Characterization of Polypeptides by Tandem Mass Spectrometry Using Complementary Fragmentation Techniques

MICHAEL LUND NIELSEN
Dissertation presented at Uppsala University to be publicly examined in B21, BMC, Husargatan 3, Uppsala, Thursday, January 11, 2007 at 14:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract


In the growing field of proteomics identification of proteins by tandem mass spectrometry (MS/MS) is performed by matching experimental mass spectra against calculated spectra of all possible peptides in a protein database. One problem with this approach is the false-positive identifications. MS-based proteomics experiments are further affected by a rather poor efficiency typical in the range of 10-15%, implicating that only a low percentage of acquired mass spectrometric data is significantly identified and assigned a peptide sequence.

In this thesis improvement in spectrum specificity is accomplished by using a combination of high-accuracy mass spectrometry and techniques that will yield complementary sequence information. Performing collision-activated dissociation (CAD) and electron capture dissociation (ECD) upon the same peptide ion will yield such complementary sequence information. Implementing this into a proteomics approach and showing the advantages of using complementary fragmentation techniques for improving peptide identification is shown. Furthermore, a novel database-independent score is introduced (S-score) based upon the maximum length of the peptide sequence tag derived from complementary use of CAD and ECD. The S-score can be used to separate poor quality spectra from good quality spectra. Another aspect of the S-score is the development of the ‘reliable sequence tag’ which can be used to recover below threshold identifications and for a reliable backbone for de novo sequencing of peptides.

A novel proteomics-grade de novo sequencing algorithm has also been developed based upon the RST, which can retrieve peptide identification with the highest reliability (>95%). Furthermore, a novel software tool for unbiased identifications of any post-translational modifications present in a peptide sample is introduced (ModifComb). Combining all the tools described in this thesis increases the identification specificity (>30 times), recovers false-negative identifications and increases the overall efficiency of proteomics experiments to above 40%. Currently one of the highest achieved in large-scale proteomics.

Keywords: Mass Spectrometry, Electron capture dissociation (ECD), Collision-activated dissociation (CAD), Proteomics, Post-translational modifications, De Novo sequencing, Bioinformatics

Michael Lund Nielsen, Department of Engineering Sciences, Box 534, Uppsala University, SE-75121 Uppsala, Sweden

© Michael Lund Nielsen 2006

ISSN 1651-6214
ISBN 91-554-6755-5

urn:nbn:se:uu:diva-7409 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7409)
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I. Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study.


II. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry.


III. New database-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse score, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques.


IV. Proteomics-grade de novo sequencing approach.

* = shared first author

V. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex mixtures.

* = shared first author
VI. Extent of modifications in human proteome samples and its effect on dynamic range of shotgun proteomics analysis.


Permission to reprint the articles was kindly granted by the publishers.

Papers not included in this thesis

1. Peptide end sequencing by orthogonal MALDI tandem mass spectrometry.
   Michael L. Nielsen, Keiryn L. Bennett, Brett Larsen, Marc Moniatte and Matthias Mann. *J. Proteome Res.*, 2002, 1, 63-71

2. Experiences and perspectives of MALDI MS and MS/MS in proteomic research.

3. Tandem MALDI/EI ionization for tandem Fourier transform ion cyclotron resonance mass spectrometry of polypeptides.


5. Shifted-basis technique improves accuracy of peak position determination in Fourier transform mass spectrometry.
6. **PhosTShunter**: A fast and reliable tool to detect phosphorylated peptides in liquid chromatography Fourier transform tandem mass spectrometry data sets.

7. Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations.

8. Hydrogen rearrangements to and from radical z fragments in electron capture dissociation of peptides.

9. *De Novo* peptide sequencing and identification with preceision mass spectrometry.


11. Long-range (>25 Å) radical transfer from N-terminal $a$-carbon to distant side chains in linear polypeptides.

    Mikhail M. Savitski, Michael L. Nielsen and Roman A. Zubarev. Submitted to *Analytical Chemistry*, 2006

13. On studying protein phosphorylation using bottom-up LC/MS/MS: the case of human $\alpha$-casein.
    Frank Kjeldsen, Mikhail M. Savitski, Michael L. Nielsen, Lei Shi and Roman A. Zubarev. Submitted to *Analytical Chemistry*, 2006
14. Liquid chromatography at critical conditions: Comprehensive approach to sequence-dependent retention time prediction.
Contents

Introduction .....................................................................................................1

1. High performance liquid chromatography .................................................4
   Reverse-phase liquid chromatography .......................................................5
   Mobile phases ........................................................................................6
   Column parameters ................................................................................7
   The effect of pH and temperature on polypeptide separation ..............8

2. Electrospray ionization ............................................................................9
   The electrospray mass spectrum ...........................................................10
   Nanoelectrospray ionization ...............................................................12

3. Mass spectrometry ................................................................................14
   Fourier transform mass spectrometer ..................................................15
   Ion motion in FTICR MS ..................................................................16
   Ion excitation and detection ..............................................................17
   Fourier transform calculations .........................................................18
   Tandem mass spectrometry ................................................................19
   Collision-activated dissociation .......................................................20
   Electron capture dissociation ..........................................................22
   Complementarity of CAD and ECD ..................................................23
   Electron Transfer Dissociation ...........................................................24

4. Protein and peptide identification ...........................................................25
   Peptide identification through database search engines .................25
   False-positive estimation using decoy database searching ..............27
   High-mass accuracy ............................................................................28
   Detection probability of tryptic peptides in FTMS ...............................29
   Improving peptide and protein identification ..................................30
   S-score for earmarking MS/MS data ................................................32
   Reliable sequence tag ........................................................................32
   Classification of MS/MS data using S-score ....................................33

5. Polypeptide De Novo sequencing .........................................................35
   Proteomics-grade de novo sequencing approach ................................36
   Rescuing false-negative identifications ..............................................38
   De Novo sequencing of endogenous peptides ....................................39
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>CAD</td>
<td>Collisionally activated dissociation</td>
</tr>
<tr>
<td>DB</td>
<td>Database</td>
</tr>
<tr>
<td>DDA</td>
<td>Data-dependent acquisition</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>ECD</td>
<td>Electron capture dissociation</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>ETD</td>
<td>Electron transfer dissociation</td>
</tr>
<tr>
<td>FN</td>
<td>False-negatives</td>
</tr>
<tr>
<td>FP</td>
<td>False-positives</td>
</tr>
<tr>
<td>FTICR</td>
<td>Fourier transform ion cyclotron resonance</td>
</tr>
<tr>
<td>HPLC</td>
<td>High performance liquid chromatography</td>
</tr>
<tr>
<td>i.d.</td>
<td>Inner diameter</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>LMW</td>
<td>Low-molecular weight</td>
</tr>
<tr>
<td>LTQ</td>
<td>Linear trap quadrupole</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass to charge ratio</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MS/MS</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>nLC</td>
<td>Nano liquid chromatography</td>
</tr>
<tr>
<td>o.d.</td>
<td>Outer diameter</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PTMs</td>
<td>Post-translational modifications</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RP</td>
<td>Reverse-phase</td>
</tr>
<tr>
<td>RT</td>
<td>Retention time</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SIM</td>
<td>Selected ion monitoring</td>
</tr>
<tr>
<td>TFA</td>
<td>Trifluoroacetic acid</td>
</tr>
<tr>
<td>TP</td>
<td>True-positives</td>
</tr>
<tr>
<td>μL</td>
<td>Microliter</td>
</tr>
<tr>
<td>μM</td>
<td>Micrometer</td>
</tr>
</tbody>
</table>
Introduction

The central dogma of molecular biology states that biological information flows unidirectional from DNA, which is transcribed to RNA and finally translated to the protein level [1]. This key assumption describes that each gene in the DNA carries the information needed to construct a protein, and therefore serves as the building template for the entire cell. Having all genomic information should therefore make it possible to predict all possible RNA transcripts and hence all possible protein sequences present in a cell. However, with the completion of the human genome [2] it became evident that the ‘one gene – one gene product (protein)’ concept does not hold. Today’s consensus puts the number of human genes in the range of 20,000 to 25,000, within a factor two of many much simpler organisms such as the roundworm (Caenorhabditis elegans) with 19,000 genes [3] and the fruit fly (Drosophila melanogaster) with 14,000 genes [4].

The chimpanzee is evolutionary the closest species to human, and various studies have aimed to understand the genetic basis of the similarities and differences between the two species [5,6]. Recently a comparison between the human and the chimpanzee genome showed more than 96 percent of the DNA sequence is identical between the two species [7]. On average, a typical human protein-coding gene differs from its chimpanzee ortholog by only two amino acid substitutions, and nearly one third of human genes have exactly the same protein translation as their chimpanzee orthologs. Therefore, since the human and chimpanzee genomes are so very similar there must be other factors which accounts for the obvious differences between man and chimp. Gene duplications [8] are a source for some of the differences between human and chimp genetic material, with about 2.7 percent of the genome representing differences produced by gene duplications or deletions during the approximately 6 million years since humans and chimps diverged from their common evolutionary ancestor.

The major difference though between man and chimp, lies in the gene products - the proteins or the ‘proteome’ [9]. Compared to the static DNA the proteome constantly changes through its biochemical interactions with the genome and the environment [10]. Studies have shown that the transcriptome and proteome differ significantly between human and chimp, especially in the brain [11]. Results that clearly show there is no strict linear relationship between genes and the protein complement of a cell.
One gene can encode several proteins through a process called alternative splicing, which increases protein diversity [12]. This occurs in eukaryotes through splicing of pre-mRNA transcribed from one gene and can lead to several different mature mRNA molecules and thereby to different proteins. It is believed that 40-60% of all human genes give rise to alternative splicing isoforms [13,14], and large scale studies on chromosomes 21 and 22 indicates that as much as 80% of all genes could undergo alternative splicing [15]. Further complexity can be added as proteins may be modified chemically after synthesis on the ribosomes so that the protein acquires a different function. This type of modification is referred to as post-translational modifications (PTMs). When combining the complexity generated by alternative splicing with that produced by PTMs, it is estimated that close to a million different protein molecules are expressed in a human cell [16].

It is therefore evident that genetic information is not enough to biologically fully understand any organism, yet alone to describe the differences between species. One of the major challenges in the post-genomic era is therefore to fully characterize and identify all proteins expressed at any given time point in cells. Large-scale protein analysis, particularly the study of structure and function of all expressed proteins has become the method of choice for a better understanding of life [17] – an approach referred to as proteomics [18]. The term was coined to make an analogy with genomics, and while it is often viewed as the "next step", proteomics is much more complicated than genomics. Not only due to the shear differences in complexity (vide supra), but unlike the scalable DNA sequencing through polymerase chain reaction (PCR), proteomics has to deal with limited and variable sample material as well as a dynamic range in protein abundance of up to 11 orders of magnitude [19].

The fundamental aspect of proteomics is protein characterization, and the first step in any proteomics analysis is determination of protein primary amino acid sequence. This basic information is essential in order to retrieve any quantitative information as to which proteins are expressed, as well as to determine localization, interaction, activity, and, ultimately protein function [20]. Once protein identification has been established localization and quantification of any PTMs present on the protein at any given time can be performed. Information that is highly valuable in e.g. signal transduction studies [21].

Mass spectrometry (MS) has emerged as the primary tool for protein characterization. This is due to the unique features of MS, which includes the ability of identifying individual proteins in complex mixtures, protein quantification and characterization of important PTMs. Over the last 15 years, the increasing versatility of MS have replaced more traditional protein characterization approaches such as Edman degradation [22].
Even though tremendous successful, MS-based proteomics still faces significant technical challenges, of which some will be discussed in this thesis. Identification of protein amino acid sequence is predominantly performed by searching acquired data in a database containing theoretical protein sequences. Although this search can be performed automatically, a problem governing this approach is the inferior ability to separate the false-positive and false-negative peptide matches [23]. Furthermore, MS-based proteomics experiments are affected by a rather poor efficiency typical in the range of 10-15%, implicating that only a low percentage of acquired mass spectrometric data is significantly identified and assigned a peptide sequence. Novel approaches are therefore needed to improve these numbers, and these methods could also be important tools in the quest for reaching the over-all goal of proteomics.

