Metabolomics Studies of ALS

A multivariate search for clues about a devastating disease.

Anna Wuolikainen

UMEÅ UNIVERSITET

Department of Pharmacology and Clinical Neuroscience
Umeå University, Sweden
2009

ISBN: 978-91-7264-885-2
ISSN: 0346-6612-1303
Front cover: Sonja Nordström, Print & Media
Printed by: Print & Media, Umeå University
Umeå, Sweden 2009

*"Study as if you were going to live forever,
live as if you were going to die tomorrow."*

Maria Mitchell

Contents

List of papers

"Not everything that can be counted counts and not everything that counts can be counted."

-Sign (said to be hanging in Albert Einstein's office)

This thesis is based on the following papers, referred to in the text by the Roman numerals in bold font:

I. **Wuolikainen A**, Moritz T, Marklund S. L, Andersen P.M and Antti H. Predictive metabolomics for detection, interpretation and validation of metabolite patterns in human cerebrospinal fluid. *Submitted.*

II. **Wuolikainen A**, Hedenström M, Moritz T, Marklund S. L, Antti H and Andersen P.M. Optimization of procedures for collecting and storing of CSF for studying the metabolome in ALS. Amyotrophic Lateral Sclerosis, 2009, 10(4), 229-236.

III. Jonsson P, Sjövik Johansson E, **Wuolikainen A**, Lindberg J, Schuppe-Kostinen I, Kusano M, Sjöström M, Trygg J, Moritz T and Antti H. Predictive metabolite profiling applying Hierarchical Multivariate Curve Resolution to GC-MS data -A potential tool for multi-parametric diagnosis. Journal of Proteome Research, 2006, 5, 1407-1414.

IV. **Wuolikainen A**, Moritz T, Marklund S. L, Antti H, and Andersen P.M. Studies of the human cerebrospinal fluid metabolome reveal alterations associated with amyotrophic lateral sclerosis and subtypes of the disease. *Submitted.*

V. **Wuolikainen A**, Moritz T, Marklund S. L, Antti H, and Andersen P.M. ALS patients with mutations in the SOD1 gene have an unique metabolomic profile in the cerebrospinal fluid compared with ALS patients without mutations. *Submitted.*

Note: **Paper II** and **III** were reprinted with kind permissions from the publishers.

Abstract

Amyotrophic lateral sclerosis (ALS), also known as Charcot's disease, motor neuron disease (MND) and Lou Gehrig's disease, is a deadly, adult-onset neurodegenerative disorder characterized by progressive loss of upper and lower motor neurons, resulting in evolving paresis of the linked muscles. ALS is defined by classical features of the disease, but may present as a wide spectrum of phenotypes. About 10% of all ALS cases have been reported as familial, of which about 20% have been associated with mutations in the gene encoding for CuZn superoxide dismutase (SOD1). The remaining cases are regarded as sporadic. Research has advanced our understanding of the disease, but the cause is still unknown, no reliable diagnostic test exists, no cure has been found and the current therapies are unsatisfactory. Riluzole (Rilutek®) is the only registered drug for the treatment of ALS. The drug has shown only a modest effect in prolonging life and the mechanism of action of riluzole is not yet fully understood. ALS is diagnosed by excluding diseases with similar symptoms. At an early stage, there are numerous possible diseases that may present with similar symptoms, thereby making the diagnostic procedure cumbersome, extensive and time consuming with a significant risk of misdiagnosis. Biomarkers that can be developed into diagnostic test of ALS are therefore needed. The high number of unsuccessful attempts at finding a single disease-specific marker, in combination with the complexity of the disease, indicates that a pattern of several markers is perhaps more likely to provide a diagnostic signature for ALS.

Metabolomics, in combination with chemometrics, can be a useful tool with which to study human disease. Metabolomics can screen for small molecules in biofluids such as cerebrospinal fluid (CSF) and chemometrics can provide structure and tools in order to handle the types of data generated from metabolomics.

In this thesis, ALS has been studied using a combination of metabolomics and chemometrics. Collection and storage of CSF in relation to metabolite stability have been extensively evaluated. Protocols for metabolomics on CSF samples have been proposed, used and evaluated. In addition, a new feature of data processing allowing new samples to be predicted into existing models has been tested, evaluated and used for metabolomics on blood and CSF. A panel of potential biomarkers has been generated for ALS and subtypes of ALS. An overall decrease in metabolite concentration was found for subjects with ALS compared to their matched controls. Glutamic acid was one of the metabolites found to be decreased in patients with ALS. A larger metabolic heterogeneity was detected among SALS cases compared to FALS. This was also reflected in models of SALS and FALS against their respective matched controls, where no significant difference from control was found for SALS while the FALS samples significantly differed from their matched controls. Significant deviating metabolic patterns were also found between ALS subjects carrying different mutations in the gene encoding SOD1.

Keywords: Amyotrophic lateral sclerosis (ALS), motor neuron disease, Lou Gehrig's disease, human disease, CSF, biomarkers, metabolomics, metabonomics, chemometrics, design of experiments, multivariate analysis.

Abbreviations

"Why is abbreviation such a long word?"-Anonymous

ALS	Amyotrophic Lateral Sclerosis
CNS	Central Nervous System
CNTF	Ciliary Neurotropic Factor
CSF	Cerebrospinal Fluid
CV	Cross Validation
DA	Discriminant Analysis
DoE	Design of Experiments
FALS	Familial Amyotrophic Lateral Sclerosis
FFD	Full Factorial Design
FrFD	Fractional Factorial Design
GC	Gas Chromatography
HMCR	Hierarchical Multivariate Curve Resolution
LC	Liquid Chromatography
LMN	Lower Motor Neuron
MND	Motor Neuron Disease
MS	Mass Spectrometry
MVA	Multivariate Data Analysis
m/z	Mass to Charge Ratio
NAA/Cr	N-acetyl aspartate/creatine
NMR	Nuclear Magnetic Resonance
OPLS	Orthogonal Projections to Latent Structures
PBP	Progressive Bulbar Palsy/Paralysis
PCA	Principal Component Analysis
PLS	Partial Least Squares (alt. Primary Lateral Sclerosis)
PMA	Progressive Spinal Muscular Atrophy
RI	Retention Index
SALS	Sporadic Amyotrophic Lateral Sclerosis
SD	Standard Deviation
SOD1	Copper- and zinc containing superoxide dismutase
TDP-43	TAR DNA binding protein
TMS	Trimethylsilyl
TOF	Time of Flight
UMN	Upper Motor Neuron

Notation

"Mathematics compares the most diverse phenomena and discovers the secret analogies that unite them." -Jean Baptiste Joseph Fourier

The following notation has been used throughout this thesis. Vectors are denoted by bold, lower case letters (e.g. $\mathbf{t}$) and matrices are denoted by bold capital letters (e.g. $\mathbf{X}$). Vectors are assumed to be column vectors unless indicated by transposition, (e.g. $\mathbf{t}^T$). A matrix inverse is denoted as $\mathbf{X}^{-1}$ for a matrix $\mathbf{X}$.

A	Number of components in model
K	Number of columns in $\mathbf{X}$
M	Number of columns in $\mathbf{Y}$
N	Number of rows in $\mathbf{X}$ and $\mathbf{Y}$
$\mathbf{B}$	Matrix of regression coefficients for $\mathbf{X}$, [K×M]
$\mathbf{E}$	Residual matrix of predictor variables, [N×K]
$\mathbf{F}$	Residual matrix of response variables, [N×M]
$\mathbf{P}$	Matrix of loading vectors for $\mathbf{X}$, [K×A]
$\mathbf{T}$	Matrix of score vectors for $\mathbf{X}$, [N×A]
$\mathbf{U}$	Matrix of score vectors for $\mathbf{Y}$, [N×A]
$\mathbf{X}$	Matrix of predictor variables, [N×K]
$\mathbf{Y}$	Matrix of response variables, [N×M]
$\mathbf{c}$	Weight vector for $\mathbf{Y}$, [M×1]
$\mathbf{p}$	Loading vector for $\mathbf{X}$, [K×1]
$\mathbf{t}$	Score vector for $\mathbf{X}$, [N×1]
$\mathbf{u}$	Score vector for $\mathbf{Y}$, [N×1]
$\mathbf{w}$	Weight or covariance loading vector for $\mathbf{X}$, [K×1]

1. Background

"We said, we said, we said. This time was gonna be different."
–Comeback Kid (Wake the Dead)

This chapter aims to acquaint the reader with amyotrophic lateral sclerosis (ALS) and explain why cerebrospinal fluid (CSF) biomarkers are of interest in ALS research. The concepts of chemometrics and metabolomics, on which this thesis is based, will also be introduced.

1.1 Amyotrophic lateral sclerosis

"Time takes us all. So why am I not just living for today?" - The Used (The Ripper)

Amyotrophic lateral sclerosis (ALS), also known as Charcot's disease, motor neuron disease (MND) and Lou Gehrig's disease, is the most common deadly, adult-onset neurodegenerative disorder. ALS strikes about 3/100 000 persons each year worldwide, men and women of all ethnical groups[1].

ALS is often described by the classical features of the disease. The clinical hallmark of ALS is a progressive degeneration of the upper and the lower motor neurons (UMN, LMN) in combination. As a result an evolving generalized paresis of skeletal muscles is observed. In later stages of ALS, patients regularly progress to become totally paralyzed. Signs from the UMNs include spasticity, while symptoms from the LMNs include atrophy, fatigue, cramps and fasciculations. Oculomotor, autonomic functions and additional brain functions are reported being relatively spared, although have been reported involved in some patients with the disease. ALS ultimately leads to death, commonly due to respiratory failure when the respiratory muscles have become involved in the disease progression.

In most cases ALS affects adults of middle age showing a rapid progression. The mean onset of ALS is 47-52 years for familial ALS (FALS) and 58-63 years for sporadic ALS (SALS)[2] but the onset may vary, ranging from juvenile forms to patients suffering from the disease in the later stage of life. The median survival time reported for ALS cases that pass without treatment is 26 months[3]. There are however some cases showing atypical progress[4]. In about 10% of the ALS cases, the patients survive more than ten years[3]. Approximately 20-50% of the patients show cognitive dysfunction of the fronto-temporal lobes and 3-5% (or more) develops dementia usually of frontotemporal type, semantic dementia or progressive non-fluent aphasia[2]. Patients diagnosed with progressive bulbar palsy (PBP, degeneration of LMN in the brainstem), progressive spinal muscular atrophy (PMA, degeneration of LMN in the ventral horn of the spinal cord) and patients diagnosed with primary lateral sclerosis[5] (PLS, degeneration of UMN) may all develop signs of both UMN and LMN degeneration and hence be diagnosed with ALS, although this is not always the case. The term motor neuron disease (MND) is often used to include diseases such as PBP, PLS and PMA[6].

ALS has been described as a heterogeneous syndrome with multiple clinical, genetic and histological subtypes with ill-defined borders[1, 7]. And despite much effort aimed at solving the mystery of ALS, it is still not known whether it is a single or multisystem disease, or several diseases associated with motor neuron death[8, 9].

ALS is commonly divided in to two major groups, SALS and FALS. Cases are classified as FALS when two, or several members in the same family have been diagnosed with ALS. The remaining cases are classified as SALS. FALS has been reported by epidemiological studies based on several different populations in 1-18% of ALS cases. Whether this reflects a hereditary disposition of the disease or populations commonly exposed by environmental factors remains unclear at present. Another fair question to ask is whether SALS cases actually are sporadic or if some cases are in fact FALS, classified as SALS due to insufficient family history.

