Methods for Mass Spectrometric Proteome Analysis

Elena Ossipova
Faculty of Natural Resources and Agricultural Sciences
Department of Chemistry
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2008

Acta Universitatis Agriculturae Sueciae

2008: 9

ISSN 1652-6880
ISBN 978-91-85913-42-8

Tryck: SLU Service/Repro, Uppsala 2008

Abstract

Ossipova, E., 2008. Methods for Mass Spectrometric Proteome Analysis.
Doctor's dissertation.
ISSN 1652-6880, ISBN 978-91-85913-42-8

The major goal of proteome analysis is structure determination, identification, estimation of expression level, and understanding of the role of any protein in an organism. In combination with genomics, proteomics can provide a holistic understanding of the biological processes occurring in any organism. Mass spectrometry-based proteome analysis typically utilizes mass spectra of peptides of digested proteins together with sequence collection searching for rapid and accurate identification of proteins. Successful proteome analysis requires good experimental design, high quality data and optimized search conditions for protein identification.

A mass spectrometry-based method for differential detection and identification of proteins in protein mixtures utilizing multivariate methods was developed. The method utilizes intensity values from matrix assisted laser desorption/ionization time-of-flight mass spectra of tryptically digested protein mixtures for the label-free identification of a protein present in different concentrations in two samples. The Probity algorithm, which assigns the statistical significance to each identification result, was applied for the protein identification.

A systematic study of the quality of peptide mass fingerprint based (PMF) protein identifications under different search constraints was performed. 2244 PMFs from 2-dimensional gel electrophoreses separated human blood plasma proteins were submitted to the Probity algorithm for protein identification under different search conditions. The number of significantly identified proteins was counted for each condition in order to find the best set of search constraints for successful outcome.

A study of how the quality of proteolytic peptide identification can be improved by optimizing the information content of tandem mass spectra and by optimizing the search constraints of the sequence collection searching was done. The X! Tandem algorithm was employed for identification of proteolytic peptides from mouse proteins. The influence of the mass accuracy of both precursor and fragment mass ions, the number of sequences included in the search, and the number of missed proteolytic cleavage sites on the number of identified peptides was explored.

Computer simulations were performed in order to investigate quantitatively the information content in tandem mass spectra of proteolytic peptides, required to identify peptides and their post-translational modification.

Keywords: Mass Spectrometry, Proteome analysis, Protein Identification, Differential Proteomics, Bioinformatics.

Author's address: Elena Ossipova, Department of Chemistry, SLU, P.O. Box 7015, SE-750 07, Uppsala, Sweden. E-mail: Elena.Ossipova@kemi.slu.se

Science may set limits to knowledge,
but should not set limits to imagination.

Bertrand Russel

To my parents

Contents

Appendix

Papers I-IV

This thesis is based on the following papers, which will be referred to by their Roman numerals.

I. Ossipova, E., Nord, L.I., Kenne, L. & Eriksson, J. 2004. Method for differential detection and identification of components in protein mixtures analyzed by matrix-assisted laser desorption/ionization time-of –flight mass spectrometry. *Rapid Communications in Mass Spectrometry* 18, 2053-2058.

II. Ossipova, E., Fenyö, D. & Eriksson, J. 2006. Optimizing search conditions for the mass fingerprint-based identification of proteins. *Proteomics* 6, 2079-2085.

III. Ossipova, E., Zhang, G., Neubert, T.A., Fenyö, D. & Eriksson, J. 2008. Improving mass spectrometry based peptide identification by optimizing the conditions for sequence collection searching. *Submitted to Journal of Proteome Research.*

IV. Fenyö, D., Ossipova, E. & Eriksson J. 2008. The peptide fragment mass information required to identify peptides and their post-translational modifications. *Manuscript.*

Reprints were made with permission from the publisher (John Wiley & Sons, Ltd.).

Abbreviations

α-CHCA	α-Cyano-4-hydroxy cinnamic acid
2D-GE	2-Dimensional gel electrophoresis
CID	Collision-induced dissociation
Da	Dalton
DHB	2,5-Dihydroxybenzoic acid
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
DTT	Ditiothreitol
ECD	Electron capture dissociation
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
FT-ICR	Fourier transform-ion cyclotron resonance
HPLC	High performance liquid chromatography
HAS	Human serum albumin
IEF	Isoelectric focusing
IG	Immunoglobulin
IPG	Immobilized pH gradient
MALDI	Matrix- assisted laser desorption/ionization
MOWSE	Molecular weight search
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
m/z	Mass-to-charge ratio
PAGE	Polyacrilamide gel electrophoresis
PC	Principal component
PCA	Principal components analysis
pI	Isoelectric point
PMF	Peptide mass fingerprint
Q	Quadrupole
QIT	Quadrupole ion trap
RNA	Ribonucleic acid
RNAse	Ribonuclease
RP	Reversed phase
SDS	Sodium dodecyl sulfate
SIMCA	Soft independent modeling of class analogy
TOF	Time-of-flight
UV	Ultraviolet

Introduction

All genetic information of any organism is stored in a genome and is encoded by the sequence of nucleotide subunits of deoxyribonucleic acid (DNA) (for some viruses, ribonucleic acid, RNA). DNA molecules contain coding and noncoding regions, and the coding regions contain the instructions required for synthesis of proteins. Regions of the DNA that can serve as a template for the formation of proteins are referred to as genes. A gene is transcribed into messenger RNA (mRNA) molecules, and proteins are synthesized according to the instructions obtained from mRNA in a process called translation.

The first genome sequencing projects in the late 1990s that yielded complete genomic sequences of the bacterium (*Haemophilus Influenzae*) [1], of yeast (*Saccharomyces cerevisiae*) [2], and of the nematode (*Caenorhabditis elegans*) [3] opened a new era of biology by demonstrating the utility of complete lists of gene products that could be present in an organism. Since then, genome sequencing of many organisms have been performed and new sequencing projects are undertaken at an increasing pace.

Proteomics is the study of structures and functions of all proteins in an organism. The goal of proteomics is global analysis of gene expression and function, which requires analytical methods to detect and quantify proteins in their modified and unmodified forms. The term “proteome” was introduced by Wilkins [4] in the middle of the 1990s as a protein complement expressed by a genome.

The proteome is a multiprotein organization in which every protein plays its own role in a larger system or network. A proteome represents all possible gene products and can exist in different forms that vary within a particular cell [5] or from cell to cell [6, 7] and most proteins can be found in several modified forms [8-11] in a wide range of abundances [12, 13]. Protein modifications can determine structure, location and function of each protein [14]. The challenge of proteomics is to detect and quantify proteins in their modified and unmodified forms. Proteomics studies can be useful for identification of peptide and protein biomarkers of disease [15, 16]. Biomarkers are molecules that indicate changes in biological processes and can be recognized or monitored. More specifically, a biomarker indicates a change in expression of a gene or state of a protein that correlates with the risk or progression of a disease. The great interest arises from the potential of biomarkers to provide earlier diagnosis and disease classification. Biomarkers have a potential to be used as a guideline in the choice of therapies, and reflect how well a treatment is working. Accurate proteome analysis is important for understanding many physiological processes occurring in an organism. Proteome analysis employs several methods, among them is the mass spectrometry (MS)-based protein identification, which attempts to identify proteins by matching mass information from proteolytically digested proteins with information of protein sequences derived from the genome. MS-based proteome analysis can potentially provide rapid and accurate identification of proteins in an organism. Profiling proteomics encompasses the description of the whole proteome of an organism (by analogy with the genome) and includes organelle mapping and differential measurement of expression levels between cells or conditions (Fig. 1).

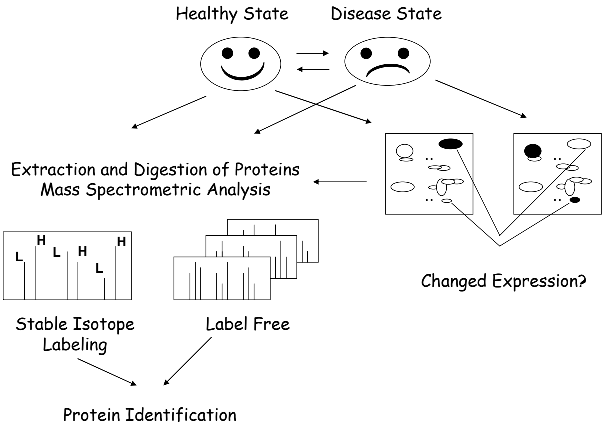

Figure 1. Mass spectrometry based proteome analysis utilizes several methods for measurement of gene expression level. i) Proteins from two samples can be separated by gel electrophoresis and the proteins that differ in amount will show the difference in spot density. Spots of interest are cut out and proteins are proteolytically digested prior to MS analysis followed by sequence collection searching. ii) Proteins from *one* sample are enriched or labelled with a heavy stable isotope, whereas proteins from another sample are treated so that the natural isotope abundance is retained. The two samples are mixed together and proteilytically digested, mass spectra of the peptide mixture are acquired, and the intensity values of unlabeled (L) and labelled (H) peaks are compared. iii) with label-free quantitation mass spectra are acquired separately from the two samples, intensity ratios at given m/z-values are compared, and mass values obtained from mass spectra are used for protein identification.

Objectives

A major objective of proteomics is to characterize in detail the entire proteome of any organism. The objective of this thesis is to develop and improve methods towards comprehensive proteome analysis, by new experimental and computational approaches that can bring the proteome analysis deeper inside of known cellular processes and help to discover unknown proteins and their functions.

Specific aims of this thesis

- Explore semi-quantitative information obtained by MALDI-TOF-MS followed by multivariate analysis of mass spectra for the identification of a protein present in different concentration in two samples of a protein mixture using the Probity algorithm.
- Evaluate the quality of proteolytic peptide mass finger print based protein identification results obtained under different search conditions using the Probity algorithm.

- Evaluate the quality of peptide fragment mass spectrometry based protein identification using the X! Tandem algorithm and investigate how the choice of data processing and search conditions impacts the successful outcome of proteome analysis.
- Investigate the information content of proteolytic peptide fragment mass spectra required to identify peptides and their post translational modifications.

Proteomics Tools

In the past few years, proteomics research has experienced a progress in experimental, instrumental, and computational approaches. Proteomics studies require protein separation and protein identification, and in many cases protein quantitation. Current proteomic technologies exploit *e.g.* gel electrophoresis, liquid chromatography, mass spectrometry and bioinformatics as tools in the proteome analysis.

Sample preparation

The study of any proteome begins with taking a biological sample: it can originate from a body fluid, a piece of tissue, cultured cells, and so on. Solid samples then need to be disrupted (pulverized, homogenized, sonicated) and contain a lot of cell components in a buffer solution. Before the analysis of proteins present in the sample can start a protein extraction is required. For proteomic analysis protein extraction with no or little contamination by other biomaterials (*e.g.* lipids, cellulose, nucleic acids, *etc.*) is desired. This can be done using detergents, reductants, denaturing agents, and enzymes.