This thesis will focus on the benefits of using high-mass accuracy MS in conjunction with complementary fragmentation techniques for polypeptide characterization. Methods are developed that improve determination of protein amino acid sequence as well as unbiased (i.e. hypothesis-free) characterization and localization of PTMs. These methods not only show a significant improvement in protein identification confidence, but additionally reduce the presence of false-positive identifications while increasing the over-all efficiency rate of proteomics experiments. The results of the research presented in this thesis have been published in six peer-reviewed papers. These papers are noted with the roman numerical I-VI. Briefly paper I describes a physicochemical study upon the detection probability of tryptic peptides. Improvement of protein identification and reduction of false-positives is demonstrated in papers II-IV, while a novel approach for unbiased analysis of protein PTMs in proteomics are presented in papers V-VI.
1. High performance liquid chromatography

High performance liquid chromatography (HPLC) is an important analytical tool for separating and quantifying components in complex liquid mixtures. HPLC is used to separate components of a mixture by using a variety of chemical interactions between the substance being analyzed (the analyte) and the chromatographic column. By choosing the appropriate equipment (i.e. column and detector), this method is applicable to samples with components ranging from small organic/inorganic molecules and ions to polymers and proteins with high molecular weight. In isocratic HPLC the analyte is forced through a column of the stationary phase by introducing a liquid at high pressure. Use of pressure gives the components less time to diffuse within the column, leading to improved resolution in the resulting chromatogram. Solvents used include any miscible combination of water or various organic liquids (the most common are methanol and acetonitrile). Water may contain buffers or salts to assist in the separation of the analyte components, or compounds such as Trifluoroacetic acid (TFA).

A further refinement to HPLC has been to vary the mobile phase composition during the analysis; this is known as gradient elution. A normal gradient in HPLC might start at 5% methanol and progress linearly to 50% methanol over 25 minutes, depending on how hydrophobic the analyte is. The gradient separates the analyte mixtures as a function of how well the changing solvent mobilizes the analyte. In this example, using a water/methanol gradient, the more hydrophobic components will elute (come off the column) under conditions of relatively high methanol; whereas the more hydrophilic will elute under conditions of relatively low methanol. The choice of solvents, additives and gradient depend on the nature of the stationary phase and the analyte. The various types of HPLC and their characteristics are summarized in the table below. In this thesis, only reverse-phase liquid chromatography will be used. The principles of reverse-phase liquid chromatography is therefore described in more detail vide infra.
### Table 1.1. Various types and applications of HPLC

<table>
<thead>
<tr>
<th>Type</th>
<th>Sample polarity</th>
<th>Molecular weight</th>
<th>Stationary phase</th>
<th>Mobile phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adsorption</td>
<td>Non-polar to somewhat polar</td>
<td>10^0 - 10^4</td>
<td>Silica or alumina</td>
<td>Non-polar to polar</td>
</tr>
<tr>
<td>Partition (reverse-phase)</td>
<td>Non-polar to somewhat polar</td>
<td>10^0 - 10^4</td>
<td>Non-polar liquid adsorbed or chemically bonded to packing material</td>
<td>Relatively polar</td>
</tr>
<tr>
<td>Partition (normal-phase)</td>
<td>Somewhat polar to highly polar</td>
<td>10^0 - 10^4</td>
<td>Highly polar liquid adsorbed or chemically bonded to the packing material</td>
<td>Relatively non-polar</td>
</tr>
<tr>
<td>Ion exchange</td>
<td>Highly polar to ionic</td>
<td>10^0 - 10^4</td>
<td>Ion-exchange resins made of insoluble, high-molecular weight solids functionalized typically with sulfonic acid (cationic exchange) or amine (anionic exchange) groups</td>
<td>Aqueous buffers with added organic solvents to moderate solvent strength</td>
</tr>
<tr>
<td>Size exclusion</td>
<td>Non-polar to ionic</td>
<td>10^4 - 10^6</td>
<td>Small, porous, silica or polymeric groups</td>
<td>Polar to non-polar</td>
</tr>
</tbody>
</table>

---

**Reverse-phase liquid chromatography**

The term ‘reverse-phase chromatography’ was first coined by Howard and Martin in 1950, and was developed due to the increasing interest in large nonpolar biomolecules [24]. As shown in table 1-1, reverse-phase HPLC (RP-HPLC) consists of a nonpolar stationary phase and a polar mobile phase. A common stationary phase in RP-HPLC is silica, and figure 1.1 shows an example of a C_{18} stationary phase commonly used. This type of stationary phase behaves like a nonpolar oil coating, because of a long hydrocarbon chain [C_{18}H_{37}^+], on the surface of solid silica particles. Therefore this finely divided solid C_{18} polymer can be considered as a liquid. When a solution consisting of a mixture of compounds passes between the C_{18} particles, the low-polarity components dissolve (partition) in the C_{18} hydrocarbon layer and therefore becomes slowed down in elution. In contrast, the more polar components in the mixture are not very soluble in the C_{18} hydrocarbon layer and are eluted through the column quickly. Polypeptides are too large...
to partition into the hydrophobic stationary phase; they adsorb to the hydrophobic surface after entering the chromatographic column and remain adsorbed until the concentration of organic modifier reaches the critical concentration to cause desorption. After desorption they interact only slightly with the surface as they elute down the column. The retention time for both polypeptides and partitioned molecules are therefore longer for molecules which are more non-polar in nature, allowing polar molecules to elute more readily. For polypeptides and small proteins less than 5,000 daltons C$_{18}$ columns are generally preferred. Proteins larger than 5,000 daltons, or small proteins that are particularly hydrophobic, are best separated using a C$_{4}$ column. Additionally C$_{8}$ columns exist, which are similar to C$_{18}$ in their application but sometimes offer a different selectivity or ability to separate particularly peptides.

![Example of common silica C18 stationary phase](image)

*Figure 1.1. Example of common silica C18 stationary phase*

**Mobile phases**

Mobile phase is the most important parameter in RP-HPLC. It refers to the solvent being continuously applied to the column, or stationary phase. The type of mobile phase used may have a big effect on the retention, and it can promote or suppress an ionization of the analyte molecules, as well as shielding an accessible residual silanol or any other active adsorption centers on the adsorbent surface. Proper selection of the mobile phase is the second most important step in the development of the separation method (the first one is the selection of the adsorbent type). The main requirement for the mobile phase is that it has to dissolve the analytes up to the concentration suitable for the detection. Variation of the eluent composition provides the great flexibility of RP-HPLC separations. The mobile phase in reversed-phase chromatography has to be polar and it also has to provide a reasonable competition for the adsorption sites for the analyte molecules. Table 1.2 shows the most important parameters of some common solvents often used as eluent components in RP-HPLC.
Table 1.2. Parameters for common solvents used in RP-HPLC

<table>
<thead>
<tr>
<th>Name</th>
<th>Molecular weight</th>
<th>Boiling point</th>
<th>UV absorption</th>
<th>Viscosity</th>
<th>Dipole moment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetonitrile</td>
<td>41</td>
<td>82° C</td>
<td>195 nm</td>
<td>0.358</td>
<td>3.37</td>
</tr>
<tr>
<td>Dioxane</td>
<td>88</td>
<td>101° C</td>
<td>215 nm</td>
<td>1.260</td>
<td>0.45</td>
</tr>
<tr>
<td>Ethanol</td>
<td>46</td>
<td>78° C</td>
<td>205 nm</td>
<td>1.190</td>
<td>1.68</td>
</tr>
<tr>
<td>Methanol</td>
<td>32</td>
<td>65° C</td>
<td>205 nm</td>
<td>0.584</td>
<td>1.66</td>
</tr>
<tr>
<td>Isopropanol</td>
<td>60</td>
<td>82° C</td>
<td>205 nm</td>
<td>2.390</td>
<td>1.68</td>
</tr>
<tr>
<td>Tetrahydrofuran</td>
<td>72</td>
<td>66° C</td>
<td>215 nm</td>
<td>2.200</td>
<td>1.70</td>
</tr>
<tr>
<td>Water</td>
<td>18</td>
<td>100° C</td>
<td>185 nm</td>
<td>1.000</td>
<td>1.84</td>
</tr>
</tbody>
</table>

*Table 1.2. Parameters for common solvents used in RP-HPLC*

Flow rate

Flow rate has little effect on polypeptide separations since polypeptide desorption is the result of reaching a precise organic modifier concentration as mentioned above. Protein resolution, therefore, is relatively independent of mobile phase flow rate. For small peptides though, the resolution may be affected by the eluent flow rate because the behavior of small peptides on the RP-HPLC column is between that of polypeptides and small molecules [25].

With smaller diameter of the RP-HPLC column the flow rate decreases and thus lowering the amount of solvent needed for polypeptide separation. Polypeptides therefore elute in smaller volumes of solvent at this reduced flow rate, and the result is an increase in detection sensitivity. As a comparison, a narrowbore column with a flow rate of 200 μL/min gives a five fold increase in sensitivity compared with an analytical column run at a flow rate of 1.0 mL/min. This is particularly useful when coupling HPLC with MS.

Column parameters

Column length.

The adsorption/desorption of proteins responsible for their separation takes place almost near the top of the column. Therefore, column length does not significantly affect separation and resolution of proteins. Small peptides on the contrary, such as those usually analyzed in proteomics, are better separated on longer columns since they adsorb/desorbs throughout the column. Consequently, columns of 15-25 cm length are often used for separation of synthetic and naturally peptides and enzymatic digest maps [25].
Column diameter.

The column diameter does not affect peak resolution, but it does affect sample loading capacity, solvent usage and sensitivity.

Column back-pressure.

Column back-pressure is directly proportional to the column length. When using more viscous solvents (see table 1.2) like isopropanol, shorter columns will result in more moderate back-pressure. Analogous, using less viscous solvents like acetonitrile is better for longer and thinner columns, hereby minimizing back-pressure.

Sample capacity.

Sample capacity is a function of column volume. For columns of equal diameter, longer columns maximize sample capacity.

The effect of pH and temperature on polypeptide separation

Since separation in RP-HPLC is based upon the hydrophobic character of the components being analyzed, it is obvious that any organic ionizable component will have a longer retention time (RT) in its neutral form compared to its ionized form. Ionization is a pH dependent process, and therefore pH of the mobile phase has a significant effect on the separation of complex organic mixtures containing basic or acidic components. Compounds in ionic forms are more hydrophilic and therefore tend to have less interaction with the hydrophobic stationary phase, and instead tend to be more solvated with the water molecules in the mobile phase. This causes a significant decrease of the RT of ionic species.

A commonly used HPLC mobile phase modifier in peptide and protein separations is TFA [26]. Its excellent ion-pairing and solvating characteristics confer unique chromatographic sensitivity on peptide separations, while its low UV cut-off (192 nm) allows detection of the peptide amide bond at 210 nm without interference from the mobile phase. Furthermore, due to its volatility, TFA can easily be removed from collected sample fractions by simple speed-vac evaporation. TFA, however causes significant signal suppression when combining LC with electrospray due to ion-pairing between TFA and the basic analytes [27], with reduced signal detection in the mass spectrometer as the result. A common substitute for TFA in LC-MS analysis is acetic acid.
2. Electrospray ionization

The idea of using electrospray dispersion of an analyte solution in a bath gas to produce solute ions for mass analysis was first introduced by Dole et al. in 1968 [28]. Dole also discovered the important phenomenon of multiple charging of molecules. The preparation of gas-phase ions from large biomolecules by solution electrospray was initially presented in 1984 [29], and it was Fenn’s work that ultimately led to the modern day technique of electrospray ionization mass spectrometry and its application to biological macromolecules. Professor Fenn was in 2002 awarded the Nobel price in chemistry for his contribution to the development of todays electrospray ionization [30]. Further improvements have lowered the continuous flow rate and thereby enabled the analysis of minute amounts of sample in a technique referred to as nano-electrospray (nanoESI) [31,32].