In 1993 the first gene associated with ALS was identified when 11 different missense mutations was found in the gene encoding for CuZn superoxide dismutase (SOD1) in 13 different FALS families[10]. To date 151 mutations have been found in the gene encoding SOD1 in patients with ALS. It is however unclear whether or not all mutations are pathogenic[11]. At present, mutations have been discovered causative of ALS in six additional genes. Genetic studies have suggested existence of eleven additional loci for ALS, but the genetic defects remain to be identified. Mutations in SOD1 is at present the most common known genetic factor and has been reported in 12-23% of FALS cases and in 1-7% of SALS cases.

Diseases striking the nervous system represent demanding diagnostic tasks even for professional neurologists. The diagnosis of ALS is challenging in particular, since no exclusive diagnostic test exists for ALS. The diagnosis of ALS is decided after an exclusion of other possible diseases that may mimic ALS[2, 12]. The El Escorial[13] /Airlie House revised criteria[14] was developed as a template for diagnosis. Numerous diseases (over 40 differential diagnoses) exhibit overlapping symptoms with ALS at an early stage. Misdiagnosis made by neurologists has been reported in 5-8% of cases, of which some patients were suffering from treatable diseases[15-17]. The average time from first symptom until receiving the diagnosis ALS has been reported to be 13-16 months by Chio, A[18, 19]. One of the problems is the increased risk of misdiagnosis at an early stage[20], while diagnosing ALS after the patient has been ill for a longer time showing generalized symptoms are more straightforward[2, 12].

There is still no cure for ALS[2]. Riluzole (Rilutek®) is the only approved drug (was approved in 1996 by the FDA) for treatment of ALS and it has been shown to prolong life modestly[21, 22]. Indications of a better effect of riluzole treatment has been shown in patients when the treatment is started early, indicating that early diagnosis is essential. Quite recent studies have pointed towards the pathogenic process of ALS to begin much earlier and the pathology to not be restricted to involve only neurons but also glial and endothelial cells[23]. However, the mechanism of action of riluzole is not yet fully understood but the drug has been suggested to antagonize the release of the excitatory neurotransmitter glutamate into the synapse (for chemical structure of riluzole see figure 1). Additional treatment of ALS consists of symptom relief efforts and ventilator support when the patient has progressed to the point where breathing is impaired[2].

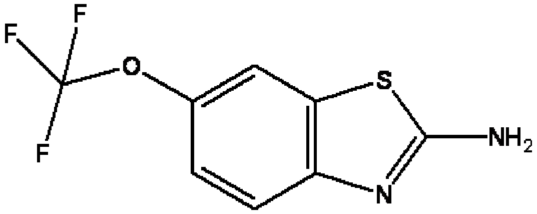


Figure 1. The chemical structure of riluzole (Rilutek®).

There are many questions remaining to be answered regarding ALS. What causes ALS, when does the disease actually start? Why do the motor neurons die? These are only a few of the important questions to be answered. But before ALS can be properly diagnosed, providing safer and earlier diagnosis by offering better diagnostic tools is of highest priority. For this reason, the area of biomarkers has gained a lot of attention.

1.2 Biomarkers

"I have not failed. I've just found 10,000 ways that won't work." -Thomas Edison

Biomarkers are substances that can function as indicators of a specific biological state, or be used to follow a biological process, e.g. a normal biological process, a pathogenic process or a response to treatment. The definition of a biomarker as proposed by the Biomarkers Definitions Working Group is that “a biomarker may be applicable as a diagnostic and/or prognostic tool, staging or classification of the extent of a disease, or to be used for prediction and monitoring of clinical response to an intervention”. Some biomarkers may also be developed and used as surrogate endpoints instead of, or in combination with, clinical endpoints (to obtain a characteristic that reflects how the patient feels, functions or survives). Biomarkers may also provide clues about the molecular and cellular bases of a disease.[24]

Finding suitable biomarkers for the diagnosis and prognosis of ALS has gained increasing momentum recently and the topic that has been reviewed thoroughly, of late by Pradat et al. [25-29]. Despite strong efforts, the search for specific biomarkers for ALS has been unsuccessful and no verified marker exists to date.

In the beginning of 2005 Rozen et al presented a global metabolomics study performed on blood from two rather small study groups[30]. Signatures related to ALS were discovered in the metabolite patterns, but so far no follow up study for validation of these potential markers has been reported. In 2008 an increased level of homocystein was reported in ALS patients in relation to healthy controls by Zoccolella et. al[31]. The results have not been confirmed by follow up studies and the increase was only seen in a few patients.

Mutations in the gene encoding for SOD1[32] may be regarded as one of the few biomarkers for ALS.

A reduction in the number of spinal anterior horn cells, presence of Bunina bodies and ubiquinated cytoplasmic inclusions in the remaining spinal motor neurons can be seen as histological hallmarks *post-mortem* in patients with ALS. These inclusions may be immunoreactive to antibodies against the TAR DNA binding protein (TDP-43) and p62. In addition, the immunoreactivity to TDP-43 has shown to be rare or absent in patients carrying mutations in the SOD1 gene[33, 34].

Till date, most studies on ALS have been performed on small and isolated sample sets. A recent study reported increased levels of inflammatory chemokines in 20 ALS patients when compared to 20 non-inflammatory, neurological disease controls[35]. A larger study from 2008 by Laaksovirta and coworkers showed elevated levels of ciliary neurotrophic factor (CNTF)[36]. Unfortunately the study compared significantly older cases to younger controls, which could likely bias the interpretation and questions the reliability of the results.

In vivo studies have been performed using proton magnetic resonance spectroscopy. Results using MRS points towards a decrease in the ratio of N-acetyl aspartate/creatine (NAA/Cr) in subjects with ALS[37]. Similar results have been shown in animal models [38]. Unfortunately, NAA/Cr cannot be used to diagnose ALS due to an overlap with other diseases[39].

The study of human disease is a complex task. This applies to ALS especially, where the site of pathology is unavailable until the patient is deceased and an autopsy can been performed. The aim of finding a biomarker, or a set of biomarkers, that could distinguish ALS from differential neurological conditions at an early stage must, therefore, rely on the theory that the pathology is reflected by the chemical composition of a biofluid as, for example, CSF[40]. A desirable feature concerning biomarkers is to be able to detect and measure them in accessible biofluids (such as blood, urine, saliva). However, searching for biomarkers in less accessible biofluids and tissues closer to the site of pathology within the central nervous system (CNS) may increase the chances of finding characteristic markers/patterns (e.g. due to MN pathology) and hence decrease the risk of finding markers remaining from symptoms throughout the body such as the breakdown of muscle tissue. Knowledge about established markers derived from the CNS may then give clues as to the identity of markers in more accessible fluids.

The large number of reported studies regarding biomarkers in ALS, without the identification of a common, specific marker, in combination with the complexity of the disease, could indicate that a pattern of markers may be a more tractable target as a diagnostic signature[29]. The assumption that it is possible to find a single, measurable disease marker may be unqualified when multiple factors could be involved in the pathology of a disease[41].

1.3 Cerebrospinal fluid

"Everything is simpler than you think and at the same time more complex than you imagine."
-Johann Wolfgang von Goethe

CSF *(liquor cerebrospinalis)* is the liquid that occupies the space between the arachnoid matter and the pia mater. Approximately 50-70% of the CSF is produced in the brain by modified ependymal cells in the choroid plexus. The volume of the space is in the range of 135-150 mL and the production rate has been estimated to 500 mL/day. The CSF has multiple functions, one is to protect the brain from physical shock but one is also to circulation chemicals and nutrients. The composition of CSF is dependent on the rate of production in the brain so analysis can offer insights of the CNS[42]. Sampling of CSF is done via a spinal tap (for method see Chapter 3.2.1) where CSF is collected into tubes (figure 2, left). The CSF is a clear (figure 2, right) low viscosity fluid with a relatively low abundance of protein constituents and a relatively high concentration of carbohydrates. An initial inquiry of patients with a suspicion of motor neuron disease usually includes cell number, albumin, glucose and lactate content to be measured in CSF (CSF tests listed for MND inquiry at Umeå University Hospital).

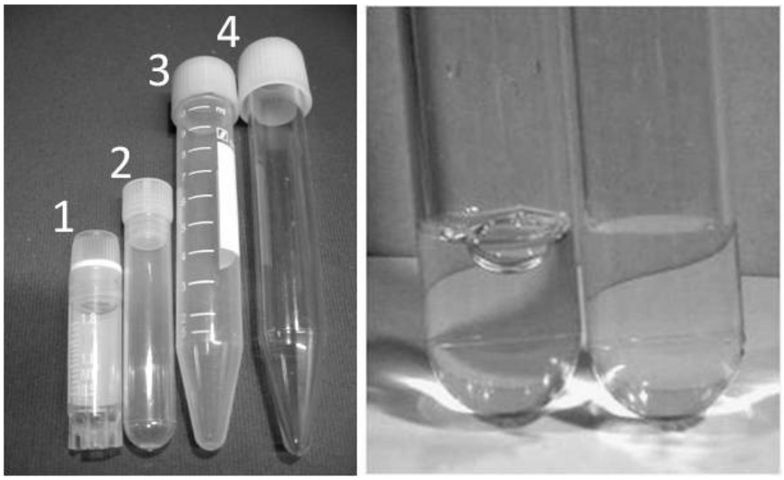


Figure 2. Tubes for collection and storing CSF (left), (1) 1.8 mL cryo-tube, (2) 3.5 mL tube for storage, (3) 10 mL polypropylene tube, (4) 10 mL polystyrene tube. CSF collected in polystyrene tubes (right).

1.4 Human metabolomics

"The perfect, normal person is rare in our civilization." -Karen Horney

The human metabolome is defined as the complete set of small molecules (e.g. amino acids, fatty acids, mono- and di-carbohydrates etc.) deriving from human metabolism[43]. It is a complex system where the chemical substances may originate from metabolism, gut micro flora[44], drugs, phytochemicals[45] etc. of various origins and environments. The metabolites present in a metabolome can be studied using two different approaches, these being; (1) A relatively small set of molecules are analyzed based on a pre-defined hypothesis, or (2) global screening to quantify and identify as many metabolites as possible, i.e. metabolomics. The former is the traditional approach within molecular biology and has successfully identified many of the components and interactions known in human metabolism today. However, the drawback is that the vast complexity of the biological system is not generally considered, so much could remain undiscovered. For this purpose, metabolomics as a part of a holistic systems biology approach may give more comprehensive information[46].

Metabonomics was introduced as a scientific field in 1999 by Nicholson et al [47] and short thereafter the term metabolomics was coined by Fiehn and coworkers[48]. These terms have similar definitions[47, 48]. However, metabonomics is often referred to as being nuclear magnetic resonance (NMR) based and used to describe multiple (but not necessarily comprehensive) metabolic changes caused by a biological perturbation. Metabolomics, on the other hand, has been established as being mass spectrometry (MS) based and places a greater emphasis on comprehensive metabolic profiling, regardless of the species being investigated. Nevertheless, the first studies performed according to the definition of metabolomics were published over 40 years ago[49, 50]. Recently a study of the human metabolome reported a major breakthrough in the field. In this study, Shrekuumaar and co-workers found sarcosine (the methylated form of glycine) as a putative biomarker for tumour progression in prostate cancer[51]. This result highlighted the potential of metabolomics as an important tool in disease diagnosis and prognosis for the future. A review of metabolomics and the search for biomarkers for use in the clinical arena was recently published by Nordström et. al.[52].

