Methods for structural studies of an antibody, screening metabolites in rat urine and analysis of spent cell cultivation media using LC/ESI-MS and chemometrics

Leila Zamani
لیلا زمانی فکری

Doctoral Thesis
Department of Analytical Chemistry
Stockholm University
2009

Stockholm University

Academic dissertation for the Degree of Doctor of Philosophy in Analytical Chemistry at Stockholm University to be publicly defended on Wednesday 16 September 2009 at 10:00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius v. 10-12, Stockholm, Sweden.
The defense will be held in English.

eleizam@hotmail.com

ISBN 978-91-7155-897-8

Cover: Fingerprints with different resolutions.

Printed by Universitetsservice US-AB, Stockholm, Sweden 2009

برای

مامان

که یه دنیا محبته!

Abstract

This thesis describes bioanalytical methods for generating fingerprints of biological systems for extracting relevant information with (protein) drugs in focus. Similarities and differences between samples can reveal the hidden relevant information, which can be used to optimize the production and facilitate the quality control of such protein drugs during their development and manufacture. Metabolic fingerprinting and multivariate data analysis (MVDA) can also facilitate early diagnosis of diseases and the effects and toxicity of drugs.

Protein-based drugs, produced from living sources (such as bacteria, yeast or mammalian cells) have been used to treat a number of diseases in humans. Currently, several protein drugs are available on the global market, and several hundred more are in clinical trials. Nevertheless, despite, the success of such biotherapeutics significant challenges remain to be overcome in maintaining their stability and efficacity throughout their production cycle and long-term storage. The native structure and functional activity of therapeutic proteins is affected by many variables from production to delivery, including variables associated with conditions in bioreactors, purification, storage and delivery. Thus, part of the work underlying this thesis focused on structural analysis of a protein drug (an antibody) using chemical labeling, peptide mapping, and evaluation of the charge state distributions of the whole protein generated by electrospray ionization. The other part focuses on non-targeted metabolomics with a view to optimizing the cell cultivation process and assessment of the drug's toxicity. A combination of appropriate analytical methods and MVDA is needed to find markers that can facilitate optimization of the cultivation system and expression of the target proteins in early stages of process development. Rapid methods for characterizing the protein drugs in different stages of the process are also required for quality control (for instance, to ensure that between-batch variations are acceptable and denaturation does not occur during storage).

In order to obtain high quality fingerprints analytical separation techniques with high resolution (such as high performance or ultrahigh pressure liquid chromatography) and sensitive analytical detection techniques (such as electrospray ionization, quadrupole or time-of-flight mass spectrometry) have been used, singly or in combination. This generates large datasets. The datasets are therefore simplified and

modeled by MVDA, after preprocessing by noise reduction, peak detection and alignment. The chemometrics methods used in this work included Principal Component Analysis, Partial Least Squares regression and Linear Discriminant Analysis.

List of papers

The following publications cover the context of this thesis:

I. Conformational studies of a monoclonal antibody, IgG1, by chemical oxidation: Structural analysis by ultrahigh-pressure LC-electrospray ionization time-of-flight MS and multivariate data analysis.
Leila Zamani, Fredrik O. Andersson, Per Edebrink, Yang Yang, Sven P. Jacobsson.
Analytical Biochemistry (2008), 380(2), 155-163.

The author was involved in developing the idea and was responsible for all experimental work (planning and performance) and writing the paper.

II. Discrimination among IgG1-κ monoclonal antibodies produced by two cell lines using charge state distributions in nanoESI-TOF mass spectra.

Leila Zamani, Jessica Lindholm, Leopold L. Ilag and Sven P. Jacobsson
Journal of the American Society for Mass Spectrometry (2009), 20(6), 1030-1036.

The author was involved in developing the idea and was responsible for all experimental work (planning and performance) and writing the paper.

III. Metabolic fingerprinting of rat urine by LC/MS. Part 1. Analysis by hydrophilic interaction liquid chromatography-electrospray ionization mass spectrometry
Helena Idborg, Leila Zamani, Per-Olof Edlund, Ina Schuppe-Koistinen, Sven P. Jacobsson.
Journal of Chromatography, B: Analytical Technologies in the Biomedical and Life Sciences (2005), 828(1-2), 9-13.

The author was responsible for analyzing the standards, and investigation and comparison of the ZIC-HILIC and C18 columns.

IV. Extracellular metabolite fingerprinting of spent mammalian cell culture medium analysis in relation to the quality of an expressed recombinant protein**.**
Leila Zamani, Emma Landegren, Eva Hedin, Anders Hagman, Nathalie Chatzissavidou and Sven P. Jacobsson.
Biotechnology Progress (2009), submitted.

The author was involved in the planning of all experimental work (except the sampling and protein measurements) and was responsible for the data evaluation and writing the paper.

Papers not included in this thesis:

Metabolic fingerprinting of rat urine by LC/MS. Part 2. Data pretreatment methods for handling of complex data
Helena Idborg, Leila Zamani, Per-Olof Edlund, Ina Schuppe-Koistinen, sven P. Jacobsson.
Journal of Chromatography, B: Analytical Technologies in the Biomedical and Life Sciences (2005), 828(1-2), 14-20

Hyphenated chromatography and chemometrics: breaking new grounds in the analysis of complex samples
Leonard Csenki, Erik Alm, Ralf Torgrip, Magnus Aeberg, Leila Zamani, Ina Schuppe-Koistinen, Johan Lindberg.
G.I.T. Laboratory Journal, Europe (2007), 11(1-2), 39-40.

Abbreviations

3Q	Triple quadrupole
BEH	Bridged ethyl hybrid
CCF	Central composite face
CD	Circular dichroism
CRM	Charge residues model
CSD	Charge state distribution
CV	Coefficient of variation
DC	Direct current
DoE	Design of experiment
ES	Electrospray
ESI	Electrospray ionization
ESI-MS	Electrospray ionization-mass spectrometry
FTICR	Fourier transform ion cyclotron resonance
H/D-MS	Hydrogen/deuterium exchange mass spectrometry
HILIC	Hydrophilic interaction liquid chromatography
HPLC	High performance chromatography
HSS	High strength silica
IEM	Ion evaporation model
Ig	Immunoglobulin
IM-MS	Ion mobility mass spectrometry
IMS	Ion mobility spectrometry
LC	Liquid chromatography
LC/MS	Liquid chromatography/mass spectrometry
LDA	Linear discriminant analysis
LOD	Limited of detection
LOQ	Limited of quantification
m/z	Mass to charge ratio
Mab	Monoclonal antibody
MALDI	Matrix-assisted laser desorption
MCP	Microchannel plate
Met	Methionine
MLR	Multiple linear regression
MS	Mass spectrometry
MVDA	Multivariate data analysis
nanoESI	Nanoelectrospray ionization
NMR	Nuclear magnetic resonance
PARAFAC	Parallel factor analysis

PAT	Process analytical technology
PC	Principal component
PCA	Principal component analysis
PLS	Partial least square
PTM	Post translation modification
Q	Quadrupole
Q-TOF	Quadrupole-time-of-flight
RF	Radio frequency
RSM	Response surface modeling
SPE	Solid phase extraction
TOF	Time of flight
TOFMS	Time-of-flight mass spectrometry
t_R	Retention time
UHPLC	Ultrahigh pressure liquid chromatography
UPLC	Ultra performance liquid chromatography
ZIC-HILIC	Zwitterionic chromatography-hydrophilic interaction liquid chromatography

Table of contents

Introduction

Human have always used substances that make them feel relaxed and stimulated. As time progressed, preparations were discovered (e.g. herbs, roots and mushrooms) that could be used to alleviate aches, pains and other ailments. These were all naturally occurring substances without refinement and isolation of specific compounds (drugs). Discovery and refinements of specific drugs to use in medicine occurred and was performed later by alchemists and medical experimenters (e.g. the discovery of ethanol and its refinement to use in medicine by a Persian chemist, physician and philosopher, *Zakariā-ye Rāzi*: (Persian: زکریـــای رازی, 865-925) [1].

There are strict requirements on today's pharmaceutical preparations in all steps of development, production, delivery etc. From the first idea to pharmaceuticals introduction on the market, many expertises in different field of science (e.g. physiologists, physicians, biochemists, organ chemists, analytical chemist, toxicologists and etc.) are involved and collaborate. Many different challenges have to be overcome on the way. One of the challenges that analytical chemists commonly meet is the analysis of complex samples.

Biological samples are generally complex, containing a wide range of compounds, in varying amounts, with widely differing chemical and physical properties, in a matrix containing salts, detergents and/or other potentially interfering substances. Thus, the analytical methods required to detect, quantify and/or characterize compounds of interest within them will depend on the physicochemical nature and quantities of both target analytes and other substances present in the samples. Typically, a single method will be insufficient to separate all types of analytes of potential interest, and interference from non-target substances, such as matrix compounds, can disturb the separation, detection and quantification of the analytes as well as the subsequent interpretation of the acquired data. Furthermore, the availability of instruments, cost, time and other practical considerations are factores that influence the choice of the analytical methods. HPLC and ESI-MS are some of the common separation and detection tools used for analyzing a wide variety of biological samples.

Ideally, the applied preparation techniques should result in samples containing only the analyte(s) of interest in appropriate forms and concentrations in a solution that is suitable for both separation and

detection. Suitable separation and detection methods can remove the bulk of the potentially interfering compounds and matrix substances, although in the real world it is often virtually impossible to eliminate all of the unwanted constituents while retaining all their constituents of interest. Further, even if the selected methods can handle the sample-related problems, there are still often problems e.g. associated with peak-shifts in LC-MS, chemical noise (caused by fluctuations in temperature, pH, concentrations of various substances etc.) and instrumental noise (with contributions from all of instrumental equipment used) [2]. The latter problems can also be overcome by suitable data processing and multivariate data analysis techniques, to some extent. Data processing and data analysis is therefore a very important part of the analytical chain in order to obtain relevant information. All steps prior to processing and analysis of the data are also of great importance for extracting reliable information, because generated data contain signals from multiple interferences, various kinds of chemical and instrumental noise and peak shifts, which may lead to loss of information, together with false positive and false negative results. Therefore, high quality data are essential for high quality information. In order to obtain high quality data, instruments with high resolution and sensitivity are essential, but not sufficient. Well-thought-out and carefully prepared experiments are also needed to extract reliable information. Therefore experimental design, analytical methodologies for fingerprinting biological samples and methods for extracting relevant information are described in this thesis.

Summary of papers

This thesis is based on four publications and is divided into three parts. Different issues with different analytical demands were addressed in each work. The first part (based on Papers I & II) is dedicated to methods for protein characterization by ESI-MS, particularly for the rapid characterization of therapeutic antibodies. The second part (based on Papers III & IV), discusses the application of two metabonomic approaches for screening, classifying and correlating the metabolite profiles of complex biological matrices. Finally, the third part (based on Papers I, II and IV) is intended to give a brief introduction to chemometric tools used in the work underlying this thesis.

This thesis illustrates how to create useful fingerprints of biological samples using appropriate mass spectrometric approaches (peptide mapping, metabolic fingerprinting and analysis of charge state

distributions of proteins generated by electrospray ionization), in conjunction with multivariate data analysis. Accordingly, the objectives of the research described in the papers were to develop methods for the structural analysis of a monoclonal antibody, metabolic profiling of rat urine samples and the analysis of spent cell cultivation media using high-resolution analytical techniques. The main foci are on extracting information regarding the structure of therapeutic proteins in cell culture production systems, and data that can facilitate optimization of their production.

The contents of the papers discussed in this thesis are briefly outlined below.

In the studies described in Paper I, chemical oxidation was used to label an IgG1 monoclonal antibody in native and denatured conformational states. Peptide mapping (fingerprinting) and PCA was then used to classify them.
In the studies reported in Paper II, the protein charge state distributions generated in nanoESI-MS, under the same experimental/instrumental conditions, were utilized as fingerprints to explore and classify, by PCA, the structural status of an antibody (in solution) produced by two different cell lines.
Paper III presents methods for screening rat urine for biomarkers, focusing on highly hydrophilic compounds. Data processing and multivariate analysis applied in this work are described in an accompanying paper by H. Idborg et al. [3] (not included in this thesis).
Paper IV reports analyses of the low molecular fraction of spent media used to cultivate cells producing a recombinant protein to identify correlations between the metabolic fingerprints and the quality of the expressed protein utilizing PCA and PLS regression.

For convenience, the studies described in Papers I-IV are referred to in the following text as Studies I-IV.

1

Part one

Protein characterization by mass spectrometry

Based on Papers I and II

1.1 Protein samples

1.1.1 Therapeutic antibodies

Antibodies (also known as immunoglobulins, Ig) comprise a class of antigen-specific, immunological proteins that are classified into five isotypes (IgA, IgD, IgE, IgG and IgM). IgG immunoglobulins, which are the major immunoglobulins in normal human serum, are monomeric and have molecular weights of ca. 150 kDa. They consist of four peptide chains, two heavy chains and two light chains, held together with a total of four disulfide bonds. Each IgG molecule has two antigen binding sites (Figure 1).

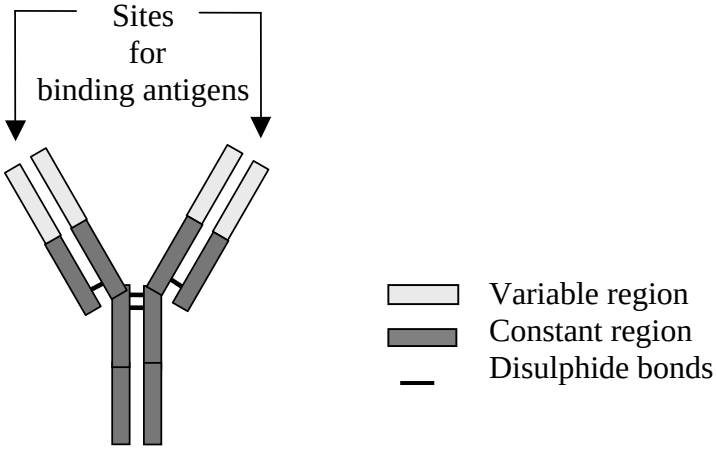


Figure 1. *Schematic diagram of an IgG-Antibody.*

Recombinant technology has allowed the production of many protein drugs, which could potentially be used to treat diseases in humans [4], by living production systems, such as bacteria, yeast or mammalian cells. Currently, several protein drugs are available on the global market, and several hundred more are in clinical trials.