Detergents, 3-([3-cholamidopropyl]dimetylammonio)-1-propane sulfonate (CHAPS), cholate) help to solubilize membrane proteins and separate them from lipids. Reductants (ditiothreitol (DTT), mercaptoetanol) reduce disulfide bonds and prevent protein oxidation. Denaturing agents (*e.g.*, urea) serve to shatter protein-protein interactions by changing the ionic strength and pH of the solution. Enzymes (*e.g.*, DNAse and RNAse) digest contaminating nucleic acids, carbohydrates (*e.g.*, glycosidases), and lipids (*e.g.*, lipases).

The resulting extract is a mixture that consists of proteins in different concentrations of varying molecular weight, solubility, and modifications. Before the mixture of proteins is subjected to mass spectrometric analysis it is usually necessary to reduce the complexity of the mixture by dividing it into fractions.

The most common techniques applied for protein mixture separation is 1D- or 2-D gel electrophoresis (1D-GE or 2D-GE). During the 1D-GE proteins are separated according to their molecular weight or isoelectric points (pI), the pH-value for which the protein has a zero net charge, while a high voltage is connected to the gel. Separation by isoelectric points uses an immobilized pH gradient (IPG) strip in which polycarboxylic acid ampholytes are immobilized on a support to create stable pH gradient [17, 18]. The strip is hydrated with a buffer and proteins are slowly loaded onto the strip under voltage. Then the voltage is increased to achieve focusing. The proteins are thus resolved into bands in order of their relative content of acidic and basic residues, whose value is represented by the isoelectric point.

Separation according to molecular weight is based on SDS-polyacrilamide gel electrophoresis (1D- or 2D-SDS-PAGE). The protein sample is dissolved in a loading buffer containing SDS. The SDS binds to proteins and imparts a negative charge in a proportion to the protein weight. The protein-SDS complexes migrate through the cross-linked polyacrylamide gel at rates based on their ability to penetrate the pores of the gel. The proteins are thus resolved into bands in order of molecular weight. The band are visualized by using different staining techniques, including e.g. silver, Coomassie, and fluorescent dyes. The degree of resolution achieved by 1D is moderate and bands may contain several proteins.

For highly complex protein mixtures 2D-SDS-PAGE can be a useful tool. This method is a combination of two different types of separation. In the first dimension, the proteins are separated according to their charge using isoelectric focusing (IEF). In the second step, focused proteins are resolved by electrophoresis on a polyacrylamide gel. Thus 2D-SDS-PAGE resolves proteins by isoelectric point and by molecular weight.

In proteomics applications, gel bands or gel spots of interest are cut out and subjected to digestion by enzymes with high digestion specificity. The goal of proteolytic digestion is to cleave proteins at certain amino acid residues to yield fragments that are suitable to MS analysis. For MS analysis protein fragments about 6-20 amino acids are desired. Trypsin is the most used protease in proteomic analysis. This enzyme digests the protein on the carboxyl side of arginine or lysine residues (except those followed by proline). The set of proteolytic peptides is unique for every protein, and hence, mass spectrometric analysis of proteolytic peptides provides a fingerprint of each protein [19].

Another method for reducing complexity of protein mixtures (*e.g*., a whole cell lysate or individual proteins) before MS analysis is proteolytic digestion of the proteins and separation of the entire proteolytic peptide mixture by multidimentional chromatographic technologies [20, 21]. This strategy involves proteolytic digestion of the proteins after their isolation from the cells or tissues. The tryptic digest of one protein yields on average 30-50 peptides, so a tryptic digest of a proteome with 5000-proteins can yield 150 000- 250 000 or even more peptides. For fractionation of proteolytic peptides a High Performance Liquid Chromatography-instrument (HPLC) is applied. Several chromatographic separations such as reverse phase (RP, based on hydrophobicity of peptides), anion and cation exchange (based on electrostatic interaction of analyte molecules with positively or negatively charged groups), and affinity (based on interactions with specific functional groups) are available. The combination of different chromatographic techniques for peptide separation can increase the efficiency of the subsequent MS-analysis. A simplistic view of the steps of a proteomics experiment – from sample via separation to MS and protein identification – is given in Fig. 2.

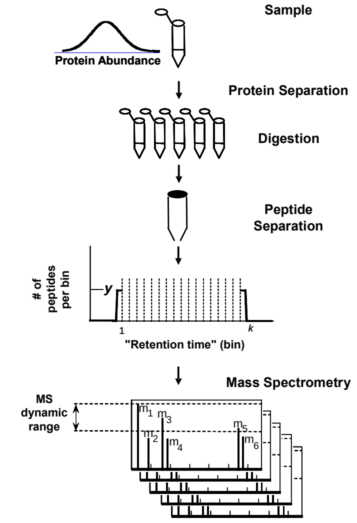

Figure 2. The proteomics workflow. Protein sample is taken from an organism. Proteins are separated into fractions and proteolytically digested. Peptides in each digested fraction are separated and subjected to mass spectrometric analysis.

Mass spectrometry

Mass spectrometry is the key source of information for protein identification in proteomics experiments. Mass spectrometry is an analytical technique based on measurement of the mass-to-charge ratio of ions. Mass spectrometry can be used for direct determination of the nominal mass of an analyte and to produce and detect fragments of the molecule that corresponds to discrete groups of atoms of different elements that reveal chemical features of the analyte molecules [22, 23]. Mass spectrometry can also, under particular circumstances, be exploited for determination of the amount of an analyte [24]. This technique was introduced by the British scientist J. J. Thomson [25], who in 1910 designed the first instrument, called mass spectrograph. Mass spectrometry's primary role was for many decades in studies of small volatile molecules [26, 27] and isotopes [28, 29].

Mass spectrometers consist of two major components: the *ion source* and the *mass analyzer.* The *mass analyzer/filter* separates ions according to their mass-to-charge ratio (*m/z*) – with the mass of the ion expressed in atomic mass units and the charge expressed as the *number* of charges that the ion possesses. The atomic mass scale definition is based on a fraction of a specific isotope of carbon. 1 mass unit is equal to 1/12 the mass of the most abundant naturally occurring stable isotope of carbon, ^{12}C. This definition of mass is synonymous with 1 Dalton (Da). A recording of the number of ions detected ("abundance") of a given *m/z* value as a function of *m/z* is a mass spectrum**.**

Ion sources

The ion source of a mass spectrometer converts the analyte molecules into ions that can be transported to and analyzed by the mass analyzer. In the early 1980s two new ionization techniques, fast atom bombardment [30, 31] and plasma desorption [32, 33] were invented and mass spectrometry became a helpful tool in the analysis of small proteins. A few years later these techniques were outperformed with respect to mass range and sensitivity by two new methods: electrospray ionization (ESI) [34] and matrix-assisted laser desoption/ionization (MALDI) [35]. ESI and MALDI opened a new era in protein analysis and are the ionization techniques used today for protein and peptide identification in proteomics.

Samples for **MALDI-MS** are introduced into the mass spectrometer in a solid state with the analyte molecules incorporated in matrix crystals. The compound of interest is mixed with small organic molecules (matrix), which have a strong absorption in the UV-region and must be capable of forming a fine crystalline solid during co-deposition with the analyte on to a plane surface. The irradiation in the mass spectrometer of this crystalline mixture by a UV laser pulse induces a large amount of energy in the condensed phase through electronic excitation of the matrix molecules. This causes the desorption of ions formed by proton transfer between the matrix and the analyte compound (Fig.3).

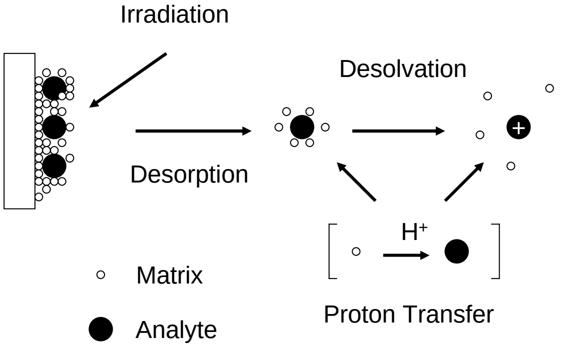

Figure 3. Ion formation in MALDI.

The matrix also serves to minimize sample damage from the laser radiation by absorbing most of the incident energy and the matrix is believed to facilitate the ionization process. The number of matrix molecules exceeds those of the analyte, thus separating its molecules and thereby preventing the formation of sample clusters which inhibit the appearance of molecular ions. Many of the commonly used matrix compounds are organic acids (Fig.4). For detection of small proteins and peptides (<10 kDa) α-cyano-4-hydroxy cinnamic acid (α-CHCA) [36] and 2,5-dihydroxybenzoic acid (DHB) [37] are used, and 3,5-dimetoxy-4-hydroxy cinnamic acid (sinapinic acid) is usually applied to detection of heavy proteins (>10 kDa).

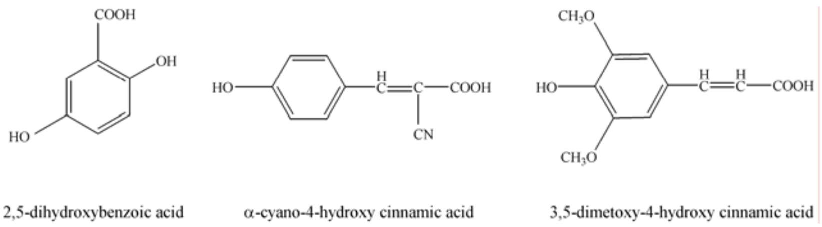
COOH
OH
HO
2,5-dihydroxybenzoic acid
HO
H
C
C
COOH
CN
α-cyano-4-hydroxy cinnamic acid
CH3O
HO
H
H
C
C
COOH
CH3O
3,5-dimetoxy-4-hydroxy cinnamic acid

Figure 4. Structures of representative compounds used as MALDI matrices.

ESI instruments allow the use of liquid sample solutions. The compound of interest is mixed with a solvent, which passes through a capillary tube with a weak flux at atmospheric pressure. An electric field is applied between this capillary and a counter electrode and induces a charge accumulation in the liquid at the end of the capillary, whereby multiply charged droplets are formed and are sprayed out from the end of the capillary. These droplets are forced by the electric field to enter a region of decreasing pressure, where evaporation of the solvent will cause explosion of the droplets and produce singly or multiply charged ions (Fig. 5).

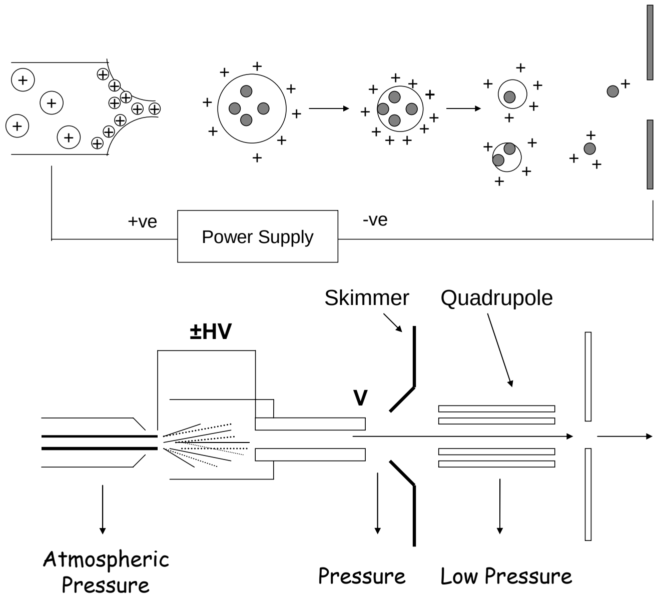

Figure 5. Top: Ion formation from electrospray ionization source. *Bottom:* ESI source coupled to a mass analyzer.