In theory human metabolomics aims to unravel the quantity and identity of all these small molecules in the human metabolome. This poses a great challenge since the biochemical species belong to diverse chemical classes and may be present in concentrations spanning a wide dynamic range [53]. Practically, human metabolomics instead strives to characterize the human metabolome by analysis of various biofluids (e.g., CSF, blood, urine, saliva, microdialysis fluid etc.) and tissues. Even though the system is dynamic, studies conducted on human subjects often consist of samples taken at one time point for each subject, providing only a snapshot of the metabolic status. There are studies where multiple samples taken over time have been investigated by metabolomics and metabonomics [54-56] but there are still not many studies performed that look in to human metabolism [54, 57-59].

Besides the potential to become an important tool for the discovery of new markers for diagnosis, prognosis and treatment of disease, metabolomics could also provide indications of pathology by highlighting specific affected pathways. However, identification of the detected metabolites is of high importance for success. Identification is still one of the great obstacles for metabolomics since few resources are available[60]. The Human Metabolomics Database (HMDB) is the most complete and comprehensive database in the world that collects metabolite and human metabolism data and was only quite recently developed by Wishart and collaborators[61]. The CSF metabolome was the first biofluid to be comprehensively characterized in HMDB[42, 61]. A minimum reporting standard regarding chemical analysis and practices related to all aspects of metabolomics has been proposed by the Chemical Analysis Working Group and Metabolomics Standards Initiative[62]. Hopefully, a combination of these efforts will result in more accessible information regarding the difference of specific metabolites in various studies and how the studies were performed.

It may be fair to say that studying the human metabolome is a difficult and challenging, but exciting task that can generate vast amounts of data to help describe a complex reality. To deal with the size and the complexity of the data, suitable tools for data analysis, interpretation and visualization are required.

1.5 Chemometrics

"We are drowning in information but we are starved for knowledge." -John Naisbitt

Chemometrics constitutes the information aspect of complex systems and should be seen as a concept for turning data into information[63]. This is achieved by combining statistical experimental design[64-66] and multivariate analysis[67-71] to extract information from complex data that has been optimized to contain the relevant information[72].

When chemometrics was introduced as a computational field it was mainly applied to chemical problems and multivariate analysis was primarily used for interpretation of spectral data[73]. The method has extended and today chemometrics is also commonly applied within biological research.

Many analytical platforms and technologies have developed over the last decades to allow samples and variables to be measured in a high-throughput manner. Due to this, it is easy to generate large amount of data potentially stored in data tables. Data tables may be useful for gathering the vast amount of data but still they will not simplify what the data means[63]. For this purpose modeling of the data may provide a better understanding. The methods commonly used within chemometrics are projection based methods. These can be used to produce models based on experimental data (also referred to as semi-empirical modeling or soft models). Even though it is today possible to measure many variables in metabolomics, the access to samples is still limited especially within the study field of human disease. This results in data tables consisting of a larger number of characterized variables than samples. The data will also contain noise and the variables being highly correlated. Missing values may also be present in the data. Chemometrics provides sophisticated multivariate statistical tools for handling such data[63].

There are three basic categories of multivariate analysis used within chemometrics being 1) exploratory analysis 2) classification analysis and 3) regression analysis. The most widely used multivariate methods are the unsupervised method principal component analysis (PCA) [69] and the supervised methods partial least squares (PLS) [70] and PLS-discriminant analysis (DA)[74], which have recently been further developed into orthogonal PLS (OPLS) [68] and OPLS-DA[67]. These methods will be further outlined in Chapter 3.

2. Aims of this work

"Nothing is impossible; the word itself says 'I'm possible.'" -Audrey Hepburn

The overall aim of this work was to find a biomarker or a set of biomarkers of diagnostic value to allow a more accurate diagnosis of ALS at an earlier stage.

To achieve this, the specific aims have been to:

- Investigate the stability of metabolites and metabolite patterns in human CSF samples in relation to variations in collection and storage procedures.

- Establish a working method for predictive metabolomics in human CSF.

- Make multivariate comparisons of the human CSF metabolome between matched control and ALS subjects to look for systematic differences in relation to ALS and ALS subtypes.

3. Methods

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

This chapter aims to present and discuss the methods used throughout this thesis. Focus will be on explaining how the chemometric thinking has been applied in all steps of the process to try to obtain tailored studies, structured in a way to allow the multivariate methods to extract information out of the complex data generated. Illustrating examples will be given from **papers I-V**.

3.1 Design of metabolomics studies

"A goal without a plan is just a wish."-Antoine de Saint-Exupery

Even though metabolomics aims to perform a hypothesis free screening of the metabolome, the designing of the studies is not. Scientific research is a process of guided learning i.e. collecting information (performing experiments, collecting information from literature etc.) about an area and further to decide whether the material supports or discards a pre-defined hypothesis. From this new hypotheses and knowledge may be generated. The object of statistical methods is to make that process as efficient as possible[64]. How a study is designed will decide the quality of the data, the information content in the data and without doubt decide what kind of conclusions may be drawn from it. Within clinical research some common study designs are the cohort studies, cross-sectional studies, case-control studies and case studies. All studies may generate useful results, but will provide different levels of interpretation. Depending on the purpose of the study, one or the other may prove more useful related to the cost, time and effort invested. However, for performing quality control of collection and storage of samples a basic statistical approach such as a factorial design can provide structure and interpretability.

3.1.1 Case –control studies

Case-control studies are often preferred when investigating rare diseases. When studying a human disease using a metabolomics approach, a well designed case-control study can provide data controlled for confounding and biases (age, sex etc.).

Case-control studies are based on subjects reported as cases of a condition (e.g.

subjects diagnosed with ALS) compared to subjects not reported as cases of the same condition (e.g. subjects not diagnosed with ALS). The most important issue in the case-control study is how the subjects are selected. The subjects should preferably be matched for all possible characteristics (age, sex, lifestyle etc.) to avoid introducing bias into the data.

3.1.2 Factorial designs

A full factorial design (FFD) is a common statistical experimental setup that was described for the first time in literature by Fisher in the early 1920's[75]. A FFD is an orthogonal design which investigates each variable at two or more discrete levels. Usually additional experiments are added in the center of the design to allow curvature and reproducibility of the model to be investigated (figure 3). FFDs often give too many experiments, hence a more convenient choice is to use a reduced FFD in terms of number of experiments i.e. a fractional factorial design (FrFD). However, using FrFD may be problematic when it comes to interpretation of the results since the investigated variables become confounded with each other. Another potential problem using FrFD is when the collection and analysis of samples is complex and samples may be lost due to technical difficulties.

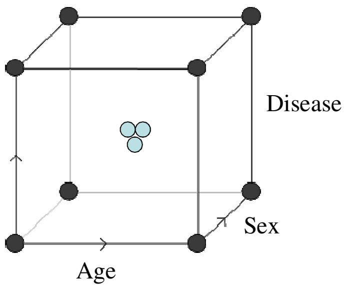


Figure 3. A FFD setup for investigating three factors (age, sex and disease) in two levels with three center points.

3.1.3 Design of studies in the presented work

Early in this project I was told "You can't design the patients in your studies". However, what can be done is to design the studies to fit the patients.

3.1.3.1 A tailored design

In **paper II**, the purpose was to investigate how alternations of collection and storage procedures may affect the concentrations of metabolites in CSF. To address factors that may influence the quality and to create an overview of time-spans to investigate, a few normal collection procedures were observed. Based on the regular procedure, factors (*i-iv*) were selected for investigation.

i) type of plastic tube (polystyrene/polypropylene)
ii) time (0, 10, 30, 90, 150 min)
iii) storage temperature (-80 °C/-20 °C)
iv) subject (age, diagnosis)

The type of plastic tube used for sampling is commonly made of polystyrene or polypropylene (figure 2, left). The time the sample was stored in room temperature differed between 10 or 30 min, if the sample was not divided into tubes for storage during this period of time, the samples were first put in refrigerator for short time storage. Therefore the samples in this study design were stored first in room temperature and samples to be stored for longer periods were first stored for 30 minutes in room temperature followed by 60 or 120 minutes in refrigerator. The possibility of freezing the sample immediately by placing them in liquid nitrogen was also included as a time factor denoted as 0 min. However, for this purpose a smaller type of cryotubes had to be used (figure 2, left). The temperature of freezer storage was also included as a factor. Samples are regularly stored at -80 °C, however in some rare cases storage at -20 °C may have been used over a shorter period of time. The combination of factors most resembling a “normal collection and storage procedure” was selected for replicates (e.g. sample collected in a polypropylene tube, kept 30 min in room temperature and stored at -80 °C). To allow for all possibilities CSF had to be collected into nine 10 mL tubes (4 polystyrene, 5 polypropylene) and four 1.8 mL cryotubes giving a total of 22 samples for each sampled subject (figure 4). The replicate tubes were collected as the fifth and last tube, allowing to check for changes from early to late collection.

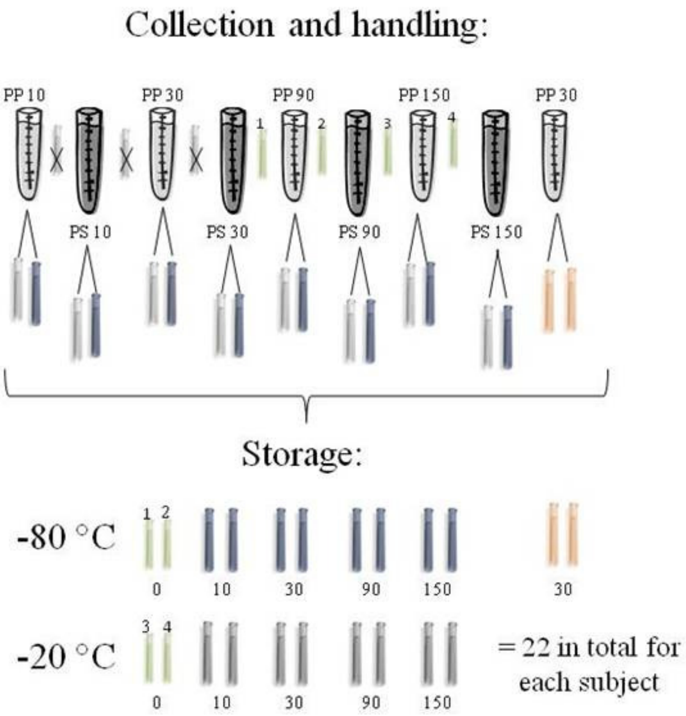


Figure 4. Tube setup for collection of CSF. PP=polypropylene tubes, PS=polystyrene tubes, numbers are given according to the time in room temperature/refrigerator storage.

To check for variability between subjects, repeatability over time and between subjects with varying diseases, we included many subjects in the study. By the time the analysis was started the total number of patients agreeing to contribute could not be fully determined. Hence the design needed to have the ability to be further extended. When samples had been collected from 13 male patients (6 ALS, 7 non-ALS) we terminated the study giving a total of 286 CSF samples.

In order to provide biological information related to disease as well as covering the complete set of variance in the investigated factors from collection and storage, the subjects were divided in subgroups of two (one ALS and one non-ALS). This allowed an adequate number of samples to be analyzed by gas chromatography coupled to mass spectrometry (GC-TOFMS) in the same run. A grouping was performed according to a scattered arrangement in the scores space of a principal component analysis [76, 77]. More details can be found in **paper II**.

