Mass spectral profiling and multivariate analysis for detection of biomarker signatures

Application to multiple sclerosis

Tarja Annikki Rajalahti Kvalheim

Dissertation for the degree philosophiae doctor (PhD)

at the University of Bergen

2010

Scientific environment

The present thesis is based upon work carried out during the years 2006-2009 and it is a result of collaboration between University of Bergen (Department of Clinical Medicine, Institute of Medicine, Department of Chemistry, and PROBE Proteomic Unit) and Haukeland University Hospital (Norwegian Multiple Sclerosis National Competence Centre, Department of Neurology, and Laboratory of Clinical Biochemistry). Pattern Recognition Systems (PRS) AS has developed and provided software for data analysis.

This project has been financed with the aid of EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, and the Norwegian Multiple Sclerosis Competence Centre, Haukeland University Hospital.

MS Nasjonalt Kompetansesenter

HELSE OG REHABILITERING
Stiftelsen for norske helse- og rehabiliteringsorganisasjoner

HELSE BERGEN
Haukeland University Hospital

Contents

Preface

"Smile – things may get worse more slowly." Unknown

Life is multivariate and full of coincidences. This story started in Finland at the Helsinki University of Technology. I wanted to become an architect but I became a chemical engineer specialized in bioprocess engineering and food technology – since all "bio" stuff was considered to ensure a great future (and work) in the beginning of the 1990's. Instead of working, I continued with post graduate studies. I spent long sleepless nights beside a bioreactor with bacteria producing a valuable enzyme. I collected lot of samples, analyzed them and produced many sheets full of numbers. At some point I had to give my data to somebody who was "just" sitting, playing around with neural networks and working from nine to five. I was so bitter. ☺

After two-three years of frustration and struggling with ill-behaving bacteria Heikki Haario mentioned the word "chemometrics" and gave an introduction to experimental design and modelling at the graduate school course on "Mathematical Tools in Biosciences". I began to see the light and managed to talk myself into the Umetrics courses "Design of Experiments" and "Multivariate Data Analysis" given by Veli-Matti Taavitsainen and Håkan Fridén. The feeling was almost schizophrenic – I was angry because nobody had taught this to us at the university and happy because, now, I (thought I) knew what to do. I would design a series of controlled experiments and optimize my enzyme production and I would use multivariate analysis for modelling and prediction. Unfortunately the reality stroke back: "We don't use these methods here" was the answer I got.

But I was converted and there was no way back to the old way of thinking. So when my friend Terhi Siimes asked me to join her at a software company representing Umetrics and its chemometrics software in Finland, I literally "left the building". No more wet chemistry for me – just data, please. I learned to know the "old" gentlemen from Umetrics (Håkan, Erik Johansson, Christer Albano, Conny Wikström, and

Lennart Eriksson) and started to give courses myself. I enjoyed teaching and spreading the message but I also realized that multivariate thinking was not as obvious to everybody as I had expected.

Through Umetrics I came in contact with Professors Svante Wold and Michael Sjöström in Umeå (Sweden). Michael arranged me a grant for post graduate studies at Umeå University; for me it was a fantastic opportunity to become a real chemometrician. I packed my things and moved from Helsinki to Umeå in February 2000. I worked with protein sequence data and bioinformatics. Yet another hot and promising "bio" stuff. I guess I was an outlier in a gang of Ume chemometricians and organic chemists since I didn't quite understood the Swedish way of doing things and the famous Jante law (though back in Finland they said that I'd been "swedished"). ☺

In August 2001 at the 7th Scandinavian Symposium on Chemometrics (SSC7) in Copenhagen I met a very special guy from Bergen. His name was Olav Martin Kvalheim and he turned my life upside down. After a few months travelling between Sweden and Norway I decided to pack my things again and move to Norway in June 2002. I continued my studies but at the same time I started my own one-woman company giving courses and consulting in chemometrics, especially in Finland.

This PhD project started literally "by accident". Olav got a neck problem during the spring 2003 and consulted a medical doctor. He was Christian Vedeler, professor in neurology at the Haukeland University Hospital. Olav got a prolapsed disc diagnosed but at the same time he managed to catch Christian's interest on one of his research areas; the use of multivariate methods in spectral profiling and medical diagnosis. Christian arranged a meeting with Kjell-Morten Myhr, the leader of the Norwegian Multiple Sclerosis National Competence Centre. He understood the potentiality and happened to have interesting spinal fluid samples in a freezer. Rune Ulvik, professor in internal medicine, had already earlier started to collaborate with Olav, so we decided to start a project with Rune as a coordinator. New co-workers came and went and the project stayed alive mainly on a hobby basis. Different methods were tested and preliminary studies were carried out. Finally, I got funding from the Norwegian

Foundation for Health and Rehabilitation and started as a PhD student in April 2006 after first giving a birth to the most important project of my life, Martin. ☺

All the persons mentioned above are acknowledged for their contribution to this story. The following persons earn an extra mentioning:

Michael Sjöström gave me a chance to enter the chemometrics research community. He is not only a rewarded chemometrian but an extremely generous person and a beekeeper as well. ☺

Kjell-Morten Myhr and Christian Vedeler have been open-minded, extremely flexible and always very positive supervisors. ☺

Reidar Arneberg has programmed everything. He has calmly listened to my frustrated complaints and offered unconditional help. ☺

Magnus Berle has collected CSF samples from healthy controls. Frode Berven and Ann Cathrine Kroksveen have been responsible for the analytical workup and MALDI analyses. ☺

Olav Martin Kvalheim has offered me his solid competence, bursting creativity and lot of (conditional ☺) help. Privately, as a husband, even he shows tendency to univariate behaviour (Figure X).

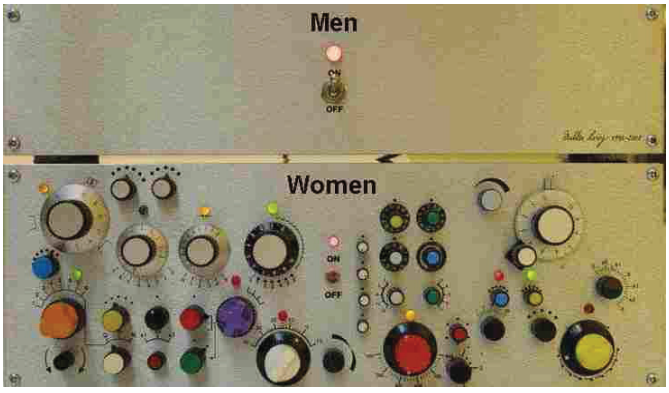


Figure X. *System complexity. Thanks to a male course participant.* ☺

List of abbreviations and notations

Abbreviations

2DE	2-dimensional gel electrophoresis
ANOVA	Analysis of variance
CCR	Correct classification rate
CE	Capillary electrophoresis
CIS	Clinical isolated syndrome
CNS	Central nervous system
COW	Correlation optimized warping
CSF	Cerebrospinal fluid
DA	Discriminant analysis
DIVA	Discriminating variables
DNA	Deoxyribonucleic acid
DoE	Design of experiments
ESI	Electrospray ionization
HPLC	High performance liquid chromatography
LV	Latent variable
MALDI	Matrix-assisted laser desorption/ionization
MCCR	Mean correct classification rate
MLR	Multiple linear regression
MRI	Magnetic resonance imaging
MRM	Multiple reaction monitoring
mRNA	Messenger ribonucleic acid
MS	Multiple sclerosis

MW	Molecular weight
NHC	Neurological healthy controls
OND	Other neurological diseases
O-PLS	Orthogonal PLS
PC	Principal component
PCA	Principal component analysis
PLS	Partial least squares
PLS-DA	PLS-discriminant analysis
PTM	Post-translational modification
RAFFT	Recursive alignment by fast Fourier transform
RNA	Ribonucleic acid
ROC	Receiver operating characteristics
SELDI	Surface-enhanced laser desorption/ionization
SR	Selectivity ratio
TOF	Time-of-flight
TP	Target projection
VIP	Variable importance on projection

Notations

Generally, bold uppercase characters (*e.g.* **X**) represent matrices, bold lowercase characters (*e.g.* **x**) represent vectors, and italic characters (*e.g. N*) represent scalars. The transpose is indicated by a superscript ^{T}. Vectors are by default column vectors and transpose transforms them into row vectors.

A	Number of LVs (PCA or PLS components)
b	Regression coefficient vector

Da	Dalton, unified atomic mass unit
E	Residual matrix
M	Number of objects/rows
m/z	Mass-to-charge ratio
N	Number of variables/columns
nM	Nanomolar, 10^{-9} mol/L
p	Loading vector
P	Loadings matrix
pM	Picomolar, 10^{-12} mol/L
t	Score vector
T	Scores matrix
w	Weight vector
X	Data matrix, spectral profiles
y	Response vector

Abstract

Mass spectrometry based protein profiling and biomarker discovery has given rise to the field of clinical proteomics. The underlying assumption is that proteins can provide information of diagnostic or therapeutic value and can thus be used in a clinical context. This thesis presents a novel approach where full mass spectral profiling and multivariate data analysis are combined to reveal biomarker signatures from complex body fluid samples. Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry is employed for acquiring full spectral profiles of intact proteins in cerebrospinal fluid (CSF) proteome. Since MALDI-TOF is relatively straightforward to use and demands relatively simple analytical work-up, it may have the potential to be used on routine basis in clinical laboratories. Multiple sclerosis (MS) is used as a model disease to demonstrate how this approach works on real proteomics data. The aim is to detect a disease signature typical for MS and use it in disease classification. MS is an example of a disease that may be difficult to diagnose at its very early stage and there is a need for diagnostic biomarkers for early diagnosis and treatment. It should however be mentioned that the developed approach is general and can be applied for other diseases, body fluids and instrumental techniques as well.

This thesis is composed of four scientific papers, each one focusing on a specific problem. In the first study (Paper I) different methods for pretreatment of spectral profiles are tested. The aim of this study is to obtain recipes for elimination of non-compositional factors and thus improve reproducibility and minimize within-group variation compared to between-group variation. Statistical experimental design is used to assess the effect of each pretreatment step and examine if there are significant interactions between these steps. An optimal pretreatment strategy is developed and applied to further work. In the second and third study (Papers II and III) new chemometric methods are developed for detection of biomarker signatures in complex spectral profiles. In Paper II a so-called selectivity ratio (SR) and accompanying SR

plot are presented for the first time and validated using spiked CSF samples. In Paper III a non-parametric discriminating variable (DIVA) test is introduced. DIVA test can be used in combination with SR plot to define statistical boundaries for biomarker detection. Both methods are in fact general and can be applied for most kind of variable selection problems. In the fourth study (Paper IV) the novel multivariate approach (including data pretreatment, SR and DIVA) is applied to real proteomic data derived from CSF samples from three different patient groups (MS, other neurological diseases and neurological healthy controls). The presented approach is able to discriminate the groups and the most important mass spectral regions (*i.e. m/z* values) contributing to separation can be found. These *m/z* values can be seen as a biomarker signature that can be used in disease classification.

List of publications

This thesis is based on following papers. They will be referred to in the text by the corresponding Roman numerals (I-IV). Papers are reprinted with kind permission from the publishers.

I Reidar Arneberg, **Tarja Rajalahti**, Kristian Flikka, Frode S. Berven, Ann C. Kroksveen, Magnus Berle, Kjell-Morten Myhr, Christian A. Vedeler, Rune J. Ulvik, and Olav M. Kvalheim, Pretreatment of mass spectral profiles: Application to proteomic data. *Anal. Chem.* **2007**, *79,* 7014-7026.

II **Tarja Rajalahti**, Reidar Arneberg, Frode S. Berven, Kjell-Morten Myhr, Rune J. Ulvik, and Olav M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35-48.

III **Tarja Rajalahti**, Reidar Arneberg, Ann C. Kroksveen, Magnus Berle, Kjell-Morten Myhr, and Olav M. Kvalheim, Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, *Anal. Chem.* **2009**, *81*, 2581-2590.

IV **Tarja Rajalahti**, Ann C. Kroksveen, Reidar Arneberg, Frode S. Berven, Christian A. Vedeler, Kjell-Morten Myhr, and Olav M. Kvalheim, A multivariate approach to reveal biomarker signatures for disease classification: Application to mass spectral profiles of cerebrospinal fluid from patients with multiple sclerosis, *J. Proteome Research* **2010**, *9*, 3608-3620.

1. Introduction

"There is no adequate defence, except stupidity, against the impact of a new idea."
Percy Williams Bridgman

A diagnosis is needed for starting any medical treatment. Once diagnosed with a certain medical condition, there may be many treatments available to help the patient. Early diagnosis and treatment will inevitably give the best prognosis for many diseases and also be cost-efficient for the society. Therefore more precise methods for early diagnosis are always needed. One way to achieve this goal is to search for disease specific markers. Biomarker discovery is currently playing a leading role in modern proteomics research.

Mass spectral profiling of body fluids, for example, blood, urine, and cerebrospinal fluid (CSF), has become a common method when searching for biomarkers. Many proteins can be screened in a single experiment and, in principle, no biological knowledge about the disease is required to be able to accomplish the analysis. Discrimination between disease affected and healthy individuals is provided by the spectral pattern itself and no identification of the individual components is needed for separating different groups of patients.

An important aspect in this context is data handling and analysis with subsequent feature selection and interpretation of the results. This is a challenging task because of the extreme complexity of the data (*e.g.* spectral profiles with many overlapping proteins). Traditionally, univariate statistics, like *t*-test, has been employed in data analysis. Unfortunately, use of univariate methods may lead to spurious results and false biomarkers. In addition, they cannot handle properly the situation with many variables contributing to the discriminatory pattern. Multivariate methods based on latent variables offer a better alternative since they take into account correlations in the data and provide tools for visualization of the data, detection of biomarker signatures and classification of samples. Furthermore, the multivariate methods are

not only applicable in the data analysis phase but also in the beginning of the study when planning and designing the experimental procedures.

Partial least squares (PLS) regression is the workhorse in multivariate data analysis. The obtained results can, however, be difficult to interpret since multi-component PLS models are usually needed to describe the variation in complex spectral profiles. Methods like target projection (TP) and orthogonal PLS (O-PLS) circumvent this problem by finding a single linear combination of variables related to the response. Even though this makes interpretation easier one question remains: how to reveal the most important spectral variables that contribute to separation between different sample groups and thus may serve as biomarker signature. By introducing a new variable selection method called selectivity ratio (SR) and an accompanying non-parametric statistical test called discriminating variable (DIVA) test, we are able to detect biomarker signatures from complex spectral profiles.

1.1 The aims

The main aim of this thesis was to develop general methods to reveal biomarker candidates in complex systems like body fluids using full mass spectral profiling and statistical multivariate data analysis. A secondary aim was to apply these methods on real proteomics data obtained from three groups of CSF samples, representing multiple sclerosis (MS), other neurological diseases (NHC), and neurological healthy controls (NHC), to detect the features distinguishing these groups.

This thesis is based on following sub-studies, the results of which are published in four scientific papers:

1. Pretreatment of mass spectral profiles. The aim in data pretreatment is to eliminate non-compositional features (*e.g.* baseline effects, shifts in *m/z* values, structured noise, and differences in absolute signal intensities due to amount of sample analyzed) from the mass spectral profiles without destroying the compositional

differences. This study shows how different data pretreatment steps influence spectral profiles obtained from matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. Factorial experimental designs, with pretreatment steps as design variables and inter- to intragroup variation as a response variable, are used to make a quantitative assessment of the effects. Interpretation of the resulting empirical models makes it possible to propose optimal schemes for pretreatment of mass spectral profiles obtained from MALDI and related techniques. (Paper I)

2. Selectivity ratio (SR) and discriminating variable (DIVA) test. The aim is to develop new quantitative tools for interpretation and variable (biomarker) selection in complex spectral profiles. Multivariate modelling approach based on PLS regression and TP is used to obtain one predictive component which in turn is the a starting point to variable selection and detection of biomarker signatures. Methods are validated using spiked CSF samples. (Papers II and III)
3. Application to mass spectral profiles of CSF from patients with MS. The developed methodology (DIVA test and SR plot) is applied for the first time to real proteomics data. The aim is to reveal the features distinguishing patients with MS from control groups OND and NHC. The detected biomarker signature can then be used for disease classification. (Paper IV)

1.2 Scope and outline of the thesis

The thesis covers some basic theory of clinical proteomics, biomarker detection, biological mass spectrometry, and multivariate analysis. The analytical chain starts from the selection of study population, and proceeds via sampling, sample preparation, instrumental analysis, data acquisition, and data analysis to the final results. This thesis has the main emphasis on the data analysis step, even though the other steps are also discussed to some extent. Figure 1 illustrates the proposed workflow starting from body fluid samples and leading to biomarker signature.

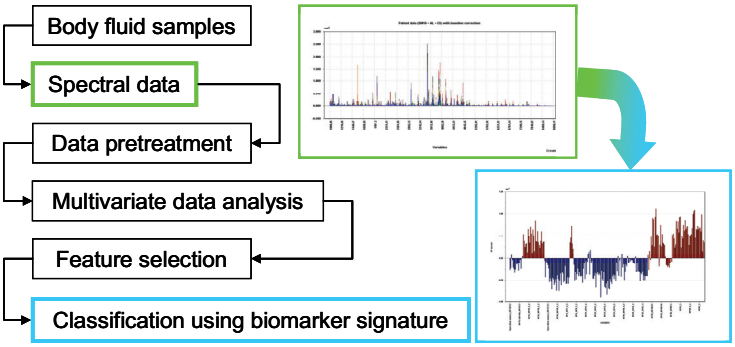


Figure 1. *Proposed workflow for pattern based biomarker detection and disease classification.*

The thesis has the following structure: Chapter 1 has given introduction to the aims and the scope of the thesis. Chapter 2 provides a background to "omics" sciences, especially clinical proteomics with emphasis on biomarker discovery. A brief introduction to multiple sclerosis is also given in this chapter. Proteomics techniques, biological mass spectrometry and MALDI-TOF instrumentation are discussed in Chapter 3. In Chapter 4 multivariate methods are described in detail, including SR plot and DIVA test, novel methods developed as part of this thesis. Chapter 5 contains a summary of the obtained results and general discussion. Conclusions and future perspectives are presented in Chapter 6.

2. Background

"I could prove God statistically. Take the human body alone – the chances that all the functions of an individual would just happen is a statistical monstrosity." George Gallup

Clinical laboratories are facing fundamental changes in the near future when it comes to methodology and instrumentation. Views of health and disease are influenced by the present post-genomic era, driven by huge advances in bioanalytical technologies and bioinformatics. The study of the human genome has been followed by the study of human proteome, providing us complementary information for the analysis and understanding of complex pathological processes. These approaches, the so-called "omics" sciences (Figure 2), have given a novel insight into human biology and a paradigm shift in healthcare is under progression.

This chapter provides theoretical background to "omics" sciences in general, and especially to the field of proteomics, with the emphasis on its most important application and the main theme of this thesis: proteomics based biomarker discovery. A short introduction is also given to MS; a chronic, inflammatory disease which has been used as a model disease in this thesis.