The charge can be positive or negative depending on the analyte and the field applied. Two mechanisms for the formation of molecular ions in ESI are accepted: the charged residue model and the ion evaporation model [38-40]. Since ESI can form multiply charged ions, it is often possible to observe very large molecules with a mass analyzer having a relatively small *m/z* range.

Mass analyzers

Mass analyzers distinguish ions according to their *m/z* ratio and influence the accuracy, range, resolution and sensitivity of an instrument. Nowadays, five types of mass analyzers are used in proteome analysis: time-of-flight (TOF), quadrupole (Q), quadrupole ion trap (QIT), Fourier transform-ion cyclotron resonance (FT-ICR), and orbitrap. These analyzers are sometimes combined in a single instrument, *e.g*. quadrupole-time-of-flight (Q-TOF).

The operating principles of the **TOF** mass spectrometer [41] involve measuring the time required for an ion to travel from an ion source to a detector usually located 1 or 2 m from the source. Ions obtain their kinetic energy by acceleration in an electric field. The ion velocities depend on *m/z* values and correspondingly ions of different *m/z* will reach the detector at different times. An important characteristic of mass analyzers is the resolution – *i.e.*, the ability of a mass spectrometer to distinguish between ions of different mass-to-charge ratios.

The resolution of TOF-analyzers is limited by the length of the flight path, but their advantage is no upper mass limitation. Improvement of resolution of TOF analyzer was accomplished by the invention of the reflectron (an electrostatic mirror), which serves to reduce the velocity distribution of ions and hence narrow the spread in time-of-flight for ions with the same *m/z* (Fig.6) [42].

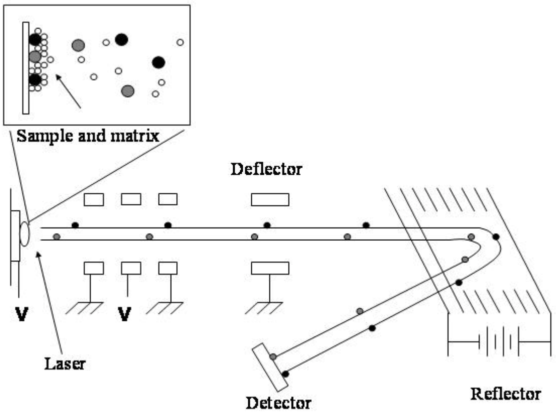

*Figure 6***.** Principle of MALDI-TOF mass spectrometer with an electrostatic reflector.

The **quadrupole analyzer (filter)** is a device that consists of four parallel rods with a direct current voltage and a superimposed radio- frequency (RF) potential applied to the rods. The field on the quadrupole determines which ions are allowed to reach the detector. As a field is imposed, ions moving into this field region will oscillate depending on their mass-to-charge ratio and, depending on the radio frequency field, only ions of a particular *m/z* are allowed to pass through. The *m/z* of an ion is determined by correlating the field applied to the quadrupole with the ion reaching the detector.

In the **quadrupole ion trap** mass analyzer, the quadrupole field is three dimensional and ions of specific *m/z* values, which depend on the level of radio-frequency voltage are stored in the device. Ions are affected in all directions so that they travel in discrete orbits within the field. The QIT operates at a relatively high pressure of ~10^{-1} Pa as opposite to ~10^{-4} Pa for the quadrupole filter and ~10^{-7} Pa for TOF-analyzers. This increased pressure allows for sufficient resolution to be achieved. The proper operating pressure is maintained by a continuous flow of helium or argon gas into mass spectrometer. The use of this buffer gas collisionally cools the ions, reducing their rotational and vibrational energies so that this damping of the ion motion extends the *m/z* range of ions that can be trapped with good efficiency and resolution. A scan sequence is applied to acquire the mass spectrum. It starts with a clearance of the ion trap and continues with accumulation when the ions are trapped in the RF field using low quadrupolar amplitude and cooling with the inert gas. During the subsequent mass analysis the field strength is increased to progressively eject ions of increasing *m/z* values out of the trap.

FT-ICR [43-45] is based on the observation that a charged particle will precess in a magnetic field at a frequency related to its *m/z* value [44, 46, 47]. The name of the technology derives from the cyclotron frequency of a precessing ion in an orbit, the plane of which is perpendicular to the applied magnetic field, and from the fact that energy can be transferred to the oscillating ion provided that the energy is available at the cyclotron frequency (*i.e.*, resonance condition). In classical ICRMS operation, energy at a *specific* frequency that corresponds to the precession frequency of the ion with the *specific m/z*- value is transmitted into the ICR cell. If an ion of this *specific m/z*- value is present in the cell, it will absorb the energy because of the resonance condition and move to an orbit of increasing radius while maintaining its characteristic precession frequency. Once the radius of the precession orbit exceed the internal dimension of the ICR cell, the ions of that *specific m/z*- value will collide with the walls of the ICR cell, producing a measurable electrical signal, the strength of which can be related to the abundance of these ions. In the FT mode a wide variety of frequencies is transmitted to the ICR cell and thus ions of many different *m/z*- values can absorb the energy at the same time. Irradiation is very brief and ions after absorbing energy will not achieve a cyclotron orbit that exceeds the dimension of the cell but become coherent within their cyclotron orbit and will induce an oscillating charge in the walls of the ICR cell as they precess. The overall induced charge oscillation in the ICR cell walls consists of an overlay of all the component frequencies of different ions oscillating in the ICR cell and FT approach allows ions of all *m/z*- values to be determined simultaneously. FTMS is unique in that an increase of the measurement time increases both sensitivity and resolution. This advantage derives from the fact that in FTMS the ions are not consumed during the detection process.

The **orbitrap** [48, 49] is a mass analyzer that consists of an outer barrel-like electrode and a coaxial inner spindle-like electrode that form an electrostatic field with quadro-logarithmic potential distribution. In an orbitrap, ions are injected tangentially into the electric field between the electrodes and trapped because their electrostatic attraction to the inner electrode is balanced by centrifugal forces. Thus, ions cycle around the central electrode in rings at the frequencies which are inversely proportional to the square root of the mass-to-charge ratio. These oscillations are detected as a time-domain signal using image current detection as sensed by the two electrically isolated components that constitute the barrel electrode. The frequencies of oscillating image current are transformed into mass spectra using a fast Fourier transform.

All the mass analyzers mentioned here have a wide spread use in proteomics. Specific priorities in a given proteomics experiment can influence the choice of mass analyzer *e.g.* FT-ICR and orbitrap mass analyzers display high performance but are unfortunately rather expensive. The experimental proteomic data employed in this thesis were generated using either MALDI or ESI ion sources, and either TOF or Q-TOF mass analyzers.

For detection of the ions, which are emerging from the mass analyzer and measurement of their *m/z* and abundances all mass spectrometers are supplied with a **detector**. The detectors most often used in modern mass spectrometers involve secondary emission of electrons. Positive or negative ions cause the emission of one or several secondary particles (usually electrons) while colliding with the detector. These secondary particles pass into an electron multiplier causing the emission of more and more electrons as they travel toward the ground potential. Thus a cascade of electrons is created that finally results in a measurable current at the end of the electron multiplier. The detector signals are transferred to a computer for recording of the signal.

Mass spectrometry based protein identification

Peptide mass fingerprinting

Protein identification using mass spectrometry is based on the information in mass spectra acquired from peptides of digested proteins. Typically enzymes that cleave proteins at known positions are used, which generate peptides whose masses can be predicted (*e.g.* trypsin digests proteins at the C-terminal of arginine and lysine). A measurement by MS of the masses of these peptides yields a peptide mass fingerprint (PMF) of the protein [19]. For PMF-based protein identification, a sequence collection containing all possible peptide sequences that can be present in an organism is needed. A computer performs a virtual digestion of entire proteins in a sequence collection in the same way as in a real experiment and creates a theoretical mass list of peptides. Mass values obtained from the acquired mass spectrum (the PMF) are then compared to the theoretical masses from the mass list in order to find matches. Since all mass spectrometers measure *m/z* values with an error, any mass value in the mass spectrum can match several mass values from the theoretical mass list. Therefore, a score that describes the degree of matching with the data is assigned to each of the proteins in the sequence collection, and a software tool is applied to rank the proteins according to their respective scores. The protein with the best score is assumed as the identification result.

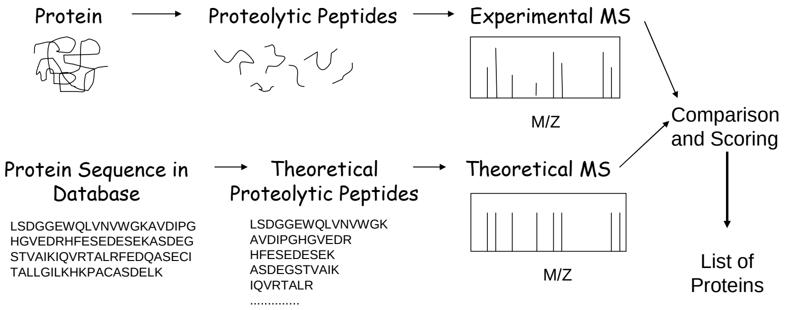

*Figure 7***.** The principle of protein identification using peptide mass fingerprinting.

MS/MS- data for protein and peptide identification

An alternative method for identifying proteins and peptides exploits tandem mass spectrometry. This technique employs *isolation* of a precursor ion which then undergoes a *fragmentation* yielding neutral fragments and fragment ions. A number of different fragmentation techniques are available that lead to the detection of different types of fragment ions. The most important techniques of fragmentation are collision-induced dissociation (CID) [50], electron capture dissociation (ECD) [51] [52] and electron-transfer dissociation (ETD) [53]. CID is an ion/neutral process wherein the ion of interest (precursor ion) is fragmented as a result of collision with inert molecules of argon or helium.

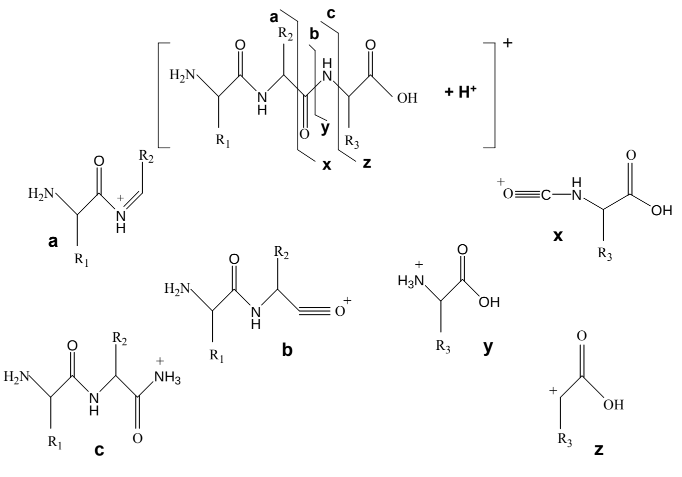

Figure 8. Main fragmentation paths of peptides. In CID b- and y- ions dominate. C- and z-ions are typically detected when using the ECD or ETD.