Monoclonal antibodies (Mabs) are produced by a type of immune cell that all are clones of a single parent cell. Humanized Mabs (i.e. Mabs designed to circumvent clinical problems associated with immune responses to foreign antigens) have promising therapeutic applications,

as a result of their high efficacity and the establishment of robust technology to manufacture recombinant proteins [5].

1.1.2 Protein conformation

The conformations of a protein are heavily dependent on its amino acid sequence (primary structure) and the physicochemical characteristics of its microenvironment. In a hydrophilic solvent, most folded proteins have a hydrophobic core and the residues of polar chains are exposed to the surrounding environment. Some proteins must undergo post translation modification (PTM) before they can fulfill their functional roles. The final conformation of a protein in solution is a result of both covalent interactions (e.g. peptide and disulphide bonds) and noncovalent interactions (e.g. ionic bonds, hydrogen bonds and hydrophobic interactions) within the molecule and between the molecule and the solvent and sometimes other molecules. Disruption of these interactions often leads to alterations in their 3D-shape, which can expose hydrophobic residues that are normally located deep within them [6-7], and can also result in the loss of functional activity [8].

1.1.3 Protein produced by different cell lines

Humanized monoclonal IgG1s are frequently subject to PTMs. The main PTMs associated with proteins are acetylation, amidation, glycosylation, phosphorylation, carboxylation and sulfation [9]. Glycosylation (i.e. the addition of polymers composed of monosaccharides) is the most widespread PTM associated with current types of therapeutic proteins [10]. The glycoprofile of glycoproteins may vary substantially depending on the cell line and cell culture protocols used to produce them, with consequent effects on their biological activity, conformation, stability and solubility [10-11].

1.2 Sample preparation prior to LC/MS analysis

Sample matrices can pose considerable challenges in attempts to identify, quantify and/or characterize compounds of interest by liquid chromatography-mass spectrometry (LC/MS), or other analytical techniques, since interfering substances in them can dramatically reduce the resolution of LC, decrease the ionization efficiency of MS

and increase the resulting chemical noise, thereby increasing the limit of detection. Therefore, sample preparation is generally required to facilitate the isolation and concentration of compounds of interest from various matrix components, and hence the separation and detection of the analytes. Sample preparation techniques that may be applied include chemical and/or physical processes, e.g. oxidation and filtration, respectively.

1.2.1 Chemical processing

Typical steps in chemical sample preparation include addition of a group, a fragment or a whole molecule to a target analyte and/or degradation or decomposition of analytes into smaller fragments. Groups may be added either for labeling (to facilitate the detection of compounds of interest), or derivatization for various purposes (e.g. to introduce specific, functionalized groups to enhance separation, to increase detection sensitivity, to protect target molecule/groups and/or for molecular structure elucidation). A technique frequently used to degrade analytes (especially proteins) into smaller fragments is enzymatic digestion. Oxidation of proteins, for labeling, and enzymatic cleavage were applied in sample preparation in Study I.

Oxidation of proteins by hydrogen peroxide

Oxidation of proteins by peroxides occurs by non-site-specific reactions. Methionine (Met) residues in the proteins are the most susceptible to oxidation by all forms of reactive oxygen species [12]. The rate of oxidation of methionine residues is affected by many factors, e.g. the accessibility of solvent containing the oxidant and the residues close to methionines. In general, the reactivity of an individual methionine residue is greater if it is exposed to the solvent rather than being buried [13-14]. Oxidation by hydrogen peroxide was used in Study I to explore the conformations of the investigated Mab.

Enzymatic digestion

Specific cleavage of a polypeptide can be achieved by chemical or enzymatic methods. Cyanogen bromide (CNBr), which cleaves polypeptide chains only on the carboxyl side of methionine residues, is an example of a reagent that allows specific chemical cleavage. Trypsin also cleaves polypeptide chains highly specifically, on the carboxyl side of arginine and lysine residues when the following residue is not a proline residue. The peptides resulting from tryptic digestion are

typically of small average size, because of the high arginine and lysine contents of most proteins. To cleave folded proteins, and obtain access to the residues hidden inside them, other steps are necessary before the cleavage. For a protein consisting of two or more polypeptide chains held together by noncovalent bonds, denaturing agents such as guanidine hydrochloride or urea are used to dissociate the chains and reducing agents such as dithiothreitol can be used to dissolve polypeptides linked by disulfide bridges. To prevent reduced cysteines reforming bridges they are alkylated with iodoacetate to form stable derivatives. Digestion is usually performed after denaturation, reduction of disulfide bonds and alkylation of the cysteine residues.

Peptide mapping

After protein molecules have been cleaved using specific enzymatic or chemical methods, the resulting peptides are usually separated chromatographically to obtain a "fingerprint" by "peptide mapping". For reproducible fingerprinting, it is very important to perform all steps carefully, otherwise peptides of varying size may be obtained, e.g. if a disulfide bond is not always cleaved, resulting in the generation of two different peptides attached to each other after the tryptic digestion and (hence) different fingerprints for the same protein. Peptide mapping by LC/MS, which generates unique fingerprints for individual proteins, is used in protein characterization, and was applied in Study I to generate reliable fingerprints of a monoclonal antibody in different conformations for classification.

1.2.2 Non-chemical processing

Commonly applied non-chemical preparative methods include filtration, adsorption, phase partitioning and/or passage through media such as "ultra-filtration gels" in which the speed of migration of compounds depends on their size. In gel filtration small molecules are retained in pores in the gel, while macromolecules pass rapidly through it and are quickly eluted. Dialysis, which is a membrane liquid-liquid extraction technique, is another type of non-chemical process that can be used to separate macromolecules from other constituents of samples. In Study II, gel filtration was applied for buffer exchange, and in Study I ultra-filtration was used to stop the oxidation process mentioned above by washing out the oxidant through the filter.

Another relatively new technique for protein sample clean-up prior to MS analysis is microfluidic electrocapture [15], in which analytes of

interest are separated from the matrix and interferents. Electrocapture has been shown to be especially useful for cleaning-up membrane proteins from detergents [16].

1.3 Mass spectrometry

Generation of ions in gas phase from compound/complex of interest is a prerequisite for any mass spectrometry (MS) experiments. Mass spectrometers are typically composed of several parts: an ion source, an analyzer, a detector, data handling system. Vacuum pumps (rotary vacuum pumps in combination with turbo-molecular pumps) are also needed to provide high vacuum to avoid ion collision. Each of these parts is discussed below, focusing on mass spectrometry of large molecules. Ions generated in the ion source are transferred, separated and resolved in the mass analyzer (e.g. a quadrupole or time of flight system, see below). The ions that survive this journey and reach the detector (e.g. a photomultiplier, MCP) will be detected. The type of detector used should, of course, be compatible with the type of analyzer. A simplified schematic diagram of a mass spectrometer is shown in Figure 2.

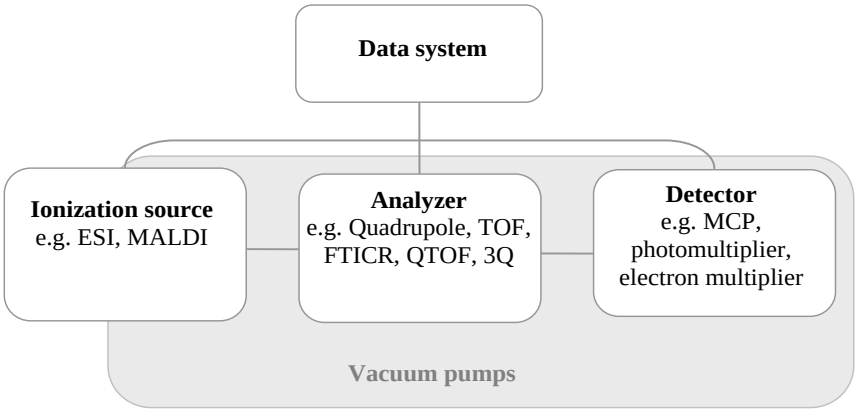


Figure 2. *Simplified schematic diagram of a mass spectrometer.*

1.3.1 Ionization (MALDI and ESI)

There are two main methods for ionizing whole proteins: electrospray ionization (ESI) [17] (which will be discussed below) and matrix-assisted laser desorption ionization (MALDI) [18], in which a laser beam vaporizes a spotted sample, dried on a metal target plate, and facilitates the ionization of compounds within it. The matrix often protects the biomolecules of interest from fragmentation by the direct laser beam, but some of the energy absorbed by the matrix helps ionize the analyte molecules (e.g. proteins) while still protecting them from the disruptive energy of the laser. These two different ionization techniques in MS (ESI and MALDI) were a part of the subject of Nobel Prize award in chemistry for 2002.

ESI is a technique for transferring ions from a liquid phase to the gas phase. The ionization process occurs at atmospheric pressure and (often) ambient temperature (when applied to proteins). In ESI, a liquid is passed through a tiny conductive capillary, which is coupled to a high potential (for example, 3-5 kilovolts) relative to a counter electrode and thus creates a spray of the solution. The liquid normally contains the analytes and a volatile buffer, which is usually much more volatile than the analytes. An inert gas such as nitrogen is usually used as a carrier, nebulizing, gas which flows in the same direction as the sample, and another stream of a neutral gas, such as nitrogen, is passed across the front of the ionization source to help evaporate the uncharged solvent molecules neutral solvent in the droplets. Furthermore, ions are created by the addition of (a) proton(s) or cation(s) such as Na^+, denoted $[M + nH]^{n+}$ or ($[M + nNa]^{n+}$), respectively, in positive ionization mode, or the removal of proton(s) or addition of anion(s) such as Cl^-, denoted $[M - nH]^{n-}$ or ($[M + nCl]^{n-}$), respectively, in negative ionization mode. Large molecules such as peptides and proteins give rise to multiply charged $[M+nH]^{n+}$ ions in positive ionization mode and $[M-nH]^{n-}$ ions in negative ionization mode. ESI is preferred to MALDI for intact non-covalent analysis of macromolecular complexes, because the sample preparation in MALDI requires acidic conditions, which may disturb the intermolecular interactions that occur in solution. Additional advantages of ESI for analyzing macromolecules include the possibility of producing multiply charged ions of large analytes. This feature makes ESI suitable for analyzing large molecules/complexes when used in combination with a mass analyzer that has a limited m/z range.

Nanoflow ESI (nanoESI)

NanoESI [19] is a low-flow version of ESI, in which a small volume (1-4 µl) of the sample dissolved in a suitable volatile solvent is sprayed through a thinner capillary (~ 1 µm inner diameter) using a sufficiently high voltage (ca. 700 - 2000 V). The flow rate of solvent using this procedure is very low (nl/min). NanoESI is a suitable choice for analysis of non-covalent complexes, because milder desolvation conditions are required as a result of the smaller sizes of droplets produced in the nanoESI source [20]. ESI and nanoESI generate similar types of spectral data for samples, so the subsequent data processing procedures are identical.

Mechanism of ESI

A critical issue concerning the ESI process is the origin of the charges carried by the gas-phase ions. It was originally suggested that the charge states observed in ESI-MS reflect the actual ionization states of the molecules in solution [21-22]. However, several lines of experimental evidence indicate that such a correlation between solution and gas phase charge states does not hold in general [23-25]. The ESI process can be described in three steps: production of charged species in droplets at the ES capillary tip as a result of oxidation-reduction reactions, shrinkage of the size of the charged droplets and formation of the gas-phase ions. Some parts of the mechanism are still not fully understood, particularly formation of gas-phase ions from charged droplets.

When a very high electric field is applied to the conductive ES capillary tip, partial separation of positive and negative electrolyte ions in solution occurs. In order to obtain positively charged ions (in positive mode), the capillary is exposed to a positive electric potential, creating a deficit of electrons, which draws the negative ions inside the capillary (against the flow), while the positive ions are enriched at the surface of the liquid at the capillary tip. The “Taylor cone” formed by the liquid jet [26] is expanded as a result of repulsion of the positive ions at the surface and the force of the electric field. The tip of the cone is less stable where the surface tension is not able to hold the surface together, resulting in formation of individually charged droplets with positive charges at the surface. Initiation of the electrospray requires an electric field E_0 with a threshold strength of [27]:

$$E_0 = \sqrt{\frac{2\gamma \cos 49°}{\varepsilon_0 r_c}} \quad (1)$$

where γ is the surface tension of the liquid (0.030 and 0.073 N/m for acetonitrile and water respectively), 49° is the half-angle of the Taylor cone, ε_0 is the permittivity of vacuum and r_c is the outer radius of the emitter tip. As can be seen from the formula, the surface tension has a substantial influence on the electric field required for the onset of ES; the higher the surface tension of the liquid, the stronger the electric field required for occurrence of the ES process, e.g. water has higher surface tension than methanol, therefore small charged droplets form more readily when methanolic samples are analyzed than when the samples are aqueous. The droplets deform (as a consequence of the electric field) and evaporate during their progress toward the MS inlet, causing them to shrink while the charges remain. Hence, the charge to volume ratio increases until the Rayleigh stability limit [28] is reached, at which point the repulsion of the charges overcomes the surface tension and the droplets undergo Coulombic fission, resulting in smaller droplets, undergo further evaporation and fission and thus become very small. The timescale of gaseous-ion production is less than 0.5 ms, and the total residence time of charged droplets in the ESI interface is a few ms [29]. Two models have been postulated to explain the formation of gas-phase ions from the very small droplets: the ion evaporation model (IEM) [30] and the charge residue model (CRM) [31]. The IEM assumes that the gas-phase ions are emitted from the small droplets, while the CRM regards the gas-phase ions as products of the evaporations and fissions resulting in single ions.
The consensus in the literature is that neither the CRM nor the IEM fully describe all experimental observations. Depending on the type of the analyte, one or the other, or both, may be involved in ion formation. For example, multiply charged globular proteins which are not denatured in the solution are probably largely produced by CRM, while ions of more open structures, such as denatured proteins, may be produced by both CRM and IEM [32].