3.1.3.2 Mining of a CSF biobank

In **paper IV** the main goal was to find a marker or a set of markers that could be developed into a diagnostic tool for ALS. The study also aimed to compare SALS with FALS and cases with and without a mutation in the SOD 1 gene in order to detect possible perturbations in the metabolite patterns in relation to these sub-groupings. For this purpose a case control study was chosen and the biobank at Umeå University Hospital was mined for samples of subjects with different subtypes of ALS and suitable control samples. The selection of subjects for the study had to be representative for the diseased population and any discovered putative marker or marker pattern should preferably be valid for patients of all ages and both genders. Alternatively, be specific for defined subsets of the population. To allow for result verification two subsets containing 79 samples each (39 ALS and 39 matched controls) were selected from the available CSF samples. The samples were selected to include both male and female subjects ranging in age between 40-80 years (figure 5). In addition, samples were selected to include patients with mutations in the SOD1 gene and other subtypes of ALS (pure UMN or LMN disease or PBP). Control samples were then matched for sex, age and time in freezer storage. The controls were selected to include differential diagnoses to ALS, other neurological conditions not mimicking ALS as well as healthy subjects.

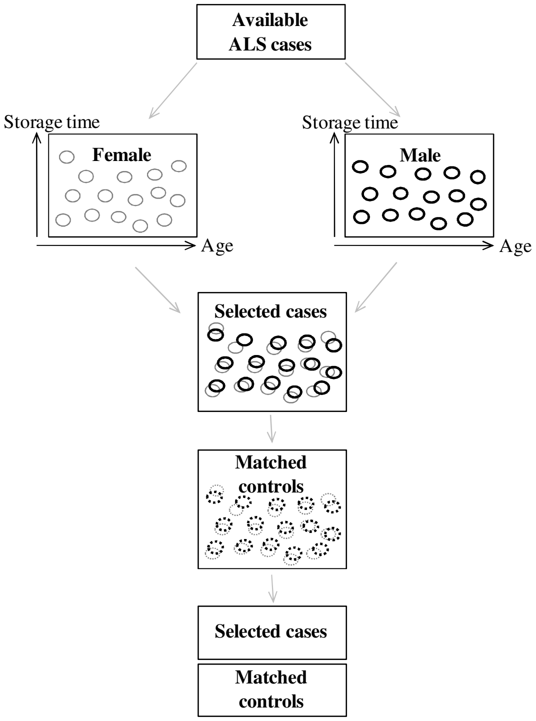


Figure 5. Scheme showing how samples from ALS patients were selected and how matched control samples were assigned.

3.2 Generation of metabolomics data

"If there is no struggle, there is no progress." -Frederick Douglass

3.2.1 CSF collection

Collection of CSF is done by performing a spinal tap. A needle is placed between the vertebras in the lower back while the patient is lying in a resting fetal position. The internal pressure can be measured prior the collection. A possible side-effect of performing a spinal tap is that some patients develop post-dural puncture headache[78]. The CSF is collected into tubes and sent for analysis or stored in -80 °C for future analysis. The procedure is most often performed for diagnostic purposes and when MND is suspected the CSF is analyzed for albumin, glucose, lactate and cell count. Depending on suspicion of disease additional test may be carried out.

CSF samples from ALS and control subjects were initially drawn as a part of the diagnostic evaluation of the patient during the visit at the Umeå University Hospital.

The ALS patients fulfilled the revised El Escorial criteria for ALS[14]. The spinal tappings were performed by the same physician at all times. A 22 G non-traumatic needle was used and the CSF was collected between the L3-L4, L4-L5 or L5-S1 while the patient was lying down on the right side. Most of the spinal taps were performed in the morning and the patients had been told to eat a large breakfast before arrival. The ALS subjects included in **papers IV** and **V** were genotyped for SOD1 gene mutations[79].

3.2.2 Extraction of metabolites from CSF

In the beginning of the first study, no extraction method for global analysis of the CSF metabolome had to our knowledge been reported. Thus, a slightly modified version of the standard in-house developed extraction method for blood[80] was tested and evaluated for CSF (**paper I**). In 2007 Pears et. al presented a protocol for GC-TOFMS of CSF[81]. No extraction procedure for CSF was though presented. In 2008 Wishart et al[42] presented work about the CSF metabolome where extraction and analysis of pooled CSF samples were performed.

3.2.2.1 A protocol for extraction of CSF

In **paper I, II, IV and V** CSF samples were thawed in room temperature (~25 °C) instead of 37 °C to avoid trigger enzyme activity. 100 µL of CSF was used for extraction and 900 µl extraction solution consisting of methanol and water (9:1, spiked with 11 stable isotope-labeled internal standards) was added. In each experiment the same batch of extraction solution was used to avoid introducing bias between samples. The amount of CSF used for extraction was varied between 100-200 µL as a test (not reported in the papers) and high concentrations of sugars in the CSF were established to cause the largest problems. The CSF samples were extracted in a bead mill without beads followed by incubation on ice. By centrifugation the supernatant could be separated and transferred to GC-vials. 200 µl of the supernatants were dried in a speedvac with heating (max 40 °C). The dry extracts were either frozen and stored at -80 °C prior GC-TOFMS or derivatized directly.

3.2.3 Derivatization prior to GC-TOFMS analysis

Prior to analysis with GC-TOFMS, derivatization is usually carried out to convert polar compounds containing functional groups, –OH, -SH or $-NH_X$, into more volatile derivatives. For this trimethylsilyl (TMS) groups are often introduced. A disadvantage associated with the reaction is that TMS-derivatization of reducing sugars usually results in five peaks (five tautomeric forms) complicating both identification and quantification. Another problem is the formation of side products (artefacts).

A two step derivatization procedure is the most widely used method for pre-treating samples prior GC-TOFMS analysis[48]. In the first step *O*-methylhydroxylamine hydrochloride is used to stabilize the carbonyl moieties in the metabolites, resulting in suppressed keto-enol tautomerism and avoidance of formation of multiple acetal-or ketal-structures. For reducing sugars this additional step limits the formation of anomers from five to two, resulting in a reduced number of eluating peaks in the chromatogram. As a second step *N*-Methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA) is added with 1% trimethylchlorosilane (TMCS) as catalyst to convert the remaining functional groups to TMS-derivatives.

The derivatization process is known to be sensitive to water. Due to this it is necessary to overview the dryness in the samples prior derivatization. For CSF high concentration of sugar compounds may be a factor causing problems when trying to obtain dry samples. The oximation reaction in the first step is a rather slow reaction, however the silylation reaction in the second step is fast. The silylation reaction will continue to proceed even though the samples are diluted with heptane prior GC-TOFMS analysis, and artefacts will continue to form as a side product during analysis. This together with fluctuations in the sensitivity of the measurements is some of the factors that may introduce systematic trends in the data seen in relation to GC-TOFMS runorder.

The derivatization procedure used in **papers I-V** has been optimized for plant extracts by Gullberg et. al.[82]. A summary of the extraction and derivatization procedures used for CSF can be found in figure 6.

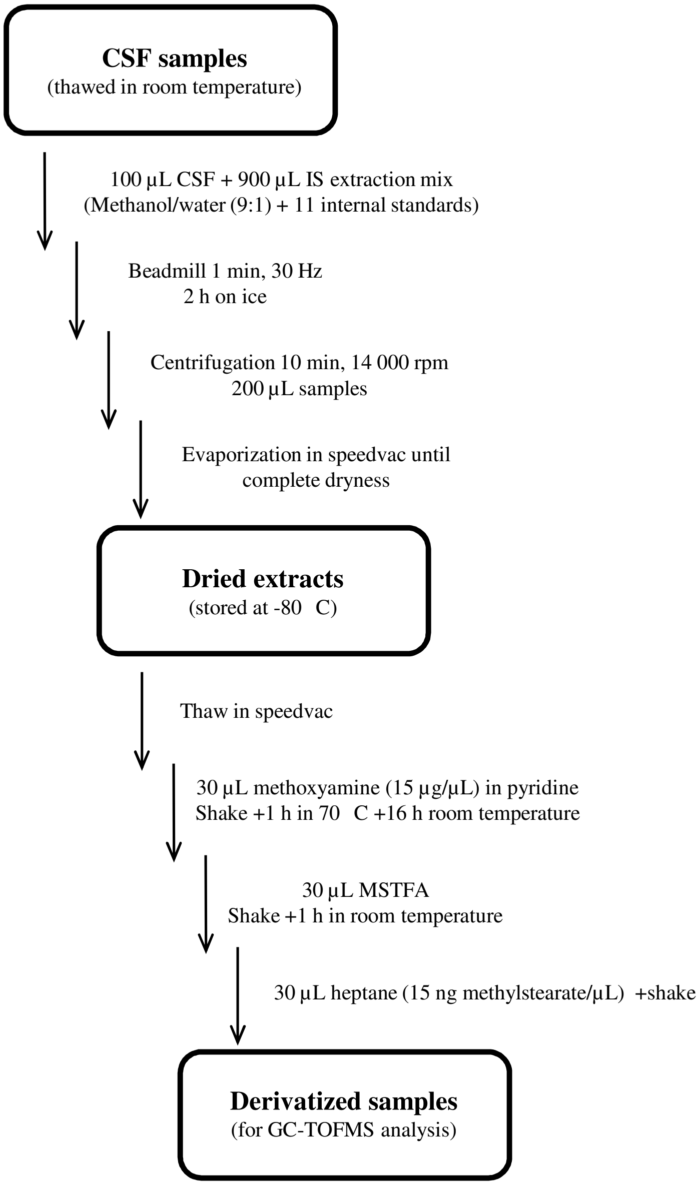

CSF samples
(thawed in room temperature)
100 µL CSF + 900 µL IS extraction mix
(Methanol/water (9:1) + 11 internal standards)
Beadmill 1 min, 30 Hz
2 h on ice
Centrifugation 10 min, 14 000 rpm
200 µL samples
Evaporization in speedvac until
complete dryness
Dried extracts
(stored at -80 C)
Thaw in speedvac
30 µL methoxyamine (15 µg/µL) in pyridine
Shake +1 h in 70 C +16 h room temperature
30 µL MSTFA
Shake +1 h in room temperature
30 µL heptane (15 ng methylstearate/µL) +shake
Derivatized samples
(for GC-TOFMS analysis)

Figure 6. A schematic overview of the extraction and derivatization procedures of CSF prior to GC-TOFMS analysis.

3.2.3 GC-TOFMS

GC-TOFMS is a common platform for generating metabolomics data. For GC-TOFMS analysis, pre-treatment of samples (i.e. extraction and derivatization) is often necessary.

3.2.3.1 Protocol for GC-TOFMS analysis

GC-TOFMS analysis in **papers I-V** was carried out according to a standardized protocol developed for various biofluids and plant extracts [80, 82]. 1 µL aliquot of the derivatized samples was injected splitless by an autosampler into the GC (10 m × 0.18 mm (inner diameter) fused-silica capillary column (chemically bonded with 0.18 µm DB 5-MS stationary phase). A temperature program was used where the temperature of the column was kept steady on 70 °C for 2 min. The temperature was then increased by 40 °C/min until 320 °C was reached. The temperature was kept at 320 °C for 2 min before the oven was allowed to return and stabilize at 70 °C. This is a rather short program developed for high-throughput analysis.

In the mass spectrometer ions were produced by hard ionization (70 eV electron beam, 2.0 mA) to allow for comparisons of the mass spectra against in-house and publicly available databases. Masses were acquired between *m/z* 50-800, at a rate of 30 spectra s^{-1} after a solvent delay of 165-170 seconds. This long delay prohibited lactate to be discovered in the samples but the delay was necessary to establish the baseline of the chromatograms.

In the beginning of the analysis a mixture of standardized alkanes (C_8-C_{40}) was analyzed to allow for retention index (RI) calculations in the system. For sensitivity assessment methylstearate (5 ng/µL in heptane) are usually analyzed between every 6th samples. For almost all of the GC-TOFMS analysis (**papers I-V**) the samples were analyzed in randomized order. However, randomizing the runorder may introduce bias of cases and controls. To avoid this, cases and controls were first randomized within each group. The runorder was then constructed by making sure that every second sample was a control.