In ESI a liquid is pushed through a very small charged, usually metal, capillary at ambient temperature and atmospheric pressure. A typical solvent for positive-ion analysis would be water/methanol or water/acetonitrile (50:50), perhaps with a small percentage of formic or acetic acid to act as a source of protons. The solution passes through a capillary tube to the ion source of the mass spectrometer, where application of a large electric field to the end of the tube (several kilovolts relative to a nearby counter-electrode) disrupts the emerging liquid surface and provides a fine spray of highly charged droplets. The net charge of each droplet has the same polarity, either positive or negative, dependent on the polarity of the applied field. A neutral carrier gas, such as nitrogen gas, is sometimes used to help nebulize the liquid and to help evaporate the neutral solvent in the small droplets. As the small droplets, now suspended in air evaporate the charged analyte molecules are forced closer together, thus concentrating the charge density. The proximity of the molecules becomes unstable as the similarly charged molecules come closer together and the droplets once again explode. This is referred to as Coulombic fission because it is the repulsive Coulombic forces between charged analyte molecules that drive it (figure 2.1). Eventually a lone ion is produced, which then continues along to the mass analyzer of a mass spectrometer.
There remains debate as to the exact mechanism of the electrospray process, especially in the later part of the process as the lone ion is formed. Two major competing theories regarding the final production of lone ions exists: The charged residue model (CRM) initially suggested by Dole [28], and the ion evaporation model (IEM) as suggested by Iribane and Thompson [33]. The CRM suggests that at the fission event, the droplet breaks in half and the output is either two identical droplets (even fission) or two disproportional droplets (uneven fission). This successive droplet fission continues until droplets are produced containing only one ion. This molecule then becomes a free gas phase ion as the last of the solvent evaporates. The IEM also assumes repeated droplet fission. However, at a certain droplet size the force of the columbic repulsion inside the droplet becomes great enough to overcome the solvation forces, and the ions desorbs directly out of the droplet and into the gas phase. It is currently not known, which of the two models that are the predominant mechanism for generating gas phase ions in ESI, but it has been suggested that both models probably occur for different analytes/solvents [34]. It is therefore assumed that polypeptide ionization using LC-MS predominantly occurs through IEM.

The electrospray mass spectrum

In positive electrospray the observed ions are quasimolecular ions, ionized by the addition of a proton (hydrogen ion) to give [M+H]+ (M = analyte molecule, H = hydrogen ion), or other cations such as sodium ion [M+Na]+.
In ESI multiply charged ions such as \([\text{M}+2\text{H}]^{2+}\) are often observed, and for a typical proteomics LC-MS analysis of tryptic peptide mixtures, 90% of all multiply charged ions are doubly charged (paper I). This multiple charging occurs if the analyte has many sites capable of being either protonated or deprotonated, such as the NH$_2$ groups in peptide/proteins or acidic groups in polynucleotides. Therefore, the choice between positive-ion or negative-ion analysis would depend on the nature of the sample. The better polarity for a given sample is the one in which the analyte are more able to accommodate the charge.

For large macromolecules, such as proteins, there will often be a distribution of many charge states and the charge on the ions can be as great as \([\text{M}+25\text{H}]^{25+}\). Note that these are all even-electron species; electrons themselves have neither been added nor removed. Figure 2.2 shows the electrospray mass spectrum of a protein, human $\alpha$-chain hemoglobin. As can be seen, the acquired mass spectrum from a solution containing a large source of protons consists of a series of \([\text{M}+z\text{H}]^z+\) ions, ranging from $z = 12$ to $z = 21$, each peak differing from its neighboring peak by one electronic charge. This allows for mass determination of molecular ions with very large masses even though the MS has a limited $m/z$ detection range.

![Figure 2.2. ESI spectrum of human $\beta$-chain hemoglobin, showing multiply charging ranging from $z=12$ to $z=21$](image)

Smaller ions, such as tryptic peptides, are less charged in the electrospray mass spectrum since these molecules contain fewer protonation sites. Still it is common that a peptide ion exists in several charge states, but tryptic peptides are predominantly detected as doubly charged species.
ESI in general is preferably performed on salt-free solution, either at low or high pH, depending on whether positive or negative ion mode is needed. Salts present a problem for ESI by lowering the vapor pressure of the droplets resulting in reduced signal through an increase in droplet surface tension, reducing volatility. Consequently, volatile buffers such as ammonium acetate can be used more effectively. Furthermore, compounds such as detergents, soluble polymers (polyethyleneglycol (PEG)) and liquids with high boiling points such as dimethylsulfoxide (DMSO) or glycerol should be avoided, since they interfere with the ionization process of analyte molecules. In summary table 2.1 shows some of the advantages and disadvantages of using ESI as ionization source in mass spectrometric analysis.

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practical mass range of up to 70,000 Da</td>
<td>The presence of salts and ion-pairing agents like TFA can reduce sensitivity</td>
</tr>
<tr>
<td>Good sensitivity with femtomole to low picomole sensitivity typical</td>
<td>Complex mixtures can reduce sensitivity</td>
</tr>
<tr>
<td>Softest ionization method, capable of generating noncovalent complexes in the gas phase</td>
<td>Simultaneously mixture analysis can be poor</td>
</tr>
<tr>
<td>Easily adaptable to liquid chromatography</td>
<td>Multiply charging can be confusing especially in mixture analysis</td>
</tr>
<tr>
<td>Easily adaptable to tandem mass analyzers</td>
<td>Sample purity is important</td>
</tr>
<tr>
<td>Multiple charging allows for analysis of high mass ions with a relatively low m/z range instrument</td>
<td>Carryover from sample to sample</td>
</tr>
<tr>
<td>No matrix interference</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2.1: Advantages and disadvantages of electrospray ionization (ESI)*

**Nanoelectrospray ionization**

Low flow electrospray was originally described by Wilm and Mann [32] and is often referred to as nanoelectrospray, nanospray, or microelectrospray. This ionization source is a variation on ESI, where the spray needle has been made very small and positioned close to the entrance to the
mass analyzer. The end result of this rather simple adjustment is increased efficiency, which includes a reduction in the amount of sample needed.

The flow rates for nanoESI sources are on the order of tens to hundreds of nanoliters per minute. In order to obtain these low flow rates, nanoESI uses emitters of pulled and in some cases metallized glass or fused silica that have a small orifice (~5 μm). The dissolved sample is added to the emitter and a pressure of ~30 PSI is applied to the back of the emitter. Effusing the sample at very low flow rates allows for high sensitivity. Also, since the emitters are positioned very close to the entrance of the mass analyzer the ion transmission to the mass analyzer is much more efficient. For instance, the analysis of a 5 mM solution of a peptide by nanoESI would be performed in 1 minute, consuming ~50 femtomoles of sample. The same experiment performed with normal ESI in the same time period would require 5 picomoles, or 100 times more sample than for nanoESI. In addition, since the droplets are typically smaller with nanoESI than normal ESI, the amount of evaporation necessary to obtain ion formation is much less. As a consequence, nanoESI is more tolerant of salts and other impurities because less evaporation means the impurities are not concentrated down as much as they are in ESI.
3. Mass spectrometry

A mass spectrometer produces ions and separates them in the gas phase according to their mass-to-charge ratio (m/z). Today a wide variety of mass spectrometers is available, ranging from bench top detectors for gas chromatography to warehouse sized instruments such as accelerator mass spectrometers. All of these share the capability to assign mass-to-charge values to gas-phase ions. Although the principles of operation and the types of experiments that can be done on these instruments differ greatly, the design resembles each other. Basically a mass spectrometer consists of an ion source, a mass-selective analyzer, and an ion detector.

Mass spectrometers are operated at reduced pressure in order to prevent collisions of ions with residual gas molecules during the flight from the ion source to the detector. The vacuum should be such that the mean free path length of an ion, i.e., the average distance an ion travels before colliding with another gas molecule, is longer than the distance from the source to the detector. At a pressure of 5x10^{-5} Torr the mean free path length of an ion is approximately one meter, almost the same length as a quadrupole instrument. This means the introduction of a sample into a mass spectrometer usually requires crossing a rather large pressure drop, and several means have been devised to accomplish this. Gas samples may be directly connected to the instrument and metered into the instrument via a needle valve. Liquid and solid samples can be introduced through a septum inlet or a vacuum-lock system. However, when connecting continuous introduction techniques like high performance liquid chromatography (HPLC) or capillary electrophoresis (CE), special interfacing becomes imperative to prevent excessive gas load.

ESI is the perfect mean to introduce liquid samples into the gas phase and ultimately the mass spectrometer. With the development of nanoESI coupling of small column RP-HPLC to MS instruments have been made feasible [35]. The low flow required for nanoESI fits perfectly to the small reverse-phase columns of todays LC-MS instrumentation, where inner column diameters of 75 micron requires a flow rate around 200-500 nL/min.
Fourier transform mass spectrometer

Fourier transform mass spectrometry (FTMS) has received considerable attention due to its ability to make mass measurements with a combination of resolution and accuracy higher than any other mass spectrometer. FTMS includes first of all ion cyclotron resonance (ICR) spectrometry, a theory of which was developed by Lawrence and Edlefsen in 1930 [36]. Lawrence built the first cyclotron accelerator, which was used to study the fundamental principles of atoms. For his discovery of the cyclotron principle and its application to problems in high energy physics Lawrence was awarded the 1939 Nobel price in physics. In 1949 the principle of ICR was first incorporated into a MS called the omegatron [37]. Over the next 15 years several instruments were designed predominantly for studying ion-molecule reactions. In 1974 Comisarow and Marshall adapted Fourier transform methods to ICR spectrometry and performed the first FTMS experiments [38]. Since then, interest in FTMS has increased exponentially and so has the number of FTMS instruments.

All FTICR MS instruments have four components in common, that are characteristic for FTICR MS compared to other mass spectrometers. One of these components is a strong magnet, which can either be a permanent magnet, an electromagnet or a superconducting magnet. Permanent magnets and electromagnets are both applicable to FTMS, but they have low field strength that limit the performance of an FTMS instrument. For that reason, very few systems have been built with these. The superconducting magnets on the other hand, are capable of producing field strengths ranging from 3 T to above 15 T.

Another component is the analyzer cell where ions are stored, analyzed and detected. The most common analyzer cell shape is cubic, introduced by

![Figure 3.1. Most commonly used analyzer shape in FTICR MS – cubic shape. The direction of the magnetic field is indicated by B.](image)
McIver [39]. Consisting of six plates arranged in the shape of a cube (figure 3.1), the cell is oriented in the homogeneous magnetic field of a superconducting magnet, so that one opposing pair of plates is orthogonal to the direction of the magnetic field, and two pairs lie parallel to the field. The two plates perpendicular to the magnetic field are called trapping plates, as they assist in trapping the ions inside the cell along with the magnetic field. The four remaining plates are used for ion excitation and ion detection.

The third feature required is an ultra-high vacuum system. FTMS instruments are more pressure sensitive to vacuum since ions being analyzed travel very long distances during experiments. Finally a sophisticated data acquisition system is required in order to perform the necessary Fourier calculations.

Ion motion in FTICR MS

In the FTICR MS analyzer cell the ion motion is governed by both a magnetic and electric field. The sum of forces acting on an ion of charge $q$ and velocity $v$ through the electric field $E$ and magnetic field $B$ is described in equation 3.1. The value of $B$ is constant over time and volume, because the magnetic field is uniform, unidirectional and homogeneous.

$$
F = qE + q(v \otimes B)
$$

The electric field arises from voltages applied to the trapping plates, and can have both r.f. and d.c. components. The combination of the magnetic and electric fields creates a three-dimensional ion trap, which allows ions to be stored in the ICR analyzer cell for seconds, minutes or even hours. Interactions by the magnetic and electric field give rise to three different types of ion motions inside the ICR cell; Cyclotron, Magnetron and Trapping motion.