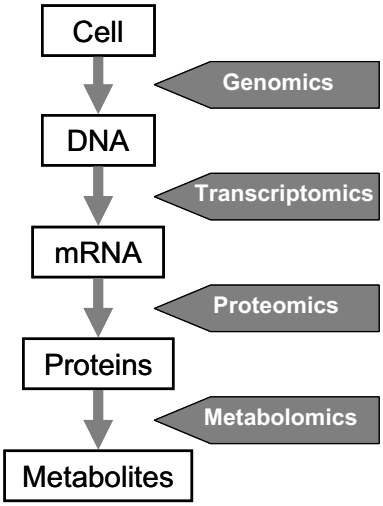


Figure 2. *Schematic presentation of "omics" sciences.*

2.1 “Omics” sciences

Every living system is built from cells. All cell functions are directed by information provided by deoxyribonucleic acid (DNA), a molecule which in turn is made of the same components in all living organisms.[1] DNA consists of nucleotides (containing a base, sugar, and phosphate) which are arranged in two long strands forming a double helix. Alternating phosphate and sugar (2-deoxyribose) residues compose the backbone of each DNA strand. One of four different bases, adenine (A), cytosine (C), guanine (G) and thymine (T), is attached to each sugar and hydrogen bonds between base pairs (A and T, G and C) hold the two strands together. The bases are arranged in a particular order and it is the sequence of these four bases that dictates the instructions required to create an organism with unique features. The complete set of DNA molecules for an organism is its genome and genomics aims to map the entire DNA sequence of a certain organism.[2] The DNA sequence is divided into different regions. A gene is a specific region of the DNA that encodes a certain protein. Genes have coding regions (exons) and non-coding regions (introns); the DNA molecule is actually mostly consisting of non-coding regions. The protein encoding process consists of transcription of coding regions of DNA into messenger ribonucleic acid (mRNA) and translation of mRNA template into a protein sequence (Figure 3). RNA is a single-stranded molecule containing ribose instead of deoxyribose and the same bases as DNA, except for thymine (T) which in RNA is substituted with uracil (U). Transcriptomics is a study of all mRNA molecules reflecting the active genes.[3]

Translation is based on the genetic code (Figure 4) where a triplet of nucleotides (a codon) in mRNA represents a single amino acid, a building block of proteins. In total there are 20 amino acids to choose from, thus the four-letter RNA code is translated to the twenty-letter amino acid code. Genes contain the recipe for the heredity by passing the traits from one generation to the other. International Human Genome Sequencing Consortium has given a recent estimate that the total number of protein-coding genes in the human body is in the range 20 000–25 000.[4] This is much less than originally expected and a surprisingly low number for our species. On the other

hand, these genes may be encoding as many as one million protein forms, that is, enzymes, antibodies, hormones, structural elements and electron carriers, which are responsible for all our vital functions. [5-7]

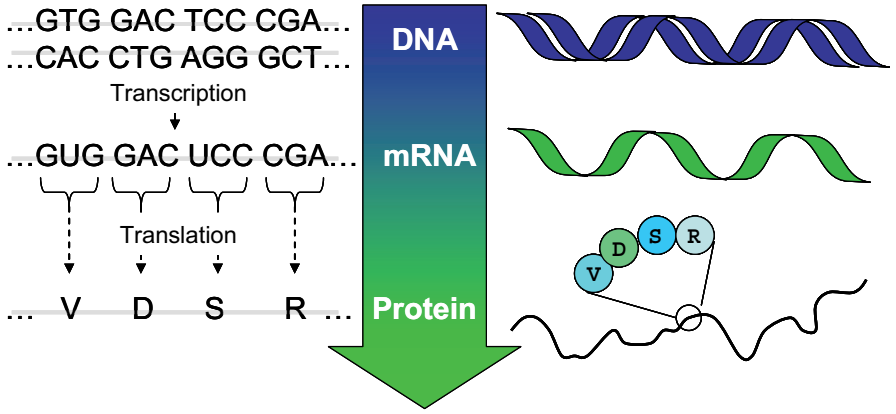
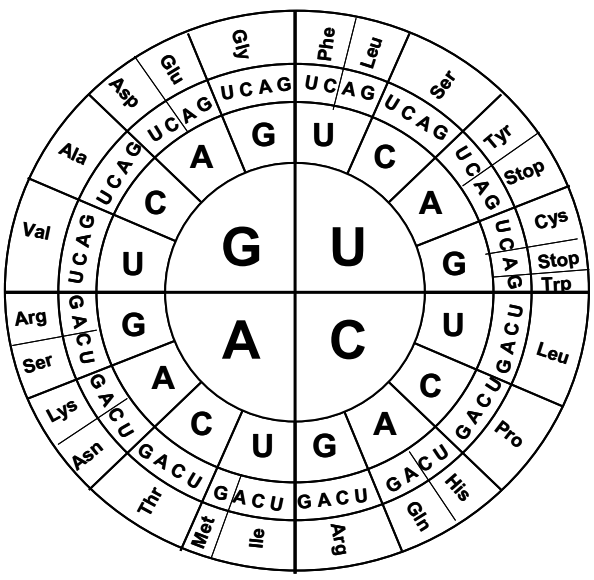


Figure 3. *Information flow in biological systems from nucleic acid into protein, the so-called central dogma of molecular biology.*[8]



Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Clutamic acid	Glu	E
Glutamine	Gln	Q
Glysine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 4. *The genetic code and the 20 amino acids with their abbreviations.*

Proteins are large biological macromolecules made up of chains of amino acids.[9] Short chains of amino acids are called peptides and proteins are therefore polypeptides. The primary structure of a protein is its amino acid sequence. Chemical properties of the amino acids cause the protein chain to fold up into a three-dimensional structure. The overall shape of the protein (tertiary structure) is consisting of secondary structure elements (most commonly beta strands/sheets and alpha helices) connected by intermediate amino acid turns. An example of this is shown in Figure 5, which illustrates a crystal structure of human cystatin C.[10] Proteins are converted to their mature form through a complicated sequence of post-translational modifications (PTM) that generate a large number of different forms of the protein, with major differences in function. Many of these PTMs are regulatory and reversible (*e.g.* protein phosphorylation) and they are responsible for protein folding, stability, cellular localization, recognition and immune reactions. [11, 12] Chemistry and behaviour of a protein are determined not only by the coding gene but also by the other proteins made in the same cell at the same time due to, for example, protein-protein interactions.

Figure 5. *The crystal structure of monomeric human cystatin C stabilized against aggregation.[10] Image from the Protein Data Bank (www.pdb.org), PDB ID: 3GAX.*

The proteome is the entire collection of proteins (including the PTMs, mutations and degradation made to a particular set of proteins) of a cell, tissue or fluid in a living organism at a given point in time. While the genome is relatively stable, the proteome is a highly dynamic system undergoing constant changes due to a large number of intra and extracellular environmental variations, for example, as a result of drug administration.

Proteomics is the study of proteomes.[13-15] Originally proteomics referred to detection and identification of a complete set of expressed proteins in an organism or living system (*e.g.* the human body). This has, however, changed during the recent years and proteomics has become a science that covers a much wider array of protein related features with experimental and computational approaches handling large amount of protein related information.[16] Large-scale analysis of complex protein expression patterns, protein-protein interactions and PTMs is now possible because of advanced proteomics technologies. The field has become very popular and despite its relatively young age (starting at 1995) huge amount of proteomic applications have been published. For instance a PubMed search with the keyword "proteomics" gave 24 344 articles (performed on April 21st 2010). The popularity has also lead to an explosion in protein databases such as European Bioinformatics Institute (EBI) Databases (www.ebi.ac.uk/Databases), Human Protein Atlas (www.proteinatlas.org), Human Proteinpedia (www.humanproteinpedia.org), National Center for Biotechnology Information (NCBI) Databases (www.ncbi.nlm.nih.gov/protein), and Worldwide Protein Data Bank (wwPDB) (www.wwpdb.org).

The last link in this chain of "omics" sciences is metabolomics.[17, 18] This is a study of metabolites; small chemical compounds produced via metabolic mechanisms representing the end products of the gene expression chain in an organism. The challenge in systems biology is to be able to integrate the information from all the different levels (from DNA to metabolites) to get the overall picture of the system itself.[19]

2.2 Clinical proteomics

Modern proteomic research is aiming for novel methodologies that can be applied directly in clinical diagnosis, monitoring and controlling of therapy and designing of drugs. The most important and widespread use of proteomics is the identification of proteins in cells, tissues or biofluids under different states (*e.g.* a specific disease). Clinical proteomics is based on the hypothesis that proteins can provide information having diagnostic or therapeutic value. Mischak *et al* defined the field of clinical proteomics as: *"The application of proteomic analysis with the aim of solving a specific clinical problem within the context of a clinical study."* [20] Clinical proteomics studies on human body fluids were first initiated in the beginning of the 2000s. A controversial study of a proteomic pattern in serum as a screening tool for ovarian cancer created huge excitement and started a flood of similar applications worldwide.[21]

Clinical proteomics aims at bringing proteomic tools into the clinical environment: understanding of patho-biological mechanisms and search for new diagnostic, prognostic or therapeutic biomarkers. The task is not easy as there are many problems that still need to be resolved before widespread implementation of proteomic techniques in routine clinical chemistry laboratories. Mass spectrometry based proteomic technologies have an increasing potential but further development in workflows and instrumentation is needed to be able to fully compete with existing techniques like protein immunoassays performed on high-throughput immuno-analyzers.[22] Proteomics and disease related discussions and opinions about the potential success of use of proteomics in clinical chemistry and especially biomarker discovery have been presented in several articles.[5, 20, 23-28] Many of them conclude that the proteomic methods are not yet ready for implementation in routine clinical laboratories, but that it seems to be possible to reach the goal in the near future.

Clinical proteomics is a highly interdisciplinary field where involvement from specialists with clinical expertise, bioanalytical chemistry, instrumental analysis and data analysis is needed. As mentioned before, the proteome is far more extensive than

the genome and measuring and analyzing proteome present a huge challenge. New technological and bioinformatic solutions are therefore needed to solve the problem. Understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

There are many important considerations related to a clinical proteomic study. General principles and rules of clinical trials should be followed when selecting study population, collecting and handling samples and analyzing them.[20] For example, it should be kept in mind that in addition to control group with healthy individuals, a control group with patients resembling the studied disease should be included to the study population. At the onset of a clinical proteomic study, a clinical problem should be clearly defined. The clinical relevance of the findings must be evaluated to be able to judge whether a new proteomic approach will result in improvements when compared to current standard procedure for diagnostics and therapy. In addition, utilizing advanced proteomic techniques in clinical environment may not always be cost-effective in practice.

Pre-analytical factors like sample collection, handling and storage may have a significant effect on the obtained results and should therefore be standardized.[20, 29] Factors affecting the instrumental analysis itself are also important to recognize. Availability of patient material is often limited and sample amounts are relatively small. Since we do not know all proteins present in a complex sample, it is desirable to keep the proteins intact. The workflow for sample handling should be fast so that time from collecting a sample to freezing it is minimized. Good reproducibility of sample preparation is important since large experimental variation introduced during this phase increases the risk of false findings as well as missing potential biomarkers.

2.3 Proteomics based biomarker discovery

Biomarkers are tests or measurements that provide information about the biological condition of the subject being tested.[30] In addition, biomarkers have an ability to segregate between different biological states, for example, disease affected patients and healthy individuals. Advances in proteomic techniques have given us the hope of discovering novel protein biomarkers. Detected and identified proteins or peptides become biomarkers after validation, that is, they must be verified and proved as reliable predictors for a certain condition.

Biomarkers can be used for monitoring disease progression, for instance by looking at the trend from visit to visit (*i.e.* biomarker velocity) within a single patient. Pharmaceutical industry has an interest to find biomarkers that reflect drug response or toxicity and can be used instead of waiting for a clinical event. Biomarkers predicting the response for the treatment (*i.e.* endpoint analysis) can be used for monitoring the efficacy of therapeutic intervention. In disease subtype classification, the aim is to find biomarkers that discriminate between known subclasses or define novel subclasses in a patient group previously treated as homogeneous.

Perhaps the most wanted application of biomarkers is disease diagnosis.[30] This is particularly the case when early intervention improves the success of treatment and current tests do not detect disease early enough. However, demands placed on biomarkers used for diagnosis are much higher than those used for monitoring disease in existing patients. For instance, quantitative values must be established to be able to set boundaries between positive and negative tests.

Proteomics has been extensively used in biomarker discovery. Comprehensive proteomic analysis and identification of proteins in a single sample does not provide us with useful biomarkers. Comparison of samples from different populations is needed to be able to reach the goal. Because of the “holistic” nature of the proteomic response to a certain disorder it is more probable that instead of one single biomarker there are actually multiple potential biomarkers that together can be used to diagnose

the disease.[20] Different diseases may also have overlapping biomarkers. This makes the task of finding biomarkers even more complicated.

Instead of striving for complete identification of the proteome one may take an alternative approach and use proteomic pattern analysis. Looking at protein patterns should be an effective method for the early diagnosis of diseases. A panel of observed biomarkers gives a signature that can be used in diagnostics without the need for identification of the biomarkers itself, even though identification is of course preferred, if achievable. Since hundreds of clinical samples per day can be analyzed utilizing mass spectral profiling, this technology has the potential to be a novel, highly sensitive diagnostic tool for the early detection of many diseases, or as a predictor of prognosis and response to therapy.

Two possible routes can be used in the proteomic analysis when comparing samples from disease affected patients with control groups and looking for differences in protein contents (Figure 6).[23]

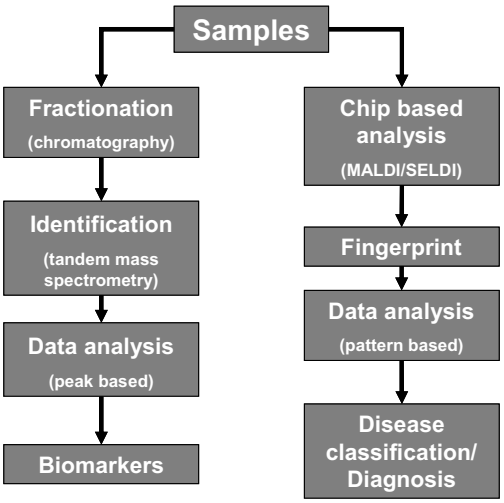


Figure 6. *Identification-based biomarker discovery versus pattern-based diagnostic proteomics.*

In identification-based biomarker discovery multidimensional fractionation of samples in combination with tandem mass spectrometry analysis is used. Hundreds of proteins that are unique or highly abundant within samples can be identified using this approach. The strategy is very time consuming and is hence setting the limit to number of samples that can be compared. Data analysis is based on peak detection and univariate statistics where peaks are compared using, for example, *t*-tests.

Pattern-based diagnostic proteomics employs high-throughput methods based on technologies like surface-enhanced laser desorption/ionization (SELDI) and MALDI mass spectrometry. A protein profile is acquired and bioinformatic data analysis methods are employed to search for differences in peak intensities between the sets of different sample groups. The method does not rely on the identification of the proteins itself, only the pattern, although selected peaks can be subjected to further analysis and identification.

Clinically relevant biomarkers have to undergo four phases: discovery, qualification, verification, and validation with assay development.[31] Mass spectrometry based techniques are suitable for the discovery phase since they can be used to measure differences in profiles between clinical samples. Full instrumental profiling of proteins generates high-dimensional datasets where the number of variables is much higher than the number of samples. This can easily give rise to false biomarkers when using traditional statistics. Validation of the potential biomarkers is the final step in the process and should be done with independent samples left out in the analysis phase or collected later for repeated analysis.

2.3.1 Some statistical considerations for biomarkers

Currently the main focus in biomarker applications is on their use as clinical tests.[30] Such tests can have two outcomes reflecting the true presence or absence of a certain clinical state, thus the result can be positive or negative, respectively. Even in cases where it is possible to obtain a quantifiable test result the common practice is often to set a threshold value which defines the borderline between negative and positive test.

Two types of errors can be defined if the test fails. False positives occur when positive test results are obtained in the absence of disease (case *b*, Figure 7). False negatives occur when negative test results are obtained in the presence of disease (case *c*, Figure 7). In both cases the consequences can be dramatic and cause unnecessary stress, further diagnostic procedures and treatment, or even mortality.

Sensitivity measures the ability of the test to find the disease when it is present, that is, the fraction correctly classified as positive in a population having the disease (*a/(a+c)*, Figure 7). Specificity measures the ability of the test to rule out the disease when it is absent, that is, the fraction correctly classified as negative in a disease-free population (*d/(b+d)*, Figure 7).

Disease / Test	Present	Absent
Positive	a	b
Negative	c	d

Figure 7. *Possible outcomes of a clinical test.*

In reality many diseases do not comply with this kind of binary black and white thinking, thus giving a grey zone in between. Let us consider the case of two populations, disease affected and healthy individuals, and calculate the mean value for the biomarker variable for each of the populations. If we then compare these means using statistical tests like *t*-test, there is a risk that the variation around the mean is so broad that the populations actually overlap to such a degree that it is impossible to define a clear cut-off value that can be used to separate the populations (Figure 8). But even if no clear separation can be provided using an individual biomarker it may be possible to achieve this goal using several biomarkers jointly, as illustrated in Figure 9 for two correlated variables. Multivariate methods take into

account correlation patterns between variables and thus enable the discrimination between samples in this type of situations as well.

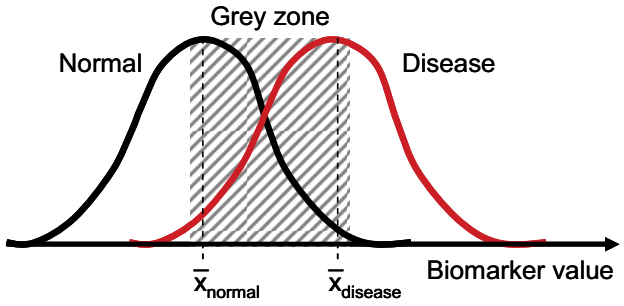


Figure 8. *Distribution around mean value for a biomarker variable when comparing two different populations. Populations are overlapping each other and there is a grey zone in between the populations.*

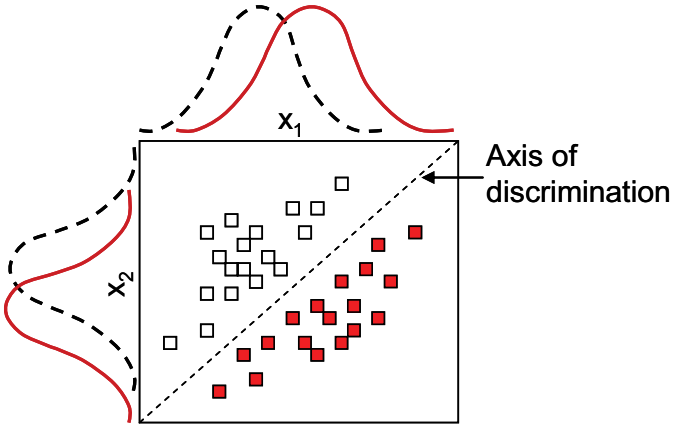


Figure 9. *A case with two groups and two correlated variables. No discrimination can be observed if the variables are considered separately.*

2.3.2 Proteomic analysis of body fluids

An adult human body consists approximately 60% of fluids.[32] Human body fluids, such as blood, saliva, CSF and urine, are believed to reflect the tissues present within a patient. Characterization of the proteomes from various body fluids has recently been under extensive interest.[23, 26, 33] Proteomics techniques, such as mass spectrometry, can be used to screen and identify hundreds of proteins in complex body fluids (see Chapter 3). The analytical tools have become more sophisticated, but discovery of biomarkers in biological fluids is still an enormous challenge because of the huge amount of proteins present in these fluids.

A 'proximal' fluid is defined as a body fluid having a direct contact with the disease affected organ.[31] Proteins or peptides secreted from diseased tissue can therefore be detected directly from proximal fluids, a property that makes them an attractive source for biomarker discovery. Proximal fluids include for instance urine (biomarkers for bladder and kidney diseases) and CSF. CSF surrounds and protects the central nervous system (CNS) from trauma. It is a clear liquid that is produced in the ventricles of the brain, which in turn are continuous with the central canal of the spinal cord. CSF is the only clinical material obtained from a living person that has a direct contact with the extracellular surface of the brain. Therefore CSF is a natural body fluid of choice for analysis of biochemical changes in the CNS and searching biomarkers for neurological diseases. A limiting factor for collecting CSF samples may be the invasive procedure of lumbar puncture, since usually a blood sample or even non-invasively collected samples are preferred.

CSF proteome resembles plasma proteome since around 80% of the proteins found in CSF are derived from blood.[34] The remaining 20% of the proteins originate from brain and many of these proteins are among the most abundant ones in the CSF (Table 1). The total protein concentration in CSF is approximately 350 mg/L, that is, about 200 times lower than in plasma. But like in plasma samples, depletion of high-abundant proteins like albumin, that mask more interesting low-abundant proteins, is required prior to proteomic profiling. Characterization of the CSF proteome from

neurologically normal individuals as well as patients with a certain neurological disease has been in focus in recent years. Several studies show that alterations in protein profiles of CSF may reflect abnormalities associated with a diverse array of neurological diseases, for example, traumatic brain injury, neurodegenerative disorders, MS, and hydrocephalus.[35-42] Even though many proteins have been identified it has not always been determined whether these proteins are causally related to the disease or not.

It is clear that a large diversity of neurological diseases could benefit from profiling of CSF proteome and subsequent identification of protein biomarkers. These markers can then be associated with onset and progression of a disease as well as predicting response to therapy.

