhang. Protein identification assisted by the prediction of retention time in liquid chromatography/tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 826:122–128, 2005.

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

[97] C. H. Becker and M. Bern. Recent developments in quantitative proteomics. *Mutat Res*, 722:171–182, 2011.

[98] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. The international protein index: an integrated database for proteomics experiments. *Proteomics*, 4:1985–1988, 2004.

[99] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, and et al. Ferro, S. The universal protein resource (uniprot). *Nucleic Acids Res*, 33:D154–159, 2005.

[100] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, and et al. Boeckmann, B. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34:D187–191, 2006.

[101] Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic Acids Res*, 41:D43–47, 2013.

[102] R. Matthiesen, L. Azevedo, A. Amorim, and A. S. Carvalho. Discussion on common data analysis strategies used in ms-based proteomics. *Proteomics*, 11:604–619, 2011.

[103] Y. F. Li and P. Radivojac. Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics*, 13 Suppl 16:S4, 2012.

[104] A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4:1419–1440, 2005.

[105] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75:4646–4658, 2003.

[106] J. Rappsilber and M. Mann. What does it mean to identify a protein in proteomics? *Trends Biochem Sci*, 27:74–78, 2002.

[107] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*, 404:939–965, 2012.

[108] A. J. Heck and J. Krijgsveld. Mass spectrometry-based quantitative proteomics. *Expert Rev Proteomics*, 1:317–326, 2004.

[109] S. E. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1:252–262, 2005.

[110] P. J. Boersema, R. Raijmakers, S. Lemeer, S. Mohammed, and A. J. Heck. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc*, 4:484–494, 2009.

[111] M. Schnolzer, P. Jedrzejewski, and W. D. Lehmann. Protease-catalyzed incorporation of 18o into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. *Electrophoresis*, 17:945–953, 1996.

[112] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, and et al. Hattan, S. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3:1154–1169, 2004.

[113] N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, and et al. Li, Y. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol*, 28:83–89, 2010.

[114] S. Hanke, H. Besir, D. Oesterhelt, and M. Mann. Absolute silac for accurate quantitation of proteins in complex mixtures down to the attomole level. *J Proteome Res*, 7:1118–1130, 2008.

[115] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proc Natl Acad Sci U S A*, 100:6940–6945, 2003.

[116] R. J. Beynon, M. K. Doherty, J. M. Pratt, and S. J. Gaskell. Multiplexed absolute quantification in proteomics using artificial qcat proteins of concatenated signature peptides. *Nat Methods*, 2:587–589, 2005.

[117] A. Keller, J. Eng, N. Zhang, X. J. Li, and R. Aebersold. A uniform proteomics ms/ms analysis platform utilizing open xml file formats. *Mol Syst Biol*, 1:2005 0017, 2005.

[118] O. Kohlbacher, K. Reinert, C. Gropl, E. Lange, N. Pfeifer, and et al. Schulz-Trieglaff, O. Topp–the openms proteomics pipeline. *Bioinformatics*, 23:e191–197, 2007.

[119] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, and et al. Frewen, B. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26:966–968, 2010.

[120] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, and et al. Pandey, A. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1:376–386, 2002.

[121] M. Mann. Functional and quantitative proteomics using silac. *Nature Reviews Molecular Cell Biology*, 7:952–958, 2006.

[122] C. H. Becker, P. Kumar, T. Jones, and H. Lin. Nonparametric mass calibration using hundreds of internal calibrants. *Anal Chem*, 79:1702–1707, 2007.

[123] J. Cox, A. Michalski, and M. Mann. Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom*, 22:1373–1380, 2011.

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

[125] L. Kall, J. D. Storey, M. J. MacCoss, and W. S. Noble. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*, 7:40–44, 2008.

[126] R Development Core Team. *R: A language and Environment for Statistical Computing*. 2010.

[127] J. Cox and M. Mann. 1d and 2d annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*, 13 Suppl 16:S12, 2012.

[128] J. V. Olsen and M. Mann. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A*, 101:13417–13422, 2004.

[129] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, and A. Nesvizhskii. The need for guidelines in publication of peptide and protein identification data: Working group on publication guidelines for peptide and protein identification data. *Mol Cell Proteomics*, 3:531–533, 2004.

[130] A. Michalski, J. Cox, and M. Mann. 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*, 10:1785–1793, 2011.

[131] A. M. Falick, W. M. Hines, K. F. Medzihradszky, M. A. Baldwin, and B. W. Gibson. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass-spectrometry. *J Am Soc Mass Spectrom*, 4:882–893, 1993.

[132] K. Ning, D. Fermin, and A. I. Nesvizhskii. Computational analysis of unassigned high-quality ms/ms spectra in proteomic data sets. *Proteomics*, 10:2712–2718, 2010.

[133] M. M. Savitski, M. L. Nielsen, and R. A. Zubarev. Modificomb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics*, 5:935–948, 2006.

[134] R. J. Chalkley, P. R. Baker, K. F. Medzihradszky, A. J. Lynn, and A. L. Burlingame. In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol Cell Proteomics*, 7:2386–2398, 2008.

[135] V. C. Evans, G. Barker, K. J. Heesom, J. Fan, C. Bessant, and D. A. Matthews. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods*, 9:1207–1211, 2012.