Cyclotron motion

The ion cyclotron motion is the basis for FTICR MS, and arises from the ions interactions with the unidirectional magnetic field. The ion experiences a force perpendicular to its velocity direction, hereby causing the ion to move in a circular motion. This force is also referred to as the Lorentz force:

$$
f_c = \frac{qB}{2\pi m}
$$

The cyclotron frequency is determined by the strength of the magnetic field ($B$), the charge present on the ion ($q$), and the mass of the ion ($m$). The cyclotron motion fall in the range of tens of kilohertz to megahertz.
**Trapping motion**

When an ion moves parallel to the magnetic field it experiences no forces from the field. It is therefore necessary to apply a static voltage to each of the two aforementioned trapping plates in order to avoid ions escaping along the z-axis (direction of the magnetic field). The trapping plates create a potential well that traps the ions inside the cubic analyzer cell. This causes the ions to undergo simple harmonic oscillation between the trapping plates along the magnetic field. By applying a small symmetric positive voltage to the trapping plates, it is possible to trap positive ions in the ICR cell, and consequently by applying a small negative voltage negative ions will be trapped.

**Magnetron motion**

Magnetron motion is due to a combination of forces derived by the magnetic field and the electric field. As mentioned above, the magnetic field constrains ion motion in the x/y-plane and the electric field constrains ion motion along the z-axis parallel to the direction of the magnetic field. The electric field therefore drives the ions away from the center of the cell, while the magnetic field prevents the ions from being accelerated into the walls of the analyzer cell. Combining these two fields gives rise to magnetron motion, a precession of the guiding center of the cyclotron motion of an ion around the center of the cell. Magnetron frequency is independent of mass-to-charge ratio, but is a function of magnitude of trapping potential (V), the magnetic field strength (B), the distance between the two trapping plates (a), as well as the geometry of the analyzer cell, represented by the geometry factor (α).

\[
f_m = \frac{\alpha V}{\pi a^2 B}
\]

Magnetron frequencies are of the order 1-100 Hz and therefore much lower than cyclotron frequencies.

**Ion excitation and detection**

The orbital radius of an ion trapped inside the ICR cell will increase when a oscillating electric field is applied on it whose angular frequency (ω’) is equal or close to the ion cyclotron frequency:
This is referred to as excitation of ions. This excitation is used to: (a) accelerate ions coherently to a larger (and thus detectable) orbital radius; (b) increase ion kinetic energy above the threshold for ion dissociation and/or ion-molecule reactions; and finally (c) accelerate ions to a cyclotron radius larger than the radius of the Penning trap, so that the ions are removed ("ejected") from the cell. All ions of the same mass-to-charge ratio are excited coherently, hereby grouping the ions more tightly. Ions of the same mass-to-charge ratio undergo cyclotron motion as a packet. If the dipolar excitation is turned off before the ions strike the cell plates, they undergo cyclotron motion with a large radius orbit. As they pass the cell’s detection plates, the coherently orbiting ion packet attracts electrons to first one and then the other detection plate through the external circuit that joins them. By doing this an alternating current is produced, also referred to as the image current. The induced current is due to the difference ΔQ in the charge induced on the conductive detection plates by the ion package:

\[ \Delta Q = -\frac{2qv}{d} \]  

(3.5)

where \( q \) is the total charge of the ion package, \( d \) the distance between the detection plates and \( v \) is the ion velocity in the \( y \)-direction (along the plate) [40]. Since the total ion charge \( q \) is proportional to the ionic charge state, the induced current and thus the FTICR signal response is proportional to the ionic charge state as well (the number of ions is assumed constant).

This sinusoidal image signal produced by the periodic cyclotron motion of the ions can be amplified, digitized and sent to a large computer for processing. The principle of detecting ions through an image current is another feature unique for FTMS. Other mass spectrometers detect ions through destructive collisions. Image current, on the other hand, is a non-destructive detection method; the ions remain inside the analyzer cell after the detection process has been completed.

Fourier transform calculations

The image current derived from ion detection is a composite of sinusoids of different frequencies and amplitudes. The frequency components of the obtained signal can be converted into a mass spectrum by applying a Fourier transform to the time domain transient (equation 3.6). The time domain is displayed as a waveform voltage versus time, whereas the frequency domain

\[ \omega = \frac{qB}{m} \]  

(3.4)
is shown as a spectrum of magnitude or power versus frequency. The transform consists of three parts; a complex exponential, which is the Fourier kernel \( e^{i\omega t} \); the weighting factor \( h(t) \) for the forward transform, and \( H(f) \) for the inverse transform; and the integration, which corresponds to summation of the separate real and imaginary components, but with a continuous variable.

\[
H(f) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f t} dt
\]  
(3.6)

As usual for a technique based on Fourier transform, the resolution improves in direct proportion to the length of the recorded transient [41]. The maximum resolution that can be achieved for a data set is shown in equation 3.7, where \( R \) is resolving power, \( f_C \) is the cyclotron frequency and \( T \) is the duration of the transient.

\[
R = \frac{f_C T}{2}
\]  
(3.7)

At 10^{-10} Torr, transients of 60 seconds or longer have been achieved demonstrating mass resolution at \( m/z > 1000 \) in excess of 10^6, much higher compared to what is possible using other mass spectrometric techniques [42].

**Tandem mass spectrometry**

The principle of MS/MS is to generate characteristic secondary fragment ions of a selected and isolated molecular precursor ion. These fragment ions are produced through activation of the particular ion of interest. In ICR MS where measurements are performed in high vacuum, ion isolation and detection usually takes place at different locations. Isolation and fragmentation is usually performed in low-vacuum region while detection of fragment masses is performed in the high-vacuum region. The nomenclature of peptide fragmentation is depicted in figure 3.2, as initially suggested by Roepstorff and Fohlman [43]. This nomenclature was later modified by Biemann [44] and extended by Kjeldsen et al. [45] and Savitski et al. [46]. Peptide bond breakage between C\(_\alpha\)-C, C-N and N-C\(_\alpha\) yields six different product ions annotated as \( a_n, b_n \) and \( c_n \) when charge retention is at the N-terminal fragment, and \( x_n, y_n \) and \( z_n \) when charge is located on the C-terminal fragment [47]. The subscript \( n \) indicates the number of residues that is retained in the fragment ion. Secondary side-chain cleavages can occur between C\(_h\) and C\(_r\) from radical \( a \).
Collision-activated dissociation

If an ion collides with a neutral atom or molecule, the kinetic energy can be converted into internal energy, called collisional activation. If there is enough excess energy to break chemical bonds, the ion will decompose. This is referred to as collision-activated dissociation (CAD) or collision-induced dissociation (CID), with both terms referring to the same principle of dissociation.

Only a portion of the ion kinetic energy, $E_{lab}$, can be converted to internal energy, which can be derived by the physics of the ion-neutral collision from a center-of-mass reference frame.

$$E_{cm} = E_{lab} \frac{m_2}{m_1 + m_2}$$

(3.8)

Here $m_1$ = the ion mass, $m_2$ = the target gas mass, $E_{lab}$ is the ion kinetic energy in the laboratory frame of reference and $E_{cm}$ is the maximum energy fraction converted into internal energy.
Increasing the ion kinetic energy and/or using a more massive target can increase the amount of kinetic energy that can be converted into internal energy. Multiple collisions with target molecules can increase the internal energy, but can also result in randomization of the internal energy and hereby increase the probability of undesirable rearrangements. Two collisional activation regimes are used in CAD, low-energy and high-energy collisions.

**Low-energy collisions**

Low-energy refers to collisions where the precursor ion is accelerated to kinetic energies in the range up to a few hundred eV. These collisions are believed to excite vibrational states, since the interaction time between an ion with mass 20 Da at an energy of 30 eV with a target of a few Å is about $10^{-14}$ s, corresponding approximately to the bond's vibration period [48]. The observed product ions depend strongly on the internal energy distribution. Increasing the collision energy shifts the center of the internal energy distribution to a higher value, and hereby changes the observed product ions. A product-ion mass spectrum resulting from 10 eV collisions can be dramatically different from one resulting from 25 eV collisions.

The mass of the target has a strong influence on the MS/MS spectrum for low-energy CAD. Heavier gases such as xenon, krypton and argon are often used as target gases to increase the center-of-mass energy, because they allow the transfer of more energy and hence increase the possibility of observing high-energy fragments.

Low-energy CAD is typically performed using triple quadrupole or hybrid instruments, where the collision chamber is a quadrupole working in the r.f. mode, allowing one to focus the ions that are angularly dispersed. The fragmentation observed through the low energy process is dominated by $b_n$ and $y_n$ ions arising from a charge-directed fragmentation pathway [49,50]. In addition, peaks are observed due to losses of ammonia (-17 Da) and water (-18 Da).

**High-energy collisions**

High-energy refers to collisions where the precursor ion is accelerated to kinetic energies of approximately 1 kilovolt or higher, which results in excitation of the electronic states. The electronic excitation rapidly converts to excess of vibrational energy, which causes fragmentation. High-energy collisions produce a broad internal energy distribution, and virtually all structurally possible fragments have some probability of occurring. The target mass does not have a large influence on the MS/MS spectrum in high-energy CAD (equation 3.8). That is also why changes in collision conditions such as collision gas, pressure and temperature do not produce large changes in the
product-ion spectrum. Helium is often used as target gas because it is inexpensive, has a high ionization potential and does not cause a large scattering of the product ions. High energy CAD is usually applied in multisector instruments, and the fragmentation pattern of singly protonated peptides is found to be dominated by $a_n$, $d_n$ and $w_n$ ions along with additional low-energy CAD fragment ions [51]. Unlike low-energy ions losses of ammonia or water are not observed. The characteristic effects of individual amino acid residues on peptide fragmentation behavior have been reviewed in detail by Papayannopoulos [52].

Electron capture dissociation

Multiply charged polypeptide ions can capture a thermal electron (<0.2 eV) and undergo fragmentation, a technique termed electron capture dissociation (ECD) [53]. Through this ion-electron recombination, the even-electron polypeptide precursor ions are charge reduced and hereby turned into odd-electron species with same nominal mass but increased m/z value. This formed radical specie is unstable and can then undergo fragmentation through a radical-initiated backbone cleavage and forming mainly $c'$ and $z'$ fragment ions (figure 3.3). The actual mechanism of ECD is still under debate.

![Figure 3.3](image)

*Figure 3.3. Suggested ECD mechanism for bond cleavage giving rise to $c'$ and $z'$ fragment ions.*

ECD has proven itself to be a very useful dissociation technique especially when looking at labile post-translational modifications on peptides and proteins. Over the last couple of years, several mass spectrometric groups around the world have shown that ECD very effectively localizes O-glycosylation sites [54], phosphorylation sites [55,56], γ-Carboxyglutamic acid [57] and internal S-S bonds [58]. Add to this the improved structure analysis of polypeptides by ECD compared to other fragmentation techniques [53,59].

The reason for these differences in ECD compared to CAD is the fragmentation occurs prior to distribution of the internal energy inside the pep-
tide molecule (non-ergodic fragmentation). This non-ergodic effect is why labile modifications are retained on amino acid residues in ECD. Other dissociation techniques, like CAD, are slow heating processes which allow molecules time to distribute all the internal energy to the weakest bonds, which in peptides are the labile modifications. These bonds will break and determination of the specific modified amino acid site is therefore difficult.

Complementarity of CAD and ECD

Complementarity of CAD and ECD fragmentation is not only observed in differences of fragment ions produced, nor the differences in ability to retain labile modifications. Even preferences as to which amino acids preferentially are cleaved show complementarity between the two fragmentation techniques. It has long been known, that ECD does not cleave N-terminal to proline whereas CAD cleaves at an increased rate [53,60]. Subtle preferences in CAD has recently been studied on large data set [61], but not until recently [62] a full study of cleavage preferential for both fragmentation techniques have been made by our group (figure 3.4). This study clearly showed the fragmentation preferences for CAD and ECD are complementary, and this complementarity can be very beneficial for proteomics studies. Comparison of the frequency preferences between CAD and ECD show a very low correlation between them (R \approx 0), clearly indicating the true complementarity of the two fragmentation techniques.