Table 1. *The most abundant proteins and their concentrations in the cerebrospinal fluid (CSF) and plasma.[34] Brain derived CSF proteins are marked with an asterisk.*

CSF proteins	Conc. (mg/L)	Plasma proteins	Conc. (mg/L)
Albumin	200	Albumin	45 000
Prostaglandin D-synthase*	26	IgG	9 900
IgG	22	α-lipoprotein	3 000
Transthyretin*	17	Fibrinogen	3 000
Transferrin	14	Transferrin	2 300
α_1-antitrypsin	8	β-lipoprotein	2 000
Apo-lipoprotein A	6	α_2-macroglobulin	2 000
Cystatin C*	6	α_1-antitrypsin	1 400

2.4 Multiple sclerosis (MS)

MS is a chronic, immune-mediated disease of the central nervous system (CNS). A triggered immune system is causing an inflammatory demyelination (loss of myelin, the protective sheath surrounding nerve fibres of CNS), axonal damage (destruction of nerve fibres) and often followed by irreversible neurological disability. MS has an unpredictable clinical course and shows usually a gradual accumulation of disability (both physical and cognitive), with major impact on normal family life and social roles. However, the outcome of the disease is heterogeneous and therefore at present impossible to predict for an individual patient.[43, 44] The cause of the disease is unknown, but it is believed that MS is a result of interplay between genetic and environmental factors. No curative treatment exists, but several disease modifying preparations are available.[44]

2.4.1 Epidemiology

MS is estimated to affect about 2.5 million individuals worldwide, and it is the most common, non-traumatic cause of disability in young adults.[43] Epidemiological studies have shown that there is a large variation in the geographical distribution of MS. The prevalence varies with latitude being higher the farther from the Equator one lives. Northern and central Europe, USA, Canada, Australia and New Zealand are considered as high-risk areas.[45] The prevalence of the disease ranges between 2 and 150 per 100 000 inhabitants, depending on the country or specific population. The prevalence rate in Norway is probably above 150 per 100 000,[46] but being lower in the northern parts of the country.[47] This difference in distribution may imply differences in both genetic and environmental risk factors.

The disease appears most often between 20-40 years of age, but it can also appear in children, and females are more affected than males.[46, 48] A ratio of 2:1 is typically reported, and recent studies suggest an increasing incidence for women, indicating gender specific changes in environmental risk factors.[49-51]

The cause of MS is unknown. Many putative environmental risk factors have been suggested but only a few have been confirmed to an increased risk of developing the disease.[52] These include sunlight exposure and vitamin D, infection with Epstein-Barr virus, and smoking.[53] About 20% of patients with MS have at least one affected relative, indicating genetic risk factors. Association to human leukocyte antigen (HLA) genes have been known for many years.[54] But recently, several other immune related genes have been identified, amongst them IL2RA and IL7RA, and the number is increasing. [55, 56]

2.4.2 Symptoms and clinical subtypes

The symptoms of MS are caused by lesions in the CNS and reflect the part of CNS which is involved. Typical clinical symptoms at the onset are numbness or paresthesia, weakness in upper or lower extremities, optic neuritis, double vision and dizziness, and coordination difficulty.[44]

The disease may be divided into relapsing remitting MS (RRMS), or primary progressive MS (PPMS) according to the initial disease course (Figure 10). RRMS account for 80-85% of the patients and is characterized by a series of unpredictable relapses with full or partial recovery in between. A relapse is defined as significant worsening of pre-existing symptoms or appearance of new neurological symptoms characteristic for MS which are developing over days or weeks, lasting at least 24 hours, and are not associated with fever or intercurrent illness.[57, 58] Approximately 50% of the patients with RRMS experience a gradual decline and convert to secondary progressive MS (SPMS) within 10-15 years of disease onset.[59] In PPMS the disease gradually progresses with steady increase in disability without any periods of relapses or recovery. In addition to these three subtypes, clinically isolated syndrome (CIS) is often referred to as a first indication of the disease.

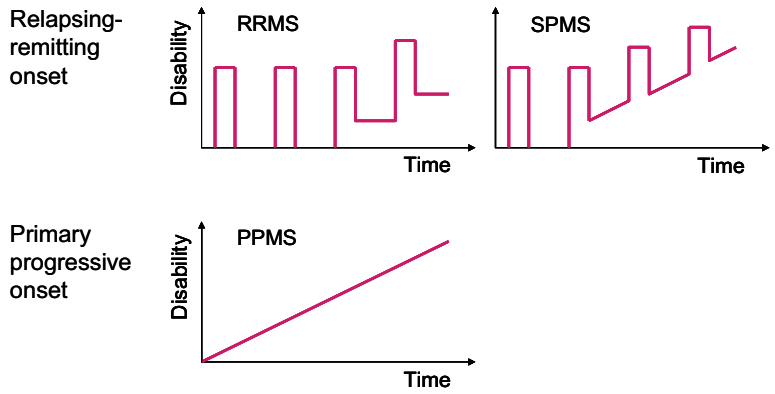


Figure 10. *Clinical subtypes of MS. The figures illustrate the two main initial (onset) courses of MS, relapsing-remitting MS (RRMS) and primary progressive MS (PPMS), and the subgroup of secondary progressive MS (SPMS).*

2.4.3 Diagnosis and treatment

Diagnosing MS may be challenging. There is no single test to make a definitive diagnosis and signs and symptoms of MS may be similar to other neurological problems. Therefore, diagnostic criteria have been created to ease and standardize the diagnostic process for physicians. Currently, the McDonald[57] and the revised McDonald criteria[58] focus on a demonstration of the dissemination of MS lesions in time and space (Table 2). The diagnosis is based on disease history and clinical examination combined with visualization of lesions with magnetic resonance imaging (MRI) (Figure 11) and detection of oligoclonal bands of IgG and barrier index in the CSF. In general, to make a diagnosis of MS, an individual must have 2 episodes of neurological symptoms referable to the CNS that are separated in space and time and that are not attributable to any other cause. Repeated MRI examinations can, however, substitute one of the clinical episodes. It may be time-consuming to establish a correct diagnosis and thus valuable time can be wasted in the first stages of the disease. In addition, subclinical disease activity with irreversible damage may occur even before the first clinical symptoms are detected. Early therapy usually slows down disease activity,[60] and might be important for the long time prognosis of the disease.[44] Thus,

there is a definitive need for sensitive and specific tests like disease biomarkers to enable an early diagnosis. No cure is available, but corticosteroids can be used to shorten relapses, various symptomatic treatments exist, and several long-term, disease modifying therapies are available.[59]

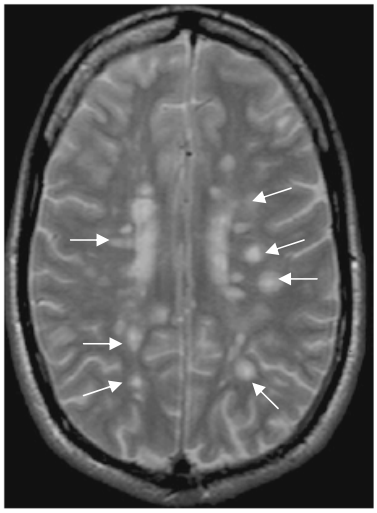


Figure 11. *T2-weighted magnetic resonance image (MRI) showing typical periventricular hyperintensive multiple sclerosis lesions.*

Table 2. *Revised McDonald diagnostic criteria for multiple sclerosis (MS).*[58]

<table>
<tr><th colspan="2">Clinical presentation</th><th rowspan="2">Additional data needed for diagnosis</th></tr>
<tr><th>Attacks</th><th>Lesions</th></tr>
<tr><td>≥2</td><td>≥2</td><td>None. But MRI and CSF analysis should be done to exclude other diagnoses. If these tests are negative, extreme caution needs to be taken before making a diagnosis of MS.</td></tr>
<tr><td>≥2</td><td>1</td><td>Dissemination in space, demonstrated by:
o MRI or two or more MRI-detected lesions consistent with MS plus positive CSF
or
o Await a further clinical attack implicating a different site</td></tr>
<tr><td>1</td><td>≥2</td><td>Dissemination in time, demonstrated by:
o MRI
or
o A second clinical attack</td></tr>
<tr><td>1</td><td>1</td><td rowspan="2">Dissemination in space, demonstrated by:
o MRI or two or more MRI-detected lesions consistent with MS plus positive CSF
and
Dissemination in time, demonstrated by:
o MRI
or
o A second clinical attack</td></tr>
<tr><td colspan="2">Monosymptomatic presentation; clinically isolated syndrome (CIS)</td></tr>
<tr><td colspan="2">Insidious nervous system progression suggesting primary progressive multiple sclerosis (PPMS)</td><td>One year of disease progression (retrospectively or prospectively determined) and two of the following:
o Positive brain MRI (nine T2 lesions or four or more T2 lesions with positive VER)
o Positive spinal cord MRI (two focal T2 lesions)
o Positive CSF</td></tr>
</table>

MRI, magnetic resonance imaging; CSF, cerebrospinal fluid; VER, visual-evoked response

3. Instrumentation

"Technology is dominated by two types of people: those who understand what they do not manage, and those who manage what they do not understand." Putt's Law

Proteomics is built on technologies that enable the analysis of a large number of proteins in a single experiment. However, there is no single technique capable of detecting all proteins in complex biological samples like body fluids or cell and tissue extracts.

This chapter provides a brief introduction to proteomics techniques used in the analysis of body fluids, and instrumentation in general. The main focus is on the use of biological mass spectrometry, and a detailed description of MALDI-TOF, the technique used in this thesis, is given.

3.1 Proteomics techniques

The overall proteomics strategy can be divided into four subsequent steps, that is, sample preparation, protein separation, identification and characterization (Figure 12).[32] Each step in this pipeline is equally important for the overall success and involves techniques with technical challenges.

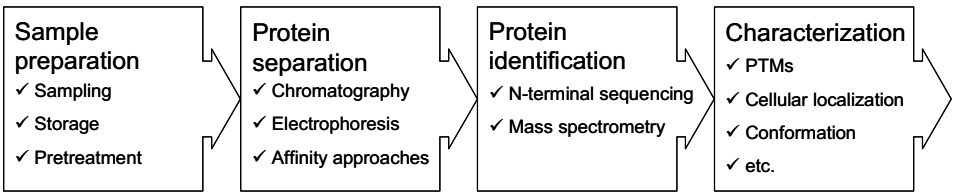


Figure 12. *Pipeline of proteomics techniques.*

Sample preparation, including preanalytical procedures like sample collection, storage, and pretreatment, may have a significant role in the overall analytical result. Preanalytical effects may alter the proteome and thus give rise to false biomarkers. Each body fluid requires an adapted sample preparation protocol tailored to the specific needs of the analytes in question.[61] Development of such a protocol is beyond the scope of this thesis. In an earlier published work, however, we have reported the preanalytical influence on the low molecular weight (MW) proteome of CSF.[29] Issues like blood contamination, different sample storage conditions, and different types of MW cut-off filters are included in this prestudy. The recommended protocol is applied to the proteomic study of CSF samples in this thesis (described in detail in Section 5.1).

When analyzing human body fluids, one challenge is how to handle the extreme complexity of the samples; for example, blood plasma may contain more than one million protein forms.[32] Another challenge is the large dynamic range of protein concentrations present in the samples.[62] For instance, in plasma proteome two clinically useful proteins, serum albumin (at the high abundance end) and interleukin 6 (at the low abundance end), differ by 10 orders of magnitude.[63] Proteins with high abundances can dominate in such a high degree that the instruments are not sensitive enough to measure the proteins with low abundances. Good separation techniques are therefore needed to make the protein mixtures less complex and thus enable mass spectral analysis and subsequent identification. The target proteins can be enriched using prefractionation methods with which the most abundant proteins masking the low abundance proteins are removed. In the actual fractionation procedure, the maximal separation between the different proteins is desired. An overview of the approaches used in proteomic analysis of body fluids is presented in Figure 13.[61]

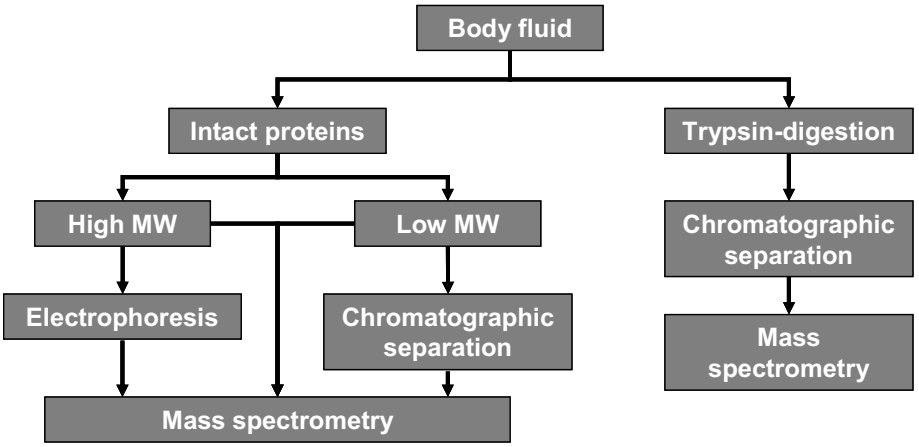


Figure 13. *Overview of different approaches in body fluid proteomics.*

The body fluid proteome can be divided into high and low MW fractions (also referred to as the peptidome) to detect small proteins and peptides with low abundances. One way to achieve this is to utilize MW cut-off filters.[64, 65] Only proteins below a chosen threshold can pass through the filter, thus giving a good enrichment of the smaller proteins and peptides.

The most widely used separation technique has traditionally been two-dimensional gel electrophoresis (2DE).[66-68] The method is a combination of two separation techniques, isoelectric focusing and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and it separates complex mixtures into single protein spots based on their isoelectric point (p*I*) in the first dimension and size (M_r) in the second dimension. After separation the protein spots are stained (*e.g.* silver or coomassie staining) to enable visual inspection. The spots can be removed from the gel by cutting and then subjected to tryptic digestion and identification by mass spectrometry (usually MALDI-TOF). Even more than 5000 proteins can be simultaneously resolved in one gel.[69] However, the loading capacity of 2DE gels is limited making the detection of low-abundance proteins very difficult without proper prefractionation. Other problems associated with 2DE are, for example, poor reproducibility and

laborious analytical work-up, which make it unsuitable for the analysis of large series of clinical samples.

In chromatographic separation methods the proteins in the sample distribute between the stationary and mobile phases. Chromatography can be used as a prefractionation tool or it can be coupled directly to mass spectrometry (so-called hyphenated techniques).[70-72] Affinity chromatography is the most commonly used prefractionation method when analyzing body fluids. In affinity chromatography the proteins are separated based on their binding to other molecules such as antibodies. The method is mainly used for removal of the most abundant proteins, for example, albumin and immunoglobulin G (IgG). Also ion-exchange chromatography and reverse phase high-performance liquid chromatography (RP-HPLC) can be utilized in prefractionation. Chromatographic separation can also be combined with 2DE giving a third orthogonal dimension for protein separation. In the so-called shotgun proteomics the entire proteome is first digested into peptides, followed by one- or multidimensional chromatographic separation prior to analysis with mass spectrometry.[73]

Several techniques are developed and can be used to detect, identify and measure hundreds or even thousands of proteins. SELDI, MALDI, capillary electrophoresis (CE) and multiple reaction monitoring (MRM, also known as selected reaction monitoring, SRM) are all mass spectrometry based methods having a potential to be applied in clinical chemistry laboratories.[22] SELDI, MALDI and CE are profiling methods providing proteomic fingerprinting of samples while SRM appears as a potential alternative to classical immunoassays by combining analytical specificity and reliable quantification.

SELDI utilizes different adsorptive surfaces to bind a certain subgroup of proteins. The bound proteins or peptides are then analyzed using the same principle as in MALDI mass spectrometry. The technique has been commercialized under the name ProteinChip®.[74] Samples are loaded onto the ProteinChip array coated with a chemically treated surface (*e.g.* hydrophobic or hydrophilic). The unbound proteins

are washed away, and the bound proteins are mixed with an energy-absorbing matrix and subjected to mass spectral analysis. In MALDI a sample is mixed with a matrix, deposited directly on a MALDI plate and targeted with the laser beam. MALDI method is described in detail in Section 3.3.

In CE mass spectrometry, fractionation using CE is combined with mass spectrometry detection.[75] In CE, high-efficiency separation is achieved by applying high voltages to generate electro-osmotic or electrophoretic flow of buffer solutions and ionic species within narrow-bore capillaries. MRM has recently become available also for proteomic analysis due to technical development of mass spectrometers (*i.e.* triple quadrupoles). In the first quadrupole mass of the parent ion is selected. In the second quadrupole the parent ion is fragmented by collision. Finally, in the third quadrupole, a specific fragment of the parent ion is selected.

3.2 Biological mass spectrometry

Mass spectrometry is one of the most powerful instrumental techniques in analytical chemistry. It can measure a panel of analytes in a single assay thus allowing, for example, full profiling of the body fluid samples. Advances in mass spectrometry technology have made today's proteomics research possible.

In mass spectrometry the measurements are carried out in a gas phase. Therefore the sample has to be first volatized and ionized using an ion source. The generated ions are guided to a mass analyzer where they are separated based on their mass-to-charge ratio (*m/z*). A detector then registers the number of ions at each *m/z* value. The output signal is acquired and presented as a spectrum of measured *m/z* values for each analyzed sample. A schematic presentation of a mass spectrometer is shown in Figure 14.

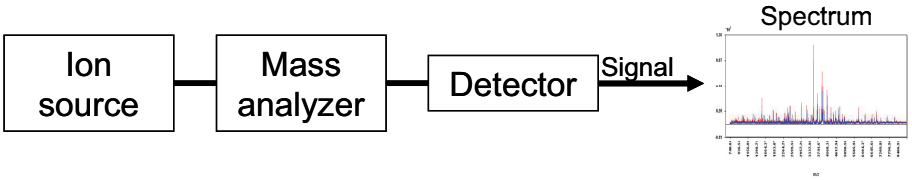


Figure 14. *A mass spectrometer consists of three parts: ion source, mass analyzer and detector. Mass spectrum is generated from the detector signal using computer.*

For the analysis of large polar biomolecules like proteins, soft ionization techniques are employed. Using electrospray ionization (ESI) and MALDI the molecules remain relatively intact. ESI ionizes proteins or peptides out of a solution and it can be coupled directly to separation techniques like liquid chromatography (LC).[76] The sample solution is supplied as a continuous flow through a needle creating an aerosol. The solvent evaporates and positively charged molecular ions are formed. MALDI sublimates and ionizes proteins or peptides out of a dry matrix using laser pulses. These ionization sources can be interfaced to different mass analyzers. In proteomics research the most commonly used mass analyzers are time-of-flight (TOF), quadrupole, ion trap, and Fourier transform ion cyclotron resonance (FT-ICR) analyzers. The operation of quadrupole and ion trap mass analyzers is based on ion motion in radio frequency electric fields.[77] The resolution and accuracy of these instruments is not very high. FT-ICR instruments, on the other hand, have extremely high performance, and separate the ions based on their cyclotron frequency in a fixed magnetic field.[78] TOF mass analyzer is described in detail in the next section.

3.3 MALDI-TOF mass spectrometry

MALDI source coupled to TOF mass analyzer is one of the most common interfaces. MALDI-TOF offers a rapid approach to the analysis of intact proteins in body fluid samples. It is a relatively simple method with excellent mass accuracy, high resolution

and sensitivity all of which makes it ideal for high-throughput profiling of proteomic samples.[11] In addition, MALDI-TOF is cost-effective and the instruments are easy to operate demanding relatively simple analytical work-up. Working mass range of TOF analyzers is large (approx. 100 Da to 250 000 Da) and only singly charged molecular ions are observed in MALDI ionization.[77]

In MALDI the biomolecules are caught within a crystalline structure, referred to as a spot, and bombarded with laser pulses. This bombardment is usually repeated several times at different positions in a spot, thus generating an average spectrum. Intensity variations, which can be observed even between the spots originating from the same sample, can be reduced in this way. Sample preparation is crucial in MALDI since the type of matrix and presence of possible impurities affect the generated ions. Features like chemical properties of the matrix, its proportion to the analyte and the way it co-crystallizes with the sample all affect the spectra and their reproducibility.[26, 79, 80]

MALDI-TOF is not regarded as a quantitative technique, but studies have shown that optimization makes it possible to reduce analytical variance substantially making the MALDI-TOF and similar mass spectrometry approaches produce more quantitative results. [81, 82] It should be noted that protein identification is not possible using standard MALDI-TOF since proteins cannot be identified based on their molecular weight only.