ECD converts a multiprotonated molecule to a radical cation, an odd-electron species as represented by following reaction:

$$MH_2^{2+} + e^-_{5\,eV} \longrightarrow MH^{+\cdot} + H\cdot$$

The radical cation formed at the time of electron capture readily fragments via a variety of pathways, but favors the formation of c- and z-type ions from proteins or peptides [54-56]. ETD adds a low- energy electron to a multiprotonated molecule via an ion-ion reaction thereby converting it to a radical cation, which dissociates in pathway analogous to those observed in ECD. Multiprotonated ions (formed *e.g.* by ESI) are guided to a reaction chamber for interaction with a beam of electron-rich anions formed in a separate ion source:

$$MH_2^{2+} + A^{-\cdot} \longrightarrow MH^{+} + AH$$

A set of proteolytic peptide fragment mass values (*e.g.* derived from b- and y-ions) provide sequence information that can be utilized for identifying the peptide that was fragmented. This identification can be done with three different methods. The dominating approach is to perform sequence collection searching [57, 58] (Fig. 9). The mass information from the fragment mass spectrum is compared with theoretical fragment mass spectra generated from a sequence collection. Theoretical fragment mass spectra are generated by *in silico* digestion of proteins, assuming the same enzyme as in the experiment, and by assuming the same fragmentation pathways as in the experiment. The degree of matching between experimental and theoretical fragment mass spectra is described by a score, and the peptide sequence in the collection that obtains the best score is assumed as the identification result.

Another approach to peptide identification using MS/MS-data is *de novo* sequencing [59]. This approach is required if a peptide of interest is not present in any sequence collection. *De novo* sequencing means that a peptide sequence is derived directly from the mass spectrum. This approach is based on permutation of amino acids giving all possible sequence combinations matching the peptide mass and generation of theoretical mass spectra for each peptide sequence. All candidates are compared with the experimental fragment mass spectrum in order to find the best match. This process is time consuming and challenging, and requires high quality data.

A third approach is *library searching* [60], where peptides are identified by comparing the experimental MS/MS spectrum with previously acquired spectra of identified peptides stored in a library [61, 62]. This approach can be very sensitive, since the comparisons involve real, already existing spectra (intensity information is included in the comparison), but of course is useless for analysis of peptides not already detected.

MS/MS- based identification has several advantages since it provides detailed information about a peptide sequence and its modifications. It allows working with complex peptide mixtures and does not require all the peptides of a given protein to be confirmed to achieve confident identification of a protein.

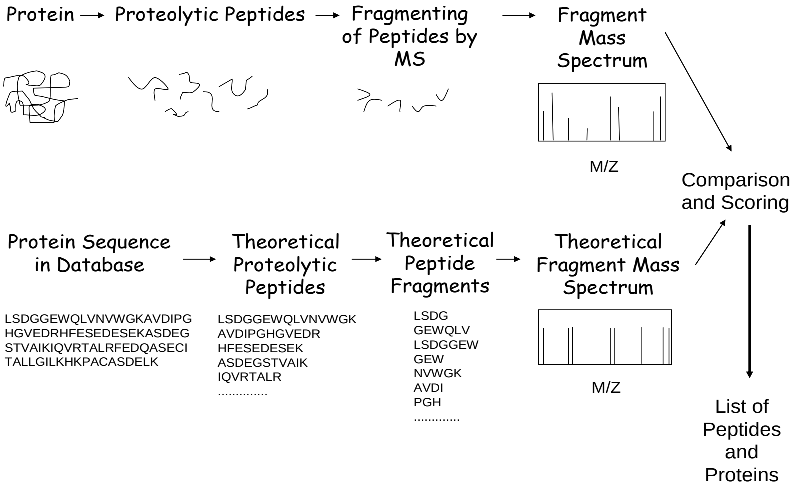

Figure 9. Protein identification using MS/MS- data and sequence collection searching.

Software tools for peptide and protein identification

Several software tools have been developed for scoring and ranking proteins or peptides in order to find the candidate in the sequence collection that gives the best match with the experimental mass spectrometric data. All these tools at some point count the number of matching mass values. The number of matches can be used directly to rank protein candidates [63]. Ranking by the number of matches favours false identification of large proteins [64], since large proteins can yield many random matches with the experimental mass spectrum. For PMF data, Mascot [65] and Profound [66] are two commercially available tools with more sophisticated scoring methods. Mascot is based on the MOWSE scoring scheme [67] and it attempts to compute a probability, P, that the number of matching fragments is random, resulting in a score $-^{10}\log$ (P). However, the details of the computation of P have not been described in the literature. Profound, which employs a Bayesian scoring function, considers individual mass errors in the scoring and ranks the protein sequences in a collection according to their probability of producing a PMF [66]. In this thesis the *Probity* algorithm was employed for PMF based protein identification. Probity ranks the protein sequences in the collection according to the risk that the matching with the experimental PMF is random, and computes a significance level for each result. The details of the underlying computations in Probity are described in references [68, 69].

Commercially available software tools for MS/MS based peptide identification include *e.g.* SEQUEST [58] and Mascot (http://www.matrixscience.com/). In this thesis, a freely available open source tool X! Tandem [70] was employed for peptide identifications. X! Tandem compares mass spectra with theoretical mass information for peptide sequences in the sequence collection searched, calculates a score for each comparison and computes an expectation value (e-value) - *i.e.*, an estimate of the risk that the score is associated with a random match. The e-value computation utilizes the score distribution of the random matches in an individual search (Fig. 10).

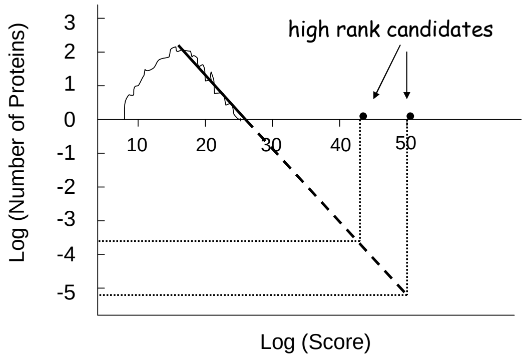

Figure 10. The expectation value (e-value) is an estimate of the risk that the score for the highest ranked peptide or protein is associated with a random match. In order to estimate the e-value, X! Tandem makes a *distribution* of all the scores for all peptides in the sequence collection that to some degree match randomly the MS/MS spectrum. E-values are calculated by extrapolating the tail of this distribution to the high scores of top ranked protein candidates. The e-value for a protein candidate is the number of protein sequences in the collection searched expected to get the same or higher score by random matching.

Differential protein quantitation and identification by mass spectrometry

(Paper I)

Quantitative proteomics investigates the changes in gene expression level in different physiological states, *e.g.* describing the differences between healthy and disease cells.

Quantitation based on 2D-GE

Several methods for differential detection are used in proteomics. Two-dimensional gel electrophoresis is a widely used approach for reflecting changes in gene expression level. At least two gels need to be compared in order to see differences between two samples. The advantage is the ability of 2D-GE to serve as image maps to allow investigators to compare changes in the patterns of spots on the gel. 2D-GE has some drawbacks. It is often difficult to perform completely reproducible 2D-GE analysis. Differences in protein migration in either dimension could be misinterpreted as changes in expression level between two samples. 2D-

GE has a relatively small dynamic range when staining is the technique for protein detection**.** The lowest concentration of protein that could be detected is approximately 1 ng/spot by using fluorescence dye [71].

Using radioactive isotope protein labeling can lead to 2D-GE detection sensitivities at attomol levels, with dynamic range of over six orders of magnitude and statistically significant quantification even of changes in the 15-20% range [72]. The low abundance proteins cannot be detected by GE, since high-abundant proteins can be 10 orders of magnitude higher in total amount of proteins in a sample [73]. This disadvantage requires steps that remove high-abundant proteins before running 2D-GE. Depletion of abundant, mostly high molecular weight proteins are often desired for loading of a much higher amount of the low copy and/or low molecular weight proteins for analysis. Various forms of immunoglobulins and human serum albumin represent the most abundant proteins in human plasma, constituting up to 80% of total plasma proteins. Depletion strategies involve the use of a chlorotriazine dye with high affinity for albumin, or the use of antibody affinity ligands for HSA and IgGs. Affinity media are made up of matrices with covalently attached antibodies to the specific abundant proteins. Immunodepletion columns allow removing on average 99.6% of high abundant proteins from serum.

Quantitation from MS-data

Stable isotope labeling

MS is in principal a non-quantitative technique; however, relative quantitative estimates may be deduced from mass spectrometric data. MS-based methods for relative quantitative determination typically use stable isotope labeling [74, 75]. Usually two protein samples are treated with some reagents to “tag” them. The tags are chemically identical, except that one is enriched with an isotope that is not the most abundant in nature. The samples are mixed together and digested with a specific enzyme. LC-MS/MS analysis of the obtained peptide mixture allows detecting duplets of molecular ions, separated by a mass corresponding to the absolute molecular mass between the isotope unlabeled and isotope labeled forms. Comparison of the relative abundances of the peaks corresponding to isotope labeled and unlabeled peptides reflect the abundance of the protein in the respective sample. This approach represents the relative quantification of a protein level in two samples. The sensitivity and dynamic range are determined by the mass spectrometer and are typically in the ranges 1-100 fmol and 10^2-10^4 respectively.

Label-free quantitation

Isotope labeling of proteins is not always practical and has several disadvantages. Labeling with stable isotopes is expensive and in complex protein mixtures experimental variation and noise can affect the quantitative value and accuracy. As an alternative, peptide and protein quantitation without isotopic tags by comparing signal intensities measured in MS analysis of two or several peptide mixtures can be applied. Mass spectra are acquired for each sample separately and mass peak intensities of peptide ions are compared in order to find correlation with protein abundances [76, 77].

The fact that intensities of individual ion signals can vary considerably between different spectra acquired from the same sample, suggests the potential usefulness of acquiring many spectra and employing statistical methods when analyzing the intensities. For example, multivariate analysis have been applied as support to MALDI-TOF proteome analysis in order to implement an easier and faster way of data handling for elucidating different protein characteristics [78] and can also be used for tracing of systematic differences between the digestion procedures [79]. Multivariate methods have also been used in comparison of 2D-gels for identification of proteins responsible for differences occurring between healthy and disease samples [80].

In this thesis a novel experimental design for MS-based differential detection was developed. It is demonstrated that differential analysis can be done without labeling procedure by using the semi-quantitative information in mass spectra together with the use of multivariate methods (principal components analysis (PCA) and soft independent modeling of class analogy (SIMCA)).