Figure 3.4. Amino acid pair preferences for ECD (left pane) and CAD (right panel) for C-terminal fragments. Columns correspond to residue B in an amino acid pair A-B.
Electron Transfer Dissociation

Although ECD is highly useful for proteomic analysis of polypeptide mixtures, the technique is currently limited to FTICR MS instruments. Attempts have successfully been made to implement ECD fragmentation in ion-traps [63,64], but the efficiency is quite low and therefore not suitable for low-abundant sample analysis as encountered in proteome studies.

Recently Hunt and co-workers introduced a fragmentation technique identical to ECD in a linear ion-trap, called electron transfer dissociation (ETD) [65]. The similarities between ECD and ETD are quite striking. Both techniques are based upon capture of an electron, which induces cleavage of the N-C backbone and results in c' and radical z· ions. The major difference between the techniques though is the source of electrons. In ECD free electrons are produced through an electron gun, a BaO2 filament that irradiates electrons. ETD on the other hand is based upon electron abstraction from anions prior to electron capture – an electron ‘transfer’. Anions are mixed in the linear ion trap with polypeptide cations, with help of sophisticated electronics, and since the cations have a higher electron affinity than the anions an electron is transferred between the two species and fragmentation occur.

Since ECD is based upon capture of a free electron the technique can take full advantage of the energy induced. ETD on the other hand has to initially use some energy for electron abstraction prior to electron capture. The overall internal energy deposited during the actual electron capture part of ETD is therefore somewhat lower compared to ECD. This energy difference is especially obvious when looking at fragmentation efficiency of doubly-charged polypeptides in both ETD and ECD. These doubly-charged species require higher internal energy deposition in order to fragment, and ETD is therefore not as efficient as ECD for fragmentation of doubly-charged ions.

For polypeptides of charges >2+ ETD seems to have higher efficiency compared to ECD. This is most probable caused by a much better overlap between cations and anions in ETD in contrast to the overlap between cations and electrons in ECD. Another difference between the two techniques is that ETD currently is only available in low-resolution instruments while ECD is only available in high-resolution instruments.
4. Protein and peptide identification

Mass spectrometry has emerged as the primary tool for characterization of proteins in complex mixtures through polypeptide sequencing. The coupling of liquid chromatography to mass spectrometry (LC-MS) was pioneered in 1987 [66,67], and has since become the method of choice for proteomic analysis. In this growing field a number of technical advances have been made over the past years, resulting in an increasingly robust and productive platform. Identification of peptides in LC-MS experiments, and therefore the protein of origin, is predominantly performed by what is referred to as the ‘bottom-up’ approach [68,69]. A protein mixture is digested using specific proteases and loaded onto a RP column separating the polypeptide mixture according to hydrophobicity, hereby reducing the overall complexity of the mixture. Fragment ion spectra of eluted polypeptides are recorded using data-dependent acquisition (DDA), where the most intense peaks in the survey scan are selected automatically, isolated, and fragmented by MS/MS [70]. The ‘bottom-up’ strategy has been very successful, as it allows for identification of thousands of proteins in complex mixtures in one run [71-75].

Peptide identification through database search engines

The cornerstone of protein identification in proteomics are search engines such as Mascot [76] and Sequest [77], which allow for automatic matching of acquired peptide MS/MS spectra against protein sequence databases. Experimental MS/MS spectra are matched against calculated MS/MS spectra for all ‘in silico’ digested protein sequences in the database [77]. The search engine then assigns to each spectrum a score indicating how well the experimental data complies with the calculated one. The assigned score is usually based upon the probability that the identification is a chance event, and in Mascot search engine derived through the formula $M = -10\log(p)$, where $p$ is the chance event probability. Higher score therefore means increased confidence in the identification, and this approach helps to discriminate between correct and incorrect peptide assignments to spectra [76].

The problem of using this database searching strategy is that the distribution of assigned scores usually is bimodal [23]. The low scoring part of this
distribution is due to false identifications (figure 4.1, dashed line) while the higher score distribution is due to correct identifications (figure 4.1 solid line). However, since the two distributions are rarely fully separated, the common approach to separate correctly and incorrectly assigned peptide sequences from each other is to apply a cut-off threshold at a certain score value (Figure 4.1, dotted line). This threshold value is chosen so that all identifications above the threshold are regarded correct with 95% certainty, while everything below the threshold is discarded. The problem with this strict division is that many correct identifications are rejected at high thresholds (the so-called false-negatives), whereas lowering the threshold increases the risk of accepting high scoring incorrect identifications (false-positives, figure 4.1, grey area). The calculated cut-off threshold is based upon the size of the database being interrogated and the stringency of the search parameter applied to the dataset.

The presence of false-positive identifications remains today one of the challenging problems in proteomics [23,78,79]. Several statistical models have been employed in an attempt to decrease the amount of false-positives and improve the search engine identification score [80-82]. A common feature of these approaches is that they try to improve the validity of the assigned score, whereas a full separation of the bimodal distribution (figure

![Figure 4.1. Theoretical distribution of peptide Mascot score values. High scoring distribution (solid line) corresponds to correct identifications, while low scoring distribution (dashed line) corresponds to false identifications. Dotted line is calculated for 95% confidence level. Grey area indicates false-positive identifications.](image-url)
4.1) is required in order to achieve complete division between correct and
incorrect peptide assignments. This can only be achieved by an improvement
of the underlying mass spectrometric data due to the statistical nature of the
distribution. Recently, attempts to improve separation of correct and incor-
rect assignments have been made by using two consecutive stages of tandem
mass spectrometry (MS/MS/MS) showing increased analysis specificity
[83]. Preprocessing of acquired data by deisotoping and charge state decon-
volution has also shown specificity improvements [84], but these procedures
have not significantly improved the quality of the acquired data.

False-positive estimation using decoy database searching

In recent years methods have been developed that use ‘decoy databases’
to empirically assess the false-positive rate associated with large-scale pro-
teomics experiments [79,85]. The principle of these databases is to search
the acquired data against a forward (normal) database, and a reversed (‘non-
sense’) database, which consists of the same proteins as the normal database
but in reversed sequence order. The false-positive rate in a dataset can
hereby be assessed by taking the number of ‘identifications’ from the re-
versed database and dividing it by the number of hits from the forward data-
base.

This approach has been very successful [86], but the question governing
the decoy-database method is whether the determined false-positive rate
might be over- or under-estimating the true rate of false identifications. Rea-
sons for this question are e.g. that since trypsin cleaves after both lysine and
arginine, reversing a database will change molecular masses of many tryptic
peptides. Additionally, for approximately half of the tryptic peptides rever-
sing the database does not destroy peptide mass or sequence tags [87], and
hereby b ion series could be mistaken for \( [y - H_2O] \) ion series. Other prob-
lems are palindromic sequences [88,89] present in databases as well as in-
versed sequence similarities between proteins [90], which interfere with the
false-positive estimation. Furthermore, studies upon recently introduced
definition homeometric peptides [91] have shown that low-resolution mass
spectrometry produces a 30% risk that an arbitrary peptide of length 10 has a
homeometric peptide in the database. For high-resolution instrument the
same risk is only ~1%, but still this percentage is in the vicinity of the false-
positive rates reported in large-scale proteomics experiments.

In general it is widely believed that the decoy database approach works
well on peptides derived from any database using tryptic specificity. Using
the same database with non-enzymatic specificity creates a much larger
search space in which the tryptic search space will be a sub-set. Due to ex-
tensive overlap and similarities between forward and decoy databases when
using non-enzymatic specificity, the decoy database approach will yield a
misleading picture of the false-positive rate (vide supra). Since the tryptic peptides are a sub-set of this non-enzymatic search space, there must exist a larger sub-set where the decoy database approach ceases to be useful. If such as sub-set do not exist, then the decoy database approach have to work either well, or not well on both the tryptic and non-enzymatic search spaces.

High-mass accuracy

The importance of high-mass accuracy for peptide identification and determination of PTMs has been known at least since 1983 [92]. Being able to measure peptide masses with high accuracy reduces the risk of false-positive identifications, but ensures no more than 1% of all peptide combinations will fall into the detected mass region [93]. Even when a 0.1 ppm mass accuracy is achieved, unambiguous identification of tryptic peptides based solely upon molecular mass is not possible [94]. Thus, the necessity of tandem mass spectrometry (MS/MS) for reliable peptide identification is evident and particularly high-mass accuracy MS/MS. Only measurements of fragment masses with high-mass accuracy decrease the rate of false-positive identifications in database searches [95,96].

Instruments, such as FTICR MS, are currently the only mass spectrometers available for achieving sub ppm mass accuracies (see table 4.1). This mass accuracy is achieved due to high temporal stability (sub-ppb/h field drift) and high spatial homogeneity (sub ppm range) of the superconducting magnet. Compared to other MS instruments, which generally rely on alternating electrostatic fields with a temporal stability in the order of sup-ppm/h and low electrical field homogeneity, the mass accuracy of FTMS is unsurpassed. Recently a magnet-free FTMS called the LTQ-Orbitrap became commercially available [97]. This instrument is capable of achieving sub ppm mass accuracies similar to ICR, but these mass accuracies are only made possible using internal calibrants [98].

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Resolution</th>
<th>Mass accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Ion Trap</td>
<td>2-15,000</td>
<td>100-300 ppm</td>
</tr>
<tr>
<td>Q-TOF</td>
<td>10-20,000</td>
<td>10-20 ppm</td>
</tr>
<tr>
<td>TOF/TOF</td>
<td>20-25,000</td>
<td>10-20 ppm</td>
</tr>
<tr>
<td>FTMS</td>
<td>50-500,000</td>
<td>1-2 ppm</td>
</tr>
</tbody>
</table>

Table 4.1. Comparison of performance characteristics for some of the tandem mass spectrometers used in proteomics.
Detection probability of tryptic peptides in FTMS

Signal response in mass spectrometry is vital for achieving successful identification of peptides. DDA selects the tallest peaks present in a survey scan and performs MS/MS. Improved signal response should therefore lead to improved peptide identification. Although proteolytic protein digests are supposed to contain equimolar amounts of peptides, the physicochemical properties such as length, mass, amino acid composition, basicity and hydrophobicity can be vastly different. The physicochemical properties of peptides are the primary reason for the vastly different protein sequence coverage obtained in bottom-up strategies from samples containing equal protein concentrations [99]. A number of studies have been performed to relate various physicochemical properties of peptides with signal response (detection probability) in ESI [100-103]. One of the important parameters determining the ESI signal response is hydrophobicity, with peptides having higher hydrophobicity yielding higher signal response compared to peptides with lower hydrophobicity, and therefore providing a better detection probability for the former species.

In paper I we investigated the influence of various physicochemical parameters of tryptic peptides on their detection probability in LC-FTMS. Conclusions were made that physicochemical properties of tryptic peptides indeed affect detection probability in LC-FTMS, and as expected hydrophobic peptides were found to be more readily detected due to a higher signal response produced in ESI. An anticorrelation was found between the pI values of peptides and their detection probability, which was explained through the peculiarities of the polypeptide charging in ESI and the non-destructive image-current detection in FTMS (equation 3.5). Acidic tryptic peptides tend to have higher masses than basic tryptic peptides, and therefore appear in higher charge states, which results in higher ICR signal response.

Although it seems quite logical that increased signal response in MS should result in a higher detection probability, no investigation has yet been made to fully confirm this. A simple way to prove this hypothesis is to look at the correlation between physicochemical properties of peptides and Mascot assigned peptide score. Figure 4.2A shows the correlation between peptide Mascot scores and peptide retention time (hydrophobicity) of 1267 tryptic peptides with masses between 1000-1100 Da. The analysis shows a very weak correlation between retention time (hydrophobicity) and Mascot score (R = 0.12). The correlation factor though is strong enough to reject the null-hypothesis that there is no correlation [104]. Figure 4.2B shows the correlation between Mascot score and peptide size of ~11,000 doubly charged tryptic peptides. A stronger correlation (R = 0.53) is observed with higher peptide masses yielding higher Mascot score. Larger peptides tend to have higher hydrophobicity compared to shorter peptides, but hydrophobicity of similar sized peptides showed a weaker correlation (figure 4.2A). This result
shows that the influence of peptide physicochemical properties upon mass spectrometric detection is not straightforward. Physicochemical properties definitely play an important role in peptide detection, and this information can be used to improve protein sequence coverage in LC-FTMS, which is vital for improved sequence verification and PTM mapping.