3.3.1 MALDI ionization

MALDI ionization was introduced for the first time in the late 1980s by Koichi Tanaka and Franz Hillenkamp, following application of use of lasers for the ionization and analysis of biomolecules.[83, 84] The decisive factor was the use of matrix, which makes it possible to perform the ionization without destroying large organic molecules.

The principle of MALDI ionization is illustrated in Figure 15. Non-volatile samples are mixed with an excess of light-absorbing matrix in an aqueous or organic solvent.

Matrices are usually small organic acids having conjugated aromatic ring structure, *e.g.* α-cyano-4-hydroxycinnamic acid (CHCA) and 2,5-dihydroxybenzoic acid (DHB). A small amount of this mixture is then introduced onto a MALDI sample plate. The solvent is evaporated and a layer of co-crystallized mixture of sample and matrix is obtained.[85, 86]

The sample plate is placed in the mass spectrometer under high vacuum and is irradiated with a pulsed laser beam (*e.g.* nitrogen laser at 337 nm).[11, 77, 87] With laser it is possible to deliver coherent and high density energy to a small space. The matrix absorbs energy at the wavelength of the laser while the sample itself remains intact. The matrix is sublimated and the mixture of sample and matrix is rapidly expanded into the gas phase. In the formed dense cloud, the energy is transferred from matrix to sample and desorption occurs by proton transfer. Singly charged molecular ions are produced since analytes in the sample usually accept a proton. Desorption process is followed by desolvation and subsequent introduction into mass analyzer. A packet of ions with different *m/z* values is generated.

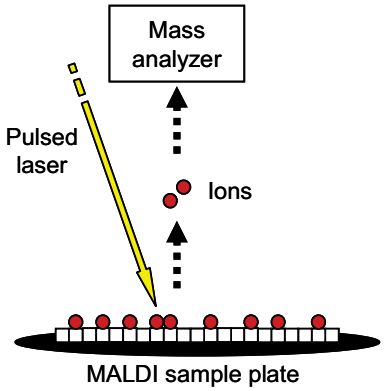


Figure 15. *Principle of matrix assisted laser desorption/ionization (MALDI). The sample and an excess amount of matrix are mixed and exposed to laser irradiation. Both neutral molecules and ions are desorbed. The gaseous ions are guided further to mass analyzer and detector.*

3.3.2 TOF mass analyzer

TOF mass analyzer measures the mass of an ion by determination of mass-to-charge ratio *(m/z)* from its flight time.[77, 85] TOF analyzer has a large mass working range, a feature that makes it ideal for analyzing *e.g.* intact proteins.

Gaseous ions produced by MALDI are guided to TOF analyzer where they are accelerated using fixed potential difference (*e.g.* 20-30 kV) into a field-free flying tube. Since all the ions are exposed to the same potential, all similarly charged ions will have same kinetic energies. When ions pass through the field-free region they are separated according to their velocities; ions with larger mass have lower velocities than ions with smaller mass. The ions hit the detector at the end of the flying tube and a signal is produced. TOF mass spectrum is the detector signal as a function of time. The flight time (*i.e.* all ions have the same start time, and the arrival time is recorded at the detector) for each individual ion is proportional to the square root of the *m/z* of a particular ion. Axis of the spectrum can be converted into an *m/z* ratio axis thus producing the conventional mass spectrum.

TOF can be used in either linear or reflector mode.[77] In linear mode the ions fly through a linear flying tube and their *m/z* is determined by the time required for the ions to reach the detector. In reflector mode, a reflecting field at the end of the flying tube is used. The ions turn around in the reflector and then hit the detector that amplifies and counts the ions. Use of reflector compensates for slight differences in kinetic energies which may occur even though the mass is the same. The result is sharper peaks in the spectrum.

4. Multivariate data analysis

"Then there is the man who drowned crossing a stream with an average depth of six inches." W.I.E. Gates

It should be kept in mind that measured data is not the same as information. Therefore an important issue in all empirical sciences, including proteomics, is how to reveal the relevant information in the data. For example, Spiegelman *et al.* [88] argue that *"Rigorous application of sound statistical and chemometric principles will benefit the overall scientific community by improving protein biomarker discovery and validation."*

Chemometrics can be defined as "information aspects of chemistry" [89] where statistical and mathematical methods are used *i)* to produce "good data" and *ii)* to extract relevant information out of measured data. The first aim can be achieved by using design of experiments (DoE) to provide a minimum number of information-rich experiments. Multivariate data analysis can be employed for the second purpose. In addition visualization of the data is an important issue and can be seen as part of chemometrics. The methods used in chemometrics are fully applicable in biosciences as well as other empirical sciences. In proteomics multivariate projection methods developed in chemometrics can be used to simplify complex proteomic data and make the visualization of spectral fingerprints easier. Furthermore, they make classification of samples and detection of biomarker signatures possible.

This chapter describes chemometric methods used in this thesis. Data pretreatment methods are presented in the first section. DoE and empirical modelling including multiple linear regression (MLR) are discussed in the second section. The third section describes latent variable methods including classification. Variable selection methods are discussed in the last section. The emphasis is placed on the two new methods developed as part of this thesis, selectivity ratio (SR) and discriminating

variable (DIVA) test. More philosophical articles about chemometrics can be found elsewhere in the literature.[89-93]

4.1 Data pretreatment

The acquired spectral profiles are arranged in a way that each row in a table represents one sample and each column one measured *m/z* number (Figure 16). There are many experimental and instrumental effects that are not related to compositional differences between samples and thus make comparison of mass spectral profiles from different sample groups difficult. Examples of sources of variation are, for example, sample collection and storage, sample preparation and instrumental artefacts. In order to remove these disturbing factors and ensure that all the collected spectra can be analyzed jointly, proper data pretreatment is necessary prior to data analysis. Pretreatment has a significant effect for the final results and should therefore be considered carefully. Crucial factors affecting the data analysis are baseline effects, shifts in *m/z* values (alignment problem), structured noise (heteroscedasticity), and differences in signal intensities caused by analytical workup and the instrumental technique (normalization problem). Other pretreatment steps to be considered are smoothing and data reduction.

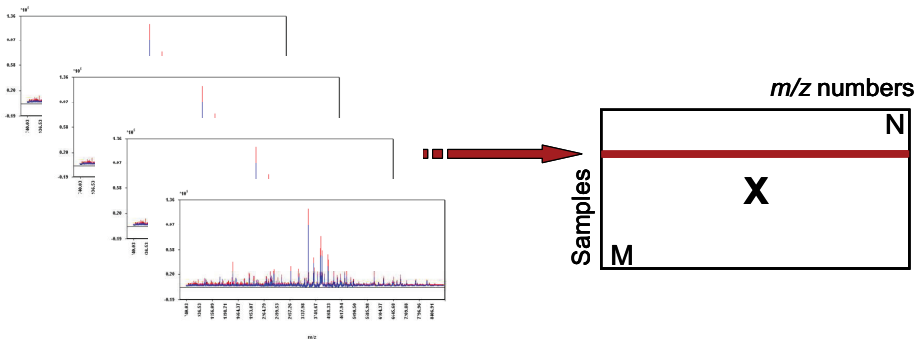


Figure 16. *A spectrum is acquired for each sample and the data (intensities) are compiled in a table (matrix* ***X****, M×N) where each row represents one sample and each column one m/z number.*

4.1.1 Baseline correction

All mass spectra have a varying baseline. The baseline is an *m/z* dependent offset which needs to be subtracted without simultaneously removing compositional information in the spectrum. Baseline effects differ a lot between different mass spectrometric techniques, and several approaches have been proposed for baseline correction.[94-96] Most instrument vendors have implemented their own algorithm for baseline correction. However, the results are not perfect and some algorithms yield negative intensities in the spectral profiles. In this thesis the problem of negative intensities has been corrected using two methods: negative spectral intensities can simply all be substituted with zero (Paper IV) or they can be corrected by assigning zero intensity to the lowest signal in each spectrum (Papers I-III). In the latter approach the whole spectrum is lifted by adding the absolute value of the largest negative intensity to the intensities at each *m/z* number throughout the entire profile and independently for each of them.

4.1.2 Smoothing

Noise in mass spectra may lead to problems when cross-correlation algorithms are used for alignment of the spectra. This problem may be reduced by smoothing the data. Smoothing acts like a filter where the aim is to increase the signal-to-noise ratio without distorting the signal. Moving average and Savitzky-Golay are the most common methods used for smoothing. Moving average creates a series of averages for subsequent subsets (with fixed window size) over the entire data set. The Savitzky–Golay method is based on local polynomial regression on a series of values to determine the smoothed value for each point.[97] In this thesis, moving average method with a window size of 10 is used.

4.1.3 Alignment

In all multivariate analysis an underlying assumption is that each column of the data matrix represents the same variable in all samples. Otherwise we cannot expect to be able to extract correct information from the data, for example, when comparing spectral profiles. Even small shifts in a series of spectral profiles, may cause large inconsistencies and serious reduction of the cross-correlation between chemically similar profiles. The problem of *m/z* shift between corresponding molecules in different spectral profiles is called the alignment or synchronization problem, and it represents an obstacle to the comparison of full spectral profiles.

One very common solution to avoid the alignment problem is to reduce spectral profiles to peaks. Unfortunately this peak picking approach leads to loss of information in complicated spectra due to overlapping peaks. A better approach is to use techniques that maximize cross-correlation between a set of spectral profiles. Andersson and Hämäläinen developed a method where cross-correlation between selected target peaks present in all profiles was maximized using Simplex optimization.[98] Entire instrumental profiles were shifted piecewise and independently according to the optimal fit with the target peaks. In correlation optimized warping (COW) linear stretching and compression is used to cross-correlate piecewise the profiles to a target profile.[99] Both these methods were originally developed for chromatographic profiles. According to our experience, the COW method is time-consuming when working with complex spectral profiles. Wong *et al.* developed cross-correlation methods based on fast Fourier transform.[100, 101] In this thesis all spectra are aligned to an average spectrum using recursive alignment by fast Fourier transform (RAFFT) cross-correlation function with window size 20. Similarly to COW, this method aligns the profiles piecewise and optimizes cross-correlation between reference spectrum and profiles to be aligned for each segment.

4.1.4 Data reduction

Full spectral profiles are truly multivariate. Each profile is described by tens of thousands of variables and huge data sets are thus provided for a set of samples. This may in turn give rise to data processing problems. Peak picking is one way to avoid this problem, but as mentioned earlier, reduction of profiles to peaks leads to loss of information. Other methods used for data reduction are maximum-entropy reduction and binning. Maximum-entropy method assumes that the information content is higher in regions with high intensity.[102] In proteomics data, however, the low intensity regions may in fact contain more information than the high intensity regions. Therefore the low intensity regions should not be reduced more than the high intensity regions.

In this thesis data reduction is achieved using binning. Binning is performed by adding adjacent *m/z* numbers throughout the spectrum using a fixed window size. A "good" description of the features in the profiles should be retained and the window size should be chosen according to the number of points needed to describe a typical peak in the profile. For instance, if a peak is originally described by 100 *m/z* numbers, a window size of 10 would usually be appropriate to balance between the time needed for data processing and still retaining the spectral features.

Both binning and maximum-entropy has an additional effect of smoothing the spectra. Therefore smoothing is not used if binning is performed.

4.1.5 Structured noise and heteroscedasticity

Structured noise represents a major problem when comparing mass spectral profiles. Increasing noise with increasing signal size is called heteroscedasticity. This type of noise influences normalization and gives false negative correlations between major peaks.[103, 104] It also impacts the minor peaks by giving false positive correlations between them. Therefore heteroscedastic noise should be transformed to homoscedastic noise prior to data analysis.

Logarithmic or n^{th}-root transformations have been used to provide a homoscedastic noise pattern.[103] Figure 17 visualizes the phenomenon and shows the effect of different mathematical transformations on heteroscedasticity. The logarithmic transformation destroys linear correlations in the profiles. This is a problem in spectral profiles where one component is described by many linearly correlated *m/z* numbers. The n^{th}-root transformation preserves perfect linear correlation, but reduces correlations in regions with only partial correlation. Furthermore, the n^{th}-root transformation reduces the intensity of major peaks compared to minor peaks. This feature may in fact turn out to be an advantage since the information content in minor peaks may in many cases be higher than in major peaks. In this thesis, the n^{th}-root transformation is used to correct for heteroscedasticity.

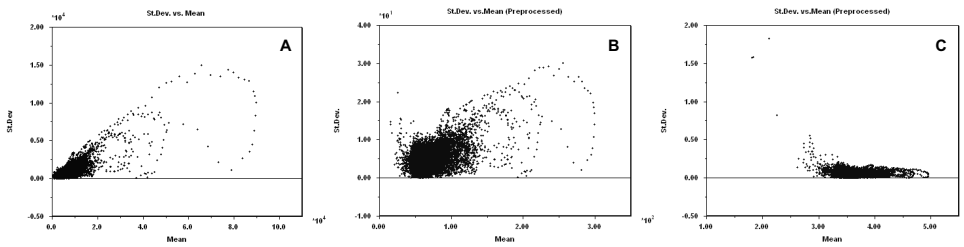


Figure 17. *Example of heteroscedastic noise and the effect of different transformations. Standard deviation versus mean intensities for baseline corrected matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry data. A) No transformation B) Square root transformation C) Log transformation.*

4.1.6 Normalization

Signal intensities between spectra derived from a certain molecule having the same concentration in different samples can vary a lot. This is due to factors like sample handling, compositional differences in the samples, distribution in the matrix and multiplicative effects. This means that the acquired profiles are not directly comparable with each other. Without internal standards, mass spectral data are usually normalized to create profiles with relative intensities. The most common methods used are normalization to constant sum or constant length.

Normalization to constant sum:

$$\mathbf{z}_i^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}} \Big/ \sum_{j=1}^{N} \mathrm{x}_{ij} \qquad \mathrm{i} = 1,2,\ldots \mathrm{M} \qquad (1)$$

In Eq. 3, $\mathbf{x}_\mathrm{i}$ and $\mathbf{z}_\mathrm{i}$ represent the profile for sample *i* before and after normalization, respectively. *N* is the number of points describing the profile and *M* is the number of instrumental profiles acquired. The transpose is indicated by a superscript T. Vectors are by default column vectors and transpose transforms them into row vectors. It should be noted that the so-called TIC normalization, as implemented by Conrad *et al.*,[105] differs from normalization to constant sum only by a scalar.

Normalization to unit length (the norm):

$$\mathbf{z}_i^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}} \Big/ \left\| \mathbf{x}_i \right\| \qquad \mathrm{i} = 1,2,\ldots \mathrm{M} \qquad (2)$$

Spectral regions with higher intensities get larger weight when the normalization to unit length is used. This means that the method is more sensitive to heteroscedastic noise compared with normalization to constant sum. To avoid this problem the profiles should be properly corrected for heteroscedasticity prior to normalization. In this thesis normalization to unit lenght is used, but calculations confirm that there are no real differences between the two procedures when heteroscedasticity is first taken care of.

4.2 Design of experiments (DoE) and empirical modeling

The purpose of an experiment is to obtain new information. To understand why certain experimental conditions give good results it is essential to realize why other conditions do not. This means that we have to introduce variation to the variables and thus perform experiments with both desired and undesired outcome. An empirical model based on the experimental data can then be estimated and used for interpretation and prediction. The quality of obtained information is dependent on the set of all experimental runs, that is, the experimental design.[106]

Models can be seen as tools to estimate reality. All models are more or less erroneous, since there is always noise in the data. Experimental error is a variation produced by both known and unknown disturbing factors that may disguise important effects wholly or partially. These confusing effects can be reduced by using experimental design and statistical analysis yielding measures of statistical qualities for estimated variables.

Confusion of correlation with causality is a common problem in all empirical research. Correlation between two variables often occurs because they are both associated with a third factor meaning that correlation does not automatically imply that the two variables have causal relationship. When data are generated using experimental design we can calculate the real effects and reveal the causality behind the studied phenomena. In addition not only linear and additive effects but also effects of the interactive and non-additive kind may be estimated by using experimental design.

The measurable result of an experiment depends on the experimental conditions, that is, variables that can be controlled by the experimenter.

$$Result = f(experimental\ conditions) \qquad (3)$$

In empirical modelling the objective is to investigate how these controlled variables influence the result. In addition it is possible to optimize the system, that is, to find the

optimal variable settings to obtain the best result. These objectives can be achieved by means of a model, where the observed result, *i.e.* response (y), is described as a function of the controlled variables, *i.e.* factors ($x_1,x_2...x_j$). The noise is left in the residual (ε).

$$y = f(x_1,x_2...x_j) + \varepsilon \qquad (4)$$

For practical purposes the function f can usually be approximated by using polynomial functions, since they give sufficiently good description of the relationship between factors and responses within a limited experimental domain. It should be noted that this type of models are local and cannot necessarily be extrapolated.

The idea behind DoE is to set up a design, in which all studied factors are varied systematically.[107] Two level factorial designs form the basis for classical experimental designs. Here each factor is investigated at two levels: a low level and a high level. For continuous variables (*e.g.* temperature) this signifies two numerical values, but for discrete variables (*e.g.* type of instrument) this signifies two different alternatives. In a full factorial design all possible settings of the factors (*i.e.* low/high combinations) are included and with N factors the number of experiments will be 2^N. Geometrical presentations of full factorial designs for two (2^2) and three (2^3) factors are shown in Figure 18.

In a fractional factorial design only a selection of all possible combinations is executed. This reduces the number of experiments and is especially useful when screening many factors. In optimization, however, more experiments are needed to be able to calculate a model with higher terms. Central composite designs can be employed for this purpose. It is always recommended to perform replicated experiments at some factor settings; this will give an estimate of experimental error.

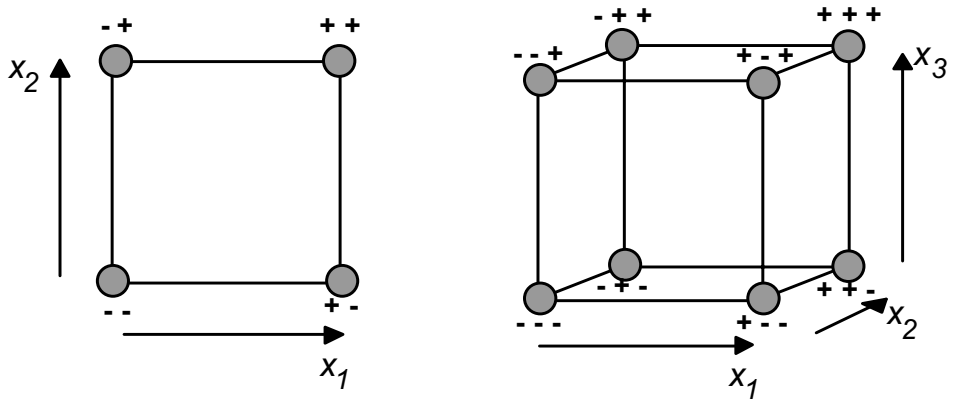


Figure 18. *Examples of two level factorial designs for two (2^2) and three factors (2^3). Sign combinations in corner points show the levels (low -/ high +) used for each factor (x_i) in that particular experiment.*

4.2.1 Multiple linear regression (MLR)

In predictive modelling the objective is to determine the relationship between several x-variables (independent variables) and one or more y-variables (dependent variables) (see Eq. 4). For instance, a model with linear, interaction and quadratic terms for two x-variables can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 + \varepsilon \qquad (5)$$

Where b_i (i = 0, 1, 2,…) are regression coefficients describing the effect of each calculated term. Eq. 5 can be written in matrix form:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \qquad (6)$$

Parameters **b** can be estimated so that the sum of the squared residuals will be as small as possible, that is, we use least squares fit. MLR is often used for estimating regression vector **b**. From the Eq. 6 we get:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (7)$$

The matrix $\mathbf{X}^T\mathbf{X}$ is always square but unfortunately it does not always have a full rank meaning that the usual inverse does not exist. The matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ in Eq. 7 is then called a generalized inverse of **X**.

Measured data are often collinear – variables are not independent and they contain the same information. MLR cannot be used in these situations. This is the reason why latent variable regression methods like partial least squares (PLS) have been developed. Instead of using strongly dependent variables in the regression we calculate a new set of (latent) variables that have a reduced dimensionality.

4.3 Latent variable methods

Characterization of complex chemical and biological systems produces multivariate, and in addition, collinear data, that is, variables describing partially or fully the same property and therefore sharing the same information content. Therefore, the original measured variables can be linearly combined to fewer, so-called latent variables, which describe the underlying structure in the data. In modelling the aim is to separate information from noise and find the patterns in the data. The concept of latent variables was first applied in psychometrics.[108] This section provides a short description of latent variable methods used in this thesis. The methods are otherwise well described in the literature.