3.2.3.2 Optimization of GC-TOFMS runorder for matched samples

To allow for result verification in **papers IV** and **V**, two sample subsets were constructed containing 78 samples each (39 ALS and 39 matched controls). The first subset was analyzed in a runorder created by the common randomization process. However, even by performing such a randomization there is a possibility to end up with a runorder bias affecting the comparison of the matched pairs (ALS vs. matched control). For this purpose the runorder for the second dataset was constructed by randomizing each pair (ALS-matched control) followed by randomizing which sample within the pairs to measure first (to avoid analyzing all ALS prior the matched control or vice versa). This procedure minimized the effect of the drift from the GC-TOFMS analysis between the matched samples as shown in figure 7.

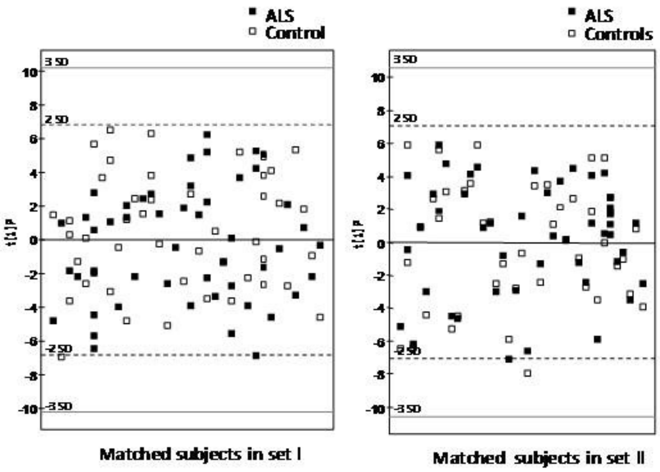


Figure 7. Optimization of the runorder for matched controls in two sets of samples. Left; samples randomized within each group and running every second sample from each group. Right: samples randomized in matched pairs with permutation of the runorder.

3.2.4 GC-TOFMS data for multivariate data analysis

In metabolomics the goal is to detect differences between samples based on their metabolite composition. For this purpose, multivariate analysis[67-70] is often used. These multivariate methods can handle two dimensional data structures. Using hyphenated methods for analysis, such as GC-TOFMS, information from both the chromatographic and the spectral dimension is provided giving a two dimensional data structure for each sample. In metabolomic studies the chromatography almost never separates all metabolites, providing chromatograms with overlapping peaks. This greatly complicates compound quantification and identification and thus also comparisons by multivariate analysis. In addition, when analyzing multiple samples as in metabolomics, the data also becomes three-dimensional. So, in order to be able to apply multivariate methods the data must be converted into a suitable format.

3.2.4.1 Hierarchical multivariate curve resolution (HMCR)

HMCR[83] is a multivariate deconvolution method able to divide the information from such three-dimensional data structures into two separate, but still linked two-dimensional parts. One part consisting of the relative concentration of resolved chromatographic profiles (e.g. metabolites) for all samples. Thus, HMCR can be regarded as a type of ‘mathematical chromatography’. For multivariate analysis it is important that the samples are characterized using the same variables, meaning that the columns in the two-dimensional data structure is in fact the concentration of the same metabolite. To check for this feature, validated-HMCR was recently developed[84]. The second part extracted by HMCR consists of the spectral information. For each resolved chromatographic profile a corresponding mass spectra is provided, which may be used for identification of the resolved compound. Multivariate curve resolution is data intensive. For this reason the data is divided into chromatographic windows and the deconvolution is performed for each window separately (hence the name hierarchical). Still, the method is rather time consuming and a limited number of samples can be resolved simultaneously.

3.2.4.2 Predictive HMCR

In order to use metabolomics as a diagnostic tool, the possibility of including and predicting the faith of new samples is crucial. In **paper III** a new feature of HMCR was implemented and tested that allowed new samples to be predictively resolved using the same parameters as used for resolving a set of representative model samples. This was achieved by resolving the new samples by utilizing the spectral information obtained for the previously resolved samples in the same defined time-windows. As a result of this the same metabolites as in the initially resolved samples will be quantified in the new samples and metabolites not present in the initially resolved sample set, will not be found in the new samples. For this reason it of great importance to use a representative subset of samples to provide valid metabolite information.

In **paper III** this was exemplified on a relatively small number of subjects but the same methodology can be used in order to, in theory resolve unlimited amounts of samples in a short time making it not only a potential tool for diagnostics but also a high throughput global metabolic screening method[85].

In **papers IV and V** the predictive feature of HMCR made it possible to combine the two datasets characterized at different points in time. This allowed modeling of all subgroups of ALS (FALS, SALS, with and without mutations in the SOD1 gene) together.

3.2.4.3 Manual calculations of metabolites

HMCR provides a matrix of resolved chromatographic profiles and mass spectra for identification that can be used for multivariate modeling. However, in some areas of the chromatograms the curve resolution performs insufficient. This is especially true for CSF where many metabolites are low abundant relative to urea, glucose (and other sugars) present in the samples at high concentrations. Smaller peaks covered by large co-eluting peaks are often missed by HMCR due to low variation between the samples. Large peaks (e.g. glucose, urea) are however split into many resolved profiles hence providing a too high number of resolved peaks. All the included IS are isotope-labeled endogenous metabolites. A problem using HMCR is an overlap of information from IS with endogenous metabolites in the resolved spectral dimension, providing un-precise calculations of the relative concentrations.

In **paper I** a method for manual calculations of small peaks in CSF was presented using a in-house developed Matlab based script (MATLAB 7.3 (R2006b), Mathworks, Natick, MA). The IS was used to locate the chromatographic window for the compounds by plotting the intensity of specific mass channels for the IS.

When the chromatographic areas were found for the IS, the area under curve was integrated for mass channels belonging to the endogenous compound. Some compounds split by HMCR were also recalculated using this method, except that the RI was used from HMCR to locate the chromatographic peaks and the mass spectrum from HMCR was used to estimate mass channels unique for the compound. This method was also used to extend the data in **papers II**, **IV** and **V**.

3.2.5 NMR data for MVA

NMR spectroscopy is a common analytical technique used to characterize biomolecules. Today it is a commonly used platform in the area of metabonomics.

In metabonomics, proton NMR (^{1}H NMR) is combined with solvent suppression techniques since biological samples contain water. Water will produce a very intense signal in the NMR spectrum that will obscure the peaks from the metabolites in the spectrum. The peak position from certain metabolites in the NMR spectrum is also sensitive to pH changes in the samples and peaks from certain metabolites may migrate for samples recorded at different pH.

3.2.5.1 NMR of CSF

In **paper II** NMR was used to analyze samples from one subject. 450 µL of CSF was combined with 50 µL D_2O. The ^{1}H NMR experiments were acquired on a 600 MHz instrument equipped with a cryoprobe to increase sensitivity. For pre-saturation of the water-signal excitation sculpting was used[86]. Quantification of metabolites was achieved using the NMR Suite software (Chenomx Inc., Canada) developed to identify and quantify metabolites in NMR data.

NMR may give complementary information to GC-TOFMS analysis. In **paper II** some peaks were found to be shifted in the NMR data between samples stored at -20°C compared to those stored at -80 °C. This suggested that a change in pH may have been introduced between CSF samples stored at different temperatures.

3.2.5.2 pH of CSF

In a publication from 2008 by Wishart et al.[42] CSF was reported to be heavily buffered by bicarbonate ions and to maintain a constant pH of 7.3.

Follow up studies of our NMR findings in **paper II** by measuring pH and pCO_2 on two samples, one stored at -20 °C and one stored at -80 °C (otherwise identically handled) showed a clear increase of the pH of the sample stored at -20 °C (The expected pH of CSF is reported to be 7.32.[87, 88]) (figure 8). Analysis of pH/pCO_2 was then performed on 25 of the CSF samples (from the 13 patients included in the study). The analysis showed clearly that samples stored at -20°C had an increased pH (above 9) and almost all CO_2 had vanished from the samples compared to the ones stored at -80 °C. Among the samples stored at -80°C the pH varied from just below 7 to slightly above 8. Some of the samples with lower pH showed a strikingly high amount of CO_2 that could not be explained. The increased pH and the lower amount CO_2 found in the samples stored at -20 °C could probably be explained by the increased storage temperature. At temperatures above -78 °C, $CO_2(s)$ has the ability to sublimate to $CO_2(g)$[89].

The literature was surveyed for information regarding CSF and pH relations and as early as in 1925 a study by MqQuarrie et. al[88] had realized this feature of CSF, however discussion about implications for metabolomics and metabonomics studies regarding variability of pH in CSF could not be found.

The body have several ways to buffer for pH changes. In blood, haemoglobin counts for the largest part of the H^+-removal, followed by other proteins and a buffering system of phosphates. Carbonic acid/bicarbonate accounts for the smallest part of the buffering in blood.[90] For CSF the buffering of pH depends largely on the carbonic acid/bicarbonate system (as suggested by Wishart et al.), and the pH in CSF will therefore depend on ventilation. The volume of the CSF space *in vivo* is constant, however when CSF is sampled the equilibrium is disturbed and CO_2 can diffuse to the surroundings. This can hence force the equilibrium towards CO_2 with implications that the pH will increase in the samples due to a lower concentration of hydrogen ions ($[H^+]$).

$$\mathbf{H^+ + HCO_3^- \rightleftharpoons H_2CO_3 \rightleftharpoons H_2O + CO_2}$$

$$\mathbf{H^+ + HCO_3^- \longrightarrow H_2CO_3 \longrightarrow H_2O + CO_2}$$

In order to measure this effect in relation to collection of CSF in tubes of different size, CSF was further collected from three patients according to a similar design setup (as the previous 13 patients). Three different tube sizes were used for the CSF collection in this batch and a time-span practically achievable was evaluated simultaneously. To mimic a worst case scenario (and a close to worst case scenario), two samples were collected in 10 mL tubes. One tube was capped and one was kept without lid to allow for free ventilation with the surrounding air. Both tubes were kept for more than 150 minutes in room temperature. The small tubes were filled with CSF, the medium tubes were filled to about half the volume and the largest tubes were only filled with a small volume in the bottom of the tube. All samples were analyzed for pH/pCO_2 and the worst case scenario samples were analyzed last. After the final analysis, the sample from the tube with closed lid was vortexed and new measurements were carried out. The results clearly showed that CSF collected in the small tubes had a pH value closer to 7.32, while the larger tubes showed increased pH (figure 8). From this we concluded that open tubes and vortexing of samples will cause the pH to rise. A take home message from this study was to control the sample handling of CSF and avoid contact with air to maintain a correct and stable pH.

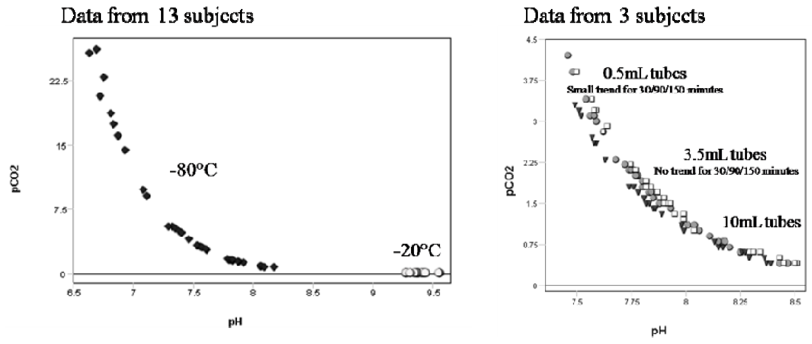


Figure 8. pH was increased in samples stored at -20 °C compared to -80 °C (left). Smaller tube sizes for collection of CSF were found to keep the pH closer to the expected pH 7.32 (right).