### Improving peptide and protein identification

High-mass accuracy is an important component in acquiring high quality data. Further improvement in spectrum specificity can be accomplished by using techniques that can yield complementary sequence information. Performing CAD and ECD upon the same peptide ion will yield such complementary sequence information. Implementing this into a proteomics approach and showing the advantages of using complementary fragmentation techniques for improving peptide identification was done in paper II. By comparison of fragment ions derived from ECD and CAD spectra of the same peptide ions, the informative ‘golden complementary pairs’ can be revealed [105]. These pairs not only verify the presence of fragment ions but additionally provide the direction of the sequence by distinguishing between
N- and C-terminal fragments. Being able to tell which fragment ions are real is highly advantageous since false-positive identifications can arise from the presence of misleading ions in the MS/MS spectra due to gas-phase reactions, stray ions, contaminants, and electronic noise. Additionally, being able to determine the direction of peptide sequences is important for polypeptide identification and de novo sequencing, since e.g. b-ions produced by CAD might be interpreted as \([y – H_2O]\) series giving rise to false-positive identifications. This is especially true when using reversed databases for determination of false-positive rates [79,85].

Paper II shows that extraction of the ‘golden complementary pairs’ derived from ECD and CAD greatly increases the analysis specificity. For a whole cell lysate an overall peptide identification confidence increase of >30 times is achieved. A significant increase of the average protein score, average number of peptides identifying a protein and total number of identified protein is also achieved (table 4.2). For comparison, the impact caused by deisotoping and charge deconvolution was investigated on the same dataset in order to be certain that the observed increases were indeed due to the complementary techniques used. Table 4.2 shows the results of this study (labeled CAD_is) along with the results from the regular use of only one fragmentation technique (CAD) and using complementary techniques. The results show that complementary fragmentation techniques in conjunction with high-mass accuracy LC-MS acquisition are very beneficial for protein identification.

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>CAD_is</th>
<th>Complementary Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average protein score</td>
<td>169</td>
<td>183 (+8%)</td>
<td>277 (+64%)</td>
</tr>
<tr>
<td>Average number of peptides</td>
<td>3.4</td>
<td>3.2 (-6%)</td>
<td>5.3 (+56%)</td>
</tr>
<tr>
<td>identifying a protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Mascot search time</td>
<td>459 s</td>
<td>402 s (-12%)</td>
<td>368 s (-20%)</td>
</tr>
<tr>
<td>Average number of identified</td>
<td>114</td>
<td>148 (+30%)</td>
<td>256 (+125%)</td>
</tr>
<tr>
<td>proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2. Results from protein identification achieved using complementary fragmentation techniques compared to deisotoped CAD and CAD.
S-score for earmarking MS/MS data

Although the Mascot score is one of the conventional validity measures used to confirm peptide identification by MS/MS data, it has a number of limitations. Due to the bimodal distribution in figure 4.1 a low Mascot score does not necessarily mean poor peptide sequence match, but rather insufficient quality of MS/MS data to secure significant statistical validation. In addition Mascot might assign to sequences an incorrect above-threshold score value simply because of a chance events, and by searching large databases invalid (false positives). Besides, the Mascot score does not fully utilize the benefit of combined use of complementary fragmentation techniques. A novel database-independent scoring method, referred to as S-score, was designed and published in paper III. The method is based upon the maximum length of the peptide sequence tag [87] derived from MS/MS spectra by combined use of CAD and ECD upon same peptide ions.

Reliable sequence tag

The reliable sequence tag (RST) is related to the S-score (paper III). As the name describes, the RST is the most reliable sequence-tag used in mass spectrometry and therefore the perfect starting point for de novo sequencing. As with the S-score a requirement for the RST is that fragment ions used to create the RST are elements of a ‘golden pair’. What separates the RST from the S-score is the requirement that fragments used also are elements of a ‘complementary pair’ and hence, a doubly confirmation (both an N-terminal and C-terminal fragment at the same cleavage site are present) of the masses is available. This double confirmation of fragment masses based upon com-

![Figure 4.3](https://via.placeholder.com/150)

Figure 4.3. (a) scatter plot showing the correlation between S-score values and Mascot score values for all spectra for which Mascot found a peptide sequence. (b) Mascot score distribution of all spectra with S-score > 7 for which Mascot found a peptide sequence. (c) scatter plot showing correlation between Mascot score values and the S-score values for which RST confirmed the Mascot suggested sequence.
Complementary fragmentation techniques makes the RST more than 98.6% reliable (paper III).

A scatter plot showing the correlation between Mascot score values and S-score values is depicted in figure 4.3A. A strong positive correlation ($R = 0.78$) is observed indicating that S-score can indeed be used as a quality measure of MS/MS spectra. A bifurcation is visibly detectable in figure 4.3A, with some high S-score data yielding low Mascot score values. Plotting the Mascot score distribution of all dta files with S-score $> 7$ showed a clear bimodal distribution (figure 4.3B). The lower part of this distribution is most probably due to wrong Mascot identifications either through novel sequences or modified peptide present in the sample, while the higher part is due to correct assignments. To test this hypothesis a scatter plot of all spectra in which RST complies with the Mascot suggested sequences is plotted in figure 4.3C. The correlation between S-score values and Mascot score values clearly improved ($R = 0.91$), and a bifurcation is no longer visible. This confirms that the S-score can be used as a means to discard poor-quality spectra prior to database searching and in combination with RST used to rescue false-negatives and filter away false-positive identifications.

Classification of MS/MS data using S-score

A study performed upon 53,105 MS/MS spectra acquired from full NCBInr search of an A431 human epidermoid carcinoma cell lysate, revealed that only 11% of all acquired data were of insufficient quality (S-score $< 2$). Using S-score as a quality measure, acquired MS/MS spectra can be divided into three quality classes; Class A being data with high S-score ($S \geq 2$) and significant Mascot identifications, comprising 22% of all acquired MS/MS spectra (figure 4.4 black bars). Class B data, which was by far the largest group, consisted of data with significant S-score ($S \geq 2$) but with no significant Mascot identification (figure 4.4 grey bars). Since this group of data (67% of total) showed large S-score properties, it most likely contains MS/MS spectra belonging to either novel polypeptide sequences not present in the database, or modified sequences. Finally as mentioned, Class C data (11%) consisted of poor-quality data with a low S-score ($S < 2$, white bars figure 4.4).

This result is in line with previous reports on efficiency rates in large-scale proteomics studies [106]. Typically 85-90% of all acquired MS/MS spectra do not get identified through standard database search, which previously has been credited to poor quality data. Using complementary fragmentation techniques it is possible to increase the success rate (22%), but still a large portion of acquired data is not being identified although the MS/MS spectra are of good quality (S-score $\geq 2$). The need for a proteomics-grade de novo sequencing algorithm to deal with sequences not present in databases,
as well as a novel software tool for unbiased PTM identification is therefore apparent. Having these tools at hand would make it possible to assess whether the large portion of Class B data indeed consists of novel and modified peptide sequences not found by a regular Mascot search approach.

Figure 4.4. Distribution of S-scores reveals three classes of data. Class A: Spectra with $S \geq 2$ and significantly identified peptide sequences by Mascot. Class B: Spectra with $S \geq 2$ but not significantly identified by Mascot. Class C: Spectra of insufficient quality with $S < 2$. 
5. Polypeptide *De Novo* sequencing

Protein identification by a standard MS proteomics approach requires that the sequence of either the protein or a highly homologous protein can be found in the interrogated database. However, of the estimated 40,000 protein families in nature, only approximately 20% are known and listed in databases. Thus sequencing peptides without searching a database, referred to as *de novo* sequencing, becomes the only possibility for retrieving a positive identification. Additionally, mutations and PTMs gives rise to false-positive identifications in a regular database search and their true identification is only achievable through *de novo* sequencing.

If the quality of the acquired data is sufficiently good, retrieval of amino acid sequence from the product ion spectrum is possible. Problems with this approach arise since it is not possible *per se* to differentiate between fragment ions derived from N-terminus or the C-terminus. One way to circumvent this problem is to partially label the C-terminal carboxyl group with $^{18}\text{O}$ [107-110]. As a result all signals from $\gamma$-ion fragments will hereby be split into doublets separated by either 2 or 4 Da (dependent on the degree of $^{18}\text{O}$ incorporation), and can easily be distinguished in the mass spectrum. Problems in this approach are that sample treatment is quite extensive as well as expensive, and additionally it is not particularly feasible in cell line analysis.

Several different commercial programs are available for *de novo* sequencing of MS/MS spectra, like PEAKS [111], Lutefisk [112,113], Sherenga [114], SeqMS [115], Compute-Q [116], PepNovo [117] and NovoHMM [118]. Most of these programs employ the so-called spectrum graph approach for generating sequence candidates, and can generate sequence suggestions on a rate comparable to one spectrum per second. The spectrum graph approach of *de novo* sequencing converts a MS/MS spectrum into a graph, where detected fragment ions make up ladders of the peptide. The approach then computes the optimal peptide by computing the optimal path through these ladders.

Unlike the spectrum graph model the PEAKS software works on the spectrum directly instead of converting it into a graph. The algorithm first generates a $\gamma$-ion matching score and a $b$-ion matching score at each mass value according to the peaks around it. If no peaks are present around a mass value, a penalty value is assigned. The software then calculates the amino
acid sequences that make up a sequence with the highest score by addition of the y-ion and b-ion matching scores. Another approach that has been used to de novo sequence product ion spectra is NovoHMM, which uses Hidden Markov Model (HMM). A HMM model is a statistical model where the system being modeled is assumed to be a so-called Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. NovoHMM carefully models the peptide sequence as the hidden parameter and the acquired spectrum as the observable parameter, and the standard algorithm can then be adopted to infer the peptide sequence from the spectrum. In general, all these software approaches try to improve the determination of peptide sequences from mass spectrometric data using various statistical tools.

As de novo sequencing is to be compared to the database searching approach, the overall efficiency and validity should be comparable (efficiency 5-15%, validity ≥95%). A recent comparison of three de novo sequencing algorithms performed on 29 known test peptides gave an average success rate of 30-35% in correctly identified amino acids [119]. The association of biomolecular resource facilities (ABRF) performed a study designed to evaluate the proficiency in sequencing unknown peptides not present in any published database. The results from 48 laboratories showed a similar success rate of 40%, meaning that from the five peptides used in the study, laboratories were on average able to correctly identify 40% of the amino acids present in these peptides [120]. The main finding of the ABRF study was that using high-mass accuracy mass spectrometry as well as complementary approaches is vital for de novo sequencing.

Proteomics-grade de novo sequencing approach

In paper IV an approach that combines hardware and software improvements to achieve a validity level corresponding to >95% correct de novo assignments of peptide sequences was presented. Here the problems of incomplete backbone fragmentation, as well as the frequent overlap of fragment masses previously associated with de novo approaches are alleviated by the use of complementary fragmentation techniques ECD and CAD in combination with high-accuracy FTMS.

The method is based upon RST as a reliable backbone for de novo sequencing. With its high reliability it is natural to use data from RST as the initial step in creating peptide sequences. Once this reliable peptide ‘backbone’ is determined the software adds the next layer of information, which in step two are fragment masses that are part of a ‘golden pair’ but not part of a ‘complementary pair’, and therefore regarded as less reliable. But since the de novo peptide sequencing is based upon the highly reliable RST, adding
the ‘golden pair’ masses doesn’t affect the over-all reliability too much. If the full sequence has not been sequenced after step two additional fragment masses are added. Overall, the proteomics-grade de novo sequencing algorithm ensures a ≥95% confidence in identified peptide sequences, which is comparable to the confidence achieved by regular database searches.