4.3.1 Latent variable projections

Each matrix can be geometrically presented in two co-existing spaces, variable space and object space, which together contain all available information in a data matrix

(Figure 19).[109] Each object (sample) is described by N measured variables thus forming an object vector, $\mathbf{x}_i^T$. Object vectors can be arranged in a matrix $\mathbf{X}$, where each row represents one object. In the same manner, each variable is described by its values for all the M objects, that is, a variable vector, $\mathbf{x}_j$. When variable vectors are arranged in a matrix $\mathbf{X}$, each column represents one variable. To visualize the data structure, object vectors can be plotted in variable space, where the number of axes is equal to the number of variables. In this way all the information in $\mathbf{X}$ regarding the relationships (similarities or differences) between objects can be displayed. Similarly, variable vectors can be plotted in object space, where the number of axes is equal to the number of objects. In this way the relationships (correlations/covariances) between variables can be shown. Since the object space shows common variation in a set of variables it also reveals underlying factors, that is, latent variables (LV). When the number of variables increases, the challenge is to find low-dimensional projections of both variable and object space. This can be achieved by projecting onto LVs. Different projections can be calculated using a generalization of NIPALS algorithm (Box 1).[110]

Box 1. Successive orthogonal projections

i) Select $\mathbf{w}_a$

ii) Project objects on $\mathbf{w}_a$:

$$\mathbf{t}_a = \mathbf{X}_a\mathbf{w}_a$$

iii) Project variable vectors on $\mathbf{t}_a$:

$$\mathbf{p}_a^T = \mathbf{t}_a^T\mathbf{X}_a/\mathbf{t}_a^T\mathbf{t}_a$$

iv) Remove the latent-variable from predictor space, i.e. substitute $\mathbf{X}_a$ with $\mathbf{X}_a - \mathbf{t}_a\mathbf{p}_a^T$.

Repeat i) - iv) for a= 1,2,..A, where A is the dimension of the model

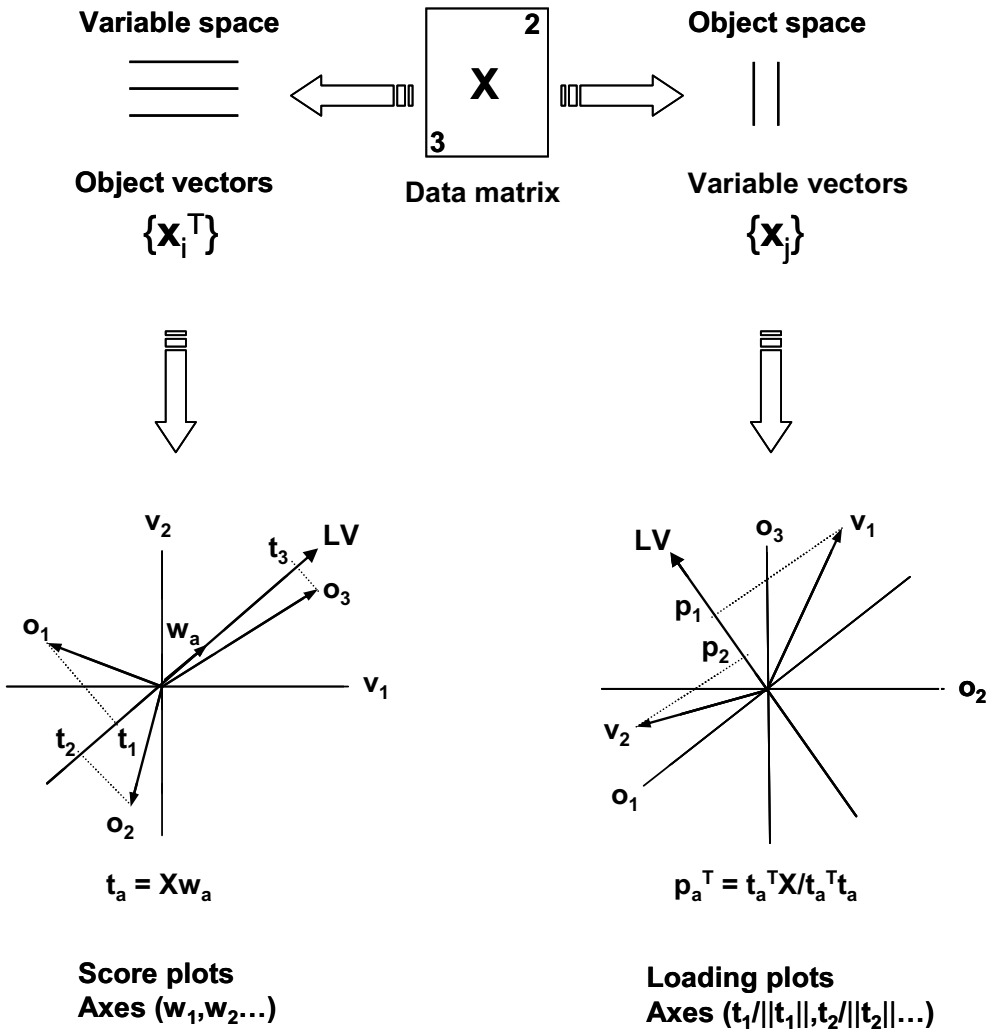


Figure 19. *The two alternative ways to look at a data matrix* ***X*** *and the principle of latent variable (LV) projections.*[111] *Three vectors,* $\boldsymbol{w}_a$, $\boldsymbol{t}_a$, *and* $\boldsymbol{p}_a$, *are needed to define the LV in the two spaces (see algorithm in Figure 20). Axes or vectors related to objects and variables are labelled with 'o' and 'v', respectively. In order to have a simple illustration, only three objects characterized by two variables are used.*

The score vector $\mathbf{t}_a$ and the loading vector $\mathbf{p}_a$ are two different presentations of the LV, in variable space and object space, respectively (Figure 19). The selection of $\mathbf{w}_a$ defines the LV uniquely and any LV method can be explained in terms of the selection of $\mathbf{w}_a$ (Box 2). Several criteria can be used for decomposition of multivariate matrices to determine the axes for projections (Box 3).

Box 2. Method overview

PCA/SVD	$\mathbf{w_a = p_a/\|\|p_a\|\|}$
PLS-DA	$\mathbf{w_a = y_a^T X_a/\|\| y_a^T X_a \|\|}$
TP	$\mathbf{w_a = b/\|\|b\|\|}$

PCA, principal component analysis; SVD, singular value decomposition; PLS-DA, partial least squares discriminant analysis; TP, target projection

Box 3. Decomposition criteria

PCA	⇒	Maximum variance
PLS	⇒	Relevant components
TP	⇒	"Real" factors

4.3.2 Principal component analysis (PCA)

The most common latent variable projection method is principal component analysis (PCA).[112, 113] The data matrix **X** is decomposed into a number of principal components (PCs) that maximize explained variance in the data on each successive component. The result is a bilinear model, a product of scores **T** and loadings **P** matrices:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E} \qquad (8)$$

X is an $M \times N$ matrix, consisting of M samples (rows) with N measured variables (columns). **T** is an $M \times A$ matrix and $\mathbf{P}^{\mathrm{T}}$ is an $A \times N$ matrix, where A is the number of calculated PCs. **T** and **P** consist of orthogonal and orthonormal vectors, respectively. **E** is an $M \times N$ matrix containing the residuals, that is, variance not explained by the PCs. Eq. 8 also shows the alternative description of the latent variable decomposition of **X** as a sum of products of score $\mathbf{t}_a$ and loading $\mathbf{p}_a$ vectors; a = 1, 2, … , A. The part of **X** explained by a pair of score and loading vectors in each step is removed before the next pair is calculated.

PCA is a valuable data visualization technique. Since each object gets a score value on each PC, objects can be presented in a score plot. Score plot can reveal patterns, such as clusters, trends and outliers, in the data. In the same manner can variables be presented in a loading plot, since each variable gets a loading value on each PC. Loading plot reveals covariance among variables and can be used to interpret findings in score plot.

Together scores and loadings map the structure in the data. The objective is to extract as many PCs as needed to explain the variation in the data; noise should be left in the residuals. Maximum number of components is equal to min [*M,N*].

4.3.3 Partial least squares (PLS) regression and PLS-discriminant analysis (PLS-DA)

One of the most common objectives in data analysis is to calculate a model which shows how one or several responses (dependent variables), **Y**, can be explained by means of predictor variables, **X**. If the number of the X-variables is low and the variables are linearly independent (as in a case of designed experiments, see Section 4.2) MLR works well. In most cases, however, there are many measured X-variables and they are not independent but correlated and noisy. This is always the case when working with spectral profiles. PLS regression is a modelling technique that takes into

account collinearity in the data.[114-116] Instead of using strongly dependent X-variables we calculate a new set of LVs that have a reduced dimensionality.

PLS decomposition can still be written similarly to PCA (see Eq. 8) but only the score vectors are orthogonal. The loading vectors are neither orthogonal nor of unit length. A normalized weighting vector **w** for PLS is calculated as the covariance between the response **y** and the data matrix **X**:

$$\mathbf{w}_{PLS,1}^{T} = \mathbf{y}^{T}\mathbf{X} \Big/ \left\| \mathbf{y}^{T}\mathbf{X} \right\| \qquad (9)$$

Scores and loadings for the PLS components are calculated successively by projecting the spectral variables **X** on $\mathbf{w}_{PLS,1}$ and by projecting **X** on the resulting score vectors. Each step is checked for predictive power by using cross validation.[117-119] The part of **X** explained by a pair of score and loading vectors in each step is removed before the next pair is calculated.

PLS regression can also be used as a supervised classification method (see section 4.3.4). The response variable is then a binary vector of zeros and ones, giving a class membership for each sample in the investigated groups. The method is called PLS-discriminant analysis (PLS-DA).[120]

4.3.4 Target projection (TP)

The PLS model can be used to predict the response from X-variables like spectral profiles. Unfortunately, numerous PLS components are usually needed to describe the variation in **X**. This makes interpretation of PLS models difficult since the information about the response is scattered between the PLS components. The TP method offers a remedy for this problem.[111] With TP, the information in the X-variables orthogonal to the response variable is removed and a single latent variable (the target-projected component) is obtained that represents the direction in the multivariate predictive space with strongest relation to the response. TP represents the

optimal way of rotating a latent variable decomposition to a known target vector (response variable).

The regression vector **b,** obtained from the PLS model, defines the direction in space with strongest relation to the response. Target-projected scores $\mathbf{t}_{TP}$, proportional to the predicted response, are obtained by projecting data matrix **X** onto the normalized regression vector $\mathbf{w}_{TP}= \mathbf{b}/\|\mathbf{b}\|$.

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{b}/\|\mathbf{b}\| \qquad (10)$$

TP loadings $\mathbf{p}_{TP}$ can then be calculated as:

$$\mathbf{p}_{TP} = \mathbf{X}^{T}\mathbf{t}_{TP}/\left(\mathbf{t}_{TP}^{T}\mathbf{t}_{TP}\right) \qquad (11)$$

The TP loadings represent the features in the X-variables explaining and predicting the response variable. But they should not be directly used for feature selection (see Section 4.4).

After target projection the PLS model is reduced to a single-component TP model:

$$\mathbf{X} = \hat{\mathbf{X}}_{TP} + \mathbf{E}_{TP} = \mathbf{t}_{TP}\mathbf{p}_{TP}^{T} + \mathbf{E}_{TP} \qquad (12)$$

With the same number of PLS components, the target-projected component is identical to the predictive orthogonal PLS component obtained from orthogonal PLS (O-PLS).[121] TP and O-PLS thus represent different algorithms to achieve the same goal.[122]

A detailed analysis of interpretation of partial least squares regression models by means of TP method can be found in a recent paper by Kvalheim.[123]

4.3.5 Supervised classification using latent variables

In supervised classification we know beforehand in which class each sample belongs to. For binary classification a response vector is created and values 1 or 0 are assigned

to the response according to the class membership of the samples. When new samples are measured and predicted using a PLS-DA/TP model response values close to 1 or 0 should be obtained. The four possible outcomes of such a binary classification, that is, true positive, false positive, false negative, and true negative, can be formulated in a 2×2 contingency table (Figure 20). If the classes have approximately the same size, the threshold in between (*i.e.* 0.5) can be used to decide the class membership for the tested samples. The threshold can of course be varied from case to case since the optimal choice is problem dependent. Balancing false positives against false negatives is used as criterion for deciding the threshold.

In multiclass problems two strategies are possible: either a single model, including all groups, or several binary models, modelling the groups pairwise. The combination PLS-DA/TP is optimal with respect to finding the most discriminatory vector for a binary classification problem. PLS-DA/TP can be extended to handle more than two groups, but the classification result can no longer be presented on a single vector for feature selection. When doing several distinct classifications, we obtain the best discriminatory features for each comparison and if one then compares the results from all classifications, the features that provide separation of all groups simultaneously can be obtained.

Class (true) / Class (predicted)	A (1)	B (0)
A (1)	TP	FP
B (0)	FN	TN

Figure 20. *The four outcomes of binary classification can be formulated in a 2×2 contingency table (confusion matrix). True positive (TP), false positive (FP), false negative (FN) and true negative (TN).*

4.4 Variable/feature selection

The number of objects is very small compared to the number of variables when full spectral profiles are acquired (*i.e.* tables are "short and fat"). This is a typical feature for all "omics" data. However, most of the variables are actually irrelevant as they represent variation not related to the response. Therefore the number of variables can be reduced drastically with minor loss of information. The challenge is to find the most significant variables. Variable selection methods aim at selecting a smaller panel of variables that are related to the response variable and thus needed for a good predictive model. When a large number of variables is measured it is impossible to test all the variable combinations in question; for instance, there are 2.46×10^{20} (500!/(490!10!) possible combinations to pick 10 variables out of 500. Variable selection methods are therefore needed to search for the best combination.

Univariate variable selection methods treat each variable (*e.g.* peaks in a spectral profile) independently. Statistical values are calculated for each variable after testing differences between profiles from two sample groups. T-statistics and analysis of variance (ANOVA) are methods often used for this purpose. However, these methods do not take into account collinearity in the data and they cannot handle properly the situation with only a few samples compared to many measured variables. It is relatively easy to find a solution by pure chance and an irrelevant model will be created. In addition the use of traditional statistical tests assumes that the data obey normal distribution, which is typically not the case in real "omics" data.[16] Due to small sample size and group heterogeneity, data in proteomic applications can not be assumed to follow a normal distribution. For these reasons *t*-test is not used in this thesis.

Several multivariate variable or feature selection methods are available based on, for example, the covariances between the response and each variable (*i.e.* PLS weights), regression coefficients, variable importance on projection (VIP), interval PLS, and genetic algorithm.[124-128] Wiklund *et al.* published recently the so-called S-plot.[129] S-

plot is a scatter plot showing covariance and correlation between the scores for the predictive O-PLS component and the spectral variables. Potential biomarker candidates should have both high covariance and high correlation to the score on the predictive component. The S-plot can be difficult to interpret since huge number of spectral variables makes it easily very crowded.

Unfortunately most of these methods will usually lead to detection of false biomarkers since they are not specific enough to select the correct variables in complex spectral profiles. In this thesis two new tools for feature selection are developed. The main tool for searching for biomarker signatures is called selectivity ratio (SR). Its statistical significance can be determined using non-parametric test called discriminating variable (DIVA) test.

4.4.1 Selectivity ratio (SR)

The variable loading vector can be misleading for selecting the regions in the spectra corresponding to potential biomarkers. Since the PLS/TP model is estimated from covariances between the X-variables and the response, spectral regions that correspond to compounds with relatively large concentration may dominate the TP loadings even if the correlation with the response itself is low.

The so-called SR plot was developed to solve this problem. SR is closely connected to the ratio of inter- to intragroup variation (see Section 5.2) and it is a measure of variable's performance to separate groups. SR can reveal regions in spectral profiles with both high explanatory and high predictive significance for the investigated response.

Explained $v_{expl,j}$ and residual (unexplained) $v_{res,j}$ variance for each variable j in the TP model can be calculated from Eq. 12. The ratio between explained and residual variance defines a selectivity ratio SR for each variable:

$$SR_j = v_{expl,j} / v_{res,j} \qquad j = 1,2,3\ldots \quad (13)$$

When calculated for spectral variables, SR can be displayed similarly to a spectrum (Figure 21). A high SR value means that the variable in question has a strong (predictive) correlation to the response, that is, the variable is highly selective. Thus, SR provides a quantitative ranking of variables (*e.g.* spectral regions) according to the response variable (*e.g.* class membership). This property makes SR a useful tool for variable selection in general and especially in applications where the number of measured variables largely exceeds the number of samples. Furthermore, by multiplying the SR with the sign of the corresponding regression coefficient or TP loading, it is possible to see if a variable increases or decreases when two groups of samples are compared.

When searching for biomarkers in complex profiles with hundreds of compounds, a boundary between significant and unimportant variables is needed. An SR threshold can be chosen by the user and it is always a compromise depending on an application. A lower SR threshold increases the risk of selecting false biomarker candidates (false positives), while a higher SR threshold increases the risk of loosing potential biomarkers (false negatives). A statistical test is therefore developed to balance between these alternative situations.

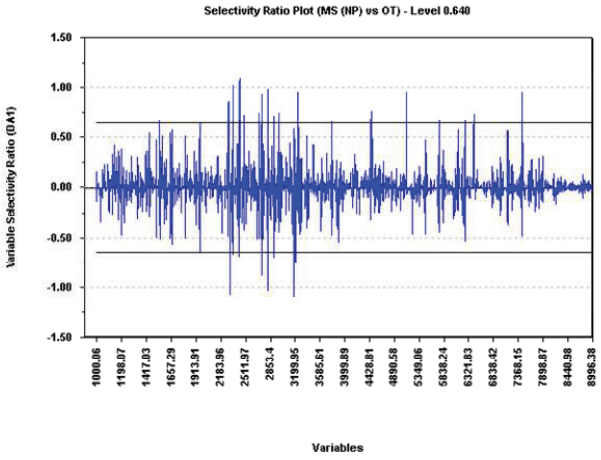


Figure 21. *Example of a selectivity ratio (SR) plot. The chosen SR limits are marked with horizontal lines.*

4.4.2 Discriminating variable (DIVA) test

Discriminatory ability of a variable is high if its explained variance on the TP component is significantly higher than its residual variance. This hypothesis can be tested using *F*-test.

$$F_{calc} = SR_i > F_{crit} = F_{(\alpha,N-2,N-3)} \qquad (14)$$

This will, however, result in a quite strict assumption that SR below one does not give significant discriminatory ability. In addition, the spectral data usually do not obey normal distribution as assumed when traditional univariate statistical tests are applied. For these reasons, a non-parametric statistical test has been developed to provide statistical boundaries for the SR method and thus make the feature selection easier. This test is called DIVA test.

In DIVA test a probability *p* is obtained directly from measuring how well variables manage to separate two groups of samples.

$$p_{DIVA} = (TP + TN)/N \qquad (15)$$

TP and *TN* represent true positive and true negative classifications, respectively. Together they represent all the correct results in binary classification (see Figure 20 in Section 4.3.5). *N* is number of samples. By ranking all the samples according to descending (or ascending) values on each variable separately and using the known class membership of each sample, the correct classification rate (CCR) can be calculated for all the variables. If a variable contributes to a perfect separation, that is, 100% CCR, samples belonging to one group get low values while samples belonging to the other group get high values on that variable. A variable contributing to completely random classification corresponds to 50% CCR with equal number of samples in each group.

Since SR is a measure of how well each variable contributes to separation of groups, increasing SR should also increase CCR. When CCR is calculated for all the

variables, suitable SR intervals are defined and mean correct classification rate (MCCR) and its standard deviation for the variables in each SR interval are calculated. When MCCR is plotted against SR for all the SR intervals, a DIVA plot is obtained (Figure 22). This plot enables an objective determination of probability based boundaries for the SR plot. The chosen probability level can be seen as a compromise to balance between inclusion of false biomarker candidates and loss of potential true biomarkers.