A number of instrumental parameters influence the stability and efficiency of the spray, and thus the MS spectra generated. The applied voltage, curtain gas flow, the position and distances between the capillary and the counter electrode, the flow of the solution and the

inner diameter and shape of the spray tip are examples of parameters that can influence the efficiency of ionization and, hence, the appearance MS spectra.

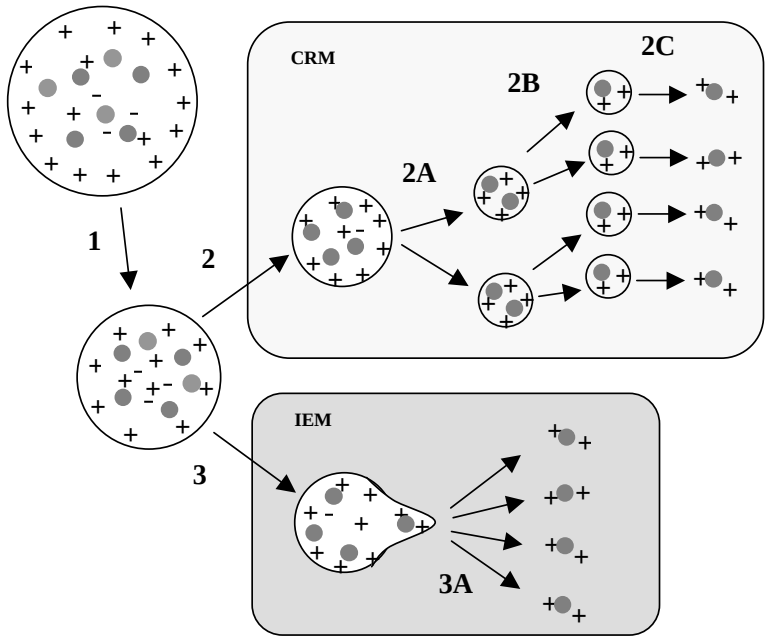


Figure 3. *Schematic diagram of the mechanisms of ion formation in ESI, according to two models for gas-phase ion generation by ESI: the ion evaporation model (IEM) and the charge residue model (CRM). 1, 2, 2A, 2B and 3 = Solvent evaporation and droplet fission in multiple steps, 2C= Solvent evaporation, 3A= ion evaporation.*

Charge state distribution

Proteins acquire many charges during the ESI process, and the multiply charged proteins give rise to a series of peaks, in which adjacent peaks differ by one proton. The mechanism controlling protein charge state distributions (CSDs) is not yet well understood [25, 28]. However, compact conformations typically result in lower charge states (higher m/z values) than unfolded conformations of the same molecule. The charge attained by protein molecules during ESI is highly affected by

the 3D-structure of the proteins in the spray solution [33]. Two hypotheses have been formulated in attempts to explain these experimental results. The first is that the lower net positive charge in positive ESI of more compact protein molecules can be explained by the lower solvent accessibility of basic residues in native proteins than in denatured proteins [33]. The second hypothesis is that higher charge states in the unfolded conformations are stabilized by increasing the distance between charges of the same polarity in the proteins [34]. However, the compactness of the protein is only one of the factors affecting the CSDs. Since the detection occurs in a solvent-free environment, various gas-phase processes may also affect them [35]. In addition, both the instrumental settings (e.g. curtain gas flow rate and desolvation temperature) and experimental settings (e.g. pH and solvent surface tension) influence CSDs [36]. The mechanism of transfer of protein ions from the solution to the gas phase is not completely understood, but understanding the influence of experimental conditions on protein CSD is important in order to extract the structural information that they can contain. The desolvation temperature and amount of organic modifier used to decrease the surface tension of the solvent are just two examples of instrumental and experimental variables, respectively, that can affect protein stability during the electrospray process, resulting in proteins unfolding. The connection between structural compactness and the apparent charge state distribution allows the folding and unfolding of proteins during ESI-MS to be tracked, provided that experiments are performed under tightly controlled instrumental and experimental conditions, since (as outlined above) both instrumental and environmental variables can strongly influence the results [36].

Protein CSDs generated by nanoESI can also be strongly affected by the geometry of the nanoflow needle, thus some needle-to-needle irreproducibility is to be expected [37], and minor conformational shifts, which cause slight shifts in CSDs, are difficult to distinguish from CSD-shifts caused by other effects [25].

CSDs were utilized in Study II to discriminate between a monoclonal antibody produced in two different cell lines, applying experimental and instrumental conditions selected to maximize the differences in the resulting CSDs of the protein.

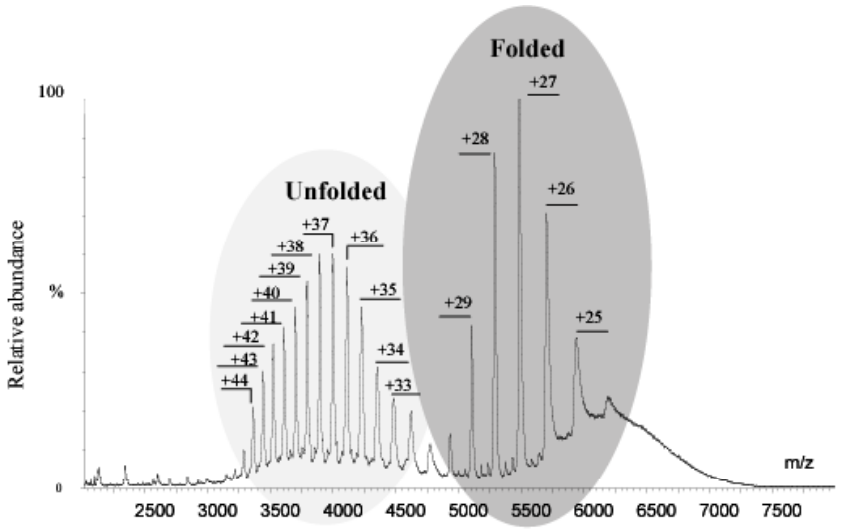


Figure 4. *Mass spectra showing the respective charge state distributions of IgG in native and unfolded states.*

1.3.1 Mass analyzer (Q and TOF)

Mass analyzers separate and resolve the ions according to their mass-to-charge (m/z) ratios using magnetic or electric fields in vacuum. Types of mass analyzer that can be used for protein structural analysis include quadrupole (Q), ion trap, time of flight (TOF), fourier transform ion cyclotron resonance and tandem (triplequadupole, Q-TOF) instruments [38-39]. These mass analyzers have different features, and they differ in terms of mass accuracy, m/z ranges that can be transferred/isolated and mass resolution. A quadrupole setup consists of four parallel rods. Each opposing rod pair is connected to electric field and radio wave generating equipment, which provides a fixed positive and negative direct current (DC) and alternating radio frequency (RF). The gas-phase ions generated in the ion source are focused and passed along the middle of these rods, which can act as a "mass filter", i.e. the combination of DC and RF potentials on the quadrupole rods only allow ions with a selected range of mass-to-charge ratios to pass through the analyzer. Ions with other m/z ratio do not have a stable pathway and collide with the quadrupole rods, never reaching the detector. In 3Q systems the first and third quadrupoles (Q1

and Q3) act as mass filter, but the second quadrupole (Q2) is an RF-only device and acts as a collision cell and can only transmit the ions. The ions passing Q1 are fragmented by colliding them with variable amounts of inert gas, such as Argon, Helium and nitrogen and differing collision energies. The degree of fragmentation is detemined by the applied voltage and type of collision gas. The resulting ions from the collision cell pass through to Q3 for filtering. A quadrupole is usually able to filter and transmit ions with m/z ratios up to 4000 Thomson (Th), and 12000 Th, respectively.

In a TOF analyzer the ions are all accelerated by a high voltage, typically 20-30 kV, giving them approximately the same kinetic energy. The accelerated ions “fly” through a long field-free tube in vacuum, and ions with different m/z ratios reach the detector at the other side of the tube at different times. Thus, the time-of-flight can be used to determine the mass-to-charge ratio of the ions. Ions with large m/z ratios take longer time to fly through the tube than ions with smaller ratios, which results in ion separation. The time ions have to separate from each other depends on the length of the flight tube. A way to increase the path length, and thus the resolution, is to use a reflector. Ions with same m/z ratio coming from same target can have different speed, due to uneven initial energy distribution, when they from the flight tube enter the reflector, which they penetrate, and are stopped and reflected. Ions with higher kinetic energies (higher speeds) will penetrate deeper into the reflector before they are stopped and turned around, thus their flight times are longer than those with lower kinetic energies, which allows more peak separation. Theoretically, TOF has no m/z-range limitation for transmission of the ions; the limitation the low efficiency of the microchannel plate detector (MCP) for low-velocity ions, such as macromolecules.

A TOF analyzer was used for separation of the peptides resulting from tryptic digestion in Study I and a Q-TOF was used for separation of whole proteins (~150 kDa) in Study II.

Collisional cooling

In ESI ions are generated in a flowing gas stream (jet) at atmospheric pressure and introduced into the mass spectrometer following staged pressure reduction. The larger ions and macromolecules (M > 50 000 Da or m/z > 4000 Th) are usually difficult to detect, partly because of transmission losses when they are accelerated in the gas jet in the ion guide chamber (Q0). The transmission of ions with very high axial

energy may result in them partially or completely missing the detector [40-41]. To improve the ion transmission of large ions, the most common strategy is to operate the ion guides at higher pressure, called collisional cooling, which improves both the resolution and sensitivity of the Q-TOF MS. In Study II collisional cooling was used, by increasing the pressure in Q0 and Q2, for focusing and transmitting the antibody ions (m/z ≈ 6000 Th). The pressure in Q0 was increased by adjusting a valve in Q0 space.

Given the higher pressure regime an additional benefit can be obtained. Ions that are not fully desolvated result in broader spectral peaks (lower mass resolution) and with increased detected masses because of the attached solvent molecules [42]. Increasing the collision gas (e.g. He) and the collision energy in Q2 results in higher resolution since collision with the gas promotes desolvation of macromolecules from solvent species, such as the buffer molecules (e.g. ammonium acetate). This occurs if the concentration of the volatile buffer is sufficiently high to efficiently out-compete other ions present in the samples, which are usually less volatile adduct-forming cations (such as Na^+ and K^+) in the ionization process. After entering the mass spectrometer, especially in the Q2 through collision with the gas molecules, the volatile molecules will separate from the macro ions, which results in higher resolution and narrower mass spectral peaks. Figure 5 shows how a volatile buffer at high concentration can promote the resolution of the ions, and the increase in the detected mass when the ion peaks are broader (less desolvated).

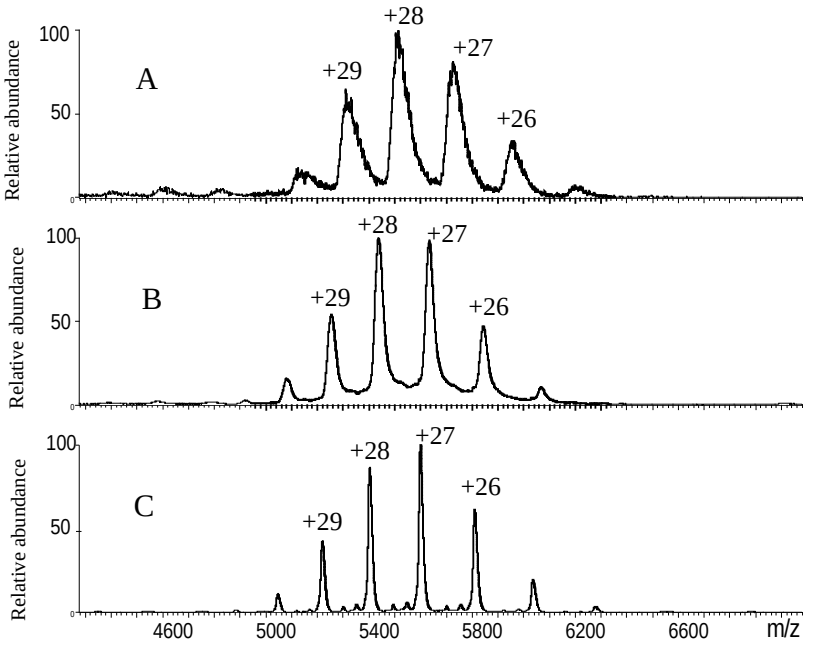


Figure 5. *Nano-ESI mass spectra of the IgG1 in (A) 0.2 M, (B) 0.5M and C) 1 M ammonium acetate buffer, pH 7.*

1.3.3 Detector (electron multiplier)

Electron multiplier [43], photomultiplier [2] and microchannel plate (MCP) [44] are some of the types of detectors used in MS instruments. A process known as secondary electron emission is used in electron multiplier detectors. Ions that survive the journey through the analyzer and reach the detector hit a surface, causing release of electrons from the outermost area of the atoms (secondary electrons). Depending on the particle type, the angle and energy at which the ion hits the surface, different numbers of secondary electrons are released, and the more that are released, the stronger the response for the particle. Different types of ions induce different responses by the detector. The mass spectrum obtained for each ion depends on the ionization efficiency, focusing, transmission and response of the detector for that particular ion. An MCP is an array of a large number of miniature electron multiplier channels parallel to each other, which has a response time of

less than 1 ns and is able to detect many ions simultaneously. To enhance sensitivity, two parallel arrays of microchannel plates can be used, but use of more than two plates is not recommended because of the associated increases in noise. MCPs are appropriate detectors for TOF analyzers, which provide high resolution of separation and hence require a fast detector.