[136] N. Neuhauser, A. Michalski, J. Cox, and M. Mann. Expert system for computer-assisted annotation of ms/ms spectra. *Mol Cell Proteomics*, 11:1500–1509, 2012.

[137] A. Michalski, N. Neuhauser, J. Cox, and M. Mann. A systematic investigation into the nature of tryptic hcd spectra. *J Proteome Res*, 11:5479–5491, 2012.

[138] A. R. Ledvina, M. M. Savitski, A. R. Zubarev, D. M. Good, J. J. Coon, and R. A. Zubarev. Increased throughput of proteomics analysis by multiplexing high-resolution tandem mass spectra. *Anal Chem*, 83:7651–7656, 2011.

[139] N. Neuhauser, N. Nagaraj, P. McHardy, S. Zanivan, R. A. Scheltema, and et al. Cox, J. High performance computational analysis of large-scale proteome datasets to assess incremental contribution to coverage of the human genome. *J Proteome Res*, submitted.

[140] J. Graumann, R. A. Scheltema, Y. Zhang, J. Cox, and M. Mann. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol Cell Proteomics*, 11:M111 013185, 2012.

Acknowledgments

I owe thanks to many people who supported and encouraged me during these last years and I can only try to acknowledge them here.

First of all, I would like to thank my supervisor Matthias for giving me the opportunity to write this thesis, for helpful input and discussions and for a lot of insight into the complete proteomics workflow.

Furthermore, I would like to thank Dmitrij Frishman for taking the responsibility as my official doctoral advisor and to Ines Antes for being the chair of my dissertation committee.

Jürgen thank you for your guidance during my life as PhD student and sharing your experience with me. For great support regarding the expert system and your patient for answering annoying questions about mass spectrometry, I am thankful to my friend and co-worker Annette. To continue I also want to thank my office mates from the Blümchen office Rochelle, Michal, Kirti for many interesting discussions on and, in particular, off topic. I also enjoyed working together with Alison on events like our lab retreats and the MaxQuant Summer Schools, which were always a great success. I also want to mention Mario, who helped me a lot with setting up web sites and is my expert for software and hardware questions.

Additionally, I would like to express my gratitude to all colleagues, especially Naga, Stefka, Korbi, Christian, Marion, Gabi, Tami, Maxi, Nina, Richard, Tom, Christoph, Johannes, for stimulating discussions, encouraging feedback and inspirations.

Last but definitively not least I would like to thank my parents who always supported me in my decisions and being proud of me.

Curriculum Vitae

Nadin Neuhauser

Geburtsdatum	16. Februar 1984 in Traunstein
Adresse	Zillertalstraße 69 81373 München
Telefon	Tel. +49 (0) 89 8578 2098
Email	neuhause@biochem.mpg.de

SCHULISCHE AUSBILDUNG

Fachhochschule	10/2004 – 03/2009	**Fachhochschule Weihenstephan**, Freising Dipl.-Ing. (FH) in Bioinformatik 2006
Fachoberschule	09/2001 – 08/2004	**Fachoberschule Traunstein**, Traunstein Fachabitur 2004 in Fachrichtung Technik
Realschule	08/1997 – 08/2001	**Maria-Ward Mädchenrealschule**, Sparz/Traunstein Mittlere Reife 2001 in Fachrichtung Wirtschaft

WISSENSCHAFTLICHE ERFAHRUNGEN

Doktorarbeit	seit 04/2009	**Max Planck Institut für Biochemie,** Martinsried Abteilung für Proteomics und Signal Transduction Computational approaches to enhance mass spectrometry-based proteomics
Diplomarbeit	09/2008 – 03/2009	**Max Planck Institut für Biochemie,** Martinsried Abteilung für Proteomics und Signal Transduction Visualisierungsaspekte der MS basierten quantitativen Proteomics
Studienprojekt	03/2008 – 06/2008	**J. Craig Venter Institute**, Rochville, USA Yeast Colony Image Recognition Server
2. Praxissemester	08/2007 – 01/2008	**European Bioinformatics Institute,** Hinxton, UK Department of Proteomics Services - IntAct Revision of software for Visualization of Molecular Interaction
1. Praxissemester	02/2006 – 06/2006	**Helmholz Zentrum München,** Neuherberg, Germany Institut für Bioinformatik /Abt. Proteomics und Metabolomics Combinatory Prediction of Splice Variants from MS/MS Data

PUBLIKATIONEN

Michalski, A., **Neuhauser, N.**, Cox, J., and Mann, M. (2012), Journal of proteome research 11, 5479-5491
A Systematic Investigation into the Nature of Tryptic HCD Spectra.

Neuhauser, N., Michalski, A., Cox, J., and Mann, M. (2012), Mol Cell Proteomics 11, 1500-1509
Expert System for Computer Assisted Annotation of MS/MS Spectra.

Nagaraj, N., Kulak, N.A., Cox, J., **Neuhauser, N.**, Mayr, K., Hoering, O., Vorm, O., and Mann, M. (2012), Mol Cell Proteomics, Special Issue 2012
System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap.

Cox, J., **Neuhauser, N.**, Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011), Journal of proteome research 10, 1794-1805.
Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment.

Gnad, F., de Godoy, L. M. F., Cox, J., **Neuhauser, N.**, Ren, S., Olsen, J. V., and Mann, M. (2009), Proteomics 9, 4642-4652
High-accuracy identification and bioinformatics analysis of in vivo protein phosphorylation sites in yeast.