A receiver operating characteristics (ROC) curve is a bivariate plot representing sensitivity, or true positives, *versus* (1 − specificity), or false positives, in a binary classification as the discrimination threshold is varied.[24] In our approach CCR is identical to sensitivity in a binary classification, and MCCR can be interpreted as a mean sensitivity for the variables within a certain SR interval. DIVA plot connects variables' classification performance to their ratio of inter- to intragroup variation and thus expands the ROC curve into a multifeature domain. ROC is described in full detail elsewhere.[130]

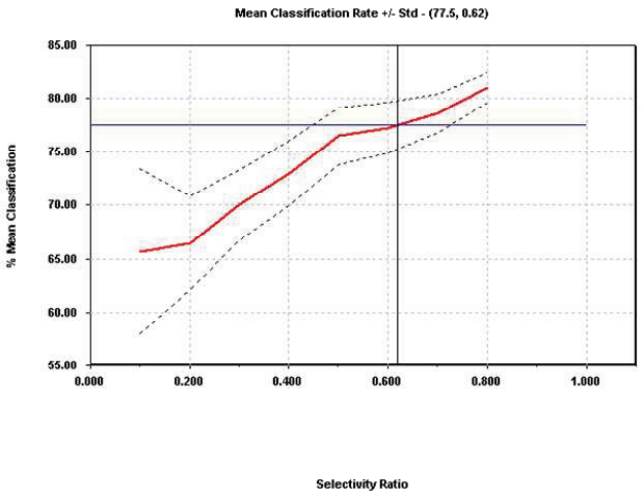


Figure 22. *Example of a discriminating variable (DIVA) plot mean correct classification rate (MCCR) and its standard deviation are plotted against selectivity ratio (SR). The chosen SR value is marked with vertical line.*

5. Summary of results

"Everything should be made as simple as possible, but not simpler." Albert Einstein

The challenges discussed in earlier chapters are addressed in the four scientific papers for which this thesis is based on. This chapter presents a summary of these papers (denoted as Papers I-IV).

5.1 Background information

5.1.1 Study population and sampling

Study population consisted of patients at Haukeland University Hospital. Clinical evaluation and diagnostic lumbar puncture of patients with MS and other neurological diseases (OND) were performed at the Department of Neurology. CSF from the neurological healthy controls (NHC) was collected from patients that received spinal anaesthesia for lower extremity orthopaedic surgery at the Department of Orthopaedic Surgery. Written informed consent was obtained from the included patients and the study was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Social Science Data Services.

CSF samples were taken by a standard lumbar puncture procedure. Few millilitres of CSF were immediately placed on ice and centrifuged at 450 x g for 10 minutes to remove cells, before freezing at -80 °C.

5.1.2 Sample preparation and instrumentation

Samples were fractionated using MW cut-off filters. The filter removed the most abundant proteins above a chosen threshold: 20 kDa (Paper I) and 30 kDa (Papers II-IV).

In MALDI-TOF analyses two matrices, α-cyano-4-hydroxycinnamic acid (CHCA) and 2,5-dihydroxybenzoic acid (DHB), were used. The low MW fraction was spotted onto a 600 μm AnchorChip® (using CHCA matrix) and the guanidinium fraction was spotted onto a steel target plate (using CHCA+DHB mix 1:1).

All the samples were analyzed using AutoFlex (Bruker Daltonics) MALDI-TOF mass spectrometer. The instrument was operated using nitrogen laser at 337 nm in ionization (laser frequency 20 Hz, ion source I 20 kV) and a positive linear mode in mass analysis. Data were acquired in two different ranges: 740-9000 Da defined as a low mass range (low MW fraction), and 6000-17500 Da defined as a medium mass range (guanidinium fraction).

5.1.3 Data sets

Paper I

Data set 1: Spiked sample

A CSF pool was created by mixing CSF from five different neurological patients. Six samples were prepared from the CSF pool: three of them were spiked with 1600 pM peptide standard (Table 3) and three were kept as reference samples. All the samples were fractionated and the low MW fraction was analyzed in triplicates. This provided a dataset of 18 spectral profiles, each described by 44 403 *m/z* values. Three profiles (representing one reference sample) were excluded from further study as outliers after visual inspection by PCA.

Data set 2: Storage data

Fresh CSF was obtained from a 30 years old male with some neurological symptoms. The sample was split into aliquots and stored under 15 different storage conditions[29] thus resulting in 15 samples. All the 15 samples were fractionated to create a low MW fraction and a guanidinium fraction. The guanidinium fraction of the samples was analyzed in triplicates to provide 45 spectral profiles, each described by 16 598 *m/z* values.

Table 3. *Peptide and protein standards used for spiking of cerebrospinal fluid (CSF) samples.*

Peptide standard		Protein standard	
m/z	Name	*m/z*	Name
1 047.20	Angiotensin II	5 734.56	Insulin
1 297.51	Angiotensin I	8 565.89	Ubiquitin I
1 348.66	Substance P	12 361.09	Cytochrome C
1 620.88	Bombesin	6 181.05	Cytochrome C ((M+2H)2+)
2 094.46	ACTH clip 1-17	16 952.55	Myoglobin
2 466.73	ACTH clip 18-39	8 476.77	Myoglobin ((M+2H)2+)
3 149.61	Somatostatin 28		

Data set 3: Replicated sample

A CSF pool was created by mixing CSF from ten different neurological patients. Five replicated samples were prepared from the CSF pool. The samples were fractionated and the low MW fraction was analyzed in triplicates. A data set of 15 spectral profiles was acquired, with each profile described by 44 403 *m/z* values.

Paper II

A CSF pool was created by mixing CSF from several neurological patients. The pool was divided into four samples (*i.e.* SP0, SP1, SP2 and SP3). Sample SP0 served as a reference sample. The other samples were spiked with the following concentrations of peptide and protein standards (Table 3): 400 pM / 2 nM (SP1), 800 pM / 10 nM (SP2) and 1600 pM / 40 nM (SP3). Three replicates for each sample were prepared and fractionated. Both fractions, the low MW fraction and the guanidinium fraction, were analyzed in triplicates thus resulting in 9 replicated spectra for each fraction of the four samples (*i.e.* 36 spectra per fraction). The peptide calibration standard should

only appear in the low MW fraction (data set 1) and the protein calibration standard should only appear in the guanidinium fraction (data set 2).

Paper III

CSF samples were drawn from NHC and randomly partitioned into five groups. One group, labelled 0 pM, was selected as reference. CSF from the other four groups was spiked with 50, 100, 200, or 400 pM of peptide standard (Table 3). Samples were fractionated in duplicates and the low MW fraction was analyzed in triplicates. Some profiles were excluded from further study as outliers after visual inspection by PCA. This provided a data set consisting of approximately 170 reference spectra and approximately 50 spectra for each of the spiked samples. Each spectrum is described by 44 403 *m/z* values.

Paper IV

CSF samples from patients with MS, OND and NHC, 18 in each group, were included (Table 4). Samples were fractionated and the low MW fraction was analyzed. Number of replicates varied between samples. In total there were 498 profiles, each described by 44 403 *m/z* values.

Table 4. *Study population in Paper IV; patients with relapsing-remitting multiple sclerosis (RRMS), other neurological diseases (OND) and neurological healthy controls (NHC).*

Group	No of patients	Females	Age range (years)	Males	Age range (years)
MS	18	15	22-73	3	30-58
OND	18	10	27-71	8	48-79
NHC	18	9	33-83	9	19-80

5.1.4 Software

MALDI-TOF sampling and preanalysis, including baseline correction, was performed using instrument's own FlexAnalysis software (Bruker Daltonics). SpecAlign version 2.3 (Cartwright Group PTCL, University of Oxford) was used for peak alignment in Paper I.[100] MATLAB version 6.5 (Mathworks Inc.) was used for calculation of the response *R* in Paper I. All additional chemometric analyses were programmed in Sirius software package from Pattern Recognition Systems AS (versions 7.0 and 8.0).

5.2 Pretreatment of mass spectral profiles (Paper I)

This study provided a recommended workflow on how to pretreat the spectral data prior to multivariate analysis. An optimal data pretreatment consist of steps that eliminate differences in profiles resulting from experimental and instrumental factors, but at the same time preserve the compositional information. Theoretical considerations of data pretreatment are presented in Section 4.1.

Factorial experimental designs, with different pretreatment steps as design variables, were employed for deciding the optimal procedure. Use of design makes it possible to evaluate the relative importance of the different steps and, in addition, their interactions since these may be expected. The investigated pretreatment steps were binning/smoothing, alignment, structured noise, and normalization. The response variable was the ratio of inter- to intragroup variation, *R*, that is, the variation between groups of samples was compared to the variation within replicated samples. Successful data pretreatment enhances the information content in the spectral profiles meaning that the ratio *R* increases. 2^4 designs were executed for three different sample sets (described in detail in Section 5.1.3) thus resulting in three different models. The most important factors and their contributions were identified from the coefficients provided by MLR. Theoretical aspects of DoE and empirical modelling are presented in Section 4.2.

The spectral profiles were baseline corrected using the instrument vendor's software (FlexAnalysis). Since this produced regions with negative intensities, the profiles were independently adjusted by the absolute value of the largest negative intensity so that the lowest intensity became zero. For spectral alignment, the SpecAlign software was used. For smoothing a moving average with a 10-point window and for binning a 10-point window was used. Several transformations to eliminate structured noise were tested and square and third root transformations were found to be appropriate. Normalization was performed using unit length (norm).

In data sets 1 and 2, a separation between different groups was expected and therefore the variation between groups should be larger than the variation within a group (*i.e.* maximize *R*). In data set 3, all profiles represented the same replicated sample and no separation was expected. Thus the optimal data pretreatment should remove the undesired effects without increasing *R*.

The results showed that the tested pretreatment steps had significant interactions and therefore they should not be interpreted solely based on their main effects. Strong interactions were observed especially between structured noise and normalization, and normalization and alignment as shown by regression coefficients (Figure 23). Normalization was undoubtedly the most important factor in all three cases. However, heteroscedastic noise should always be transformed to homoscedastic noise prior to normalization. Otherwise there might be a risk of false biomarker discovery. Use of n^{th} root transformation, with n equal to two or three, is recommended. Log transformation is too rough and it also destroys linear correlations in the profiles. Alignment was necessary in most cases and should be performed routinely. No difference between smoothing and binning was observed. Binning reduces the number of data points and speeds up the data analysis. It was shown that even reduction with one order of magnitude can be done without loss of information. In addition, use of binning made smoothing unnecessary, since binning acts as a filter.

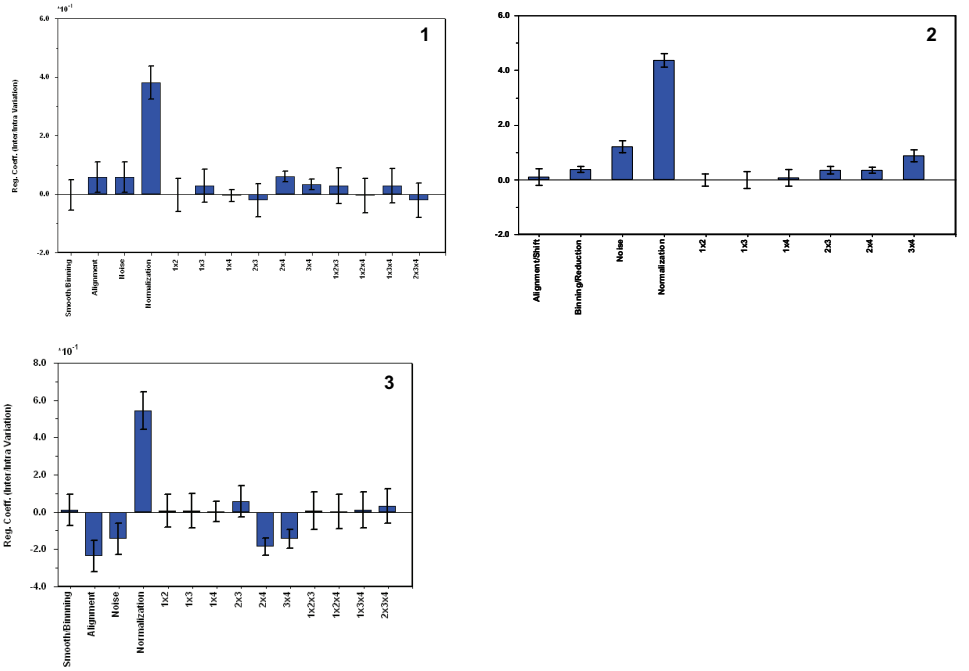


Figure 23. *Regression coefficients and their confidence intervals for data sets 1-3. Positive bar means that the ratio of inter- to intragroup variation, R, is increasing and negative bar means that R is decreasing when changing a variable from -1 to +1 level. For data sets 1 and 2, increasing R implies that profiles within a group become more similar without destroying the compositional correlation pattern. For data set 3, the undesired effects should be removed without increasing R, since this is a replicated sample and no separation should be observed.*

The analysis also demonstrated the need for replicated analysis; both to be able to reduce heteroscedastic noise, and to assess experimental errors.

The recommended order of pretreatment steps was as follows: smoothing/binning, alignment, transformation of structured noise and normalization. The pretreatment scheme was applied to further work in Papers II-IV (Figure 24).

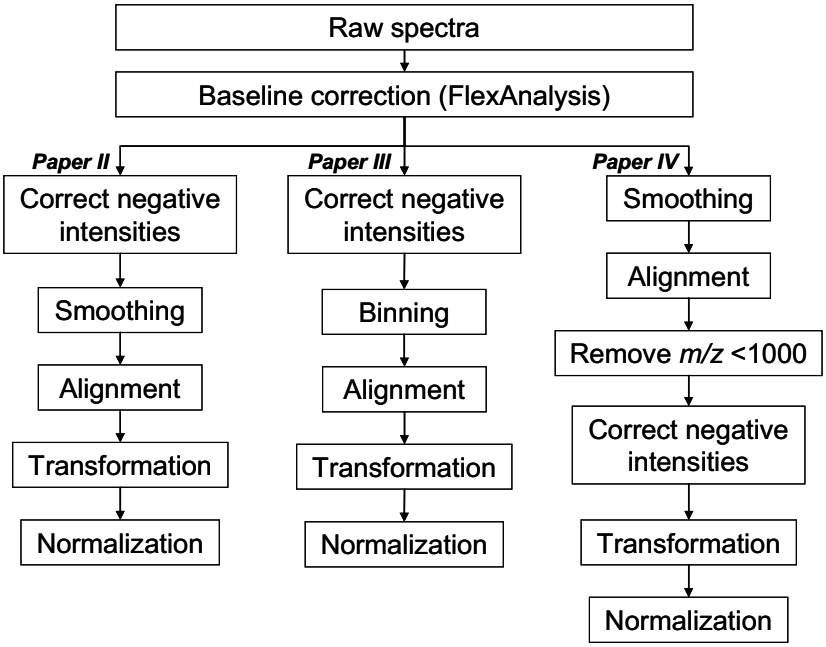


Figure 24. *Pretreatment scheme used in Papers II-IV. Specifications for different steps are: smoothing using moving average with window size 10, binning with window size 5, alignment using recursive alignment by fast Fourier transform (RAFFT) with window size 20, transformation using square root, and normalization using unit length. Negative intensities were corrected by either assigning zero to the lowest signal in each profile independently (Papers II and III) or setting such intensities to zero (Paper IV).*

5.3 Variable selection using selectivity ratio (SR) (Paper II)

A new method for variable selection in complex spectral profiles was presented in this study. The method uses so-called selectivity ratios (SR) and it is based on TP modelling. The SR for all the spectral variables on the target-projected component is obtained by calculating the ratio between explained variance and residual variance. The SR plot is a quantitative display of all the SRs showing all important features for interpreting the target component and ranking the discriminating variables (*e.g.*

biomarkers). Variables with a high SR value have a good discriminatory ability and are therefore contributing most to the separation between different groups. The theory of SR is presented in detail in Section 4.4.1.

The method was validated using two data sets (described in detail in Section 5.1.3) consisting of pooled CSF samples spiked with known concentrations of peptide and protein standards. The acquired MALDI profiles were pretreated prior to data analysis according to the scheme presented in Figure 24. Three binary classification models (*i.e.* reference samples SP0 *versus* spiked samples SP1, SP2 and SP3) were then created using a combination of PLS-DA and TP modelling. A binary response variable, giving the class membership of the samples, got values 0 and 1 for the reference samples and spiked samples, respectively. SR threshold was chosen by visual inspection; here SR value 3 was used, that is, at least 75% of the original variance was explained by the selected variables.

Spiking imitates the pathogenesis of a disease where abundances of certain proteins increase over a period of time. The results showed that in the low MW range (740-9000 Da) it was possible to detect spiked peptides at least down to 400 pM level using SR method without severe problems with false biomarker candidates. At this level we were able to detect six of the seven added peptides in the spiked sample (Figure 25). Three false candidates were also detected with SR values just above the chosen SR limit. With higher concentrations the detection became less vulnerable to false candidates. In the medium mass range (6000-17500 Da) it was possible to detect spiked proteins at least down to 2 nM level. At this level only two of the six added proteins were detected. However, it was possible to classify the samples correctly in both data sets using the selected *m/z* regions, meaning that the diagnostic power of the reduced profiles (*i.e.* biomarker signature) was high.

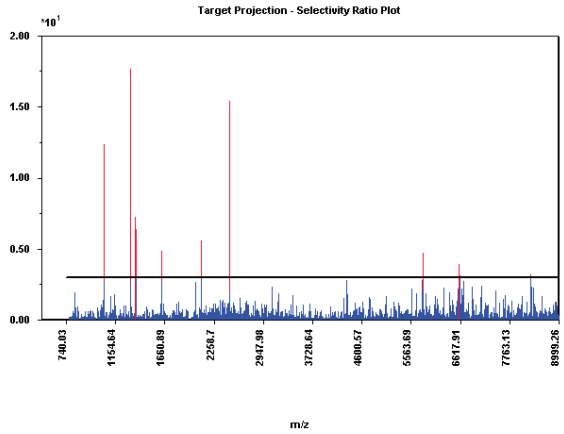


Figure 25. *Selectivity ratio (SR) plot for low mass range (740-9000 Da) for the reference samples (SP0) versus spiked samples (SP1) classification. The chosen SR threshold is marked with horizontal line.*

Comparison with some commonly used tools for variable selection (*e.g.* X-weights, regression coefficients, and VIP) showed that SR had the best performance (Figure 26). This is probably due to the fact that target projection utilizes both the predictive ability (regression coefficients) and the explanatory ability (spectral variance/covariance matrix) for the calculation of the SR.

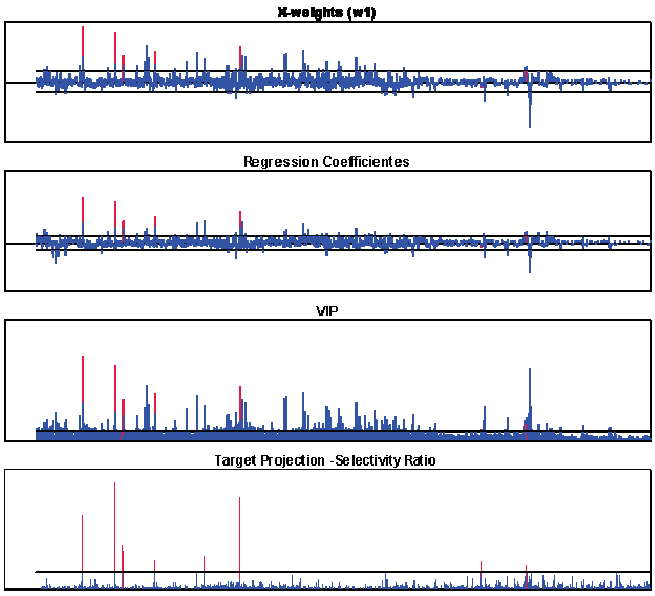


Figure 26. *Comparison between different variable selection methods applied to spectral profiles in low mass range. From the top: Covariances between spectral variables and group membership (i.e. X-weights, w), regression coefficients, variable importance on projection (VIP), and selectivity ratio (SR) plot.*

5.4 Discriminating variable (DIVA) test for defining probability based boundaries for the SR plot (Paper III)

In this study the SR method was further improved by introducing a nonparametric discriminating variable (DIVA) test to obtain probability based boundaries for the SR plot. Furthermore, interpretability of the SR plot was enhanced by multiplying each SR value with the sign of the corresponding regression coefficient (or TP loading as in Paper IV). Introduction of signs to SR plot makes it possible to see which variables are more and less abundant when comparing different groups. The combination of SR plot and DIVA test provides an objective and quantitative tool for variable selection

in complex spectral profiles. The theory of DIVA is presented in detail in Section 4.4.2.