The main reasons for achieving this high confidence level are the complementary fragmentation techniques used, but more importantly the high-mass accuracy achieved by the mass spectrometer. In order to retrieve reliable complementary fragments from CAD and ECD a mass accuracy of <20 mDa is needed. The importance of mass accuracy in de novo sequencing is even more pronounced, as shown in figure 5.1. By comparison between Mascot identified and de novo suggested sequences the effect of mass accuracy on the false positive rate could be estimated (paper IV). The false-positive rate increases dramatically from 4% to 14% when the mass accuracy requirement for fragment ions is relaxed from 0.02 Da to 0.1 Da. This illustrates that mass accuracies better than 0.04 Da is required in order to reliably perform proteomics-grade de novo sequencing, and that only high-resolution mass spectrometers are suitable for this task at the same confidence and sensitivity level as a regular database search. This is well in-line with the observations and conclusion of the ABRF study.

Figure 5.1. Relation between mass accuracy and percentage of false-positives in Proteomics-grade de novo sequencing.
Rescuing false-negative identifications

The ability of RST to rescue false-negatives and being a backbone sequence tag for de novo sequencing was tested on the A431 cell lysate sample. Searching the full NCBI database with tryptic specificity yields an significant M-score threshold of 42 (≥ 95% confidence). The complete M-score distribution of all Mascot suggested polypeptide sequences is depicted in figure 5.2 (grey distribution). RST was able to rescue 1871 polypeptide sequences, initially assigned a below threshold M-score by Mascot search engine. This number corresponds to an increase of ‘Class A’ data by 15%, and increases the overall efficiency to 27% (figure 5.2, black distribution).

Applying proteomics-grade de novo sequencing to the same A431 cell lysate further identified 2278 sequences not initially found by a standard Mascot search, increasing the amount of assigned MS/MS data to 31%, (Class A data). Still many good quality MS/MS spectra remained unexplained, and these results indicate that a large portion of Class B data could be due to extensive modifications of polypeptides.

In identifying novel polypeptide sequences through de novo approaches, caution has to be made when trying to identify modified sequences. An over-

![Figure 5.2. Distribution of Mascot score values for all spectra assigned peptide sequence by Mascot from A431 cell lysate (grey bars) and spectra significantly identified by Mascot and rescued by RST (black bars).](image)
lap between novel peptide sequences and modified sequences is highly possible, since some PTMs simply convert one amino acid into another, like e.g. deamidation (Asparagine into Aspartic acid) and methylation (Glutamic acid into Aspartic acid). Furthermore, ubiquitination adds two glycine residues to a backbone lysine amino acid during tryptic cleavage, and could potentially be regarded as backbone glycine residues during de novo sequencing.

**De Novo sequencing of endogenous peptides**

Rescuing false-negative database identifications becomes even more important when analyzing endogenous peptides, often referred to as peptidomics [121]. Although this definition includes all peptides in the system, peptidomics studies have mostly been focused on the analysis of bioactive endogenous peptides. Prior to the development of MS approaches, identification of these peptides were performed using Edman degradation [22], but with the emergence of MS/MS technology MS-based approaches for poly-peptide identification has largely replaced Edman degradation. This approach is similar to the proteomics-based protein identification, where MS/MS spectra are searched in a database containing all known protein/peptide sequences. One problem with this approach is the lack of known enzyme specificity that governs endogenous peptides, which increases the risk for false-positive identifications through a standard database search approach. Harsher restrictions (higher Mascot score threshold) must therefore be applied, compared to a standard tryptic data search, in order to achieve significant identification. In doing so the identification success rate decreases dramatically and drops well below the 5-15% that usually is achieved from a regular tryptic proteomic experiment [23]. Applying proteomics-grade de novo sequencing approach to human saliva samples should therefore be able to significantly improve the number of peptides identified.

Human saliva contains a large number of proteins and peptides that help maintain homeostasis in the oral cavity. These salivary components play a crucial role in digestion, lubrication and formation of a pellicle that coats and protects teeth and other oral surfaces, hence human saliva is a potential source of novel diagnostic markers, therapeutic agents as well as antibiotics [122-124]. One such group of antibiotics is the low molecular weight compounds referred to as cationic peptides, an important group of endogenous peptides. These peptides exhibit antimicrobial activity and are able to kill a broad variety of gram-negative and gram-positive bacteria, as well as fungi and certain viruses [125]. The continuous use of antibiotics around the world has resulted in multi-resistant bacterial strains, and consequently there is an urgent need for identification of alternatives to the commonly used antibiotics – thus cationic peptides are good candidates for such a new generation of antibiotics to treat these resistant bacteria.
The result of this human saliva peptide analysis is depicted in figure 5.3. A total of 207 unique endogenous peptide sequences (< 3 kDa) were identified (figure 5.3, black bars). 77 of the sequences (37%) were originally identified with a below threshold score (M-score > 62) by Mascot search engine and would therefore have been discarded. These below-threshold peptides could be fully validated and thereby rescued using RST confirmation, increasing the number of identified peptides by 60%. Adding to this information, a total of 44 endogenous polypeptide sequences were completely sequenced by proteomics-grade de novo approach. Combining these de novo sequenced peptide sequences with the rescued RST sequences resulted in a 93% increase of significantly identified endogenous peptides, hereby increasing the efficiency from 7.7% to 15%.

Proteomics-grade de novo sequencing shows a great potential in peptidomics, both as a compliment to the standard database searching approach where false-negative identifications can be rescued, but also as a potential for discovering novel sequences. This could particularly be useful in e.g. analysis of the low-molecular-weight (LMW) range of the circulatory proteome [126]. It is believed that the LMW ‘peptidome’ of the serum proteome could be a rich source of cancer-specific biomarkers, since tissue proteins

Figure 5.3. Distribution of Mascot score values for all Human saliva spectra with a endogenous peptide sequence assigned by Mascot (grey bars), and spectra significantly identified by Mascot, RST and de novo sequencing combined (black bars).
that are too large to passively diffuse into the circulation could still be presented as fragments of the parent molecule (endogenous peptides) [127].
6. Post-translational modifications

PTMs of proteins play a critical role in the regulation and function of many known biological processes [128,129]. Proteins can be post-translationally modified in many ways; protein chains can be cleaved, many different chemical groups being attached to them (e.g. phosphorylation, acetylation, methylation, glycosylation and ubiquitination), and finally, proteins can be internally or externally cross-linked by e.g. disulfide bonds.

The extent of modifications is very important for shotgun proteomics which is based on identification of peptides derived from enzymatically digested complex protein mixtures [130]. The shotgun approach is known to face the so-called dynamic range challenge arising from the fact that concentrations of proteins in whole proteomes or complex mixtures such as blood plasma differ by many orders of magnitude [131]. The challenge therefore is to detect low-abundance peptides in the presence of much more abundant competitors [132], but extensive modifications can exacerbate this problem significantly. As traditional database search approaches allow for only few modifications to be included in the search, protein identification in shotgun proteomics is largely based on the detection of unmodified polypeptides. Abundant modifications could reduce the detection probability, as modified peptides from abundant proteins might mask peptides from low-abundant proteins. This could result in an elevated rate of false positive identifications and/or affect the reproducibility of the proteome analysis, which is already rather poor [19]. Full sample analysis with identification of all acquired MS/MS spectra is therefore necessary in order to achieve complete and significantly correct characterization of an entire proteome.

Furthermore, identification of all PTMs present in a sample is highly relevant from the biological point of view. Traditionally PTM studies have followed the thought that one modification site corresponds to one regulatory function, and therefore focused on specific amino acid residues or a small number of residues in a specific protein. However, instead of single-site modifications being responsible for protein activity and stability, it is now evident that proteins can be modified at several sites functioning together – a phenomenon referred to as multisite modifications [133]. One such modification that is widely recognized as being responsible for regulation of protein functional activity and involved in multisite signaling is phosphorylation [134]. Phosphorylation affects an estimated one-third of all proteins and is the most widely studied PTM [135]. The importance though,
of other key modifications involved in signaling processes has only been appreciated relatively recently such as methylation of arginines [136,137], ubiquitination [138], SUMOylation [139], acetylation [140] and O-GlcNAc glycosylation [141] and they have been studied much less by proteomics. Moreover, in phosphoprotein analysis phosphorylated proteins are predominantly present in substoichiometric amounts in the cell compared to their non-phosphorylated counterparts. Therefore, it is advantageous to perform an up-front enrichment of phosphopeptides prior to MS analysis. The most successful strategies for achieving this enrichment have been the use of metal affinity- and antibody-based methods [86,142-144]. Although this enrichment approach has been very successful in identifying thousands of regulatory phosphorylations sites in a single experiment, it is by design highly biased towards detection of phosphorylation sites [145,146]. In order to achieve a thorough understanding of cellular signaling, identification of all components in the entire signaling network is required, including all proteins and PTMs. Information regarding regulatory multisite modifications other than phosphorylation will be lost during phospho-enrichment strategies. Although combinations of non-phospho modifications together with phosphorylations on the same tryptic polypeptide chain will be enriched by phospho-specific treatment, identification by MS will still not be achieved by phosphoenrichment-based strategies simply because the bias of the analysis is towards phosphorylation only.

The extent of modifications in the same peptide has recently been shown by our group with a tryptic polypeptide derived from human α-casein protein [147]. Analysis of the tryptic peptide MESSISSSSEEMSLSK showed that at least 12 different modified isoforms existed in the analyzed shotgun sample (figure 6.1), arising from different types of in vivo and in vitro modifications including phosphorylation. These modified isoforms produced similar signal response and eluted of the chromatographic column at vastly different RT values, illustrating that sample complexity is tremendously enhanced due to modifications.

As a result, an unbiased approach to identify the full extent and functional importance of all protein modifications involved in cell signalling needs to be realized. Such an approach may also provide stochiometric information to phosphoproteome experiments, a dimension that is missing so far.
With the development of the ModifiComb approach, a fast and reliable method for identifying hundreds different types of PTMs at a time was published in paper V. The ModifiComb program analyzes MS/MS spectra and assigns to base peptides (unmodified, known peptide sequences) so called dependent peptides (modified versions of the base peptide). The software then creates a ΔM histogram that contains information on all modifications present in the analyzed sample above a certain level. Using this unbiased approach unexpected and novel modifications are revealed by ModifiComb in a straightforward manner, as presented in paper V where a novel +12,000 Da modification of proline residues is identified. This result shows that ModifiComb could be a valuable tool in detection of e.g. novel modifications involved in signal transduction.

Other scientific laboratories have published programs similar to ModifiComb [148,149]. These approaches do not use the benefits of neither high-mass accuracy MS (<20 mDa) nor complementary fragmentation techniques.
Having high-mass accuracy makes differentiation between modifications with similar masses possible, illustrated by CO₂, N₂ and C₂H₄ which all have the same integer mass. The same differentiation between tri-methylation and acetylation is not possible with lower resolution instruments. Both MS-alignment and VEMS are generating excellent results, but due to analysis speed requirements they are limited in their ΔM region of possible modifications. ModifiComb on the other hand is fast, uses unlimited ΔM values and is capable of distinguishing between modifications with similar masses (figure 6.2). This is illustrated through the peak denoted by ‘G+H’ which is a doublet, shown as an inset around +28 Da. The lighter peak G is mainly due to + CO contribution (formylation), whereas the heavier peak H is due to + C₂H₄ (ethylation or di-methylation). The most abundant modifications observed are listed below.