1.4 Mass spectrometry and protein structural analysis

Mass spectrometry has major biotechnological (e.g. studies of proteins, peptides [45], pharmaceutical (e.g. drug discovery and drug metabolite analysis; [46], clinical (e.g. drug testing; [47], environmental (e.g. water quality and food contamination monitoring; [48-49] and geological (e.g. oil composition analysis; [50] applications. The speed, specificity and sensitivity of mass spectrometry make it particularly attractive for rapid protein identification and characterization [51-52]. Here, however, the focus of the discussion is on the application of mass spectrometry for studying the characteristics and structures of therapeutic IgG1 proteins.

A number of other structural techniques, such as X-ray crystallography [53], nuclear magnetic resonance (NMR) spectroscopy [54] and circular dichroism (CD) spectroscopy [55] are often currently used for structural and conformational analysis of proteins. These techniques provide complementary information regarding the structure and conformation of proteins, but they all have certain practical drawbacks. X-ray crystallography requires high quality crystals, which can only be formed from proteins of very high purity. NMR can be used to characterize protein molecules in concentrated solutions, but has only been applicable to relatively small proteins of at most 30-40 kDa. CD is an established technique for studying the secondary structure of proteins. The application of mass spectrometry in protein conformational analysis is relatively new, but MS can be applied in either of two ways for studying protein folding. One approach is to label proteins in a way that results in chemical differences depending on differences in their conformation. An example of this approach is hydrogen/deuterium exchange mass spectrometry (H/D-MS), in which amide protons in the protein undergo isotopic exchange depending on

solvent accessibility, which is directly related to protein structure [56]. Exchange of an amide proton with a solvent deuterium, referred to as hydrogen/deuterium (H/D) exchange, requires both solvent accessibility and breaking of the amide hydrogen bond [57]. Therefore, such sites that are involved in secondary structure and/or are not accessible by the solvent exchange more slowly, due to H-bonding interactions and shielded from the solvent, than those that are exposed and unbounded [58]. Thus, kinetic studies of protein folding and unfolding can be performed using ESI-MS in conjunction with on-line labeling [59-60]. In Study I two conformational variants of the same protein were probed using hydrogen peroxide oxidation in conjunction with peptide mapping by LC/MS analysis, and fingerprints with distinct differences were obtained for them.

Another approach is to directly exploit protein CSDs generated by ESI-MS to obtain information on the compactness and conformation of proteins in solution before they transfer to the gas phase. Compact conformations typically result in lower charge states (higher m/z values) than unfolded conformations of the same molecule. In a very general sense this technique (protonation during ESI) can be regarded as analogous to the covalent labeling methods mentioned previously. The previous methods (e.g. H/D exchange) cause modifications that can be precisely located in the protein by peptide mapping and MS/MS. In contrast, protonation of a protein during ESI does not yield regio-selective information about changes in the conformation. CSD was used as a probe for classifying the same protein produced in two different cell lines in Study II.

Mass spectrometry combined with ion mobility spectrometry (IMS), ion mobility-mass spectrometry (IM-MS), has recently been introduced as a new tool for structural analysis of protein complexes [61]. In IM-MS ions are separated based on both their drift times in a gas-filled ion mobility chamber and their m/z ratios. Different molecules with different shapes, charge and volumes have different cross-sections; Molecules of the same charge with larger cross-sections have longer drift times. IM-MS offers valuable data that cannot be obtained from mass spectra alone, allowing (for instance) the separation of isomers and conformers. The latest version of an IM-TOFMS, which is commercially available from Waters, is a traveling-wave ion mobility spectrometer [62], called the Synapt HDMS.

2

Part two

Metabolic fingerprinting
(Non-targeted metabolomics)

Based on Papers III and IV

2.1 Metabolic fingerprinting

In a living organism diverse chemical reactions occur that allow the organism to grow, maintain life processes and interact with its environment. These chemical reactions (metabolism) result in large numbers of intermediates and end-products (metabolites) in widely varying concentrations. Metabolomics, which is the study of the metabolite complement of a living system, provides valuable indications of what has actually happened in the living system. The words “metabolomics” and “metabonomics” are often used interchangeably, although the former has been defined as comprehensive profiling, in which attempts are made to identify and quantify all the metabolites of a biological system [63], while 'metabonomics' has been defined as the quantitative measurement of dynamic multiparametric metabolic responses of a living system to changes in endogenous metabolite levels that result from disease or therapeutic treatments [64]. Metabonomics is an effective tool and has great potential in toxicological studies and searching for biomarkers of diseases. Useful information can be obtained from defining different metabolomes and understanding how changes in the concentrations of metabolites relate to health and disease. The main challenges in metabolomic studies are posed by the dynamic changes in levels of the metabolites. The composition of small molecules in biological samples (constituent metabolites) is a metabolic fingerprint that is unique for a specific system at a specific time [65]. Such fingerprints are influenced by both the genome of the organism and exposure to environmental variables. NMR and LC/MS are typical analytical techniques that have been used for producing metabolic fingerprints [66]. The investigation of metabolites can be performed in either of two ways: targeted or non-targeted metabonomics. Targeted analysis includes identification and quantification of pre-known metabolites or metabolite classes, while non-targeted metabonomics involves non-biased determination of as many metabolites as possible in biological samples. This part of the thesis is focused on analytical applications of non-targeted metabonomics.

2.1.1 Biomarker

A biomarker [67] is an indicator (in the form of a specific compound or combination of specific compounds) of a particular state of a biological

system. Biomarkers have been used (*inter alia*) in diagnosis, prognosis and toxicity studies. Utilizing biomarkers for these purposes is highly convenient, since instead of examining the entire spectrum of components of complex biofluids, a single compound, or group of compounds, can be determined.
A lot of work is involved in finding a validated potential biomarker, and collaborations between many researchers with expertise in different fields are generally required to identify a "good" (reliable and convenient) biomarker. Key objectives from an analytical perspective are to find robust relationships between relative levels of the compounds and (for instance) a particular disease, and to ensure that the variations are related to the disease rather than variations between individuals, or other factors such as the stress caused by the sampling or dosing if, for example, a drug has been administered to animals.

2.2 Mammalian cell cultivation

Cell cultivation is the process of growing cells under controlled conditions, for diverse purposes, but here it refers to the cultivation of bacterial, yeast or mammalian cells to produce a particular protein from a cloned gene. Various devices and systems can be utilized for this biotechnological process, called bioreactors. Mammalian cell lines have become increasingly important for producing recombinant therapeutic glycoproteins. The quality of the expressed protein in a bioreactor depends on many factors including (*inter alia*) the nutrients supplied (glucose, amino acids, vitamins, trace elements, compounds required for nucleotide and lipid biosynthesis, etc.), and many other environmental variables (such as temperature). The growth and production limitations depend on the availability of these nutrients and the accumulation of toxic metabolic by-products from the cells (such as ammonia and lactate) [68-69]. In order to design and control cell cultivation processes the pharmaceutical industry applies the process of analytical technologies (PAT) [70], the aims of which are to ensure the final product quality in an effective way. For this purpose, a good understanding of factors that contribute to productivity and the variability in cell cultures is important, and combinations of appropriate analytical chemistry and multivariate data analysis techniques are

important tools for screening or identifying important parameters for optimizing the cultivation process.
Metabolic fingerprinting yields sample-specific patterns, which can be used: to classify proteins produced by cultures grown under different conditions, to facilitate characterization of samples, to identify appropriate markers that are indicative of the conformation of biopharmaceutical proteins during cell cultivation, and to acquire a better understanding of cellular metabolism. This methodology was used in Paper IV.

2.3 Samples

The biological samples used in the metabolomics studies this thesis is partly based upon consisted of rat urine and spent cell culture medium. Biological samples are usually complex and difficult to analyze because they contain large numbers of compounds, such as small metabolites, salts, proteins and lipids at various concentrations. Further, different biological samples carry different kinds of information. For metabolomics studies urine is a relevant biofluid to analyze, because the water-soluble waste of the body is filtered from the blood and excreted in the urine, hence urine contains various metabolites. The other advantages of using urine samples are that urine is produced in large amounts relative to most other biofluids, it is easy to collect and the amounts of proteins and lipids in urine are negligible. The main constituents of urine are water (which generally accounts for 95% of its mass), mineral ions and several organic compounds, including urea, sodium chloride and creatinine. In addition, there are many other compounds, mostly polar small organic compounds at substantially lower concentrations [71]. A difficulty associated with analyzing urine is that the composition of urine produced by a given individual varies diurnally and also depends much more strongly on the individual's liquid-intake than blood or plasma, which have more stable compositions over time. Spent-cell culture medium is also a biological fluid; the difficulty with this matrix is that it represents both a source of nutrients and a waste matrix, containing both nutrients and eliminated toxic compounds. Proteins, lipids, salts, nutrients and metabolites are all likely to be present in such a matrix at varying concentrations.

2.4 Preparation of biological samples

Various techniques have been used to prepare samples of diverse tissues and matrices composition prior to LC/MS analysis. The samples used in Study III all consisted of rat urine, and the objective was to separate as many compounds (both polar and nonpolar) as possible for screening. For this purpose, solid phase extraction (SPE) [72] was initially used to fractionate the urine samples into two fractions (polar and nonpolar), which were separately analyzed. The samples used in Study IV all consisted of spent cell culture medium and again the objective was to separate as many possible metabolites as possible, in this case using centrifugal filtering as a preparative technique.

2.4.1 Solid phase extraction

In solid phase extraction (SPE) liquid samples are passed through a pre-conditioned solid sorbent bed, and compounds retained in the solid phase bed are thus separated from the other constituents, which pass through to the waste. More liquid is usually added to wash out the matrix remaining in the bed without disturbing the analytes of interest. The last step is eluting the analytes from the solid bed, which should be done effectively with low volumes of a suitable solvent. Otherwise the analytes will be diluted or present in an unsuitable solvent for subsequent analysis, e.g. by LC/MS. In such cases, the problems are addressed by additional steps for concentrating the analytes by solvent evaporation and/or solvent exchange. Wide ranges of sorbents and formats for SPE applications are available, and the technique can be used for either off-line or on-line preparation. SPE is usually used for clean-up, concentration and fractionation of samples. In Study III, SPE was used to fractionate the rat urine samples into two – retained (eluate) and non-retained (wash) – fractions. The stationary phase used in this work was a hydrophilic-lipophilic balanced copolymer SPE cartridge, which retains more polar compounds than C18 columns. The reason for analyzing both fractions (elute and wash) was that many polar compounds are expected to be present in urine. The fractions from the eluates and washes were subsequently separated on C18 and ZIC-HILIC analytical columns, respectively, which are discussed below.

2.4.2 Centrifugal filtering

To separate the macromolecules from a sample, one of the options is ultrafiltration [73-74], in which hydrostatic pressure (generated by centrifugal force) is used to force species of a smaller than threshold size through a filter with pores of a suitable size, which retains particles/molecules larger than the pores on its surface. This technique provides a convenient way to remove the proteins and other macromolecules from samples. It is also suitable for concentrating macromolecules and/or for buffer exchange. In Study I (as mentioned before) ultrafiltration was used for washing out the oxidation reagent (to stop the oxidation) then the protein retained by the filter was recovered and analyzed further, while in Study IV the fraction that passed through the filter (containing small molecules) was collected and analyzed further. The fraction containing macromolecules (>10 kDa) retained by the filter was also collected for further analysis, but these results are not reported here.

2.4.3 Protein precipitation

Another way to remove the proteins from biological samples is by protein precipitation [73], in which proteins are denatured directly in the initial biological samples and form non-soluble aggregates that precipitate. To denature and precipitate the proteins a water-miscible organic solvent (e.g. methanol, ethanol or acetonitrile) or a strong acid (e.g. trichloroacetic acid) can be used and the precipitant is removed by centrifugation. The drawback is that some small molecules that interact with the proteins may be removed together with the proteins. Protein precipitation was one of the protein removal methods evaluated in Study IV. However, small molecules were lost to a higher extent when protein precipitation was used compared with ultrafiltration. Therefore ultrafiltration was the method of choice for that work.

2.5 LC/MS and metabolomics

For metabolites in a state of flux, there is currently no ideal single detector. NMR and MS (often coupled to LC) have been frequently used for this purpose [75], but each has specific strengths and weaknesses. One of the main strengths of NMR is that the biological samples do not require treatment prior to the analysis, while MS can only be used to detect metabolites after they have been separated. On

the other hand, MS is very sensitive for metabolite detection and identification. Indeed, in terms of sensitivity MS is probably the best method, though absolute quantification of all individual compounds requires calibration curves for each compound, which is not realistic due to the large numbers of metabolites present in most samples, and the fact that many metabolites have not yet even been identified. LC/MS methods only provide relative values for each of the detected compounds (increases or decreases relative to a control or reference sample). Furthermore, there is no single separation method that works for all metabolites. However, combining separation and detection methods with high sensitivity and resolution can result in high-resolution fingerprints, facilitating the detection of subtle metabonomic changes.