Principal components analysis [81, 82] is a computational method that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component (PC)), the second greatest variance on the second coordinate, and so on. Aims of PCA are to discover or reduce the dimensionality of the data set and to identify new underlying variables. The reduction of a data set is found by new orthogonal axes, which describe the direction of maximum variance in the data set. The first PC lies along the direction of maximum variance in the data set. The second PC will lie along a direction orthogonal to the first PC and in the direction of the second largest variance. Each PC is characterized by two pieces of information, the scores and the loadings. The loadings define the orientation of the computed PC plane with respect to the original variables and provide information on how the old variables are linearly combined to form the new variables, the PC scores. The loadings unravel the magnitude (large or small correlation) and the manner (positive or negative correlation) in which the measured variables contribute to the formation of scores. A set of scores represents the position of the object in the new coordinate system, and is calculated for each object and loading.

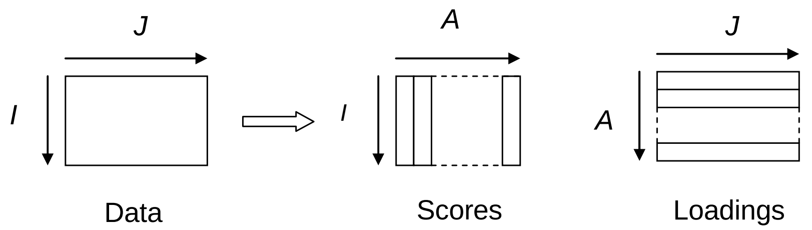

Figure 11. Principal component analysis.

Soft independent modeling of class analogy (SIMCA) is useful when two classes of data overlap (Fig. 12) [83].

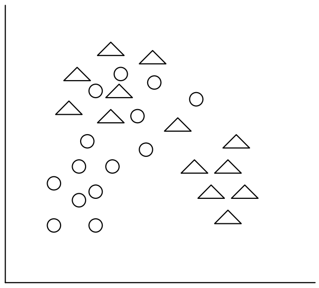

Figure 12. Two overlapping classes.

The main idea of this technique is that objects in one class show similar rather than identical behaviour because they possess a particular class pattern, which makes all these objects more similar to each other than to the objects of any other class. The actual classification technique is not applied in this investigation but the discriminatory power, $D_{j,}$ related to SIMCA was calculated by calculating the PCA models for each class independently. The number of PCs needed to describe the model of each respective class is determined. Each class could be described by a different number of PCs. SIMCA can measure how well a variable, *j*, discriminates between two classes by calculating its discriminatory power D_j. In order to determine D_j, it is necessary to compare the objects of each respective class to both class models by computing the residual matrix, *E*, for each combination of class and class model. The residual matrices describe the differences between the objects and the model.

$$E = Data - Scores * Loadings'$$

For each *E* computed the standard deviation S_{jresid} of the residual of the variable *j* is derived. Knowing S_{jresid} for each combination of class and class model allows computation of the discriminatory power D_j that reveals the variables that discriminate between two classes:

$$D_j = \sqrt{\frac{{}^{\text{classAmodelB}}S_{jresid}^{\;2} + {}^{\text{classBmodelA}}S_{jresid}^{\;2}}{{}^{\text{classAmodelA}}S_{jresid}^{\;2} + {}^{\text{classBmodelB}}S_{jresid}^{\;2}}}$$

The bigger *Dj* value, the higher is the discriminatory power. This feature of *Dj* is suitable for the problem of detecting what is causing the differences in signals in two different samples- *e.g.* what protein is present in a different concentration in two protein mixtures.

The mass spectrometry based method demonstrated in this thesis involved acquisition of multiple MALDI-TOF mass spectra of PMF of protein mixtures. Three protein mixtures consisting of four different proteins, where the concentration of three of those proteins was held constant and the concentration of one of the proteins was varied were tryptically digested. Peptide mixtures obtained after digestion were subjected to mass spectrometric analysis. On average, 200 MALDI mass spectra were acquired for each protein mixture. PCA and SIMCA were applied for data analysis in order to trace differences in intensity values between protein mixtures. Three classes were defined and each class was associated with a different protein mixture. All combinations of two classes were analyzed using the SIMCA method. The discriminatory power D_j obtained for each combination of two classes as a result of SIMCA described the differences among two samples. Plotting D_j values against *m/z* values resulted in a transformed proteolytic peptide mass spectrum (a *D*- spectrum, Fig. 13). *M/Z* values obtained from the *D*-spectrum were used for sequence collection searching in order to identify the compound present in a different amount in two samples.

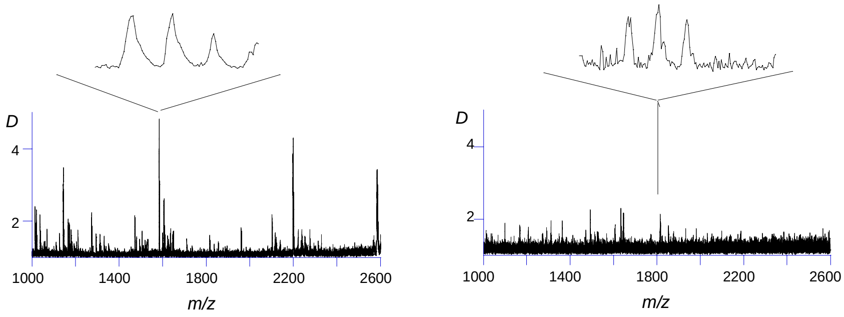

Figure 13. Left: D-spectrum describing the difference between two real protein mixtures. The concentration of one protein was four times higher in one of the samples. *Right*: *D*-spectrum for two artificial classes generated by randomly dividing the spectra obtained for the *same* protein mixture into two classes.

The mass values obtained from a *D*- spectrum were organized in descending order of their peak area and submitted to the Probity algorithm [68] starting with the mass value corresponding to the strongest peak followed by the strongest together with next strongest peak and so on. Probity determines how well any particular sequence in a collection matches to the data and assigns a statistical significance level. Figure 14 shows the results of protein identification using Probity. It is seen that carbonic anhydrase is the first protein to yield a highly significant result and therefore it is identified as the protein whose concentration was changed in the two protein mixtures.

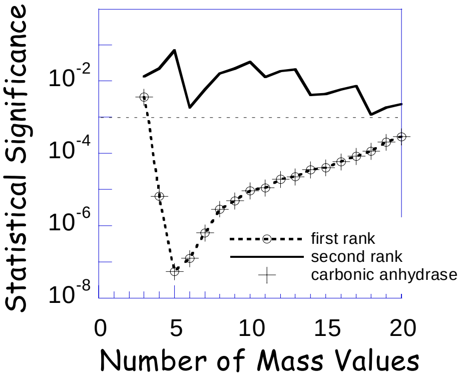

*Figure 14***.** The protein identification results obtained by the Probity algorithm utilizing information from *D*-spectrum.

This demonstrated label-free mass spectrometry based method for detecting quantitative changes has *e.g.* the potential to support 2D-GE-based differential analysis by allowing identification of the protein component that differs in cases where a gel-spot is poorly resolved and contains several proteins.

The influence of search conditions on the quality of protein and peptide identification

(Paper II and III)

The ability to identify a protein present in a sample depends on three factors: (i) the experimental design, (ii) the data quality, and (iii) the choice of protein identification algorithm (including search conditions). A good experimental design for protein identification must handle the complexity and wide range of protein abundances [84]. Algorithms used for protein identification should maximize the number of true results, minimize false results, and assign a significance level to all results. Various algorithms have their default search conditions, *e.g.* for mass accuracy or the assumed digestion efficiency, but these conditions are not necessarily optimal in any given experiment. In this thesis, the effect of search conditions was explored for PMF-based as well as MS/MS-based protein identification.

The study of the influence of search conditions for PMF-based protein identification was done using 2D-GE separated and proteolytically digested human plasma proteins analyzed by MALDI-TOF. The Probity algorithm [68] was applied to examine the impact of data processing and different search constraints. The influence of the mass accuracy *Δm*, the number of missed cleavage sites, *u*, and the size of a sequence collection on identification results was investigated at three significance levels (0.05, 0.01, and 0.001). The relation between the significance level and the number of identified proteins for a given search condition is displayed in Figure 15. Data processing in the form of a mass correction procedure utilizing deviation between experimental and theoretical mass values for matching albumin peptide masses was applied in order to improve the mass accuracy. This procedure leads to a minor but measurable improvement of the number of significant results (Fig. 16).

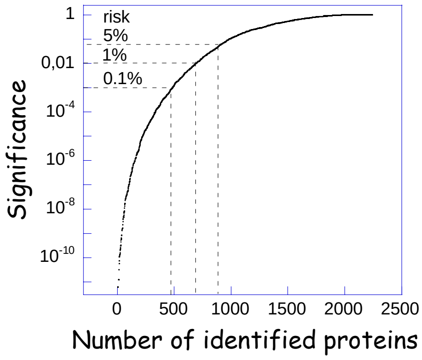

*Figure 15***.** The significance level as a function of the number of identified proteins for a given search condition.

The dependence of the number of significant results on different values of *Δm* was studied by varying of *Δm*-values from ±0.01 to ±0.3 Da. The maximum number of significant results was achieved at *Δm*= ±0.14. This value was employed as an optimum value for further searches of optimal settings of other constraints.

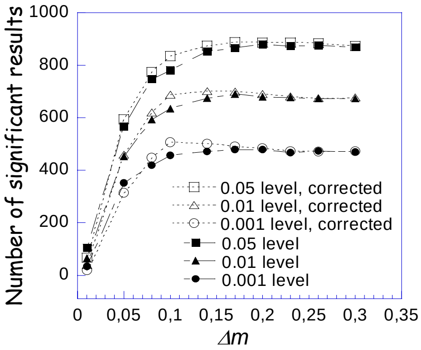

Figure 16. The number of proteins identified by Probity at different significance levels as a function of the *Δm* employed in the sequence collection search.

The number of missed cleavage sites, *u*, can influence protein identification results since the number of potentially matching mass values varies with *u*. It is seen in Figure 17 (left) that the number of significant results decreases as a function of an increasing number of missed cleavage sites for this dataset. However, the number of missed cleavage sites can vary between different PMFs depending on the cleavage efficiency of identified proteins present in the mixture. Figure 17 (right) represents a comparison of results obtained for two different ways of choosing *u*-values in the search. The number of identified proteins is increased when using the *u*-value that gives the best significance level for each respective PMF compared to the use of a fixed *u*-value for *all* PMFs.

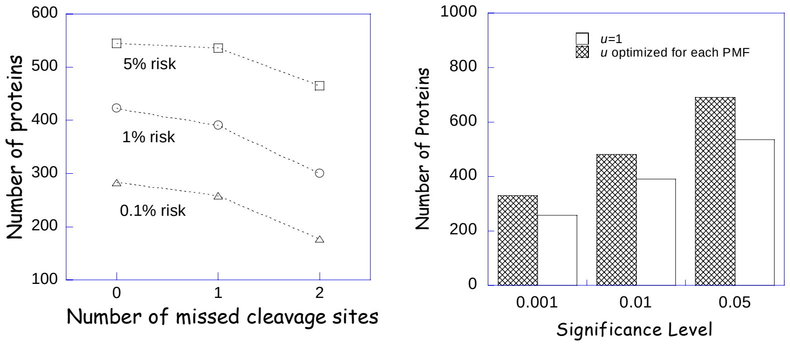

*Figure 17***.** *Left:* The influence of different *u*-values on the number of significant results for three different significance levels. *Right*: Comparison of the number of statistically significant identification results for different *u*-value conditions. The white bars represent the use of *u*=1 (fixed) for all PMFs. The cross-hatched bars represent results obtained by using the *u*-value that yielded the lowest risk of obtaining a false result for each respective PMF.