3.3 Multivariate analysis of metabolomics data

"Failure is success if we learn from it." -Malcolm Forbes

PCA and PLS are multivariate projection methods commonly used within chemometrics. In this work PCA [69] was used to explore the data and OPLS/OPLS-DA[67, 68] a further development of PLS, was used to detect systematic patterns between sample groups and to extract information regarding metabolites causing the separation between these groups.

3.3.1 PCA

PCA was first described by Karl Pearson in 1901[91] and has since then become one of the most widely used latent variable based method. Latent variables are variables that may not be directly observed or measured, but can be described or measured in variables directly affected by an underlying factor. For metabolomics this may be a perturbation of several metabolites remaining from the same pathway, perturbed by an underlying biological mechanism (such as a disease).

PCA compresses correlated variables in a multivariate data matrix **X** into A orthogonal principal components (latent variables). A is the number of components used to describe the variation within the data, hence the number of model components. PCA can be calculated using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm or alternatively by using Singular Value Decomposition (SVD)[92, 93].

The data (for instance metabolomics data) is stored in a matrix **X**, where each column describes a measured variable (e.g. a metabolite) and each row describes the concentrations of the measured variables (metabolites) for a sample. In PCA, the matrix **X** will be approximated by a matrix product of lower rank than **X**, $\mathbf{TP}^T$, where **T** is a matrix of scores that summarizes the variables (metabolites) in **X**. **P** is a matrix of loadings that reflects the importance of the variables in **X** for describing the principal components (latent variables) **T**. The first principal component will describe the largest variation found in the data. The second principal component will be orthogonal to the first component and describes the second largest variation within the data and so on. In order to estimate the number of principal components to use in the model, cross validation (CV) is the most commonly used method[94]. The variation not captured by the principal components is stored in a residual matrix **E**. The residual matrix should explain only low-variance stochastic events (i.e. noise) if the CV has performed sufficiently. A PCA model can be summarized as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

Since metabolomics data consist of multi-collinear variables (i.e. metabolites) the number of extracted scores **T** is often much less than the number of measured variables in **X**.

A useful feature of PCA is the possibility to overview the multivariate data by plotting the principal components. The data may be viewed as one, two- or three-dimensional pictures to map relationships between observations and variables (i.e. scores and loadings) to uncover clusters, groupings and trends and to discover deviating observations and variables (figure 9).

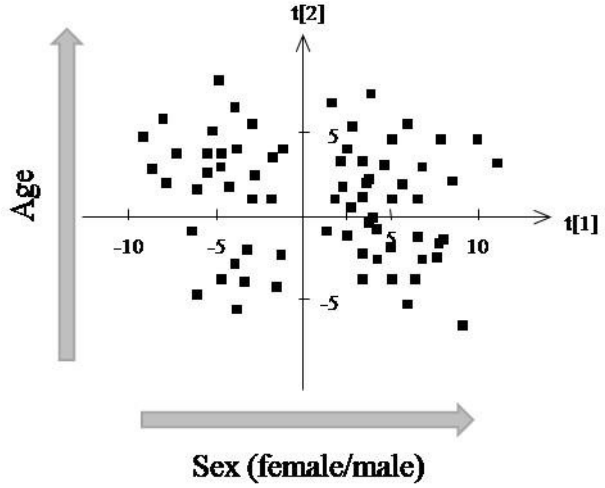


Figure 9. PCA can unravel groups, trends and deviating samples in the data. Here a plot shown of the scores from the two first principal components (t1/t2). The first score describes the largest direction of variation in the metabolomics data and the second score the second largest variation. The latent variables may for instance hold information related to sex, age etc.

3.3.2 PLS and OPLS

Multivariate regression methods are useful for relating measured variables (such as metabolites) in a matrix **X** to properties in a vector **y** or matrix **Y** (for instance continuous variables such as the age, weight or blood pressure of patients or discrete variables such as e.g. disease). These methods are called supervised methods since information about the response is used to find the linear relationship between **X** and **Y**. It can be shown that the multiple linear regression (MLR) method, also called Ordinary Least Squares (OLS)[73] provides the optimal solution to find the maximum fit of **X** to **Y** (i.e. minimizing the sum of squares of the residuals (the mismatch between the predicted and measured values of **Y**). The linear relationship between **X** and **Y** can be described as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{F}$$

Where **B** is the regression coefficients and **F** represents the residuals (the variation that cannot be explained by the model). When **B** is known, applying this solution to new samples measuring the same variables in **X, Y** can be predicted for new samples. The key is then to find **B**, and for MLR the solution to **B** becomes

$$\mathbf{B} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y}$$

However, if the matrix **X** contains co-linearity among the variables, or if the number of columns is larger than the number of rows, the inverse $(\mathbf{X}^{T}\mathbf{X})^{-1}$ does not exist and hence the equation for **B** cannot be solved. This is more or less always the case for metabolomics data where samples are characterized by a number of variables (e.g. metabolites) that far exceeds the number of characterized samples and the variables are co-linear and noisy. Therefore alternative methods are needed to model metabolomics data.

PLS and OPLS are latent variable regression methods based on the same assumption as made for PCA, namely that **X** can be described by a smaller number of latent variables. PLS starts by finding a set of latent variables, scores **T**. The scores are used to solve the problem of finding **B** by replacing $(\mathbf{X}^{T}\mathbf{X})^{-1}$ with $(\mathbf{T}^{T}\mathbf{T})^{-1}$. The score vectors ($\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, …, $\mathbf{t}_A$) are by definition linearly independent (orthogonal) and the inverse exist. PLS and OPLS can hence handle data structures where **X** contains co-linear variables and noise, such as metabolomics data.

In PLS the latent variables are calculated to maximize the co-variation between **X** and **Y** in relation to PCA where the latent variables are calculated to maximize the variation explained in **X**. PLS is hence focused on describing the variation in **X** that can be used to predict the response **Y** (e.g. the metabolites in **X** that contains information about the response **Y**). The PLS score vectors are formed as linear combinations of the original variables in **X** by

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

Where $\mathbf{t}_a$ represents the a th score vector and $\mathbf{w}_a$ represents the corresponding a th weight vector or the co-variance loading.

The models for **X** and **Y** can be summarized as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{T}\mathbf{C}^{T} + \mathbf{F}$$

The PLS components can be extracted using iterative algorithms (e.g. in this work the PLS-NIPALS algorithm)[73, 93, 95].

For PLS to be able to predict properties of new samples, all systematic effects in **X** must be incorporated into **T**. Systematic effects may include both variation that is linearly dependent to **Y** (correlates to **Y**) and linearly independent to **Y** (does not correlate to **Y**). When **Y** consists of a single response (**y**) the scores are calculated to be good estimators of both **X** and **y**, meaning that the residuals from both should be small. In order to obtain good predictions it is necessary to deal with systematic variation in **X** unrelated to **Y** so that only one PLS component is needed for modeling a single **y**. Recently PLS was further developed into OPLS dealing with such variation. OPLS splits the variation that is correlated to **Y** and the variation that is uncorrelated (i.e. orthogonal) to **Y** into two blocks. This provides OPLS with the advantage, compared to other generalized inverse regression models, that it facilitates model interpretation and visualizations of both types of variations.

A special case of PLS and OPLS is when the response **Y** is constructed as dummy variables holding information about the sample class. This is called discriminant analysis (DA) and variables in **X** can be extracted that separates (or discriminates) between the sample classes (the methods are then called PLS-DA/OPLS-DA).

3.3.3 MVA in this work

The combination of MVA methods used in this work was selected based on the data structures provided by the underlying designs.

In **papers I** and **III** the purpose was to model the data and check for possibilities to predict new samples, hence OPLS-DA was applied to a training data set and new samples belonging to a test data set were predicted into the model.

In **paper II** the aim was to find trends in the data caused by different collection and storage of CSF. The design of the study (Chapter 3.1.3.1) made it possible to extract variations in the metabolite patterns caused by the investigated factors (recall i-iv) for each subject using OPLS/OPLS-DA. The purpose of using multiple patients was to avoid interpreting metabolites altered by chance and instead focus on metabolites that were commonly affected to provide information of which metabolites were more prone to change due to non-biological factors. Hierarchical modeling using PCA to describe the overall trends in the factors was therefore applied (figure 10).

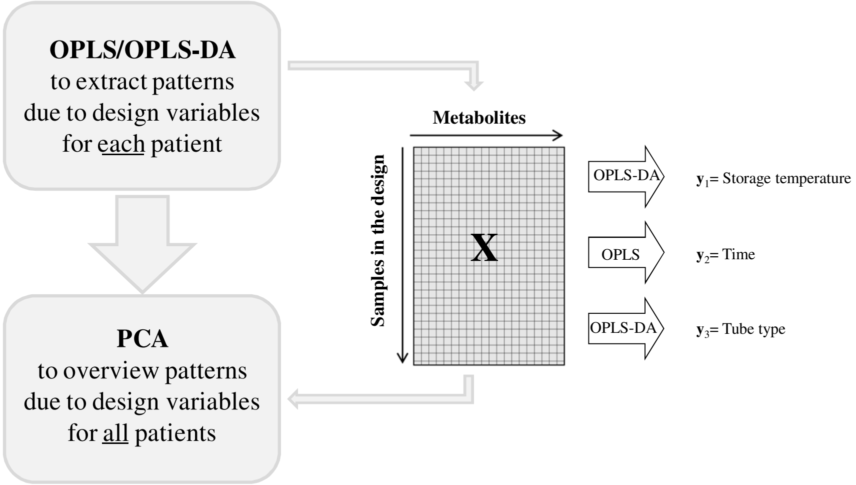


Figure 10. A schematic view of the hierarchical modeling (OPLS/OPLS-DA-PCA) of the designed metabolomics data in **paper II**.

Trends related to all the investigated factors (i-iv) could be found. The largest effect was found to be related to the storage temperature separating in the first component in the PCA model (figure 11, left). In the second and the third component differences related to tube type was found. A trend related to time and the collection order of the samples was also found in these components (figure 11, right). Both the samples collected as replicates (time 30, polypropylene tubes) and the samples collected in liquid nitrogen (time 0) were found to be more similar to samples collected before and after (recall figure 4), suggesting that the order of sampling may be likely to cause the trend rather than the factor time. A possible interaction of the tube type and the time may due to an increased in pH as seen from NMR and pH/pCO_2 measurements. The features of PCA to detect interactions and latent variables that may not be directly measured in the experiment setup can hence provide clues about underlying factors responsible for the trends seen in the data.

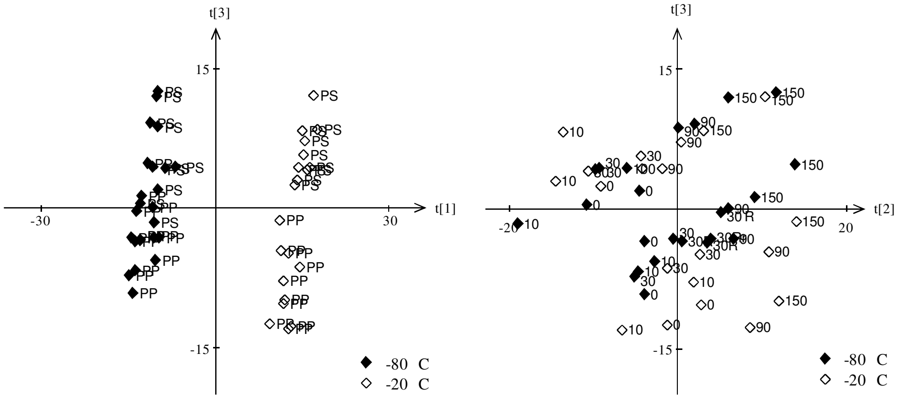


Figure 11. Hierarchical modeling (OPLS/OPLS-DA-PCA) revealed storage temperature to be the largest factor introducing perturbations in the metabolite data (left), trends in the data was also found in relation to tube type and the time the sample had been kept in room temperature/refrigerator (right).