LC is a suitable technique for separating analytes in an aqueous sample prior to quantitative analysis. Separation of compounds in LC is based on variations in the partitioning of the compounds between the stationary and the mobile phase, which affects the time (i.e. retention time, t_R) each analyte takes to migrate from injector to the detector. It is very important to ensure that the migration path from injector to the LC column (where separation occurs) and from the column to the detector is as short and thin as possible to avoid peak broadening (which compromises quantification, causes peak overlap, and increases both the limit of the detection, LOD, and limit of quantification, LOQ). The t_R is longer for an analyte that tends to partition more into the stationary phase than for other analytes that partition more strongly into the mobile phase. Use of a combination of different stationary phases and mobile phases of different compositions offers various separation modes and selectivity and t_R, respectively. Other parameters that may affect results include the injection volume, flow rate and column dimensions (particle size, length and diameter). Normal phase, reversed phase (RPLC), ion exchange and hydrophilic interaction liquid chromatography (HILIC) are four different modes of separation [76]. Mobile phases with differing compositions are compatible with each of them. A higher pressure is needed to push the mobile phase through the packing of the column when the particle size is smaller, but the smaller the particles the more efficient the column will be. The pressure in a standard high performance liquid chromatography (HPLC) system can go up to approximately 5800 psi. By decreasing the particle size, speed and peak capacity can be extended to new limits, but the system must be able to handle substantially higher pressures. Hence, manufacturers

have developed “ultra high pressure chromatography (UHPLC)” instruments that are capable of handling back-pressures of up to 15000 psi. [77-78]. UHPLC columns packed with particles smaller than 2 µm create high back-pressures that exceed the limit of traditional HPLC systems. One of the advantages of UHPLC over conventional HPLC is the capacity to increase the speed of analysis without compromising efficiency because the Van Deemter curves tend to flatten out at higher linear velocities, when using the smaller particles [79].

In Study III, two HPLC setups were used for the analysis of urine samples and in Study IV, UHPLC was used to analyze the low-molecular weight fraction of the spent cell culture medium.
The stationary phases used in Study III were C18 (RPLC) and zwitterionic chromatography-hydrophilic interaction liquid chromatography (ZIC-HILIC) columns [80]. The ZIC-HILIC column, which has a silica-based stationary phase with highly polar permanently zwitterionic substituents (thereby retaining a permanent aqueous layer on the stationary phase surface), has reversed selectivity compared with C18, i.e. the strongest RPLC solvents (nonpolar organic solvents) are the weakest HILIC solvents. Highly polar compounds that C18 columns cannot retain can be retained by HILIC columns, hence they allow complementary separations to C18-based separations in metabonomic analyses.
The stationary phase used in Study IV was an Acquity UPLC C18 column, with a stationary phase based on high strength silica (HSS) T3 (1.8 µm) [81] for separation of metabolites in the spent cell culture medium. This kind of stationary phase retains more polar compounds than Acquity UPLC C18, based on Bridged Ethyl Hybrid (BEH) particles [82]. The latter kind of stationary phase was used in Study I for the separation of tryptic-digested peptides.
The number of metabolites detected by MS is increased when both positive and negative modes are used, and when metabolites are separated using both RPLC and HILIC.

3

Part three

Chemometrics

Based on Papers I, II and IV

3.1 Chemometrics

Chemometrics, which refers to the application of multivariate mathematical and statistical tools to data obtained from chemical measurements [83], was used in several of the studies. The aim of the work underlying this thesis was to build a detailed picture of complex biological samples (fingerprinting), employing analytical methods and instruments with high sensitivity and resolution, as discussed in parts one and two of the thesis and the accompanying papers. Chemometrics was subsequently used for capturing the relevant information from the data, and this part is dedicated to the data analysis and chemometric methods applied.

In a chemometric approach the experiments and analysis must be planned in such a way that the acquired data provide (modelable) information about the explored phenomenon. Hence, the measurements must in some way reflect the phenomenon of interest. It is important to remember that not even chemometrics can help if we seek information in an inappropriate place or in an inappropriate way! However, it is not always easy to be sure *a priori* if the data reflect the properties of interest; sometimes we want to know *if* there are relationships between the obtained data and the properties of interest. In such cases, validation of resulting models (which is always important) is especially vital.

Figure 6 describes the common chemometric procedure, from defining a problem to drawing conclusions. For reliable conclusions much work and knowledge is required at every step. This part of the thesis describes design of experiments, preprocessing of the data prior to multivariate data analysis and classification methods employed in the studies included in this thesis.

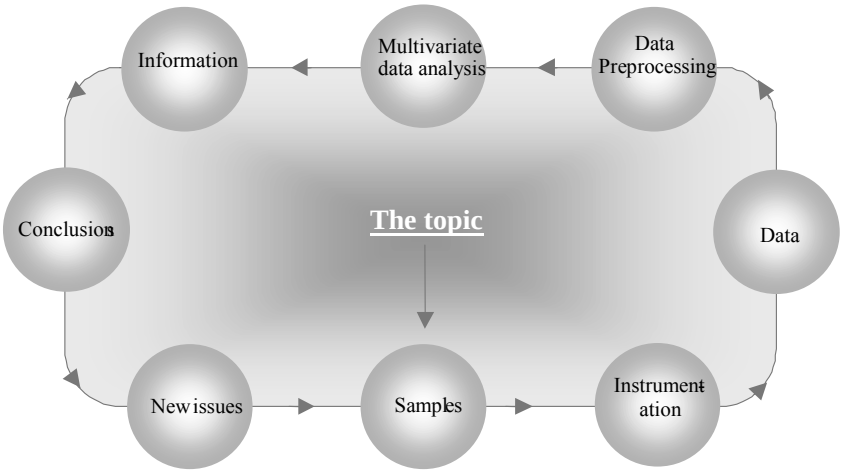


Figure 6. *Schematic description of a common procedure for solving complex problems from an analytical chemistry point of view.*

3.1.1 Design of experiment

The aim of Design of Experiment (DoE) [84] is to obtain as much information as possible from as few experiments as possible. DoE is a good way to identify interactions between significant factors and optimal values of variables of the investigated system. The chosen experimental domain must span the important known variation. So, some knowledge about the experimental system is required in order to obtain a reasonably complete and reliable model. However, an ideal design is not always viable for practical reasons, particularly when a living organism is involved; since some settings of some variables will be impossible to apply in real life. In order to confounding by the non-controlled phenomena, randomization of the experiments (collecting samples, instrumental analysis etc.) is needed.

Important aspects to consider in each DoE are: experiments, factors, responses, the experimental domain and experimental design. A factor (X) is a controllable variable that is varied in a predefined interval (low and high level settings, denoted by "-" and "+", respectively, with or without a center-point, denoted "0"). A response (Y) is a dependent variable that is measured. The experimental domain is the overall domain spanned by all the factors under investigation. There are many different types of experimental designs, and those used for screening

purposes differ from those used for optimization purposes. The aim of screening is to sift the important factors from the non-important ones, while the aim of optimization is to find settings of the important factors that will yield an optimum response. Therefore, screening designs generally include fewer experiments than optimization designs for "response surface modeling" (RSM). The experiments to perform, in terms of the permutations of varied factors, are selected according to a "factorial design", of which there are various types. Examples of such designs for three-level RSM, including Box Benhken and Central Composite Face (CCF) full or fractional factorial designs are shown in Figure 7. Different types of experimental design are described further elsewhere [85-86]. Experiments included in the Box Benhken and CCF designs are indicated by white circles and grey circles, respectively. Center points are experiments in which all the design factors are at their mean values. The factors (here pH, temperature and concentration) are varied in each experiment, and the response (signal) depends on the factors. The run order is randomized, since this is the best way to minimize systematic errors due to handling, instrument drift, etc. In some cases randomization cannot be applied, for example if changing design factors is very time-consuming, costly and/or difficult. This may be considered when analyzing the results and seeking systematic patterns. In experiment 1, the factor settings for pH and temperature are at their lowest respective levels in the experimental domain, while the concentration is at the middle of the interval. The signal (response) is then measured and so on.

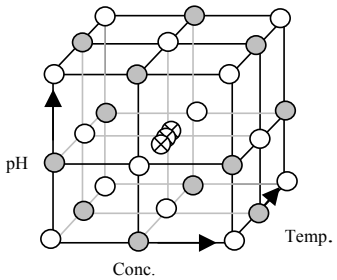
The results of the experiments are used to calculate a regression model of the relationships between the factors (X) and the response (Y), which can be expressed by the following second degree polynomial equation:

$$Y = A_0 + A_1X_1 + A_2X_2 + \ldots + A_nX_n + A_{12}X_1X_2 + A_{13}X_1X_3 + A_{23}X_2X_3 + \ldots + A_{11}X_1X_1 + A_{22}X_2X_2 + \ldots + A_{nn}X_nX_n$$

where the regression variance is partitioned into independent (A_0), linear (A_1, A_2, …, A_n), quadratic (A_{11}, A_{22}, …, A_{nn}) and cross-product (A_{12}, A_{13}, A_{23}, …) components to determine the fit of the function and the relative significance of each component. The ideal type of regression for fitting the data depends on the aim of the study and the dataset, e.g. if there are missing data and a large number of factors,

partial least squares (PLS) regression [83] is preferred to multiple linear regression (MLR) [87].
A reduced central composite face (CCF) experimental design was used in Study II, for finding optimal experimental settings in order to distinguish conformational states of a protein produced by two different cell lines by their ESI-MS charge state distributions. Four variables were varied over three levels. Since variations in distribution of ESI-generated charge states of proteins are to be expected, because of factors such as needle-to-needle variations, using different capillaries is very important to reduce the effect of this variation.

A)



B)

		Factors (X)			Response (Y)
Exp.No	RunOrde	pH	Tem	Conc	Signa
1	13	-	-	0	?
2	2	+	-	0	?
3	7	-	+	0	?
4	12	+	+	0	?
5	6	-	0	-	?
6	14	+	0	-	?
7	15	-	0	+	?
8	4	+	0	+	?
9	10	0	-	-	?
10	1	0	+	-	?
11	9	0	-	+	?
12	11	0	+	+	?
13	5	0	0	0	?
14	3	0	0	0	?
15	8	0	0	0	?

Figure 7. A) *Graphical illustration of two experimental designs: Box Benhken design (○) and central composite face (CCF) (●). Center points (⊗) belong in the both designs, they are in triplicates.* ***B)*** *Experimental design with factor settings for each experiment by CCF design.*

3.1.2 Preprocessing of LC/MS data prior to multivariate pattern recognition modeling

The data obtained from an observation in LC/MS analysis have three dimensions: retention time (t_R), m/z and intensity. An observation with a specific m/z and t_R can be regarded as the identity of an ion in LC/MS

data. The produced data (raw data) is a sum of a *structured data* and *noise*. The fraction of the raw data that correlates with the property of interest is *structured data (contains information)*; everything else is *noise*. The amounts of data produced from LC/MS analyses have increased greatly over the years as a result of instrumental developments (which have provided higher sensitivity, accuracy and resolution). In non-targeted analysis, when every compound in a complex sample is of interest, it is more difficult to optimize the experimental methods regarding the matrix effects, thus stronger matrix effects on the resulted data are to be expected.

The reasons for pretreating the data generally are to reduce the effects of the noise, simplify the data by reducing the dimensionality of the data and to improve the predictive ability of the model. The data interpretation is more time-consuming than the data acquisition and thus has traditionally been the bottleneck in LC/MS-based metabolomic studies. The data are often pretreated before modeling the dataset.

For initial preprocessing of LC/MS data software is needed to compress the raw LC/MS datasets obtained (which are of mega- or giga-byte scale) and structure the data in a suitable format for applying pattern recognition methods; NetCDF is an example of a suitable format for further MVDA of the data. The next step is to reduce both chemical noise (caused by fluctuations in temperature, pH and concentrations etc.) and instrumental noise (which is a composite of noise from all of the instrumental components) [2]. Data points with the same m/z-value that are consistently observed may be contributions from the mobile phase. In Studies I and III an in-house program was used to produce a two-dimensional matrix from LC/MS data in NetCDF format. The software filters the data with a "two-dimensional finite impulse response filter" to improve the Gaussian shade of the peaks and reduce high-frequency noise. A peak list is then produced by position (m/z and t_R) and intensity via automatic peak extraction.

The complexity of the samples poses another great challenge in data processing for reliably comparing LC/MS data acquired from different samples. Various methods can be applied, for example peak detection and de-noising of LC/MS data [3], but all of the methods involve sacrifice of some information, and compromises that simplify the complexity of the datasets. Problems and state-of-the-art methods related to processing metabonomic datasets have been discussed in recent reviews by Åberg et al. [88], Listgarten and Emili [89] and Vandenbogaert et al. [90].

Before comparing samples to find trends and correlations, peaks need to be aligned, i.e. the samples need to be synchronized, by sorting the data from each sample in such a way that datapoints originating from the same ion have the same identity (in terms of t_R and m/z). This is not straightforward since shifts in m/z and the t_R axis commonly occur in LC/MS analyses , thus the shifts may occur in both the t_R and detected m/z of a given ion time because of instrument drift, variations in the chemistry of the separation system and stochastic (uncontrolled) variation. Analyses of complex samples generate large numbers of peaks, with many overlaps, making peak alignment very challenging, especially when samples of a biofluid with a highly dynamic composition, such as urine, are analyzed.