<table>
<thead>
<tr>
<th>Mass</th>
<th>Comp.</th>
<th>Mass</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A  -48.0033</td>
<td>- CH₄S</td>
<td>F   15.9949</td>
<td>+ O</td>
</tr>
<tr>
<td>B  -17.0265</td>
<td>- NH₃</td>
<td>G   27.9949</td>
<td>+ CO</td>
</tr>
<tr>
<td>C  -2.0156</td>
<td>- H₂</td>
<td>H   28.0310</td>
<td>+ C₂H₄</td>
</tr>
<tr>
<td>D   0.9840</td>
<td>- NH + O</td>
<td>I   31.9898</td>
<td>+ O₂</td>
</tr>
<tr>
<td>E   14.0156</td>
<td>+ CH₂</td>
<td>J   57.0214</td>
<td>+ C₃H₇NO</td>
</tr>
</tbody>
</table>
Extent of modifications in proteomics experiments

The extent of modifications in a proteome sample is of significant importance not only for biology, but also for analytical sciences. In paper VI a study was undertaken to quantify the extent of modifications present in samples within the dynamic range of 1:100 using ModifiComb approach. Results from analysis of highly abundant proteins yielded approximately 1 modification per amino acid, a much higher number than initially expected [16]. Several modifications are in vitro modifications due to sample treatment and preparation. In an extensive analysis of PTMs of individual proteins [150], the ratio between in vivo and in vitro modifications was approximately 1:3. Assuming this ratio to be typical, a typical shotgun proteome sample would yield 0.4 in vivo + in vitro modifications per amino acid. The observations made in paper VI estimate this number to be 2.5 times higher, but it is likely that the real number is significantly higher since e.g. glycosylations were not included in this study.

Glycosylations further increase the complexity problem of shotgun proteome samples, since the ‘glycome’ is several orders of magnitude more complex than the proteome. The glycome exceeds the complexity of the proteome as a result of the even greater diversity of the glycome’s constituent carbohydrates, and is further complicated by the sheer multiplicity of their possibilities of combination and interaction of the carbohydrates with each other and with proteins [151].

The results published in paper VI shows without a doubt that modifications are a significant source of sample complexity in shotgun proteomics analysis. Not taking all modifications into account during data analysis (unbiased analysis) will result in an elevated rate of false-positive identifications, and reduces the reproducibility and efficiency of proteome analysis.

This scenario is best illustrated through figure 6.3, where the theoretical distribution of unmodified tryptic polypeptide concentrations in a complex mixture is depicted (solid line). The distribution is assumed to be normal on a logarithmic scale [152]. Assuming that 10 modified isoforms exist for every tryptic polypeptide at substoichiometric range of 1:2 to 1:100 a distribution of modified polypeptides is obtained (dashed line). When measurements are performed with the dynamic range of 1:1000 (concentrations above 0.01 mg/mL), the number of unmodified polypeptides exceeds that of modified peptides. However, at concentrations down to $10^6$ mg/mL several modified polypeptide will be detected per each unmodified one, and in lower concentration regions the ratio rapidly grows in favor of modified peptides (exceeding 100 at concentration 1:109).
Overall efficiency

ModifiComb analysis of A431 cell lysate yielded totally 3640 dependent peptides, when no size limitations were set for detected PTMs. The aforementioned possible overlap between modified peptides and de novo sequenced peptides was investigated. The result is depicted in figure 6.4 as a Venn diagram. In total 433 spectra where commonly identified by both ModifiComb and de novo sequencing, corresponding to 8% of all identified MS/MS spectra by both ModifiComb and de novo sequencing. The result showed that ~20% of all de novo sequenced peptides were indeed due to modifications converting one amino acid into another. The biggest contribution to this conversion was deamidation.
In total 5495 spectra were identified as being either novel or modified sequences. Combining this result with the spectra assigned as being part of Class A data, a total of 19,588 spectra could be uniquely and significantly identified using the approaches presented in this thesis. This corresponds to an efficiency of 42% for all acquired MS/MS spectra with an S-score \( \geq 2 \), but still a large portion of acquired data needs to be identified.

A requirement for successful identification by ModifiComb is that base and dependent peptides have minimum four complementary fragments in common. Many spectra with lower S-scores \( (2 \leq S \leq 3) \) do not contain enough fragment masses to fulfill this requirement. This is illustrated by plotting the S-score distribution of all dependent peptides identified by ModifiComb (figure 6.5), which show that dependent peptides predominantly have S-scores \( \geq 4 \), as expected. Comparing this result to the S-score distribution in figure 4.4 of all Class B data, it is visible that there exist approximately 12,000 unidentified spectra with S-score \( 2 \leq S \leq 3 \). These spectra do not get identified by ModifiComb and they probably do not contain enough information for full de novo sequencing either. It is therefore likely to assume that only spectra with S-scores \( \geq 3 \) (or maybe even \( S \geq 4 \)) are of adequate quality for successful ModifiComb and full de novo sequencing analysis. Therefore, taking only spectra with S-score \( \geq 3 \) into account an efficiency of \( \sim 50\% \) is achieved, and for data with S-score \( \geq 4 \) the efficiency further increases to \( \sim 55\% \).

Currently a new version of ModifiComb is under development, which only requires 3 common peaks between base and dependent peptides for positive identification by including comparison of S-score values derived from the base and dependent peptide. This improvement will make it possible to dig deeper into the pool of unassigned MS/MS spectra and therefore
further improve the overall efficiency. The potential increase in sensitivity by lowering the required number of common fragments can be illustrated by looking at the number of identified phospho-peptides by ModifiComb. Due to the neutral loss of the labile phosphorylation upon CAD fragmentation, MS/MS spectra often lack backbone fragmentation, and very few complementary fragment ions are therefore detected between ECD and CAD of serine-/threonine phospho-peptides. Due to this ModifiComb rarely finds serine- or threonine-phosphorylated peptides in cell lysate samples with a requirement of 4 common fragments. Performing phosphorylation analysis upon the A431 cell lysate using the proprietary software PhosTShunter [153], a total of 80 serine-/threonine phosphorylated peptides were identified with very high significance. These were not identified by ModifiComb and this shows that a large portion of unidentified Class B data might still belong to modified sequences.

Further improvements in efficiency might also be achieved by searching unassigned data with $S$-score $\geq 2$ in a database containing expressed sequence tags (EST) [154,155]. Although this approach is very time consuming, searching spectra of not sufficient quality for de novo sequencing might identify sequences not present in current protein databases.

![Figure 6.5. Distribution of $S$-score values for all dependent peptides identified by ModifiComb.](image-url)
Concluding remarks

In the post-genomic era MS-based proteomics has emerged as the method of choice for peptide and protein identification. Being able to significantly identify which proteins are present in a given biological sample is of vital importance. This thesis presents, for the first time, the utility of complementary fragmentation techniques in conjunction with data-dependent LC-MS/MS on a FTICR MS, an instrument which allows for high-mass accuracy measurements. Using high-mass accuracy MS in combination with complementary fragmentation techniques CAD and ECD improves the validity of the interrogated peptide mixture through the use of 'golden complementary pairs', which not only verify the presence of fragment ions but also identify the terminal direction of these. This ion filtering not only improves the identification significance but also improves the overall speed of database searching and increases the average number of peptides used for protein identification (paper II). The use of complementary fragmentation techniques further makes it possible to validate the spectrum quality prior to database searching, and can be used as backbone identification for proteomics-grade de novo sequencing (paper IV). This RST backbone is not only useful as a measure for de novo sequencing but also as a complement to standard database searching, rescuing previously discarded false-negative identifications (paper III).

The complementary fragmentation techniques CAD and ECD are now routinely available on commercial FTMS instruments, and some ion traps provide a combination of CAD and ETD. Therefore, the combined use of these techniques will continue to increase as the efficiency and robustness also increases. Furthermore, the need for de novo sequencing in proteomics experiments will increase as reliable and efficient approaches become available. Therefore, the current shortcomings of CAD-alone approach simply require an improvement in identification confidence, which can only be achieved by combined use of complementary fragmentation techniques in conjunction with high-mass accuracy instruments.

Modifications are a significant source of sample complexity in shotgun proteomics experiments (paper VI), increasing the risk of false-positive identifications. This risk is further elevated if not all possible modifications are
included in the database search. The novel software tool ModifiComb (paper V) can be used for a database-independent search and reveals thousands of unique modifications present in a single cell lysate. Performing unbiased PTM analysis not only yields any type of modification present in the sample being analyzed, it also reveals occupancy rates [147] of identified PTMs, information that currently is lost during enrichment strategies.

Applying all of the above methods to a single human cell lysates increases the efficiency (percentage of identified MS/MS spectra) of the whole proteomics experiment from ~10% to an unsurpassed 42% (of all MS/MS spectra with S-score ≥ 2). Through this process thousands of novel sequences are reported using proteomics-grade de novo sequencing and even a larger number of MS/MS spectra are reported as being modified species (vide supra). Improvements to the presented techniques are still possible, as e.g. the threshold requirements for ModifiComb can be lowered from four common fragments to three by software improvements. Furthermore, our recently reported hydrogen losses from radical z ions in ECD [46], as well as informative w and u ion series [156] can be implemented into the methods presented in this thesis. This should increase the sensitivity of the presented methods, and hereby improve the overall efficiency significantly, and will provide a mean, to dig deeper into the analyzed proteome.
Acknowledgements

This thesis concludes four years of graduate student training, and many people have been involved in this project during these years. People whom I am tremendous grateful to:

A very special and heartfelt thank you to my supervisor Roman Zubarev, who has shared his enormous knowledge in mass spectrometry with me and allowed me to do this work in his laboratory. Always encouraging and supporting me, and never to busy to help in any given situation. Words can not describe my gratitude.

СПАСИБО ЗА ВСЁ, РОМАН.

Also a special thanks to current and former member of ’the Roman empire’ at BMMS: Misha Savitski, Frank Kjeldsen, Chris Adams, Oleg Silivra, Sasha Misharin and Thomas Köcher, for making days at the office so very special for me. For great collaborations, fruitful discussions, fun, friday beers and simply just being absolutely the greatest colleagues possible.

Everyone else at BMMS: Per Andrén, Anna Nilsson, Marcus Svensson, Maria Fälth and Karl Sköld, for enjoying your company and discussions in the coffee room. For your help and assistance whenever we needed some parts or guidance. A special thanks also goes out to the secretaries involved: Paula Almqvist, Åsa Hammarström and Erica Johansson, without your help and support I would never have been able to manage through the bureaucratic paper-jungle that comes with being a graduate student.

I would also like to send a thank you til Jesper V. Olsen for his help and his great friendship. Always available whenever I had a question or needed some small tips and tricks. I am definitely looking forward to be working with you again.

A very warm and special thanks goes out to my parents Marianne and Poul, as well as my sister Mette. Thank you so much for your everlasting support and love, for always pointing me in the right direction whenever I needed advise and for always believing in me.
Finally, and most importantly, special thanks to my beloved Rita, without you this journey would never have been so great. Thank you so much for everything that you have given me over the years, for making everyday so colorful. You mean everything to me.
Karakterisering av polypeptider med tandem masspektrometri genom användning av komplementära fragmenterings tekniker


I den här avhandlingen har det åstadkommts en stor förbättring i identifikation av masspektra, tack vare kombinationen av hög massnoggrannhet och komplementära fragmenterings tekniker. Användning av kollisoin inducerad dissociation (CAD) och elektron fångst dissociation (ECD) på varje peptid kan åstadkomma detta. Tillämpning av den här metodiken i standard proteomik experiment och den medföljande förbättringen är visade. Ett nytt databas oberoende betyg (S-score) är införd för att separera dåliga och bra spektra. Också en pålitlig sekvens-tag är utvecklad som gör det möjligt att rädda spectra med låga sökmotor betyg och fungera som bas för de novo sekveneringen.

En ny de novo algoritm är utvecklad som för första gången kan sekvenera peptider med hög (>95%) tillförlitlighet. Ett annat mjukvara redskap är utvecklad, Modificomb som kan identifiera alla modifieringar oavsett om dem är kända eller okända. Kombinationen av alla dessa verktyg förbättrar tillförlitligheten av peptid identifikation 30 gånger och leder till att 40% av all MS/MS kan identifieras. Detta är den högsta siffran uppnådd i proteomiken så här långt.
References


139. Johnson, E. S. Protein modification by SUMO. Annu. Rev. Biochem., 2004. 73, 355-382.


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 252

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology*. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)