The study of the influence of search conditions for MS/MS-based protein identification was performed under different search constraints using the X! Tandem algorithm [70]. X! Tandem computes an expectation value (e-value) for each identified peptide (Fig. 10). A large set of tandem mass spectra from 1D-GE separated mouse proteins was subjected to X! Tandem searching and the number of statistically significant results at three different significance levels (e-value < 0.05, < 0.01, and < 0.001) was counted for each data processing step and search constraint.

Peptide identification was performed for different settings of the mass accuracy of the precursor ion (Δm_p), the mass accuracy of fragment ions (Δm_f), the number of missed cleavage sites (*u*), and the number of peptide sequences searched.

Using default settings of constraints for X! Tandem, it was found that for peptides identified at a good significance level, the deviation between theoretical and experimental mass values was pronounced. It was found that the magnitude of mass deviations increases with increasing mass and that the distribution of mass deviations varies between different LC-MS/MS runs. Therefore, a procedure to describe mass deviations for precursor and, fragment mass ions using least-squares-fits to linear functions was employed for each respective LC-MS/MS run. The resulting functions were applied to correct the measured mass values. Figure 18 displays the mass deviations before and after applying the mass correction procedure. Applying the mass correction procedure noticeably increased the number of results identified (Fig. 19).

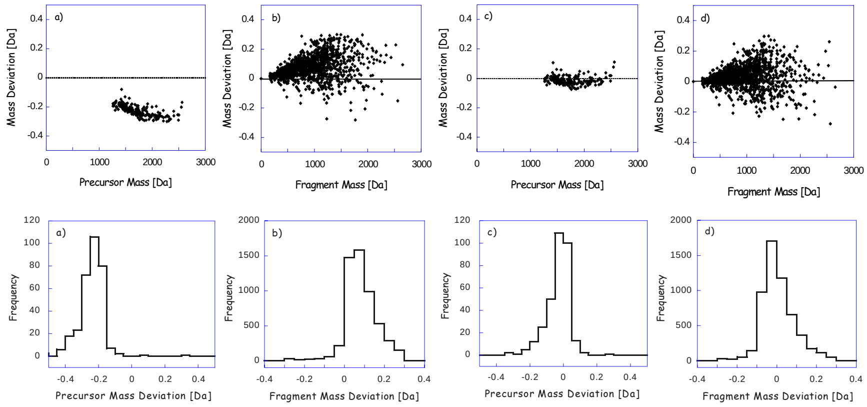

*Figure 18***.** Mass error distributions for one MS/MS run for precursor mass ions and fragment mass ions before (a, b) and after (c, d) applying a mass correction procedure.

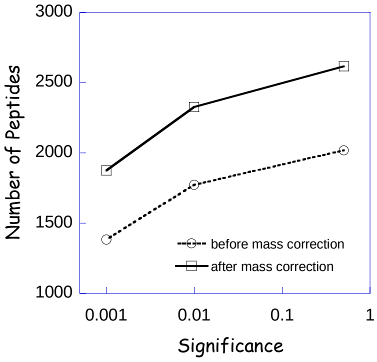

Figure 19. The number of unique peptides identified at three significant levels before and after applying a mass correction procedure for precursor and fragment ions.

The impact of the assumed mass accuracy of the precursor ion Δm_p and the assumed mass accuracy of the fragment ions Δm_f was studied. As can be seen in Fig. 20 (left) it was found that the number of peptides identified increased with increasing Δm_p up to the point where the true error distribution (Fig. 18) is covered and unnecessarily high settings of Δm_p does not reduce the ability of the algorithm to identify peptides. This observation is supported by simulations employing *in silico* generated *S. cerevisiae* peptide MS/MS spectra and searching the *M. musculus* sequence collection. It is seen in Fig. 20 (*right*) that the Δm_p employed in the search has negligible influence on the number of randomly matching fragments. In contrast, it is seen in Fig. 21 that increasing the value of the fragment mass error Δm_f, beyond the point where it covers the error distribution of the data, considerably increases the number of randomly matching fragments. Therefore, the maximum number of peptides is identified when Δm_f corresponds with the errors of the data and then decreases when increasing Δm_f (Fig. 21) .

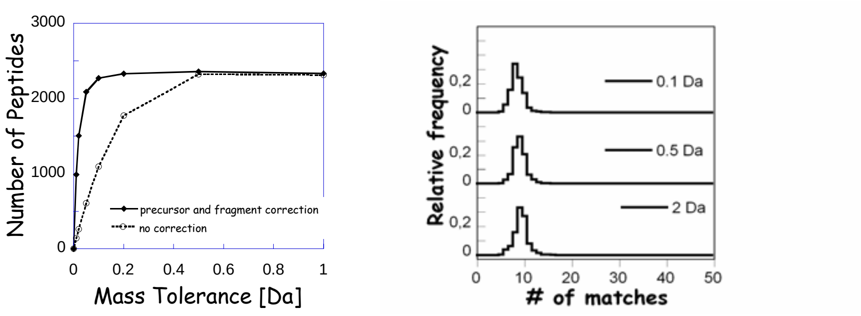

Figure 20. Left: The number of peptides identified at the 0.01 significance level as a function of the precursor ion mass tolerance when fragment mass tolerance was set to 0.2 Da for mass corrected data and data with no correction. *Right:* Distribution of the number of matching fragment masses for *in silico* generated *S. serevisiae* peptide MS/MS spectra randomly matching *M. musculus* peptides for different *precursor* ion mass tolerances.

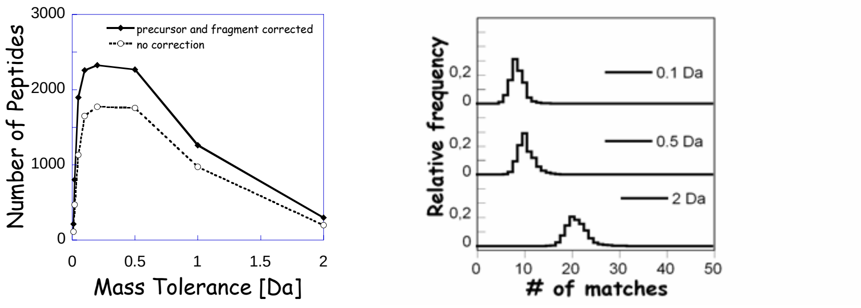

Figure 21. Left: The number of peptides identified at the 0.01 significance level as a function of the fragment ion mass tolerance when the precursor mass tolerance was set to 0.2 Da for mass corrected data and data with no correction. *Right:* Distributions of the number of matching fragment masses for *in silico* generated *S. serevisiae* peptide MS/MS spectra randomly matching *M. musculus* peptides for different *fragment* ion mass tolerances.

The peptide fragment mass information required to identify peptides and their post-translational modifications

(Paper IV)

Current proteomic methods typically utilize information from fragment ion mass spectra of proteolytically digested proteins followed by sequence collection searching to identify proteins and their post-translational modifications (PTMs). PTMs can not be predicted from the genome, but PTMs are important for determining protein activity, protein localization and interactions with other proteins [14]. Proteins may undergo various types of PTMs [85]. Phosphorylation is probably one of the most widespread and better understood PTMs. The modification of the side chain of serine, tyrosine, and threonine by a phosphate moiety (H_3PO_4) results in phosphoproteins, which are involved in *e.g.* metabolic pathways, membrane transport, cell growth and signaling processes [11].

MS/MS represents a general method for analysis of protein modifications. The analysis is however challenging, *e.g.* because the ionization of modified peptides compared with unmodified species is often poor and the dynamic range of the mass spectrometers is limited. To overcome existing problems good experimental design and protocols and at the same time understanding of the role of the information content in MS/MS spectra are needed. In this thesis the examination of what features of tandem mass spectra for modified and unmodified peptides are important for successful identification of peptide and their PTMs, was performed by computer simulation.

Sets of proteolytic peptide fragment mass spectra were generated *in silico* from *S. cerevisiae* proteins assuming exposure to trypsin. The peptides were selected randomly in a mass region $m_p \pm \delta$, where $\delta = 2$ Da and $m_p = 1000$, 1500, 2000, and 2500 Da. 50 peptides were selected for each m_p and for each randomly selected peptide a set of peptide fragment mass spectra were generated. The *number* of randomly selected *b* and *y* fragments in each spectrum was varied over a broad range. By adding randomly selected fragment mass values sets of spectra containing various numbers of background peaks were generated. Spectra of *modified* peptides were generated by assuming a 100% probability of exactly 1 phosphorylation in proteolytic peptides containing S, T, and Y amino acid residues.

These various types of *in silico* generated mass spectra were employed to investigate the outcome of *S. cerevisiae* sequence collection searching with X! Tandem under different search constraints. The *critical number of fragment masses* was defined as the number of fragment masses in the spectra that yields a 50% chance of success (Figure 22).

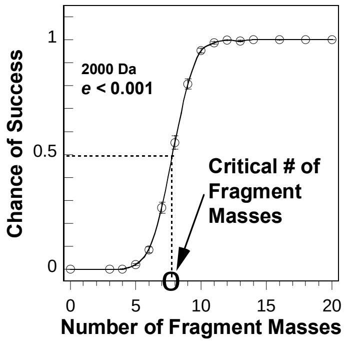

*Figure 22***.** The chance of success – i.e., the fraction of the spectra that yield a true result and an *e*-value below a desired threshold – as a function of the number of fragment masses in the spectra. The *critical number of fragment masses* is the number of fragment masses in the spectra that yields a 50% chance of success.

The critical number of fragment masses was studied as a function of the precursor ion mass error (varied in the range 0.01-100 Da) and fragment ion mass error (varied in the range 0.01-2 Da). The results showed a relatively small influence of precursor ion mass error on the outcome of peptide identification results (Figure 23, *left*), whereas the fragment mass error had a much stronger influence on the critical number of fragment masses (Figure 23, *right*).

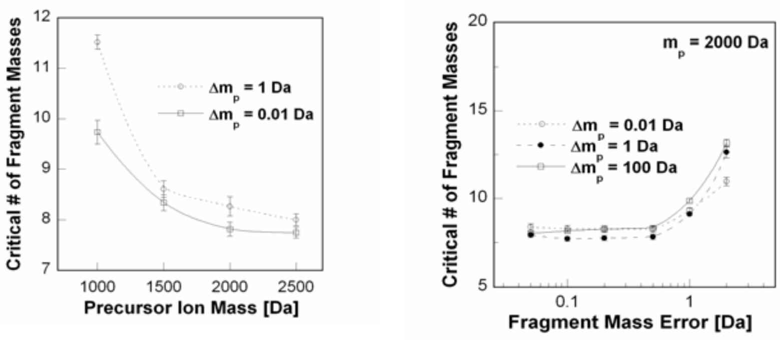

Figure 23. Left: The critical number of fragment masses as a function of the *precursor ion mass* for two different settings of the precursor ion mass errors; the mass accuracy of fragment ion was ±0.5 Da and the e-value threshold was set to 0.001. *Right:* The critical number of fragment masses as a function of the *fragment ion mass error* for three different settings of the precursor ion mass errors at e <0.001.