There are several methods for aligning LC/MS data [88],. A common way for separating data of interest from the noise in LC/MS analysis is to detect the peaks (peak detection) and compare the peak list from each sample to each other. In peak detection the three-dimensional LC/MS data are converted into a peak list (in which each point indicates the intensity of a specific peak with a specific identity in terms of t_R and m/z) consisting of one vector. Alignment can be applied to either raw data or the peak list. In the studies underlying this thesis peaks were aligned using peak lists. Retention times and mass windows can be used to align peak lists in such a way that peaks with the same retention time and m/z window in different peak lists are defined as the same peak. The chosen retention time window should be larger than the retention shift and the m/z window should be larger than the mass accuracy of the instrument. A complicating factor is that the shifts in retention time and m/z may vary at different places in the chromatograms and mass spectra, respectively. Further, overlapping peaks in both windows will be defined as the same peak, providing a further source of confusion and potentially increasing the number of apparently false negative results. In addition, using too small windows can result in false positive results. The width of the windows is optimal when replicates end up at the same place, while the distances between the replicates are shorter than those between the samples of different sample groups in a Principal Component Analysis (PCA) score plot, which will be discussed later. In Studies I and IV the LC/MS data were acquired using instruments with high resolution (UPLC-TOFMS), yielding many peaks along the m/z axis, since multiple peaks were generated for each ion arising from their ^{13}C isotopes, which complicate the pretreatment and interpretation of the data. Peaks that

arise from the same ions may contribute to the modeling in the same way, and in the loading plot(s) resulting from this kind of data the corresponding peaks will most likely have the same angle in the loading plot.
Various other data preprocessing methods can also be applied, all of which influence the results of chemometric analyses, and different types of pretreatment may give different score plots. Depending on the types of variables involved, different types of preprocessing can be used. The signal intensities in an LC/MS dataset depend on both response factors and concentrations of individual constituents. The response factors are affected by instrumental variation and chromatographic conditions (e.g. temperature and injection parameters). The variation in the data due to these factors is likely to obscure the compositional information in the following data analysis, and hence fingerprinting which is based on variations in concentrations of specific constituents. Normalization [91] is then required to be able to compare the metabolite levels in different samples, which is especially important when comparing samples of biofluids, in which the metabolite levels are not regulated, such as urine. In Study I, peptide mapping (fingerprinting) was utilized in the classification, and the repeatibility of the process was estimated by calculating coefficients of variation (CV) for certain peptides. Since the CVs were low for all samples, normalization was not used. In Study II, distribution of the charge states was used, and the relative intensities provided a way to normalize the data. In Study IV, in which spent cell culture medium was fingerprinted, normalization was not applied, since the appropriate normalization approach was unknown (and to avoid further increasing the complexity in data interpretation). In addition, the repeatability of replicates was sufficient for the purposes of the study. The other common preprocessing approach involves scaling and mean centering [91]. The scaling and normalization can be performed in several ways. Thus, understanding of their function is essential for correct interpretation of multivariate data systems. For an in-depth discussion of pretreatment of mass spectral profiles, see Arnenerg et al. [92]. Use of appropriate data preprocessing and reliable peak alignment methods facilitates subsequent multivariate analysis of the dataset.

3.1.3 Chemometric models

Experiments in chemistry are generally of multivariate nature, i.e. many variables are measured in a single experiment. For example, a

single LC/MS observation carries three measurements for each of many compounds: t_R, m/z and intensity values. Some important terms in multivariate data analysis are variables and objects (samples). The variables characterize the objects and are generally divided into factors and responses denoted the data matrix **X** and dependent matrix **Y**, respectively.

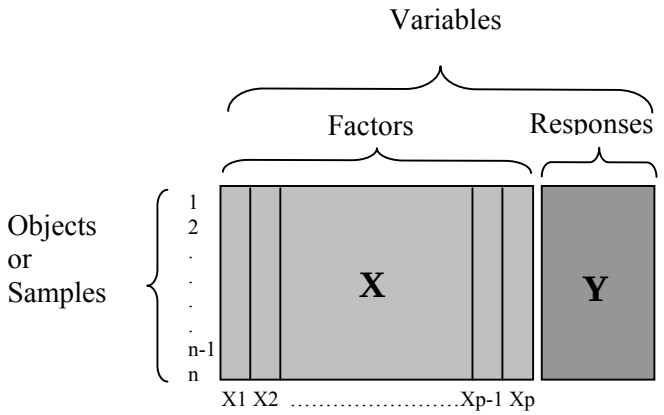


Figure 8. *Illustration of the terms variables, factors, responses and objects. 1, 2, …, n-1 and n are the object-number. X1, X2, …, Xp-1 and Xp are the factor-number.*

One of the multivariate data analysis-based procedures that is commonly used in chemometrics is pattern recognition, for which several types of methods are available.

Principal component analysis (PCA)

PCA [93] is an unsupervised method for recognizing patterns in two-dimensional data matrices; the corresponding method used for higher dimensional data matrices is parallel factor analysis (PARAFAC) [94]. The overall aims of these methods are to identify trends and highlight the similarities and differences in high dimension (multivariate) data matrices. PCA can only be performed on a single n x p data matrix

(denoted **X**), i.e. a matrix consisting of n objects (rows) and p variables (columns), where n < p. PCA is a linear mathematical transformation method that converts the data to a new coordinate system and projects the maximal variance of the data onto the first coordinate (first Principal Component, PC1), and the second greatest variance in the dataset to a second Principal Component (PC2), which is orthogonal to the PC1, and so on. If all variables are uncorrelated and there is no noise in the data, which is unlikely in chemical observations, the number of PCs required to explain all the variance is p. However, in a LC/MS dataset, for example, in addition to the huge amount of noise that will be present, many peaks will be attributable to the same compound and there will be correlations in signals from some compounds in a biofluid. In PCA the **X** data are decomposed into PCs consisting of a score vector (**t**) and a loading vector (**p**). One score value is assigned to each sample related to each PC and one loading value for each variable related to each PC. If two PCs are sufficient to describe the information in an n x p **X** data matrix adequately, the scores matrix (**T**) will be an n x 2 set and the loadings matrix (**P**) will include 2 x p calculated loadings:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}' + \mathbf{E}$$

where **T** is the scores matrix and has as many rows as the original data matrix, while **P** is the loadings matrix and has as many columns as the original data matrix. The number of the columns in the matrix **T** is the same as the number of rows in the matrix **P** and equals the number of significant PCs. **E** is the residual matrix (unexplained part), which has the same size as the original data matrix.

Figure 9 illustrates the decomposition of matrix **X** by PCA using two PCs. Scores and loadings values in PC2 are calculated from the residuals matrix $\mathbf{E_1}$.

$$\mathbf{X} = \mathbf{t_1 p_1}' + \mathbf{t_1 p_1}' + \mathbf{E_2}$$

PCA was used in Study I for the visualization and classification of the oxidized Mab in different conformations with respect to its peptide mappings, and it appeared to be a good method for this purpose, when it is (in broad outline) a rotation of the coordinate system. PCA was

also used in Study II for classification of the protein of interest in different conformations with respect to their charge state distributions in ESI-TOFMS and in Study IV for studying the metabolic fingerprinting of spent cell culture medium sampled on different days to observe clustering or separation in the sample set. The application of PCA in this dynamic study (media from different days) showed that samples obtained from different days could be significantly separated in order, from the earliest day until the last day in the series, along the first PC.

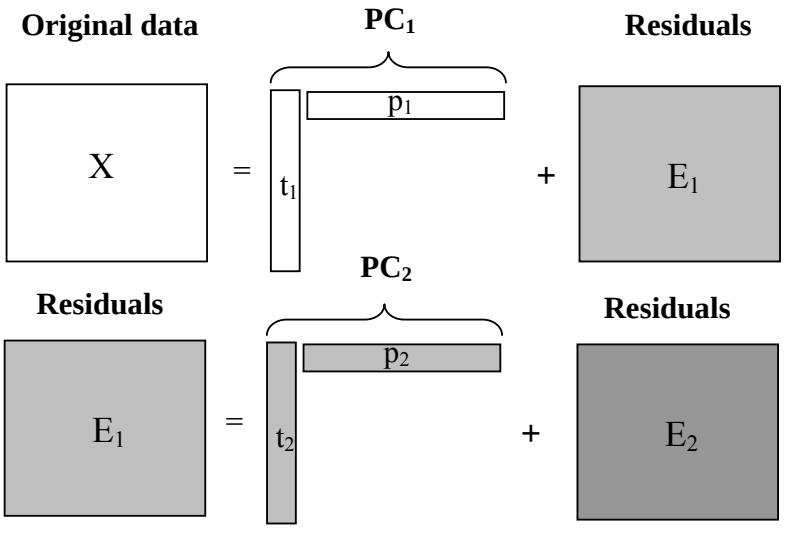


Figure 9. *Schematic diagram of the decomposition of matrix X by PCA using two PCs.*

Partial least squares (PLS)

Multivariate methods that find relationships between two data matrices, **X** and **Y**, are generally called regression methods, e.g. partial least squares (PLS) regression , which is a supervised pattern recognition method. The objective is to predict the y-variables from the x-variables, thus the maximum covariance between **Y** and **X** is sought. PLS involves regressing the variances of **X** and **Y** in such a way that the correlations between **X** and **Y** are maximized. PLS regression has been

described in detail in the literature [83], so only a short description is presented here. In PLS, as in PCA, dimensionality is reduced by decomposition of the datasets into scores and loadings for both **X** and **Y**. In PLS calibration, the **X** matrix results in a scores matrix (**T**), loadings matrix (**P**), loading weights (**W**) and a residual matrix (**E**). The **Y** matrix results **in** a scores matrix (**U**), a loadings matrix (**C**) and a residual matrix (**F**).

$$\mathbf{Y = BX + F}$$

where **B** is the PLS regression coefficients.

$$\mathbf{B = W\,(PW)^{-1}\,C}$$

The process is schematically illustrated in Figure 10. PLS was used in Study IV for finding relationships between the metabolic fingerprints of the spent cell culture medium and the concentration of the active target protein, glucose, lactate and number of viable cells. Despite the complexity of the data and overlapping of some compounds in either m/z or t_R, the PLS regression provided good correlations and high predictive ability for concentrations of active protein, glucose and lactate. However, the predictive ability for the model with respect to the number of viable cells was weak.

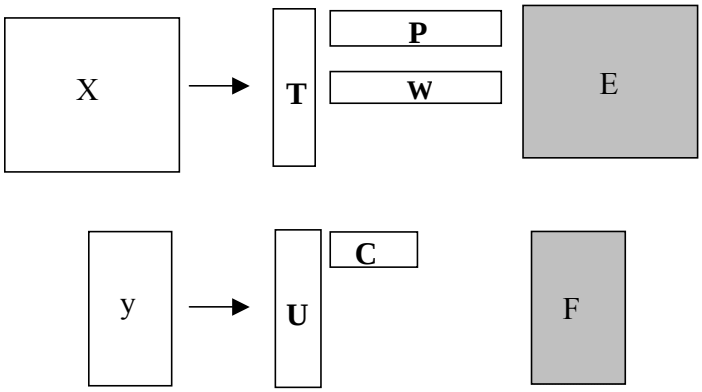


Figure 10. *Schematic diagram of matrix X and Y by PLS regression.*

Linear discriminant analysis (LDA)

Linear discriminant LDA, also called Fisher-LDA, is a supervised technique for recognizing patterns in, and classifying, a set of observations into predefined classes by finding linear combinations of the initial variables that optimally separate two or more data classes. The objective of this classification technique is to maximize the proportion of between-class variance relative to the within-class variance in the dataset and thus maximize the separation by projecting the data obtained from samples into a space of low dimensions, i.e. to extract characteristics that only contain information that is relevant to the classification. Fisher-LDA has been thoroughly described elsewhere [95], but the procedure is schematically illustrated in Figure 11 for classifying classes A and B. LDA was utilized in Study II for classifying the same proteins produced in different cell lines.

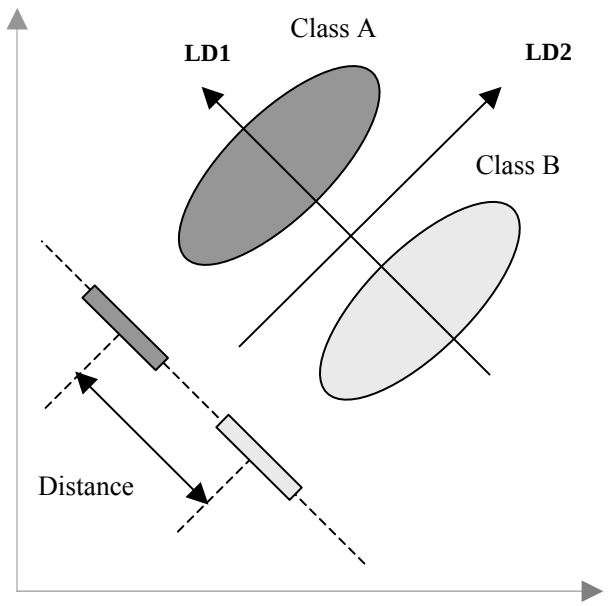


Figure 11*. Linear discriminant analysis of two classes with equal covariance matrices.*

3.1.4 Validation of the models

It is very important to be aware that mathematical models provide a simplification of complex realities, rather than truly portraying the modeled realities. Hence, validation is a crucial step of modeling in chemometrics, since it provides indications of the reliability of conclusions that can be drawn from the modeling. Without validation, acquired models cannot be applicable in real life. Chemometric models can be validated in either of two ways. External validation is one approach, in which new objects (test sets) that have not been used in the model are applied. The test set should be of known and suitable content. Use of test sets is recommended when large numbers of samples are available, and 30-50 % of the samples can be used as a test set, while the rest of the objects can be used in the training set for modeling. Internal validation is the other approach, in which objects that have been used in the modeling are also used in validation of the model e.g. cross-validation [96]. In cross-validation only a training set is required, but one (or a group) of objects is removed at a time. The object(s) that was/were removed is/are used to test the performance of the model on "new" data. Each time another object(s) is/are removed and used as "new" data, a new test of the predictive ability of the model is performed. The validation is performed by comparing the predicted and true values and used to estimate the goodness of the predictive ability of the model.

Concluding remarks and future outlook

I started my PhD studies in Professor Sven Jacobsson's research group by hunting for potential biomarkers for disease and drug toxicity by metabolic fingerprinting of biological samples, which resulted in two publications (one included in this thesis, Paper III). However, shortly after I started we also began characterizing protein drugs, prompted by rapid advances in the production and potential use of protein-based drugs.

This section presents some concluding remarks concerning the characterization of protein drugs, e.g. monoclonal antibodies, and optimization of cell cultivation for the efficient production of protein drugs.