The improved SR plot combined with DIVA test was validated using a data set (described in detail in Section 5.1.3) consisting of CSF samples spiked with known concentrations of peptide standard. Samples were not pooled but used separately, thus giving a more realistic imitation of a real diagnostic situation where patient to patient variation is present. The acquired MALDI profiles were pretreated according to the scheme presented in Figure 24. In this study binning (window size 5) was used instead of smoothing. This reduced the number of variables from 44 403 to 8 881 *m/z* numbers. Four binary classification models (*i.e.* reference samples 0 pM *versus* spiked samples 50, 100, 200 and 400 pM) were then created using a combination of PLS-DA and TP modelling. A binary response variable, giving the class membership of the samples, got values 0 and 1 for the reference samples and spiked samples, respectively.

An SR threshold was chosen with the help of DIVA test. DIVA plot (Figure 27) shows MCCR plotted against SR. It was obtained by first calculating the SR and the percent CCR for all the binned *m/z* numbers. Variables were then sorted according to their SR values and MCCR and its standard deviation were calculated for specified SR intervals. In 0 pM *vs.* 100 pM classification the SR threshold of 0.5 (corresponding to approx. 80% MCCR) was chosen. The results showed that at the concentration level 100 pM, three of the seven added peptides were visible above the threshold (Figure 28). However, an excellent separation between the sample groups was obtained with PCA when using only the selected *m/z* regions (less than 0.3% of the original profiles). This demonstrates again the high diagnostic power of this methodology.

The results were very similar for the three other classifications as well. However, for the lowest concentration (50 pM) one of these three peptides was under the chosen SR threshold. This resulted from depletion due to smaller and smaller amounts of peptide standard.

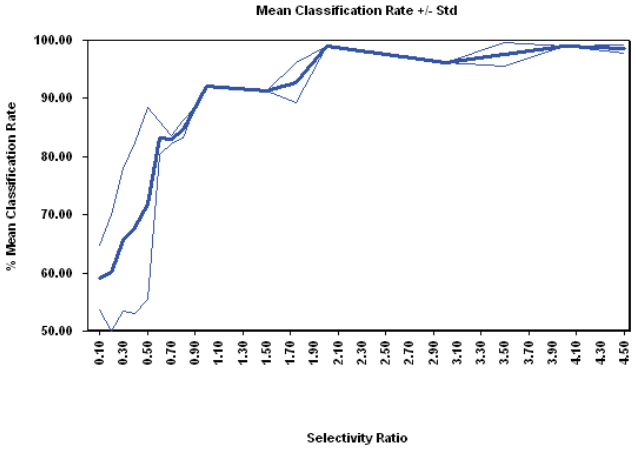


Figure 27. *Discriminating variables (DIVA) plot for the 0 pM versus 100 pM classification.*

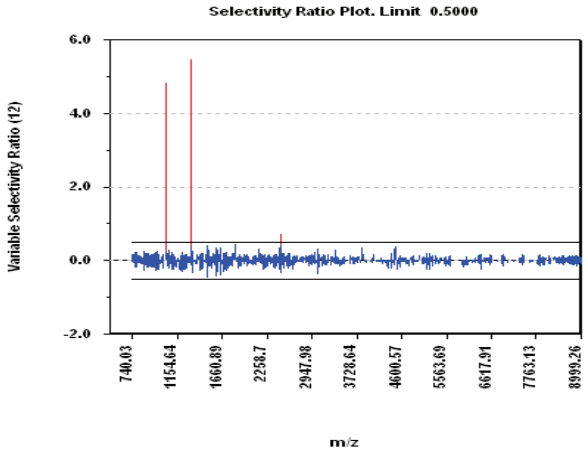


Figure 28. *Selectivity ratio (SR) plot for the 0 pM versus 100 pM classification.*

Comparison with other methods showed that our approach had the best performance providing only correct candidates (Figure 29).

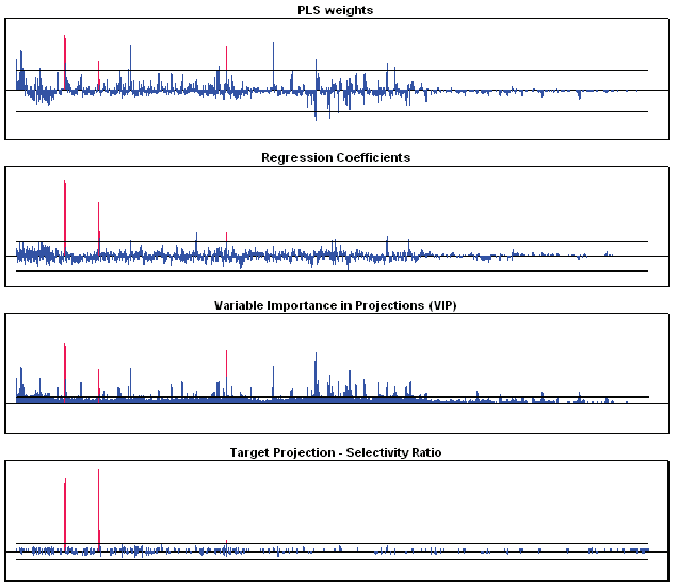


Figure 29. *Comparison between different variable selection methods applied to spectral profiles in low mass range. From the top: Covariances between spectral variables and group membership (i.e. X-weights, w), regression coefficients, variable importance on projection (VIP), and selectivity ratio (SR) plot.*

5.5 Application: Biomarker signatures for disease classification (Paper IV)

In this work the low MW fraction of the CSF proteome of samples from patients with MS, OND and NHC (described in detail in Section 5.1.3) was characterized using MALDI-TOF spectral profiling. Our novel targeted multivariate approach was applied to the profiles to reveal the features distinguishing different groups. The detected biomarker signature was then used for disease classification. The complete workflow is shown in Figure 30. This is the first application of the developed methodology to the analysis of real proteomics data. The approach is general and can be applied for other diseases and instrumental techniques as well.

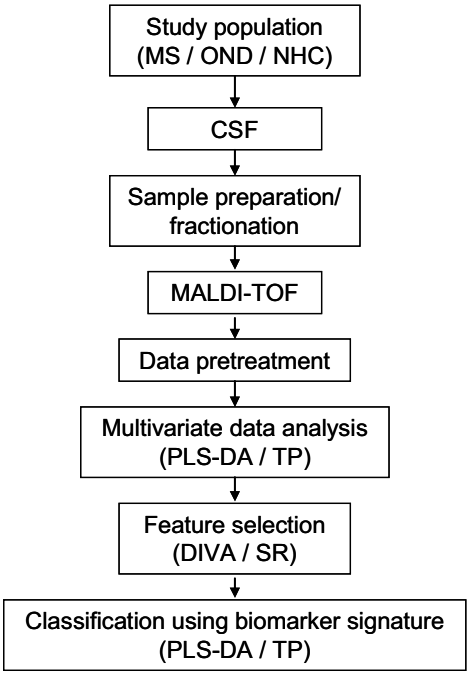


Figure 30. *Workflow used in Paper IV.*

The acquired MALDI profiles were pretreated according to the scheme presented in Figure 24. Three binary classification models (*i.e.* MS *vs* OND, MS *vs* NHC, and OND *vs* NHC) were created using a combination of PLS-DA and TP modelling. A binary response variable, giving the class membership of the samples, got value 0 for all the NHC samples and 1 for all the MS samples. OND samples got the value 0 in MS *vs* OND classification and 1 in OND *vs* NHC classification. An appropriate SR threshold was chosen with the help of DIVA test for each classification separately; the value varied between 0.5 and 0.6.

Several discriminating spectral regions were found using this approach. The selected *m/z* regions corresponded to 0.2-1.8% of the original spectral profiles. These regions were not necessarily whole peaks but fractions of them. The most interesting regions were those common to several classifications; we observed eight *m/z* regions that

were common to two subset classifications. Two of the regions were relatively more abundant in MS patients (around *m/z* 2299 and 2498). Four regions were relatively more abundant in patients with neurological diseases, that is, MS and OND (around *m/z* 2430, 3237, 5121, and 7162) while two regions were less abundant in the samples with neurological diseases (around *m/z* 2318 and 2821).

Despite the huge reduction in spectral variables, the loss of information in the classification pattern was surprisingly small. In binary classification excellent separation between the two groups in each subset was observed (Figure 31).

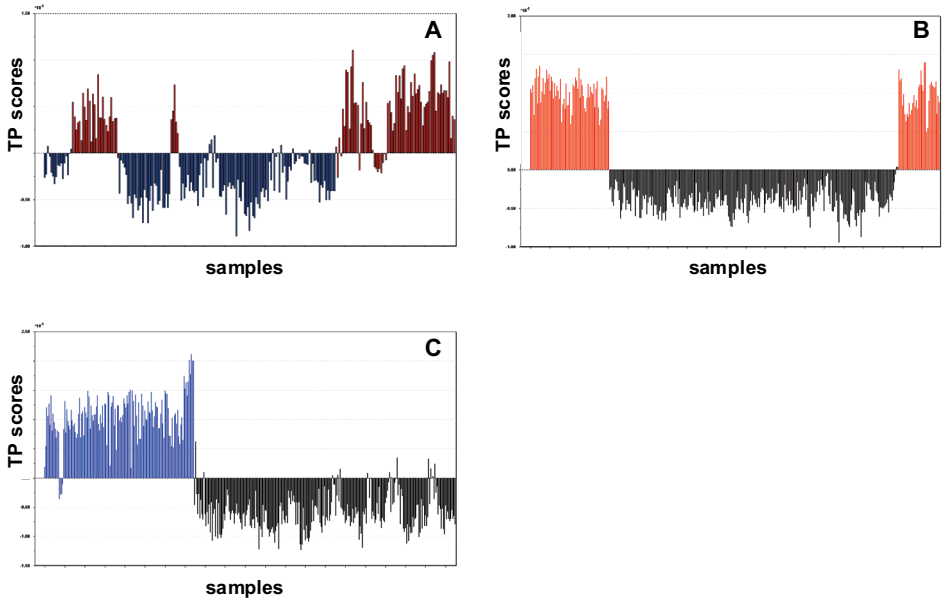


Figure 31. *Scores on the target-projected (TP) component for the detected biomarker signatures in classifications (A) MS vs. OND, (B) MS vs. NHC, and (C) OND vs. NHC. MS (red), OND (blue) and NHC (black).*

A few of the replicated spectra from some samples comparing MS *vs.* OND and OND *vs.* NHC were falsely classified. In MS *vs.* OND classification only one MS and one OND patient were falsely classified when taking into account the replicate variation in the samples. In OND *vs.* NHC classification only one NHC patient was systematically falsely classified. The separation of MS from NHC was perfect after variable reduction.

6. Conclusions and future perspectives

"By three methods we may learn wisdom: First, by reflection, which is noblest; second, by imitation, which is easiest; and third by experience, which is the bitterest."
Confucius

This thesis has discussed the detection of biomarker signatures using mass spectral profiling and multivariate analysis. The main results of the thesis can be summarized as follows:

1. Traditional way to analyze spectral profiles is to reduce the full profiles to peaks. Doing this, however, may also reduce the information content in the data. In addition, most of the problems encountered using full spectral profiles are also valid for a peak based approach. Data pretreatment is needed to eliminate non-compositional features from the spectral profiles without destroying the compositional differences. The problem with heteroscedastic noise is an example of this dilemma. Heteroscedastic noise can be transformed to homoscedastic noise using the n^{th}-root transformation, but this transformation impacts the correlation between variables. On the other hand, it provides better opportunities for variables with low concentrations to be detected in the biomarker discovery process. According to our experience a choice of n equal to two is reasonably robust, but this can be verified by comparing replicates after the transformation. Strong interactions exist between different pretreatment steps. Normalization of the mass spectral profiles without prior transformation of structured noise may give rise to false biomarker candidates. Replicated analyses are always recommended; both to be able to reduce heteroscedastic noise, and, to assess experimental errors.

2. Several methods exist for feature selection but most of them have not given satisfying results when applied to biomarker detection. Univariate methods, such as *t*-test and ANOVA, should not be used when data are highly collinear and not normally distributed. In addition, it is relatively easy to find biomarkers by pure chance when

the number of variables is high compared to the number of samples. Multivariate methods are mainly based on looking at variable loadings or regression coefficients. They will usually lead to detection of false biomarkers since these methods are biased towards selection of variables with large variance. In this thesis a new feature selection tool called selectivity ratio (SR) plot is developed. The SR plot shows quantitatively all important features present in profiles needed for interpretation of the target component, and thus makes an objective selection of discriminating variables possible. Accompanied with a new nonparametric discriminating variable (DIVA) test, to obtain probability based boundaries for SR, the method is shown to outperform currently available graphical tools for feature selection. This method can be utilized in all multivariate methods providing a single predictive component for each y-variable.

3. The combination of PLS-DA and TP modelling of spectral profiles, followed by feature selection with SR plot and DIVA test, provides a workflow for detecting biomarker signatures from complex spectral profiles. With this approach both narrow and broad discriminatory *m/z* regions are located without the need for the assumption that variables must represent whole single separated peaks. These regions can be combined into disease patterns with the best possible performance with respect to separating groups from each other. Furthermore, these disease patterns provide the most promising *m/z* regions for biomarker discovery in the investigated fraction. These features of our approach represent an advantage over the conventional peak based methods.

Future perspectives

In this study we have developed new methods to reveal *m/z* regions with discriminating ability and thus a pattern serving as a potential biomarker signature. Follow-up analysis is necessary to provide the identification of the candidate biomarkers. Multivariate resolution techniques can for example be used to separate overlapping peaks. These techniques are time-consuming and not easy to automate, since they need more expertise and interference from the operator. A better approach

is therefore to use other mass spectroscopic techniques that can utilize the information already obtained about promising *m/z* regions.

Although only used on MALDI-TOF spectral profiles in the present work, our data-analytical approach is much more general. The generalization to surface-enhanced laser desorption/ionization (SELDI) mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy data is obvious, but by using the technique of unfolding, also data from hyphenated instruments like liquid chromatography-mass spectrometry (LC-MS) can be analyzed by our approach.

The combination of SR plot and DIVA test represents a variable selection method that can be applied to all data analytical problems with many correlated variables and complex patterns. For example, gene expression data and metabolomic data can be analysed using this approach.

References

1. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P., *Molecular Biology of the Cell*. 4th ed.; Garland Science: New York, 2002.
2. Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, J. C.; Hutchison, C. A.; Slocombe, P. M.; Smith, M., Nucleotide sequence of bacteriophage ϕχ174 DNA. *Nature* **1977,** 265, (5596), 687-695.
3. Wang, Z.; Gerstein, M.; Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009,** 10, (1), 57-63.
4. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **2004,** 431, (7011), 931-945.
5. Omenn, G. S., From Human Genome Research to Personalized Health Care. *Issues Sci. Technol.* **2009,** 25, (4), 51-56.
6. Pennisi, E., Bioinformatics - Gene counters struggle to get the right answer. *Science* **2003,** 301, (5636), 1040-1041.
7. Stein, L. D., Human genome: End of the beginning. *Nature* **2004,** 431, (7011), 915-916.
8. Crick, F., Central dogma of molecular biology. *Nature* **1970,** 227, (5258), 561-563.
9. Whitford, D., *Proteins - structure and function*. Wiley: Chichester, 2005.
10. Kolodziejczyk, R.; Michalska, K.; Hernandez-Santoyo, A.; Wahlbom, M.; Grubb, A.; Jaskolski, M., Crystal structure of human cystatin C stabilized against amyloid formation. *FEBS Journal* **2010,** 277, (7), 1726-1737.
11. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** 422, (6928), 198-207.
12. Penque, D., Two-dimensional gel electrophoresis and mass spectrometry for biomarker discovery. *Proteomics - Clinical Applications* **2009,** 3, (2), 155-172.
13. Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J.-C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphery-Smith, I.; Williams, K. L.; Hochstrasser, D. F., From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis. *Nat. Biotechnol.* **1996,** 14, (1), 61-65.
14. Wilkins, M. R.; Sanchez, J. C.; Gooley, A. A.; Appel, R. D.; HumpherySmith, I.; Hochstrasser, D. F.; Williams, K. L., Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. In *Biotechnology and Genetic Engineering Reviews, Vol 13*, Intercept Ltd: Andover, 1996; Vol. 13, pp 19-50.
15. James, P., Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly Review of Biophysics* **1997,** 30, (4), 279-331.
16. Wilkins, M. R.; Appel, R. D.; Van Eyk, J. E.; Chung, M. C. M.; Gorg, A.; Hecker, M.; Huber, L. A.; Langen, H.; Link, A. J.; Paik, Y. K.; Patterson, S. D.; Pennington, S. R.; Rabilloud, T.; Simpson, R. J.; Weiss, W.; Dunn, M. J., Guidelines for the next 10 years of proteomics. *Proteomics* **2006,** 6, (1), 4-8.
17. Dunn, W. B.; Ellis, D. I., Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* **2005,** 24, (4), 285-294.
18. van der Greef, J.; Smilde, A. K., Symbiosis of chemometrics and metabolomics: past, present, and future. *J. Chemometr.* **2005,** 19, (5-7), 376-386.

19. van der Greef, J.; Martin, S.; Juhasz, P.; Adourian, A.; Plasterer, T.; Verheij, E. R.; McBurney, R. N., The art and practice of systems biology in medicine: Mapping patterns of relationships. *J. Proteome Res.* **2007,** 6, (4), 1540-1559.
20. Mischak, H.; Apweiler, R.; Banks, R. E.; Conaway, M.; Coon, J.; Dominiczak, A.; Ehrich, J. H. H.; Fliser, D.; Girolami, M.; Hermjakob, H.; Hochstrasser, D.; Jankowski, J.; Julian, B. A.; Kolch, W.; Massy, Z. A.; Neususess, C.; Novak, J.; Peter, K.; Rossing, K.; Schanstra, J.; Semmes, O. J.; Theodorescu, D.; Thongboonkerd, V.; Weissinger, E. M.; Van Eyk, J. E.; Yamamoto, T., Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteom. Clin. Appl.* **2007,** 1, (2), 148-156.
21. Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A., Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* **2002,** 359, (9306), 572-577.
22. Lescuyer, P.; Farina, A.; Hochstrasser, D. F., Proteomics in clinical chemistry: will it be long? *Trends Biotechnol.* **2010,** 28, (5), 225-229.
23. Veenstra, T. D.; Conrads, T. P.; Hood, B. L.; Avellino, A. M.; Ellenbogen, R. G.; Morrison, R. S., Biomarkers: Mining the biofluid proteome. *Mol. Cell. Proteomics* **2005,** 4, (4), 409-418.
24. Vitzthum, F.; Behrens, F.; Anderson, N. L.; Shaw, J. H., Proteomics: From Basic Research to Diagnostic Application. A Review of Requirements & Needs *J. Proteome Res.* **2005,** 4, (4), 1086-1097.
25. Gillette, M. A.; Mani, D. R.; Carr, S. A., Place of Pattern in Proteomic Biomarker Discovery *J. Proteome Res.* **2005,** 4, (4), 1143-1154.
26. Villar-Garea, A.; Griese, M.; Imhof, A., Biomarker discovery from body fluids using mass spectrometry. *J. Chromatogr. B* **2007,** 849, (1-2), 105-114.
27. Fung, E. T., Strategies in Clinical Proteomics: From Discovery to Assay. *Preclinica* **2004,** 2, (4), 253-258.
28. Fehniger, T. E.; Marko-Varga, G., Proteomics and Disease Revisited: The Challenge of Providing Proteomic Tools into Clinical Practice. *J. Proteome Res.* **2010,** 9, (3), 1191-1192.
29. Berven, F. S.; Kroksveen, A. C.; Berle, M.; Rajalahti, T.; Flikka, K.; Arneberg, R.; Myhr, K.-M.; Vedeler, C.; Kvalheim, O. M.; Ulvik, R. J., Pre-analytical influence on the low molecular weight cerebrospinal fluid proteome. *Proteomics - Clinical Applications* **2007,** 1, (7), 699-711.
30. LaBaer, J., So, you want to look for biomarkers. *J. Proteome Res.* **2005,** 4, (4), 1053-1059.
31. Rifai, N.; Gillette, M. A.; Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **2006,** 24, (8), 971-983.
32. Ahn, S.-M.; Simpson, R. J., Proteomic Strategies for Analyzing Body Fluids. In *Proteomics of Human Body Fluids*, Thongboonkerd, V., Ed. Humana Press Inc.: Totowa, NJ, 2007; pp 3-30.
33. Meng, Z.; Veenstra, T. D., Proteomic analysis of serum, plasma, and lymph for the identification of biomarkers. *Proteomics - Clinical Applications* **2007,** 1, (8), 747-757.
34. Ramström, M.; Bergquist, J., Proteomics of Human Cerebrospinal Fluid. In *Proteomics of Human Body Fluids*, Thongboonkerd, V., Ed. Humana Press Inc.: Totowa, NJ, 2007; pp 269-284.
35. Blennow, K., CSF biomarkers for mild cognitive impairment. *J. Intern. Med.* **2004,** 256, (3), 224-234.
36. Blennow, K.; Hampel, H., CSF markers for incipient Alzheimer's disease. *Lancet Neurol.* **2003,** 2, (10), 605-613.