The simulations revealed that spectra may contain a lot of background peaks and still yield significant peptide identification results. Figure 24 displays the comparison of results from spectra containing no background with results for spectra containing 50% and 80% randomly added background peaks. It is seen from figure 24 that fragment mass spectra with 80% background require more fragment masses in tandem mass spectra for successful identification than spectra with 50% background or with no background. Addition of 50% background to fragment mass spectra leads to moderate increase of the critical number of fragment masses when fragment mass error are in the region 0.1< Δm_f<0.5 Da.

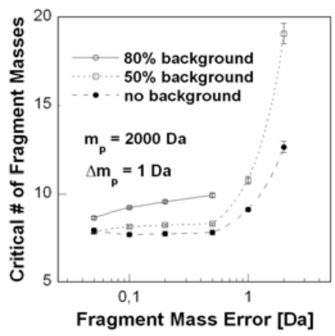

*Figure 24***.** The critical number of fragment masses as a function of the fragment mass error adjusted in the search, for three different levels of background.

The efficiency of identification of singly phosphorylated peptides under different search conditions displayed minor differences from identification results for unmodified peptides (Fig. 25). Hence, phosphopeptides can be identified under similar search constraints as unmodified peptides.

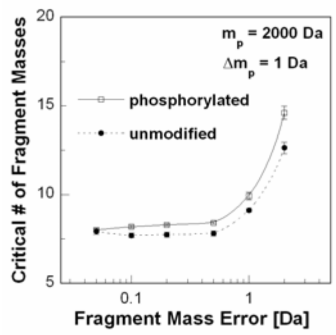

*Figure 25***.** The critical number of fragment masses as a function of the fragment mass error, for singly phosphorylated peptides and unmodified peptides.

The results of the computer simulations suggest that the number of fragment ions in tandem mass spectra and fragment mass errors < ±0.5 Da are critical features for successful peptide identification. Therefore, experimental designs yielding many fragments (*e.g.*, by employing different fragmentation techniques) are desirable.

Conclusions

A new mass spectrometry based method for differential detection and identification of components in mixtures of proteins has been developed. The method utilizes semi-quantitative information in MALDI-TOF spectra acquired from proteolytically digested protein mixtures. The mass spectra are analyzed by multivariate methods followed by protein identification using the Probity algorithm. The method detects and identifies a component present in two protein mixture in different concentrations and has the potential to be used for differential detection of unresolved up- or down-regulated proteins separated by 2-DE.

The impact of different sequence collection search constraints on the statistical significance of peptide mass fingerprint (PMF) based protein identification results was demonstrated. Using the Probity algorithm for the first time on a large experimental data set revealed that optimizing the number of missed cleavage sites for each respective peptide mass fingerprint has a strong influence on the number of significant results.

The influence of different search constraints on successful peptide identification results was studied using a large data set from the mouse proteome and the X! Tandem algorithm for sequence collection searching. A mass correction procedure based on least-squares-fits to linear functions was employed in order to improve the successful outcome of identification result. It was found that the precursor ion mass tolerance has a minor influence on the number of identified peptides while the fragment ion mass tolerance has a strong influence, since large fragment mass errors yield a lot of random matching and reduces the ability of the search engine to identify peptides.

A simulation-based study of the peptide fragment mass information required for successful identification of peptides and their PTMs was performed. This study demonstrates that the number of fragments generated by MS/MS and the mass accuracy of the fragment mass ions are the keys for successful identification of peptides and their PTMs.

Outlook

The goal of proteomics to comprehensively analyze the proteins present in an organism and to understand the function of any protein, is far from reached. Despite the enormous steps which have been taken in proteome research during the last 15 years the number of problems that needs to be solved are still vast.

Challenges for the future include developing new sensitive methods for MS and MS/MS based proteomics for resolving and analyzing complex protein mixtures, establishing experimental protocols for higher sequence coverage of proteins, elaborating gel- and label-free methods for quantitative proteomics for detection of even minor changes in gene expression levels, inventing methods for quantitation of post-translational modifications, and developing robust bioinformatics tools [86] for handling of the enormous volumes of proteomics data.

References

1. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., and et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd [see comments]. *Science* 269(5223), 496-512.
2. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 genes. *Science* 274(5287), 546-&.
3. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium [published errata appear in Science 1999 Jan 1;283(5398):35 and 1999 Mar 26;283(5410):2103 and 1999 Sep 3;285(5433):1493]. *Science* 282(5396), 2012-2018.
4. Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., and Humphery-Smith, I. 1995. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* 16(7), 1090-1094.
5. Ross, P.L., Huang, Y.L.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D.J. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics* 3(12), 1154-1169.
6. Krogan, N.J., Cagney, G., Yu, H.Y., Zhong, G.Q., Guo, X.H., Ignatchenko, A., Li, J., Pu, S.Y., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ui, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., and Greenblatt, J.F. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440(7084), 637-643.
7. Barile, M., Pisitkun, T., Yu, M.J., Chou, C.L., Verbalis, M.J., Shen, R.F., and Knepper, M.A. 2005. Large scale protein identification in intracellular aquaporin-2 vesicles from renal inner medullary collecting duct. *Molecular & Cellular Proteomics* 4(8), 1095-1106.
8. Savitski, M.M., Nielsen, M.L., and Zubarev, R.A. 2006. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational

modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* 5(5), 935-948.

9. Nielsen, M.L., Savitski, M.M., and Zubarev, R.A. 2006. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics* 5(12), 2384-2391.

10. Li, H.J., Sethuraman, N., Stadheim, T.A., Zha, D.X., Prinz, B., Ballew, N., Bobrowicz, P., Choi, B.K., Cook, W.J., Cukan, M., Houston-Cummings, N.R., Davidson, R., Gong, B., Hamilton, S.R., Hoopes, J.P., Jiang, Y.W., Kim, N., Mansfield, R., Nett, J.H., Rios, S., Strawbridge, R., Wildt, S., and Gerngross, T.U. 2006. Optimization of humanized IgGs in glycoengineered Pichia pastoris. *Nature biotechnology* 24(2), 210-215.

11. Hunter, T. 1998. The Croonian Lecture 1997. The phosphorylation of proteins on tyrosine: Its role in cell growth and disease. *Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences* 353(1368), 583-605.

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

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

14. Mann, M. 2003. Proteomic analysis of post-translational modifications. *Nature biotechnology* 21(3), 255.

15. McDonald, W.H. and Yates, J.R., 3rd. 2002. Shotgun proteomics and biomarker discovery. *Dis Markers* 18(2), 99-105.

16. Kennedy, S. 2002. The role of proteomics in toxicology: identification of biomarkers of toxicity by protein expression analysis. *Biomarkers* 7(4), 269-290.

17. Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Gorg, A., Westermeier, R., and Postel, W. 1982. Isoelectric-Focusing in Immobilized Ph Gradients - Principle, Methodology and Some Applications. *Journal of Biochemical and Biophysical Methods* 6(4), 317-339.

18. Bjellqvist, B., Basse, B., Olsen, E., and Celis, J.E. 1994. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* 15(3-4), 529-539.

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

20. Wolters, D.A., Washburn, M.P., and Yates, J.R., 3rd. 2001. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 73(23), 5683-5690.

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

22. Hoye, T.R., Dvornikovs, V., Fine, J.M., Anderson, K.R., Jeffrey, C.S., Muddiman, D.C., Shao, F., Sorensen, P.W., and Wang, J. 2007. Details of the structure determination of the sulfated steroids PSDS and PADS: New components of the sea lamprey (Petromyzon marinus) migratory pheromone. *Journal Of Organic Chemistry* 72(20), 7544-7550.

23. Feng, X. and Siegel, M.M. 2007. FTICR-MS applications for the structure determination of natural products. *Analytical And Bioanalytical Chemistry* 389(5), 1341-1363.

24. Burlina, F., Sagan, S., Bolbach, G., and Chassaing, G. 2006. A direct approach to quantification of the cellular uptake of cell-penetrating peptides using MALDI-TOF mass spectrometry. *Nature Protocols* 1(1), 200-205.

25. Thomson, J.J. 1910. Rays of Positive Electricity. *Philosophical Magazine* 20, 752.

26. McLafferty, F.W. 1962. Mass Spectrometric Analysis - Aromatic Halogenated Compounds. *Analytical Chemistry* 34(1), 16-&.

27. Gohlke, R.S. and McLafferty, F.W. 1962. Mass Spectrometric Analysis - Aliphatic Amines. *Analytical Chemistry* 34(10), 1281-&.

28. Howard, O.H. 1968. Simultaneous Determination Of Uranium Its Isotopes And Alpha Activity In Urine By Mass Spectrometry. *American Industrial Hygiene Association Journal* 29(4), 355-&.

29. Amarel, I., Bernas, R., Foucher, R., Jastrzeb.J, Johnson, A., Teillac, J., and Gauvin, H. 1967. Half Life Determination Of Some Short-Lived Isotopes Of Rb Sr Cs Ba La And Identification Of 93 94 95 96rb As Delayed Neutron Precursors By On-Line Mass-Spectrometry. *Physics Letters B* B 24(8), 402-&.

30. Morris, H.R., Panico, M., Barber, M., Bordoli, R.S., Sedgwick, R.D., and Tyler, A. 1981. Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis. *Biochem Biophys Res Commun* 101(2), 623-631.

31. Barber, M., Bordoli, R.S., Sedgwick, R.D., and Tyler, A.N. 1982. Fast atom bombardment mass spectrometry of the angiotensin peptides. *Biomed Mass Spectrom* 9(5), 208-214.

32. Sundqvist, B., Roepstorff, P., Fohlman, J., Hedin, A., Hakansson, P., Kamensky, I., Lindberg, M., Salehpour, M., and Sawe, G. 1984. Molecular weight determinations of proteins by californium plasma desorption mass spectrometry. *Science* 226(4675), 696-698.

33. Sundqvist, B., Kamensky, I., Hakansson, P., Kjellberg, J., Salehpour, M., Widdiyasekera, S., Fohlman, J., Peterson, P.A., and Roepstorff, P. 1984. Californium-252 plasma desorption time of flight mass spectroscopy of proteins. *Biomed Mass Spectrom* 11(5), 242-257.

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

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

36. Beavis, R.C., Chaudhary, T., and Chait, B.T. 1992. Alpha-Cyano-4-Hydroxycinnamic Acid As A Matrix For Matrix-Assisted Laser Desorption Mass-Spectrometry. *Organic Mass Spectrometry* 27(2), 156-158.