Advances in genetics and genomics, including the sequencing of the human genome, have resulted in the development of new biological drugs to treat cancer and other grave diseases. Recombinant therapeutic protein technology is one of the fastest growing areas of the pharmaceutical industry. Mammalian cells are the dominant cultivation system, due to their ability to perform complex PTMs required for the stability and functionality of the proteins. However, PTMs also include unwanted modifications, e.g. oxidation, deamidation, variable glycosylation, misfolding and aggregation. These deviations from desired structures not only pose challenges for bioprocessing, but may also have undesirable consequences for the patient and can lead to immune responses to the protein drug. Detecting and preventing these modifications has become a major challenge for the biotechnology industry.

In the papers included in this thesis, two different methodologies are presented for characterizing recombinant antibodies in different conformations and produced by different cell lines, respectively.

Antibodies in different conformations were characterized using data from UHPLC/ESI-MS-based peptide mapping and PCA in Study I, exploiting differences in the rates of oxidation of specific methionines in two extreme protein conformations (native and denatured). This probing and all the other analytical methods used seem to provide appropriate methods for classifying conformational variants of the protein. Such investigation of levels of unfolding provides interesting indications of this method's ability (and limitations) for detecting conformational changes. Comparison of the information acquired with

data generated by other structural analysis methods (e.g. CD and X-ray crystallography) can provide further indications of the potential utility and complementary information that this type of probing, in combination with peptide mapping (using UHPLC/MS) and MVDA, can supply.

The other method, developed for characterization of the antibody produced by different cell lines using ESI charge state distributions combined with MVDA, also proved to be appropriate for distinguishing the proteins. Other researchers in the field of gas phase ions of macromolecules have found that CSD of macromolecules in ESI can be used to provide meaningful information on large-scale unfolding [97], but is not suitable for detecting subtle structural changes, therefore it would be interesting to combine our PTM profiling results from Study II with CD and X-ray crystallography data to obtain further, complementary information.

Another interesting research area is investigation of the interactions that result in dimerization and polymerization of the protein drug molecules during purification and storage. These interactions could be studied using high mass ESI-Q-TOF or, even better, by a high mass ESI-Q-IM-TOFMS/MS.

Optimizing and controlling parameters in a cell cultivating process in terms of costs and the quality of expressed proteins etc. are highly laborious and costly. Extracellular metabolic fingerprints in a spent-cell culture medium showed good correlations with the protein quality data presented in Paper IV. Identifying the variables that strongly contribute to this correlation may be the next step toward finding relevant biomarkers. Since the cultivated cells produce proteins other than the target protein, fingerprinting of the high molecular fraction and including information regarding this fraction in cell profiles in more comprehensive analysis can provide useful information. Analyzing highly polar metabolites by HILIC and combining mass spectra data acquired in both negative and positive modes, can also offer more detailed and extensive picture of the reality. In addition, combining data from NMR and MS analyses could improve metabolic fingerprinting, since they are complementary techniques (NMR can detect compounds that either are not retained by the LC system or are not ionized in ESI, and MS can detect compounds at lower concentrations than NMR). Different techniques for fusing LC/MS and NMR data have been discussed in an article by Forshed at al. [98].

Working in fields such as cell cultivation, metabolomics and the characterization of large proteins has assured me about the usefulness and necessity of multivariate data analysis for extracting relevant information. The multivariate projection methods PCA and PLS regression played a central role in these analyses, for finding trends and groups in data based on fingerprints and for relating protein quality in cell culture media to changes in extracellular metabolite fingerprints.

Mass spectrometry combined with multivariate data analysis proved to be a powerful approach for characterizing proteins and metabolomic analyses. Hyphenating high-resolution techniques such as in UHPLC/ESI-Q-TOFMS increases the power of the method. Protein charge state distributions in ESI also carry structural information that can be used for characterization of proteins.

To conclude, I see a bright future given the many potential applications and innovations in mass spectrometry coupled with other analytical tools for protein conformational studies as well as for fingerprinting of biological systems. For conformational studies of macromolecules, construction of MS instruments with more controlled parameters such as vacuum, needle position, temperature and gas flow rates, and etc. are needed, due to significant influence of these parameters on the quality of MS spectra. Utilizing fingerprinting of dynamic systems (e.g. biological systems) in pattern recognition is, in general, very useful and valuable. Online monitoring and measurements of the biological systems can therefore lead to more reliable and representative characterization of living systems.

Acknowledgement

Jag vill tacka alla som på något sätt har hjälpt mig att sätta prefixet Dr. före mitt namn. Vägen dit har varit en lång och krokig berg- och dalbana, med få trafiksskyltar och utan säkerhetsbälte!!! Som tur är, ÄLSKAR jag äventyr!

Först och främst vill jag tacka min handledare, Sven Jacobsson. Din skarpa blick för HELHETEN, din positiva attityd och multivariata tänkande har guidat mig rätt under resans gång.

I would also like to thank my second supervisor, Leopold Ilag, who shared his enthusiasm for proteins and mass spectrometry, and always kept up-to-date with this technique and its applications. Thank you for introducing me to “collisional cooling”!

Helena Idborg, tack för din proffsiga handledning av mitt exjobb – en härlig tid med mardrömmar om ”super-standarden” och dina komplexa matriser.

Tack till alla andra som jag har samarbetat med, mött eller haft mailkontakt med, på SU, AstraZeneca och andra företag: Yang Yang, för all hjälp med MS:en och ditt tålamod med en nybörjare på ditt fina labb; Per Edebrink, för din kompetens och lugn, det var ett privilegium att jobba med er båda; Fredrik Andersson för ditt välgjorda program, MzExplorer; Eva Hedin för din noggranna läsning av peket; Nathalie Chatzissavidou för din proffsighet och dina tydliga och snabba svar på mina e-mails; Anders Hagman för samarbetet i cellkultur projektet; Per-Olof Edlund för den korta tiden du var min biträdande handledare, du utstrålar kunskap i ämnet; både små och stora molekyler.

Jag vill också tacka alla andra på AstraZeneca R&D som fick mig att känna mig välkommen. Ett särskilt tack till Malin, Yang, Anna och Maria for alla luncher och fika-raster. Tack Anna för tipset om proteinet.

Tack till alla duktiga ex-jobbare jag har samarbetat med: Kajal, för ditt djup i allt du gör, för att du är så smart och äkta, för alla gånger vi brottades med µLC:en och för fikat på Avenyn; Jessica, för din

livsglädje och förmåga att hitta bra referenser, hur många gånger öppnade du egentligen kapillären? Emma, för din envishet, för att du hade stenkoll på dina stora Excel-filer och är grym på samarbete - tack för din vänskap.

Och TACK till alla doktorander (före mig och samtida): Helena H., jag kan säkert fylla sidor om allt jag vill tacka dig för: ALLA våra diskussioner, allvarliga och tokiga. Mina tre sista år på IAK skulle ha varit tråkigare och tommare utan dig. Vilka roliga stunder vi hade när vi såg cyklande kossor flyga!!! Du är en vanlig kanin, fast *ovanligt fin*... Shahram, for being a great QToF-playmate, for our discussions, covering just about anything from moleculars to politics to, of course, Iran – you are a really good friend; Malin för att du ser igenom en och är en bra lyssnare, samt för dina avhandlingstips; Thuy, for being so smart and nice, for the interesting discussions about your CD:s and one-µl-eluates; Nana, för farsi-snack och din nyfikenhet, för mysiga frukostar i Colorado; Axel, för bio-följd-av-drink-traditionen (speciellt Bio på Rio och grannkrogen), ser fram emot fler tillfällen; Caroline för ditt skratt, för att du är sällskaplig och för dina svampkunskaper. Magnus, för morgonfika och hjälp med kemometri och Matlab (speciellt med Fisher-m-filen).

Tack till Jenny, för Reykjavik, Islandshästarna, alla fester som blir extra festliga med dig på dansgolvet och all tips och uppmuntran. Tack till alla andra doktorer som har delat med sig av sin glädje, kunskap och erfarenheter på våra traditionella träffar: Stina M (för att du är originell och enkel); Sindra (för din tuffa och spännande stil); Johanna (För din lugn som överaskar med bus); Stina C. (för att du delar med dig av dina erfarenheter); Yvonne (för mitt första disputationsdeltagande).

Tack till Jonas R., inte för layouten av denna avhandling men för programmet ”LeilaMatch” och alla pratstunder om trevliga och otrevliga ämnen, för skjuts till Göteborg på analysdagarna; Petter för allt skoj, samt idén om att tvätta ”konen” med myrsyra; Gunnar, för att du också gillar att diskutera samhällsfrågor, och för PTM-referenserna; Christoffer for all skratt på kräftskivan, du blev nästan spårlöst försvunnen; Ralf, för att du bryr dig om hur det går och grattade varje gång en artikel accepterades; Theres, för din passion för allt som kittlar smaklökarna; Erik, for Matlab-hjälp och Tequilan; Yasar, för att du är så hjälpsam och alltid väljer spännande resmål; Fredrik, för ”disco

musik”; Tuula, för tips med husråttorna, det funkade; Gianluca, for Italien cheese; Silvia, for your will to learn; Zdenek, for your mobile GPS on our way to the Uppsala conference; Anna Lisa, for your kindness and hard work in the mass-lab; duktiga doktoranderna som ändrade riktning halvvägs: Ragnar för att du var så mån om den globala orättvisan; Leo, för ditt sätt att vara och samarbetet i jämställdhetskommiten.

Anders för att jag fick vara ”språkslang” på julfesten; Conny som snabbt tillgodosåg mitt akuta behov av extern hårddisk till den enorma mängden rådata; Bosse, vad glad jag blir varje gång jag ser dig i korridoren, du är en stor tillgång för IAK; Anita, för ett gott samarbete på grundkursen och AKII - du är så snäll och strukturerad; Ulrika för diskussionen om skillnaden mellan ”mass accuracy” och precision”, hur var det nu igen? Ann-Marie, för all praktisk hjälp och för London-snacken; Ulla, för hjälp med svåra blanketter; Ninni, för samarbetet på bioanalyskursen; Lena, för att du har jobbat längst av alla på IAK, samt din hjälpsamhet; Björn, för all uppmuntran samt vanan att titta förbi; Roger, för att vi fick ta del av era opublicerade resultat om etanolbilar.

Tack till Hillevi och alla andra på kemibiblioteket för hjälp med referenser.

Ett extra tack till alla som har korrekturläst manuset och kommit med värdefulla kommentarer: Sven, Shahram, Axel, Malin, Pol, Magnus, Ulrika, Gunnar, Roger, Ralf och Anders.

I would also like to thank those special ones in my real world, outside the department of Analytical Chemistry: my family (my mother, sisters and brothers and their families) and friends all around the world, for giving me strength and courage to meet the challenges of my life. My very special thanks goes to my mother:
سپاس مامان، برای اینکه همیشه باورم داشتی. خیلی دوستت دارم
I can’t name you all, but you are all invaluable. I would like to give special thanks to Bahar, for you are open-minded, for your warmth, spontaneity. Every time we meet, no matter after how long, you are the same lovely “Bahar”; to Shora, for your big world, your new-thinking, honesty, your youthfulness and those enormous cups of tea you served me. You girls are both life itself. To Ghiyam for being such a brother,

always standing by me and making me laugh. You are a real fighter. To Reza for your warm hugs- you are a source of goodness.

Maria, thanks for our long phone conversations about relationships and our children, for showing me London in 4 days. Thank you, Mina, for that high-quality friendship we share, for our spontaneous trip to Paris. That sunny day, outside the Notre-Dame, I knew that I had found a friend for life.

Jag vill speciellt tacka mina barn: Simon, för din unika blick på världen som aldrig upphör att inspirera mig, för ditt smittsamma skratt, för att du gillar historier om ”när mamma var barn”, för din passion för fotboll. Du ”äger”! David, för att du sprider så mycket glädje, för dina oändliga frågor som får mig att inse hur lite jag kan om allt i universum. Dina fotbollsfinter är bäst! Tack. Ni får mig att känna att vad jag är bäst på är inte analytisk kemi utan mamma rollen. TACK killar.

Sist men inte minst vill jag tacka dig Reza, min älskade livskamrat, for ditt mod att hoppa på mitt berg-och-dalbanetåg. Tack för ditt stöd och din kärlek, för att du ställer krav. Vissheten att huset stod, att barnen var lyckliga och att samma kärleksfulle man väntade mig när jag äntligen kom hem från jobbet, möjliggjorde allt detta...

The work underlying this thesis is financed by AstraZeneca R&D Södertälje, which is gratefully acknowledged.

References

1. Bosworth CE, Asimov MS, editors. History of Civilizations of Central Asia. Volume IV, Part II. : Delhi, Motilal; 2003.
2. Skoog DA, Holler FJ, Nieman TA, editors. Principal of Instrumental Analysis (5th Edition). Philadelphia: Harcourt Brace & Company; 1998.
3. Idborg H, Zamani L, Edlund PO et al. Metabolic fingerprinting of rat urine by LC/MS Part 2. Data pretreatment methods for handling of complex data. J Chromatogr B Analyt Technol Biomed Life Sci 2005;828(1-2):14-20.
4. Wilkins DK, Mayer A. Development of antibodies for cancer therapy. Expert Opin Biol Ther 2006;6(8):787-96.
5. Liu X, Pop LM, Vitetta ES. Engineering therapeutic monoclonal antibodies. Immunol Rev 2008;222:9-27.
6. Pierce A, deWaal E, Van Remmen H et al. A novel approach for screening the proteome for changes in protein conformation. Biochemistry 2006;45(9):3077-85.
7. Pierce A, de Waal E, Van Remmen H et al. A novel approach for screening the proteome for changes in protein conformation. [Erratum to document cited in CA144:327224]. Biochemistry 2006;45(17):5686.
8. Creighton TE, editor. Proteins Structure and Molecular Principles. New York: Freema; 1984.
9. Matsumoto M, Nakayama K. Post-translation modification proteomics: fusion with reverse genetics. Idenshi Igaku Mook 2005;2:298-307.
10. Walsh G, Jefferis R. Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol 2006;24(10):1241-52.
11. Yamauchi E, Taniguchi H. Analysis of post-translation modification. Idenshi Igaku Mook 2005;2:99-104.
12. Vogt W. Oxidation of methionyl residues in proteins: tools, targets, and reversal. Free Radical Biol Med 1995;18(1):93-105.
13. Griffiths SW, Cooney CL. Relationship between Protein Structure and Methionine Oxidation in Recombinant Human Î±1-Antitrypsin. Biochemistry 2002;41(20):6245-52.