In **papers IV** and **V** the aim was to find metabolites separating ALS from controls, thus OPLS-DA analysis was applied.

3.3.3.1 MVA in the search for biomarkers of ALS

Unraveling information regarding disease in data acquired from human samples is a complex task. A complicating factor is the many sources of variation is present in the data unrelated to the status of disease. By not considering and taking such variation into account severe misinterpretations of the data can occur.

Looking specifically at ALS, samples have often been stored for many years before an adequate number to include for statistical comparisons is obtained. Furthermore the age for disease onset in the patients tends to vary and the diagnosis itself is uncertain. Thus patients classified as ALS may in fact have a differential diagnosis and subjects classified as controls may have ALS although showing no symptoms or not yet been diagnosed. Besides the aspect of having false positive and false negative subjects, the subjects may also have other diseases (multiple diseases, diagnosed or not diagnosed). Patients with ALS may also have been symptomatic for longer or shorter times before a spinal tap is performed. In addition ALS is known to be a heterogeneous disease showing various rates of progression and degeneration of UMN and LMN.

From a statistical point of view these systematic variation sources can be divided into related metabolite patterns (e.g. disease) and unrelated metabolite patterns (e.g. analytical drift, different times of sample storage etc.). Heterogeneous factors may be disease subgroups better suitable for modeling. Possible misdiagnoses are regarded as errors in **y** and poor sample characterization are regarded as errors in **X**.

A controlled selection of samples (Chapter 3.1.3.2) can help in making factors known to introduce bias in the data unrelated to the disease variation. However, MVA should still be utilized to identify and overview patterns in relation to such variation sources. Because even though precaution has been taken to avoid bias, confounding factors can still exist in the data.

In **paper IV** PCA was used to screen for groupings, trends and deviations among samples. OPLS-DA was used to find patterns of metabolites separating between ALS and controls, FALS and controls as well as SALS and controls. To look for patterns in relation to ALS, cases diagnosed with ALS were modeled against their matched controls. All subjects were plotted in the model scores to allow visualization of the difference between the matched subjects. From this it was clear that the majority of the ALS cases differentiated from their matched control in the same direction. This was better visualized by subtracting the matched control score values from their corresponding ALS case score values (figure 12). Interestingly, the results showed a larger number of SALS cases not separating from the control group as compared to FALS cases. These samples were complicating the separation in the model and resulted in a model with poor predictive ability.

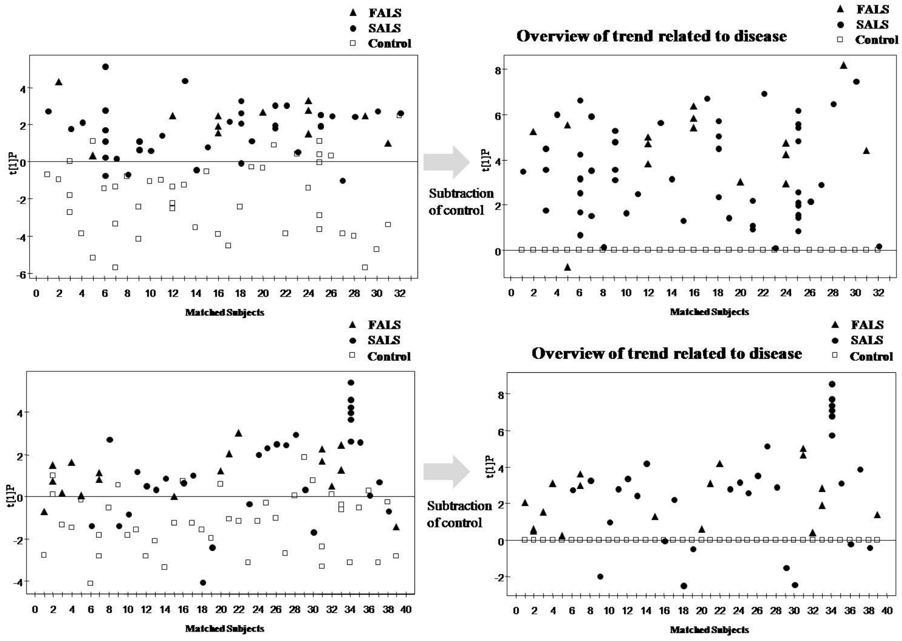
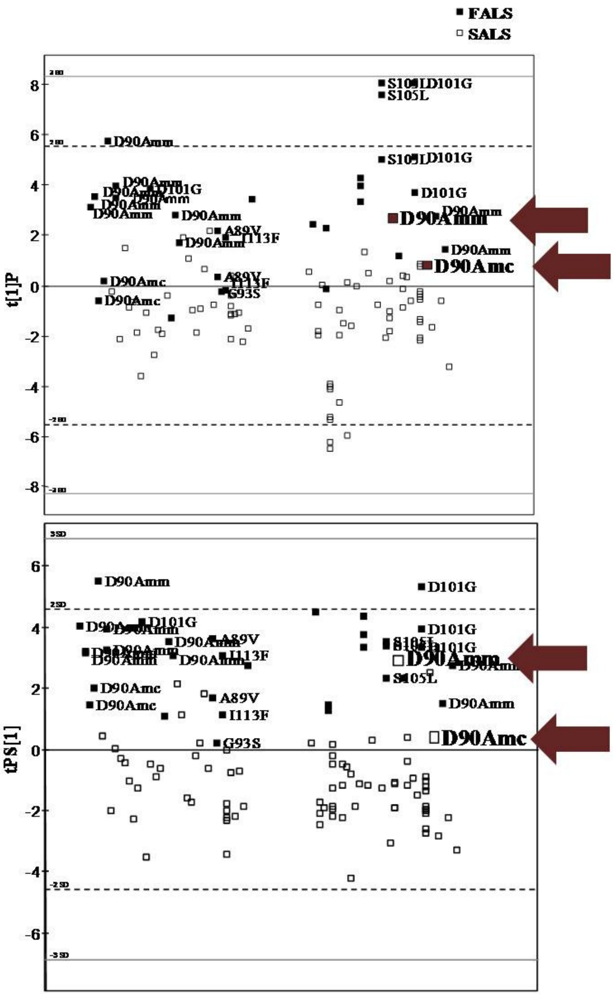


Figure 12. OPLS-DA scoreplots. Matched pairs (ALS-control) versus the first score vector (**t**[1]). ALS was modeled against controls (left) for the two datasets (above/below). Matched controls subtracted from the ALS in **t**[1] (right).

The ALS group was further divided into FALS and SALS and remodeled against their matched controls. This modeling revealed a significant separation for FALS versus controls. This was however not the case for SALS where a significant model between SALS and controls could only be obtained for one of the two datasets. This division into ALS subgroups made it possible to detect FALS as a more homogeneous group as compared to SALS.

Two datasets were analyzed at different points in time (six months in between). Drifts in the GC-TOFMS instrumentation were a possible complicating factor for the interpretation of the data after samples in the first subset had been predictedly resolved by HMCR into the other. This problem with additive and multiplicative deviations between sets of samples analyzed with a large time-span in between has been recognized previously (in other types of samples such as blood) and efforts have been made for solving this problem. In **papers IV** and **V** this deviation between the datasets could be accounted for by using the matched controls for normalization of the data allowing modeling of all ALS cases together.

OPLS-DA modeling between FALS and SALS revealed a significant difference between the two ALS subgroups. Two SALS cases were carriers of a SOD1 mutation and modeling was therefore also performed after excluding these subjects. These subjects were instead predicted into the new model. After exclusion of these two subjects a stronger model was obtained and the predictions into the model placed them closer to the FALS group as compared to SALS (figure 13). This indicated that a mutation in the SOD1 gene may be a possible cause for the separation between FALS and SALS or the patients could in fact be FALS rather than SALS (not yet discovered).



SALS (SOD1 positive) included in model.

SALS (SOD1 positive) predicted into model.

Figure 13. OPLS-DA model between FALS and SALS (above) showing two SALS cases with mutation in the gene encoding SOD1 included in the model and (below) showing the same cases predicted into the model.

Modeling FALS and SALS carriers of a mutation in the SOD1 gene (here called SOD1 positive) against SALS cases negative for SOD1 mutations (here called SOD1 negative) also resulted in a significant separation. Prediction of FALS (SOD1 negative) samples into the model resulted in four out of six cases being predicted into the group of SALS (SOD1 negative) while two samples were predicted on the border between the groups. OPLS holds the feature of summarizing the disease related variation of the metabolite data in the first model component which makes it possible to combine information from several models to check for consistency in the metabolite patterns between models. The large bias between FALS cases and carriers of a mutation in the SOD1 gene could be visualized by plotting the first score vector (**t**[1]) from the FALS versus SALS model against **t**[1] from the model of SOD1 mutation (negative versus positive) (figure 14). Here, two subjects regarded as FALS without a mutation in the SOD1 gene were clearly classified into the SALS group.

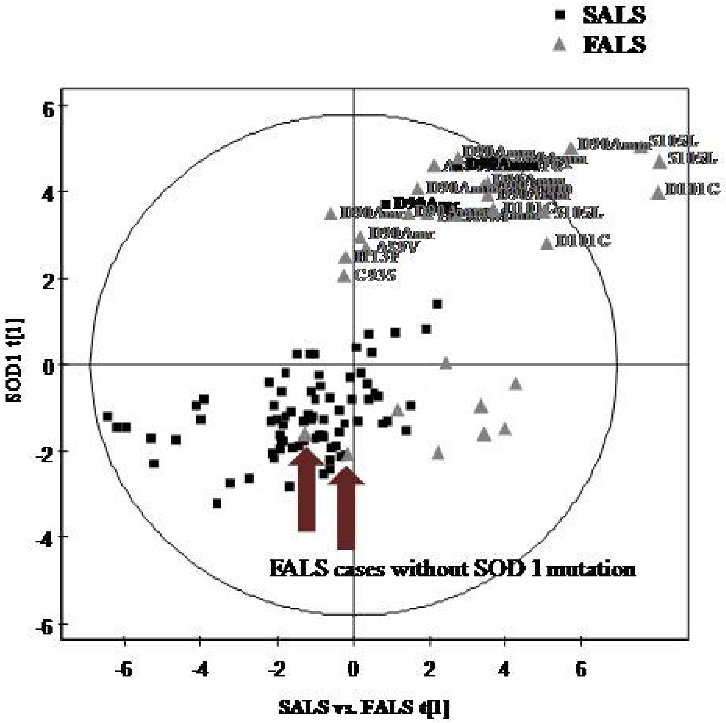


Figure 14. Combining information from two OPLS-DA models (SALS-FALS versus SOD1 mutation-non-SOD1 mutation) indicated two FALS cases without mutations in the SOD1 gene to fit into the SALS group.

To extract single metabolites or patterns of metabolites related to the separation between subjects, the importance of metabolites in **X** can be interpreted from the first correlation loading vector ($\mathbf{p}_{corr}$[1]) alternatively the first covariance loading vector ($\mathbf{w}$*[1]) of the OPLS-DA models. The interpretation will then reflect the metabolites separating the groups seen in the scores **t**[1]. To sort out which metabolites are common for the separation between matched samples, an additional approach was used in this work. Contribution of metabolites according to separation between pairs (ALS versus matched control or FALS versus matched control) were summarized and divided into metabolites showing an increase/decrease in relative concentration in subjects with ALS in relation to the control in 2/3 of the matched pairs (figure 15).