37. Strupat, K., Karas, M., and Hillenkamp, F. 1991. 2,5-Dihydroxybenzoic Acid - A New Matrix For Laser Desorption Ionization Mass-Spectrometry. *International Journal Of Mass Spectrometry And Ion Processes* 111, 89-102.

38. Kebarle, P. 2000. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of Mass Spectrometry* 35(7), 804-817.

39. Gaskell, S.J. 1997. Electrospray: Principles and practice. *Journal of Mass Spectrometry* 32(7), 677-688.

40. Cole, R.B. 2000. Some tenets pertaining to electrospray ionization mass spectrometry. *Journal of Mass Spectrometry* 35(7), 763-772.

41. Mamyrin, B.A. 2001. Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal Of Mass Spectrometry* 206(3), 251-266.

42. Mamyrin, B.A. 1994. Laser-Assisted Reflectron Time-Of-Flight Mass-Spectrometry. *International Journal Of Mass Spectrometry And Ion Processes* 131, 1-19.

43. Marshall, A.G., Wang, T.C.L., and Ricca, T.L. 1984. Ion-Cyclotron Resonance Excitation De-Excitation - a Basis for Stochastic Fourier-Transform Ion-Cyclotron Mass-Spectrometry. *Chemical Physics Letters* 105(2), 233-236.

44. Marshall, A.G. 2000. Milestones in Fourier transform ion cyclotron resonance mass spectrometry technique development. *International Journal Of Mass Spectrometry* 200(1-3), 331-356.

45. Comisarow, M.B. and Marshall, A.G. 1996. Fourier transform ion cyclotron resonance spectroscopy (Reprinted from Chemical Physics Letters, vol 25, pg 282-283, 1974). *Journal of Mass Spectrometry* 31(6), 586-587.

46. Marshall, A.G., Hendrickson, C.L., and Jackson, G.S. 1998. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews* 17(1), 1-35.

47. Amster, I.J. 1996. Fourier transform mass spectrometry. *Journal of Mass Spectrometry* 31(12), 1325-1337.

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

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

50. Hayes, R.N. and Gross, M.L. 1990. Collision-Induced Dissociation. *Methods In Enzymology* 193, 237-263.

51. Zubarev, R.A., Kelleher, N.L., and McLafferty, F.W. 1998. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* 120(13), 3265-3266.

52. Baba, T., Hashimoto, Y., Hasegawa, H., Hirabayashi, A., and Waki, I. 2004. Electron capture dissociation in a radio frequency ion trap. *Analytical Chemistry* 76(15), 4263-4266.

53. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. 2004. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 101(26), 9528-9533.

54. Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L., Kruger, N.A., Lewis, M.A., Carpenter, B.K., and McLafferty, F.W. 2000. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem* 72(3), 563-573.

55. Syrstad, E.A. and Turecek, F. 2005. Toward a general mechanism of electron capture dissociation. *J Am Soc Mass Spectrom* 16(2), 208-224.

56. Kjeldsen, F., Haselmann, K.F., Budnik, B.A., Sorensen, E.S., and Zubarev, R.A. 2003. Complete characterization of posttranslational modification sites in the bovine milk protein PP3 by tandem mass spectrometry with electron capture dissociation as the last stage. *Anal Chem* 75(10), 2355-2361.

57. Mann, M. and Wilm, M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 66(24), 4390-4399.

58. Eng, J.K., McCormack, A.L., and Yates, J.R. 1994. An approach to correlate mass spectral data with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976.

59. Wilm, M. and Mann, M. 1996. Analytical properties of the nanoelectrospray ion source. *Anal Chem* 68(1), 1-8.

60. Craig, R., Cortens, J.C., Fenyo, D., and Beavis, R.C. 2006. Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research* 5(8), 1843-1849.

61. Hummel, J., Niemann, M., Wienkoop, S., Schulze, W., Steinhauser, D., Selbig, J., Walther, D., and Weckwerth, W. 2007. ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *Bmc Bioinformatics* 8.

62. Liu, J., Bell, A.W., Bergeron, J.J.M., Yanofsky, C.M., Carrillo, B., Beaudrie, C.E.H., and Kearney, R.E. 2007. Methods for peptide identification by spectral comparison. *Proteome Science* 5.

63. Mann, M., Hojrup, P., and Roepstorff, P. 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* 22(6), 338-345.

64. Eriksson, J., Chait, B.T., and Fenyo, D. 2000. A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal Chem* 72(5), 999-1005.

65. Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18), 3551-3567.

66. Zhang, W. and Chait, B.T. 2000. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* 72(11), 2482-2489.

67. Pappin, D.J.C., Hojrup, P., and Bleasby, A. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 3, 327-332.

68. Eriksson, J. and Fenyo, D. 2004. Probity, A Protein Identification Algorithm with Accurate Assignment of the Statistical Significance of the Results. *Journal of Proteome Research* 3(1), 32-36.

69. Eriksson, J. and Fenyo, D. 2004. The statistical significance of protein identification results as a function of the number of sequences searched. *Journal of Proteome Research* 3, 979-982.

70. Craig, R. and Beavis, R.C. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20(9), 1466-1467.

71. Patton, W.F. 2000. A thousand points of light: The application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis* 21(6), 1123-1144.

72. Cahill, M.A., Wozny, W., Schwall, G., Schroer, K., Holzer, K., Poznanovic, S., Hunzinger, C., Vogt, J.A., Stegmann, W., Matthies, H., and Schrattenholz, A. 2003. Analysis of relative isotopologue abundances for quantitative profiling of complex protein mixtures labelled with the acrylamide/D-3-acrylamide alkylation tag system. *Rapid Communications in Mass Spectrometry* 17(12), 1283-1290.

73. Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., and Sanchez, J.C. 2000. The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis* 21(6), 1104-1115.

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

75. Moritz, B. and Meyer, H.E. 2003. Approaches for the quantification of protein concentration ratios. *Proteomics* 3(11), 2208-2220.

76. Chelius, D. and Bondarenko, P.V. 2002. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *Journal of Proteome Research* 1(4), 317-323.

77. Wang, W.X., Zhou, H.H., Lin, H., Roy, S., Shaler, T.A., Hill, L.R., Norton, S., Kumar, P., Anderle, M., and Becker, C.H. 2003. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry* 75(18), 4818-4826.

78. Gottlieb, D.M., Schultz, J., Petersen, M., Nesic, L., Jacobsen, S., and Sondergaard, I. 2002. Determination of wheat quality by mass spectrometry and multivariate data analysis. *Rapid Communications in Mass Spectrometry* 16(21), 2034-2039.

79. Backstrom, D., Moberg, M., Sjoberg, P.J.R., Bergquist, J., and Danielsson, R. 2007. Multivariate comparison between peptide mass fingerprints obtained by liquid chromatography-electrospray ionization-mass spectrometry with different trypsin digestion procedures. *Journal Of Chromatography A* 1171(1-2), 69-79.

80. Marengo, E., Robotti, E., Righetti, P.G., Campostrini, N., Pascali, J., Ponzoni, M., Hamdan, M., and Astner, H. 2004. Study of proteomic changes associated with healthy and tumoral murine samples in

neuroblastoma by principal component analysis and classification methods. *Clinica Chimica Acta* 345(1-2), 55-67.

81. Jackson, J.E., *A User's Guide to Principal Components*. 1991, New York: John Wiley.

82. Wold, S., Esbensen, K., and Geladi, P. 1987. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 2(1-3), 37-52.

83. Brereton, R.G., *Chemometrics Data Analysis for the Laboratory and Chemical Plant*. 2003: Wiley.

84. Eriksson, J. and Fenyo, D. 2007. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol* 25(6), 651-655.

85. Krishna, R.G. and Wold, F. 1993. Post-translational modification of proteins. *Adv Enzymol Relat Areas Mol Biol* 67, 265-298.

86. Palagi, P.M., Hernandez, P., Walther, D., and Appel, R.D. 2006. Proteome informatics I: Bioinformatics tools for processing experimental data. *Proteomics* 6(20), 5435-5444.

Acknowledgements

First of all, I would like to thank my supervisor **Jan Eriksson** for introducing me to the field of bioinformatics, for your passion to science, for all knowledge you shared with me, for all help and support I've got, for **all** your **time** you spent with me in discussions about mass spectrometry, proteomics, statistics and life. You are an excellent teacher and nice person, thank you so much!

I would also like to thank my assistant supervisor, **Lennat Kenne**, for giving me the opportunity to be member of your group, for all your optimism, enthusiasm, sharp questions, and interesting points of view.

I am grateful to all people who made this work possible and especially:

Dr. **David Fenyö** for fruitful collaboration and interesting discussions.
S**uresh Gohil**, the person, who keeps mass spectrometers in work, for the help and assistance with mass spectrometry and for showing me the soul of mass spectrometers.
Dr**. Lars Nord** for introducing me to the field of chemometrics and writing excellent Matlab scripts.
Bernt Andersson for making a magic transformation of the difficult teaching process into a real pleasure.
Lena Andersson for, with a smile, taking care of all administrative work.
Dr. **Corine Sandström** for creating pleasant atmosphere during seminars at the department.
Dr. **Anders Sandström** for your enthusiasm in the teaching of our undergraduate students.
Prof. **Vadim Kessler** and Dr. **Gulaim Seisenbaeva** for our Russian discussions about everything.
Cristian Marcu for solving all possible problems in computer world.
Johan Bankefors, my ever room mate, for scientific discussions and funny conversations, for sharing with me your good mood and sweetmeats.

I wish to express my deep gratitude to the former and present colleagues at the Department of Chemistry:
Olesya Nikonova, Natallia Torapava, Rolf Andersson, Anders Broberg, Lasse Andersson, Nils Nyberg, Susanna Broberg, Semiha Dag, Birgit Hakkarainen, Jörgen Sjögren, Anton Pohanka, Per Hagback, Gunnar Almqvist, Marta Kaszovska, Pia Werndrup and many other people for pleasant and sociable environment.

I am really thankful to my best Swedish friends**, Line Strand** and **Joel Sohlberg**, for supporting me during all this years, for all your help to my family, for all time we spent together during nice dinners and cheerful vocations and, of cource, **Emil**, **Maja** and **Abbe** for taking care of Daniel and Polina.

I want to express my gratitude to my Russian friends in Sweden during all this years: **Konstantin**; **Alex** and **Olga**; **Artem** and **Lena**; **Igor'**, **Natasha**, **Anton** and **Alisa**; **Olesya** and **Theo**; **Natasha**; **Ruslan** and **Oksana**; **Igor'** and **Marina** for creating wonderful relaxing atmosphere.

I would like to thank my friends in Russia and my class mates from our 11 "B".

I am deeply thankful to my **Mother** and **Father** for their love, understanding, and support. Дорогие мои мама и папа, огромное вам спасибо за ту неоценимую помощь, которую вы всегда готовы нам оказать, за вашу любовь, заботу и поддержку, за то, что вы всегда рядом, несмотря на расстояние, разделяющее нас.
My grandmother, brother, sisters-in-low, mother-in-low, all my relatives for love and support.

Last, but not least, I would like to thank **Dmitri**, **Danila** and **Polina** for being the joy of my life, the light in my tunnel, my past, my present and my future.