14. Chu JW, Yin J, Brooks BR et al. A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. J Pharm Sci 2004;93(12):3096-102.
15. Astorga-Wells J, Joernvall H, Bergman T. A microfluidic electrocapture device in sample preparation for protein analysis by MALDI mass spectrometry. Anal Chem 2003;75(19):5213-9.
16. Shariatgorji M, Astorga-Wells J, Jornvall H et al. Microfluidic Electrocapture-Assisted Mass Spectrometry of Membrane-Associated Polypeptides. Anal Chem (Washington, DC, U S) 2008;80(18):7116-20.
17. Yamashita M, Fenn JB. Electrospray ion source. Another variation on the free-jet theme. J Phys Chem 1984;88(20):4451-9.
18. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 1988;60(20):2299-301.
19. Wilm M, Mann M. Analytical Properties of the Nanoelectrospray Ion Source. Anal Chem 1996;68(1):1-8.
20. Benesch JLP, Ruotolo BT, Simmons DA et al. Protein Complexes in the Gas Phase: Technology for Structural Genomics and Proteomics. Chem Rev (Washington, DC, U S) 2007;107(8):3544-67.
21. Guevremont R, Siu KWM, Le Blanc JCY et al. Are the electrospray mass spectra of proteins related to their aqueous solution chemistry?. J Am Soc Mass Spectrom 1992;3(3):216-24.
22. Gaskell SJ. Electrospray: principles and practice. J Mass Spectrom 1997;32(7):677-88.
23. Wang G, Cole RB. Disparity between solution-phase equilibria and charge state distributions in positive-ion electrospray mass spectrometry. Org Mass Spectrom 1994;29(8):419-27.
24. Mansoori BA, Volmer DA, Boyd RK. Wrong-way-round electrospray ionization of amino acids. Rapid Commun Mass Spectrom 1997;11(10):1120-30.
25. Grandori R. Electrospray-ionization mass spectrometry for protein conformational studies. Curr Org Chem 2003;7(15):1589-603.
26. Taylor GI. Disintegration of Water Droplets in an Electric Field. Proc Roy Soc London 1964;280(1382):383.

27. Smith DPH. The electrohydrodynamic atomization of liquids. IEEE Trans Ind Appl 1986;IA-22(3):527-35.
28. Kebarle P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. J Mass Spectrom 2000;35(7):804-17.
29. Kebarle P, Tang L. From ions in solution to ions in the gas phase - the mechanism of electrospray mass spectrometry. Anal Chem 1993;65(22):972A-86A.
30. Iribarne JV, Thomson BA. On the evaporation of small ions from charged droplets. J Chem Phys 1976;64(6):2287-94.
31. Dole M, Mack LL, Hines RL et al. Molecular beams of macroions. J Chem Phys 1968;49(5):2240-9.
32. Fernandez de la Mora J. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. Anal Chim Acta 2000;406(1):93-104.
33. Chowdhury SK, Katta V, Chait BT. Electrospray ionization mass spectrometric peptide mapping: a rapid, sensitive technique for protein structure analysis. Biochem Biophys Res Commun 1990;167(2):686-92.
34. Fenn JB, Mann M, Meng CK et al. Electrospray ionization-principles and practice. Mass Spectrom Rev 1990;9(1):37-70.
35. Kaltashov IA. Gas phase ion chemistry and measurements of macromolecular properties in solution. Abstracts of Papers, 229th ACS National Meeting, San Diego, CA, United States, March 13-17, 2005 2005:ANYL-405.
36. Samalikova M, Matecko I, Mueller N et al. Interpreting conformational effects in protein nano-ESI-MS spectra. Anal Bioanal Chem 2004;378(4):1112-23.
37. Sharon M, Robinson CV. The role of mass spectrometry in structure elucidation of dynamic protein complexes. Annu Rev Biochem 2007;76:167-93.
38. Volmer DA, Sleno L. Mass analyzers: An overview of several designs and their applications, Part I. Spectroscopy (Duluth, MN, U S) 2005;20(11):20-6.
39. Volmer DA, Sleno L. Mass analyzers: An overview of several designs and their applications, Part II. Spectroscopy (Duluth, MN, U S) 2005;20(12):90-5.
40. Krutchinsky AN, Chernushevich IV, Spicer VL et al. Collisional damping interface for an electrospray ionization

time-of-flight mass spectrometer. J Am Soc Mass Spectrom 1998;9(6):569-79.
41. Chernushevich IV, Thomson BA. Collisional cooling of large ions in electrospray mass spectrometry. Anal Chem 2004;76(6):1754-60.
42. McKay AR, Ruotolo BT, Ilag LL et al. Mass Measurements of Increased Accuracy Resolve Heterogeneous Populations of Intact Ribosomes. J Am Chem Soc 2006;128(35):11433-42.
43. Busch KL. The electron multiplier. Spectroscopy (Eugene, Oreg) 2000;15(6):28-33.
44. Laprade BN, Labich RJ. Microchannel plate-based detectors in mass spectrometry. Spectroscopy (Eugene, Oreg) 1994;9(5):26,7, 30.
45. Burbaum J, Tobal GM. Proteomics in drug discovery. Curr Opin Chem Biol 2002;6(4):427-33.
46. Cowan DA, Laidler P. Mass spectrometry in drug metabolism. Biomed Health Res 1998;25(Drug Metabolism: Towards the Next Millennium):159-74.
47. Smith D, Spanel P. The challenge of breath analysis for clinical diagnosis and therapeutic monitoring. Analyst (Cambridge, U K) 2007;132(5):390-6.
48. Songsermsakul P, Razzazi-Fazeli E. A Review of Recent Trends in Applications of Liquid Chromatography-Mass Spectrometry for Determination of Mycotoxins. J Liq Chromatogr Relat Technol 2008;31(11 & 12):1641-86.
49. Xu X, Zhang X, Peng Y. Application of ICP-MS in water quality analysis. Zhongguo Weisheng Jianyan Zazhi 2006;16(6):763-6.
50. Roussis SG, Fedora JW, Fitzgerald WP et al. Advanced molecular characterization by mass spectrometry: applications for petroleum and petrochemicals. Anal Adv Hydrocarbon Res , [Symp] 2003:285-312.
51. Wagner DS, Anderegg RJ. Conformation of cytochrome c studied by deuterium exchange-electrospray ionization mass spectrometry. Anal Chem 1994;66(5):706-11.
52. Cai X, Dass C. Conformational analysis of proteins and peptides. Curr Org Chem 2003;7(18):1841-54.
53. Kartha G. Picture of proteins by x-ray diffraction. Accounts Chem Res 1968;1(12):374-81.

54. Wuethrich K. High-resolution NMR experiments for studies of protein conformations. NATO Adv Study Inst Ser , Ser A 1982;45(Struct. Mol. Biol.):215-35.
55. Gratzer WB. Optical methods for studying protein conformation. Tech Life Sci , [Sect]: Biochem 1978;B108:1-43.
56. Katta V, Chait BT. Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. Rapid Commun Mass Spectrom 1991;5(4):214-7.
57. Englander SW, Mayne L, Bai Y et al. Hydrogen exchange: the modern legacy of Linderstrom-Lang. Protein Sci 1997;6(5):1101-9.
58. Chung EW, Nettleton EJ, Morgan CJ et al. Hydrogen exchange properties of proteins in native and denatured states monitored by mass spectrometry and NMR. Protein Sci 1997;6(6):1316-24.
59. Pan J, Rintala-Dempsey AC, Li Y et al. Folding kinetics of the S100A11 protein dimer studied by time-resolved electrospray mass spectrometry and pulsed hydrogen-deuterium exchange. Biochemistry 2006;45(9):3005-13.
60. Wilson DJ, Rafferty SP, Konermann L. Kinetic unfolding mechanism of the inducible nitric oxide synthase oxygenase domain determined by time-resolved electrospray mass spectrometry. Biochemistry 2005;44(7):2276-83.
61. Ruotolo BT, Giles K, Campuzano I et al. Evidence for Macromolecular Protein Rings in the Absence of Bulk Water. Science (Washington, DC, U S) 2005;310(5754):1658-61.
62. Iacob RE, Murphy JP,III, Engen JR. Ion mobility adds an additional dimension to mass spectrometric analysis of solution-phase hydrogen/deuterium exchange. Rapid Commun Mass Spectrom 2008;22(18):2898-904.
63. Fiehn O. Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol 2002;48(1-2):155-71.
64. Nicholson JK, Lindon JC, Holmes E. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999;29(11):1181-9.

65. Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp Funct Genomics 2001;2(3):155-68.
66. Fiehn O, Kind T. Metabolite profiling in blood plasma. Methods Mol Biol (Totowa, NJ, U S) 2007;358(Metabolomics):3-17.
67. BiomarkerDefinitionsWorkingGroup, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. 2001;69:89-95.
68. Butler M. Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. Appl Microbiol Biotechnol 2005;68(3):283-91.
69. Moreira AR. The evolution of protein expression and cell culture. BioPharm Int 2007;20(10):56, 58,59, 62, 64-65, 68.
70. http://www.fda.gov.
71. Long C, King EJ, Sperry WM, editors. Biochemist' Handbook. : Van Nostrand (Princeton, N.J); 1961.
72. Fritz JS, editor. Analytical Solid-Phase Extraction. New York: WILEY-VCH; 1999.
73. McDowall RD. Sample preparation for biomedical analysis. J Chromatogr , Biomed Appl 1989;492:3-58.
74. McDowall RD, Doyle E, Murkitt GS et al. Sample preparation for the HPLC analysis of drugs in biological fluids. J Pharm Biomed Anal 1989;7(9):1087-96.
75. Moco S, Bino R, De Vos RCH et al. Metabolomics technologies and metabolite identification. TrAC, Trends Anal Chem 2007;26(9):855-66.
76. Hemstroem P, Irgum K. Hydrophilic interaction chromatography. J Sep Sci 2006;29(12):1784-821.
77. MacNair J, Patel K, Tolley L et al. Ultra high pressure liquid chromatography. Book of Abstracts, 216th ACS National Meeting, Boston, August 23-27 1998:ANYL-180.
78. Grumbach ES, Mazzeo J, Diehl D. Ultra performance LC: A new dimension in chromatography. Not Tec Lab 2004;12(4):20,22.
79. Durham DK, Hurley TR. Effect of Sub-2-Micron Particle Size on Peak Efficiency, Capacity, and Resolution in Preparative Liquid Chromatography. J Liq Chromatogr Relat Technol 2007;30(13):1895-901.

80. Jonsson T, Appelblad P. Separation of polar and hydrophilic compounds using a zwitterionic stationary phase in hydrophilic interaction liquid chromatography. LC-GC Eur 2004(Applications Book):57-8.
81. New L, Chan ECY. Evaluation of BEH C18, BEH HILIC, and HSS T3 (C18) column chemistries for the UPLC-MS-MS analysis of glutathione, glutathione disulfide, and ophthalmic acid in mouse liver and human plasma. J Chromatogr Sci 2008;46(3):209-14.
82. Sievers D. New HPLC-column for chromatography. New possibilities in method development. GIT Spez Sep 2005;25(1):22-4.
83. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 2001;58(2):109-30.
84. Lundstedt T, Seifert E, Abramo L et al. Experimental design and optimization. Chemom Intell Lab Syst 1998;42(1,2):3-40.
85. Eriksson L, Johansson E, Kettaneh Wold N et al, editors. Design of Experiments Principles and Applications. Umeå, Sweden: Umetrics AB; 1996.
86. Brereton RG, editor. Chemometrics Data Analysis for Laboratory and Chemical Plant. United Kingdom: Wiley; 2003.
87. Bro R, editor. Håndbog I Multivariable Kalibrering. Copenhagen, Denmark: Jordbrugsforlaget; 1996.
88. Aberg KM, Alm E, Torgrip RJO. The correspondence problem for metabonomics datasets. Anal Bioanal Chem 2009;394(1):151-62.
89. Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 2005;4(4):419-34.
90. Vandenbogaert M, Li-Thiao-Te S, Kaltenbach H et al. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. Proteomics 2008;8(4):650-72.
91. Craig A, Cloarec O, Holmes E et al. Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. Anal Chem 2006;78(7):2262-7.
92. Arneberg R, Rajalahti T, Flikka K et al. Pretreatment of Mass Spectral Profiles: Application to Proteomic Data. Anal Chem (Washington, DC, U S) 2007;79(18):7014-26.

93. Jackson JE, editor. A User's Guide to Principal Components. New York: Wiley; 1991.
94. Bro R. Multiway calibration. Multilinear PLS. J Chemom 1996;10(1):47-61.
95. McLachlan GJ, editor. Discriminant Analysis and Statistical Pattern Recognition. New York: John Wiley and Sons; 1992.
96. Baumann K. Cross-validation as the objective function for variable-selection techniques. TrAC, Trends Anal Chem 2003;22(6):395-406.
97. Kaltashov IA, Abzalimov RR. Do Ionic Charges in ESI MS Provide Useful Information on Macromolecular Structure?. J Am Soc Mass Spectrom 2008;19(9):1239-46.
98. Forshed J, Idborg H, Jacobsson SP. Evaluation of different techniques for data fusion of LC/MS and 1H-NMR. Chemom Intell Lab Syst 2007;85(1):102-9.