37. Mase, M.; Yamada, K.; Shimazu, N.; Seiki, K.; Oda, H.; Nakau, H.; Inui, T.; Li, W. D.; Eguchi, N.; Urade, Y., Lipocalin-type prostaglandin D synthase (beta-trace) in cerebrospinal fluid: a useful marker for the diagnosis of normal pressure hydrocephalus. *Neurosci. Res.* **2003,** 47, (4), 455-459.
38. Conti, A.; Sanchez-Ruiz, Y.; Bachi, A.; Beretta, L.; Grandi, E.; Beltramo, M.; Alessio, M., Proteome study of human cerebrospinal fluid following traumatic brain injury indicates fibrin(ogen) degradation products as trauma-associated markers. *J. Neurotrauma* **2004,** 21, (7), 854-863.
39. Lescuyer, P.; Allard, L.; Zimmermann-Ivol, C. G.; Burgess, J. A.; Hughes-Frutiger, S.; Burkhard, P. R.; Sanchez, J. C.; Hochstrasser, D. F., Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. *Proteomics* **2004,** 4, (8), 2234-2241.
40. Ramstrom, M.; Ivonin, I.; Johansson, A.; Askmark, H.; Markides, K. E.; Zubarev, R.; Hakansson, P.; Aquilonius, S. M.; Bergquist, J., Cerebrospinal fluid protein patterns in neurodegenerative disease revealed by liquid chromatography-Fourier transform ion cyclotron resonance mass spectrometry. *Proteomics* **2004,** 4, (12), 4010-4018.
41. Zhang, J.; Goodlett, D. R.; Quinn, J. F.; Peskind, E.; Kaye, J. A.; Zhou, Y.; Pan, C.; Yi, E.; Eng, J.; Wang, Q.; Aebersold, R. H.; Montine, T. J., Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. *J. Alzheimers Dis.* **2005,** 7, (2), 125-133.
42. Hammack, B. N.; Fung, K. Y. C.; Hunsucker, S. W.; Duncan, M. W.; Burgoon, M. P.; Owens, G. P.; Gilden, D. H., Proteomic analysis of multiple sclerosis cerebrospinal fluid. *Mult. Scler.* **2004,** 10, (3), 245-260.
43. Compston, A.; Coles, A., Multiple sclerosis. *Lancet* **2002,** 359, (9313), 1221-1231.
44. Compston, A.; Coles, A., Multiple sclerosis. *The Lancet* **2008,** 372, (9648), 1502-1517.
45. Kurtzke, J. F., Reassessment of distribution of multiple-sclerosis. Part one. *Acta Neurol. Scand.* **1975,** 51, (2), 110-136.
46. Grytten, N.; Aarseth, J. H.; Nyland, H.; Midgard, R.; Myhr, K. M., A 50-year follow-up of the incidence of multiple sclerosis in Hordaland County, Norway. *Neurology* **2006,** 66, (2), 182-186.
47. Gronlie, S. A.; Myrvoll, E.; Hansen, G.; Gronning, M.; Mellgren, S. I., Multiple sclerosis in North Norway, and first appearance in an indigenous population. *J. Neurol.* **2000,** 247, (2), 129-133.
48. Pugliatti, M.; Riise, T.; Sotgiu, M. A.; Sotgiu, S.; Satta, W. M.; Mannu, L.; Sanna, G.; Rosati, G., Increasing incidence of multiple sclerosis in the province of Sassari, northern Sardinia. *Neuroepidemiology* **2005,** 25, (3), 129-134.
49. Orton, S. M.; Herrera, B. M.; Yee, I. M.; Valdar, W.; Ramagopalan, S. V.; Sadovnick, A. D.; Ebers, G. C.; Canadian Collaborative Study, G., Sex ratio of multiple sclerosis in Canada: a longitudinal study. *Lancet Neurol.* **2006,** 5, (11), 932-936.
50. Ramagopalan, S. V.; Byrnes, J. K.; Orton, S. M.; Dyment, D. A.; Guimond, C.; Yee, I. M.; Ebers, G. C.; Sadovnick, A. D., Sex ratio of multiple sclerosis and clinical phenotype. *European Journal of Neurology* **2010,** 17, (4), 634-637.
51. Koch-Henriksen, N.; Sørensen, P. S., The changing demographic pattern of multiple sclerosis epidemiology. *The Lancet Neurology* **2010,** 9, (5), 520-532.
52. Ebers, G. C., Environmental factors and multiple sclerosis. *Lancet Neurol.* **2008,** 7, (3), 268-277.

53. Pugliatti, M.; Harbo, H. F.; Holmoy, T.; Kampman, M. T.; Myhr, K. M.; Riise, T.; Wolfson, C., Environmental risk factors in multiple sclerosis. *Acta Neurol. Scand.* **2008,** 117, 34-40.
54. Jersild, C.; Fog, T.; Svejgaar.A, HL-A antigens and multiple-sclerosis. *Lancet* **1972,** 1, (7762), 1240.
55. Hafler, D. A.; Compston, A.; Sawcer, S.; Lander, E. S.; Daly, M. J.; De Jager, P. L.; de Bakker, P. I. W.; Gabriel, S. B.; Mirel, D. B.; Ivinson, A. J.; Pericak-Vance, M. A.; Gregory, S. G.; Rioux, J. D.; McCauley, J. L.; Haines, J. L.; Barcellos, L. F.; Cree, B.; Oksenberg, J. R.; Hauser, S. L.; International Multiple Sclerosis Genetics Consortium, Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **2007,** 357, (9), 851-862.
56. Lill, C.; McQueen, M.; Roehr, J.; Bagade, S.; Schjeide, B.; Zipp, F.; Bertram, L. The MSGene Database. Alzheimer Research Forum. Available at http://www.msgene.org/
57. McDonald, W. I.; Compston, A.; Edan, G.; Goodkin, D.; Hartung, H. P.; Lublin, F. D.; McFarland, H. F.; Paty, D. W.; Polman, C. H.; Reingold, S. C.; Sandberg-Wollheim, M.; Sibley, W.; Thompson, A. J.; van den Noort, S.; Weinshenker, B. Y.; Wolinsky, J. S., Recommended diagnostic criteria for multiple sclerosis: Guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Ann. Neurol.* **2001,** 50, (1), 121-127.
58. Polman, C. H.; Reingold, S. C.; Edan, G.; Filippi, M.; Hartung, H. P.; Kappos, L.; Lublin, F. D.; Metz, L. M.; McFarland, H. F.; O'Connor, P. W.; Sandberg-Wollheim, M.; Thompson, A. J.; Weinshenker, B. G.; Wolinsky, J. S., Diagnostic criteria for multiple sclerosis: 2005 Revisions to the "McDonald Criteria". *Ann. Neurol.* **2005,** 58, (6), 840-846.
59. Myhr, K. M., Diagnosis and treatment of multiple sclerosis. *Acta Neurol. Scand.* **2008,** 117, 12-21.
60. Stueve, O.; Bennett, J. L.; Hemmer, B.; Wiendl, H.; Racke, M. K.; Bar-Or, A.; Hu, W.; Zivadinov, R.; Weber, M. S.; Zamvil, S. S.; Pacheco, M. F.; Menge, T.; Hartung, H. P.; Kieseier, B. C.; Frohman, E. M., Pharmacological treatment of early multiple sclerosis. *Drugs* **2008,** 68, (1), 73-83.
61. Govorukhina, N.; Bischoff, R., Sample Preparation of Body Fluids for Proteomics Analysis. In *Proteomics of Human Body Fluids*, Thongboonkerd, V., Ed. Humana Press Inc.: Totowa, NJ, 2007; pp 31-69.
62. Corthals, G. L.; Wasinger, V. C.; Hochstrasser, D. F.; Sanchez, J. C., The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis* **2000,** 21, (6), 1104-1115.
63. Anderson, N. L.; Anderson, N. G., The human plasma proteome - History, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002,** 1, (11), 845-867.
64. Shores, K. S.; Knapp, D. R., Assessment approach for evaluating high abundance protein depletion methods for cerebrospinal fluid (CSF) proteomic analysis. *J. Proteome Res.* **2007,** 6, (9), 3739-3751.
65. Georgiou, H. M.; Rice, G. E.; Baker, M. S., Proteomic analysis of human plasma: Failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **2001,** 1, (12), 1503-1506.
66. O'Farrell, P. H., High-resolution 2-dimensional electrophoresis of proteins. *J. Biol. Chem.* **1975,** 250, (10), 4007-4021.
67. Klose, J., From 2-D electrophoresis to proteomics. *Electrophoresis* **2009,** 30, (S1), S142-S149.
68. Klose, J., Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues - novel approach to testing for induced point mutations in mammals. *Humangenetik* **1975,** 26, (3), 231-243.

69. Gorg, A.; Weiss, W.; Dunn, M. J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004,** 4, (12), 3665-3685.
70. Horvatovich, P.; Hoekman, B.; Govorukhina, N.; Bischoff, R., Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. *J. Sep. Sci.* **2010,** 33, (10), 1421-1437.
71. Evans, C. R.; Jorgenson, J. W., Multidimensional LC-LC and LC-CE for high-resolution separations of biological molecules. *Anal. Bioanal. Chem.* **2004,** 378, (8), 1952-1961.
72. Issaq, H. J.; Conrads, T. P.; Janini, G. M.; Veenstra, T. D., Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis* **2002,** 23, (17), 3048-3061.
73. Wu, C. C.; MacCoss, M. J., Shotgun proteomics: Tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther.* **2002,** 4, (3), 242-250.
74. Vorderwülbecke, S.; Cleverley, S.; Weinberger, S. R.; Wiesner, A., Protein quantification by the SELDI-TOF-MS–based ProteinChip® System. *Nature Methods* **2005,** 2, (5), 393-395.
75. Shen, Y. F.; Smith, R. D., Proteomics based on high-efficiency capillary separations. *Electrophoresis* **2002,** 23, (18), 3106-3124.
76. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** 246, (4926), 64-71.
77. Kicman, A. T.; Parkin, M. C.; Iles, R. K., An introduction to mass spectrometry based proteomics - Detection and characterization of gonadotropins and related molecules. *Mol. Cell. Endocrinol.* **2007,** 260, 212-227.
78. Bergquist, J., FTICR mass spectrometry in proteomics. *Curr. Opin. Mol. Ther.* **2003,** 5, (3), 310-314.
79. Villanueva, J.; Philip, J.; Chaparro, C. A.; Li, Y. B.; Toledo-Crow, R.; DeNoyer, L.; Fleisher, M.; Robbins, R. J.; Tempst, P., Correcting common errors in identifying cancer-specific serum peptide signatures. *J. Proteome Res.* **2005,** 4, (4), 1060-1072.
80. Callesen, A. K.; Christensen, R. D.; Madsen, J. S.; Vach, W.; Zapico, E.; Cold, S.; Jorgensen, P. E.; Mogensen, O.; Kruse, T. A.; Jensen, O. N., Reproducibility of serum protein profiling by systematic assessment using solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **2008,** 22, (3), 291-300.
81. Liland, K. H.; Mevik, B.-H.; Rukke, E.-O.; Almøy, T.; Skaugen, M.; Isaksson, T., Quantitative whole spectrum analysis with MALDI-TOF MS, Part I: Measurement optimisation. *Chemometrics Intell. Lab. Syst.* **2009,** 96, (2), 210-218.
82. Forshed, J.; Pernemalm, M.; Tan, C. S.; Lindberg, M.; Kanter, L.; Pawitan, Y.; Lewensohn, R.; Stenke, L.; Lehtio, J., Proteomic data analysis workflow for discovery of candidate biomarker peaks predictive of clinical outcome for patients with acute myeloid leukemia. *J. Proteome Res.* **2008,** 7, (6), 2332-2341.
83. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T.; Matsuo, T., Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** 2, (8), 151-153.
84. Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Anal. Chem.* **1988,** 60, (20), 2299-2301.
85. Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T., Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. *Anal. Chem.* **1991,** 63, (24), A1193-A1202.
86. Wysocki, V. H.; Resing, K. A.; Zhang, Q.; Cheng, G., Mass spectrometry of peptides and proteins. *Methods* **2005,** 35, (3), 211-222.

87. Bakhtiar, R.; Tse, F. L. S., Biological mass spectrometry: a primer. *Mutagenesis* **2000,** 15, (5), 415-430.
88. Spiegelman, C. H.; Pfeiffer, R.; Gail, M., Using chemometrics and statistics to improve proteomics biomarker discovery. *J. Proteome Res.* **2006,** 5, (3), 461-462.
89. Wold, S.; Sjostrom, M., Chemometrics, present and future success. *Chemometrics Intell. Lab. Syst.* **1998,** 44, (1-2), 3-14.
90. Kvalheim, O. M., The latent variable. *Chemometrics Intell. Lab. Syst.* **1992,** 14, (1-3), 1-3.
91. Kvalheim, O. M., The Latent-Variable (Factor) Approach to the Analysis of Multivariate Data: History, Philosophy, and Scientific Implications. In *Understanding and History in Arts and Science*, Skarsten, R.; Kleppe, E. J.; Finnestad, R. B., Eds. Solum Forlag AS: Oslo, 1991; Vol. 1, pp 161-173.
92. Wold, S., Personal memories of the early PLS development. *Chemometrics Intell. Lab. Syst.* **2001,** 58, (2), 83-84.
93. Wold, S., Chemometrics; What do we mean with it, and what do we want from it? *Chemometrics Intell. Lab. Syst.* **1995,** 30, (1), 109-115.
94. Wagner, M.; Naik, D.; Pothen, A., Protocols for disease classification from mass spectrometry data. *Proteomics* **2003,** 3, (9), 1692-1698.
95. Williams, B.; Cornett, S.; Dawant, B.; Crecelius, A.; Bodenheimer, B.; Caprioli, R., An algorithm for baseline correction of MALDI mass spectra. In *Proceedings of the 43rd Annual ACM Southeast Regional Conference*, Kennesaw, Georgia, 2005; Vol. 1, pp 137-142.
96. Liu, Q.; Krishnapuram, B.; Pratapa, P.; Liao, X.; Hartemink, A.; Carin, L., Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra. In *ASILOMAR Conference: Biological Aspects of Signal Processing*, 2003.
97. Savitzky, A.; Golay, M. J. E., Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964,** 36, (8), 1627-1639.
98. Andersson, R.; Hämäläinen, M. D., Simplex focusing of retention times and latent variable projections of chromatographic profiles. *Chemometrics Intell. Lab. Syst.* **1994,** 22, (1), 49-61.
99. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **1998,** 805, (1-2), 17-35.
100. Wong, J. W. H.; Cagney, G.; Cartwright, H. M., SpecAlign - processing and alignment of mass spectra datasets. *Bioinformatics* **2005,** 21, (9), 2088-2090.
101. Wong, J. W. H.; Durante, C.; Cartwright, H. M., Application of Fast Fourier Transform Cross-Correlation for the Alignment of Large Chromatographic and Spectral Datasets. *Anal. Chem.* **2005,** 77, (17), 5655-5661.
102. Karstang, T. V.; Eastgate, R. J., Multivariate calibration of an X-ray diffractometer by partial least-squares regression *Chemometrics Intell. Lab. Syst.* **1987,** 2, (1-3), 209-219.
103. Kvalheim, O. M.; Brakstad, F.; Liang, Y. Z., Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise *Anal. Chem.* **1994,** 66, (1), 43-51.
104. Rietjens, M., Reduction of error propagation due to normalization - effect of error propagation and closure on spurious correlations. *Anal. Chim. Acta* **1995,** 316, (2), 205-215.
105. Conrad, T. O. F.; Leichtle, A.; Hagehulsmann, A.; Diederichs, E.; Baumann, S.; Thiery, J.; Schutte, C., Beating the noise: New statistical methods for detecting signals in MALDI-TOF spectra, below noise level. In *Computational Life Sciences Ii, Proceedings*, Berthold, M. R.; Glen, R.; Fischer, I., Eds. Springer-Verlag Berlin: Berlin, 2006; Vol. 4216, pp 119-128.

106. Box, G. E. P.; Hunter, W. G.; Hunter, J. S., *Statistics for Experimenters*. John Wiley & Sons: New York, 1978.
107. Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nystrom, A.; Pettersen, J.; Bergman, R., Experimental design and optimization. *Chemometrics Intell. Lab. Syst.* **1998,** 42, (1-2), 3-40.
108. Horst, P., *Factor Analysis of Data Matrices*. Holt, Rinehart and Winston, Inc: New York, 1965.
109. Kvalheim, O. M., Interpretation of direct latent-variable projection methods and their aims and use in the analysis of multicomponent spectroscopic and chromatographic data. *Chemometrics Intell. Lab. Syst.* **1988,** 4, (1), 11-25.
110. Kvalheim, O. M., Latent-structure decompositions (projections) of multivariate data. *Chemometrics Intell. Lab. Syst.* **1987,** 2, (4), 283-290.
111. Kvalheim, O. M.; Karstang, T. V., Interpretation of latent-variable regression models. *Chemometrics Intell. Lab. Syst.* **1989,** 7, (1-2), 39-51.
112. Jackson, J. E., *A Users' Guide to Principal Components* Wiley: New York, 1991.
113. Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometrics Intell. Lab. Syst.* **1987,** 2, (1-3), 37-52.
114. Geladi, P.; Kowalski, B. R., Partial least-squares regression - a tutorial. *Anal. Chim. Acta* **1986,** 185, 1-17.
115. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., The collinearity problem in linear-regression - the partial least-squares (PLS) approach to generalized inverses. *Siam Journal on Scientific and Statistical Computing* **1984,** 5, (3), 735-743.
116. Wold, S.; Sjöström, M.; Eriksson, L., PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* **2001,** 58, (2), 109-130.
117. Stone, M., Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B-Methodol.* **1974,** 36, (2), 111-147.
118. Filmoser, P.; Liebmann, B.; Varmuza, K., Repeated double cross validation. *J. Chemometr.* **2009,** 23, (3-4), 160-171.
119. Smit, S.; van Breemen, M. J.; Hoefsloot, H. C. J.; Smilde, A. K.; Aerts, J.; de Koster, C. G., Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* **2007,** 592, (2), 210-217.
120. Sjöström, M.; Wold, S.; Söderström, B., PLS discriminant plots. In *Pattern Recognition in Practice II*, Elsevier Science Publ. B. V.: Holland, 1986; pp 461–470.
121. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *J. Chemometr.* **2002,** 16, (3), 119-128.
122. Kvalheim, O. M.; Rajalahti, T.; Arneberg, R., X-tended target projection (XTP)-comparison with orthogonal partial least squares (OPLS) and PLS post-processing by similarity transformation (PLS plus ST). *J. Chemometr.* **2009,** 23, (1-2), 49-55.
123. Kvalheim, O. M., Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *J. Chemometr.* **2010,** 24, (7-8), 496–504.
124. Hoskuldsson, A., Variable and subset selection in PLS regression. *Chemometrics Intell. Lab. Syst.* **2001,** 55, (1-2), 23-38.
125. Centner, V.; Massart, D. L.; deNoord, O. E.; deJong, S.; Vandeginste, B. M.; Sterna, C., Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* **1996,** 68, (21), 3851-3858.
126. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S., *Multi- and Megavariate Data Analysis: Principles and Applications*. Umetrics Academy: Umeå, Sweden, 2001; p 533.

127. Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B., Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* **2000,** 54, (3), 413-419.
128. Lavine, B. K.; Davidson, C. E.; Rayens, W. S., Machine learning based pattern recognition applied to microarray data. *Comb. Chem. High Throughput Screen* **2004,** 7, (2), 115-131.
129. Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J., Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal. Chem.* **2008,** 80, (1), 115-122.
130. Brown, C. D.; Davis, H. T., Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics Intell. Lab. Syst.* **2006,** 80, (1), 24-38.