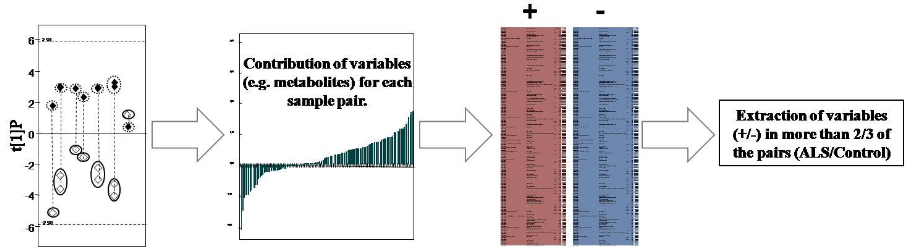

t[1]P
6
4
2
0
-2
-4
-6
Contribution of variables (e.g. metabolites) for each sample pair.
+
-
Extraction of variables (+/-) in more than 2/3 of the pairs (ALS/Control)

Figure 15. A scheme for extracting variables separating the groups based on contribution of variables (i.e. metabolites) between individual pairs, decided by applying a 2/3 cut-off for significance.

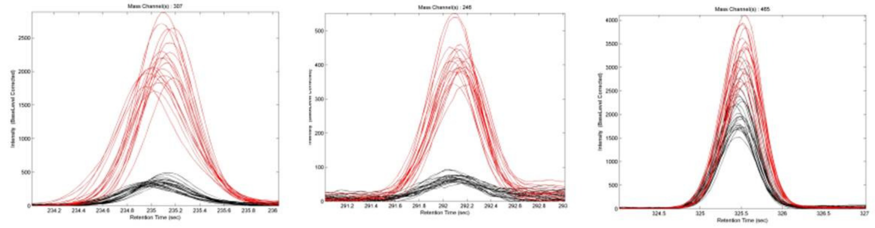
In **papers IV** and **V** glutamic acid was found to decrease in patient with ALS and FALS compared to controls. In addition the FALS group showed a larger decrease than SALS. Glutamic acid have been reported to increase in patient with ALS compared to controls [96, 97] although normal levels have also been reported[98]. In **paper II** glutamic acid was found increased in samples stored at -20 °C. It is therefore important to consider such factors as bias when interpreting the data.

The question is whether metabolites such as glutamic acid and other more prone to change in relation to factors like storage are suitable to draw disease related conclusion from. To do so, confounding factors must be clearly overviewed and controlled to not misinterpret the results. This is one of many aspects that may be regarded as important when looking to find markers for disease. In addition, the metabolites and patterns of metabolites suggested for diagnosis should preferably not be strongly correlated to factors such as sex and age. Chemometrics provides tools for extraction of such information from the data[99, 100].

3.4 Interpretation of metabolomics data

"All meanings, we know, depend on the key of interpretation."- George Eliot

MVA modeling is one way to extract information from metabolomics data. The meaning of the information must however be interpreted to allow for a better understanding and validation of the underlying biological processes. MVA may be used to interpret data but may also be used to generate information and input for further interpretation (e.g. pathway analysis, analysis by complementary techniques of highlighted compound classes i.e. target analyses, raw data inspection). The input metabolite data for such analyses should however be statistically validated, preferably in multiple studies, to avoid misinterpretation (which may regard pathway analysis in particular). However, depending on the system under investigation different interpretations must be performed. In **paper II** the changes in metabolite pattern originated from storage (e.g. chemical reactions in samples rather than metabolism). Methods used within chemometrics are considered transparent since the results in the latent variables can easily be traced back to the original variables (e.g. metabolites) in the raw data. In this study the raw data was inspected to confirm significant changes of concentration between samples in glyceric acid, glutamic acid and citric acid (fig. 16). Glyceric acid was found to be most affected by the different storage temperature. These results were found to be in accordance with the study made by Levine et. al[101].



GLYCERIC ACID **GLUTAMIC ACID** **CITRIC ACID**

Figure 16. Three acids prone to change in concentration due to storage at different temperatures (-20 °C/-80 °C).

3.4.1 Bioinformatics

When studying human disease the area of bioinformatics may provide complementary tools and software specifically developed for highlighting associated metabolic pathways in relation to metabolites found deviating in classification modeling. Pathway analysis can help understanding the biological relevance of metabolite deviations and clues may hence be provided regarding systems biology (the connection between genes, proteins and metabolites) and new hypotheses may be generated regarding mechanisms (e.g. disease). Bioinformatics also provides tools for visualization and for creating maps of the metabolic pathways, thereby results may become more clear for interdisciplinary collaborators and enhance interpretability.

3.5 Predictive metabolomics

"The only person who is educated is the one who has learned how to learn and change." -Carl Rogers

Predictive metabolomics is a an approach based on implementing the chemometric concept to metabolomic studies to allow for screening of large sample sets, without compromising the quality of the data. The aim of predictive metabolomics is to enable diagnostic modelling and pattern verification in independent samples and sample cohorts. The development of the approach has been largely dependent on the progress of HMCR (**paper III**) and its ability to perform curve resolution on new sample sets together with the use of predictive MVA modelling.

An important factor for predictive metabolomics is the application of the chemometric concept already from the beginning of the studies and throughout its duration. This means utilizing experimental design protocols for designing studies and for deciding on inclusion of samples or optimal conditions (e.g. sample handling and analytical procedures). In this way the probability to generate informative data without systematic bias is increased. Furthermore, chemometrics is applied throughout the studies for data processing, analysis and evaluation to allow for modelling, validation and visualization of the results.

The methodology we use for predictive metabolomics is to our knowledge the only approach were both data processing and multivariate classification allows predictions of new samples in a high throughput fashion. Predictive metabolomics are today continuously developed and applied routinely within the metabolomics facility at Umeå University for data processing in a variety of projects ranging from plants to humans[56, 58, 99, 102-105]. The current focus is on further development of the methodology to obtain robust diagnostic systems based on whole metabolite profiles or specific metabolite patterns.

Judging from the research aiming to find metabolic markers for disease or treatment of disease it is highly likely that specific metabolite patterns rather than single metabolites will make up the predictive signatures for ALS, ALS subtypes or treatment response. With this reasoning it is evident that predictive metabolomics, with its described properties, could be valuable for finding and verifying marker patterns that could later be developed into diagnostic or prognostic systems. Furthermore, a future use of the methodology will be to use the validated metabolite patterns and link it to pathway predictions. In this way possibilities for a deeper understanding of disease progression and clues for development of new targeted therapies based on validated metabolic signatures may be offered.

4. Results and discussion of the papers

"The beginning of knowledge is the discovery of something we do not understand." -Frank Herbert

This chapter aims to outline the objectives of, discuss and summarize each of the **papers I-V** included in the thesis.

4.1 Paper I

"When it is not necessary to make a decision, it is necessary not to make a decision." –Lord Falkland

Objective: To provide a working method for global analysis of CSF according to the predictive metabolomics concept.

Results:

(a) A working method for predictive metabolomics was presented for CSF based on GC-TOFMS data.
(b) Data extension by manual calculations of metabolites not fully resolved by HMCR to increase the information prior multivariate analysis was implemented.
(d) Data processing and data analysis was performed predictively to check for the ability of performing external validation, multiple study verifications and future diagnostics.

Comments and Discussion: When the first work of this thesis was initiated no protocol for global metabolite profiling of CSF using GC-TOFMS or similar methods could be found in the literature. In order to trust our results and to be able to use and apply the general metabolomics approach on a routine basis for CSF, the methodology had to be tested and adjusted to suit some of the unique features of CSF. CSF is often mistaken to be easier to analyze than blood and general reviews of metabolomics often mentions CSF as a biofluid suitable for metabolomics studies. Hence, there is a common belief that methods are readily reported in the literature. However, most studies have been performed measuring only a few metabolites or globally analyzed by NMR. The fact is that global analysis struggles to detect, quantify and identify a large set of metabolites derived from a range of chemical classes (i.e. sugars, amino acids, fatty acids, etc.) in CSF. The obstacles of analyzing CSF are the high levels of sugar compounds in combination with the low abundance of other metabolites. NMR can offer robust measurements of many metabolites yet have drawbacks such as low sensitivity. GC-TOFMS is commonly used for global screening of metabolites[106]. Within our group protocols for GC-TOFMS based metabolomics have been developed for blood (plasma, serum)[80], tissue, urine, saliva and a protocol for erythrocytes was previously reported by A, J et al[107]. In addition, fellow scientists have on repeated occasions addressing the issue how we perform our global analysis on CSF using GC-TOFMS as the analytical platform.

Paper I shows how the predictive metabolomics method can be used for CSF. The strategy was based on HMCR uniquely combined with manual integration of metabolites from GC-TOFMS data. The HMCR method used for resolving chromatographic profiles could not produce sufficient results for the whole chromatogram so in order to extend the data in such areas HMCR was complemented with manual calculations of metabolites to extend the information of the dataset prior multivariate analysis. This was followed by multivariate sample classification by means of OPLS-DA. An example was given on data from a real study from which the method was tested on a subset of samples from two patients. The data was divided into a training set (one patient; 22 CSF samples) and a test set (one patient; 22 CSF samples). The training set was resolved using HMCR with default settings, and the test set was predictively resolved by HMCR using the same settings as for the training set. OPLS-DA was the used to build a model from the training set data with regards to the storage temperature (-20°C or -80°C). The test set was then predicted into the model and the predictive accuracy was shown to be 100%. This was an important step to assure that the predictive metabolomics approach could be modified and applied successfully to model the CSF metabolome.

Other groups have recently presented studies where global analysis of CSF has been successfully performed [42, 81]. However, details are missing about the robustness of the methods and the ability to include new samples for external validation. Recently Crews et. al reported a study where CSF was characterized by LC-MS. The study aimed to investigate analytical and biological variance in CSF as compared to blood, where CSF was shown to be less variable[108].

4.2 Paper II

"In every walk with nature one receives far more than he seeks." -John Muir

Objective: To explore possible variations in the CSF collection procedure and storage conditions and investigate their effect on metabolite concentrations. One aim was to provide more reliable interpretations of metabolic changes in CSF samples from biobanks. A further aim was to give information and guidance on how to minimize such variation for future collection and storage of CSF.

Results:

(a) Changes in CSF metabolite patterns could be linked to all the investigated factors.
(b) Significant alterations in metabolite concentrations were found between samples stored in -20 °C compared to samples stored in -80 °C.
(c) Glyceric acid was found to be the metabolite most affected by storage temperature.
(d) Glutamic acid, pyroglutamic acid and citric acid were also found to be affected by storage temperature, although more moderately altered.
(e) pH was found to be increased in samples stored at -20 °C.
(f) Use of larger tubes for collection and careless treatment of the CSF could potentially increase pH in samples.
(g) Guidelines for collection and storage of CSF for metabolomics studies were provided.

Comments and Discussion: Before drawing conclusions about a change in concentration of one or several metabolites, confounding effects should always be considered. In this study different procedures for collection of CSF together with varied storage conditions were investigated. Studies concerning quality of samples and sampling have been performed in blood[109], urine[109], amniotic fluid[110]. However the literature regarding CSF was sparse and nothing was found looking at of the effects on metabolites.

The results showed that storing CSF samples at -20 °C instead of -80 °C resulted in perturbations of the metabolite composition. pH was increased in samples stored at -20 °C compared to samples stored at -80 °C. pH-fluctuations in CSF have been known for a long time[88]. However, the effect of sample quality for metabolomics has not been considered. When storing CSF samples in biobanks for future analysis, -80° storage is recommended and precaution should be taken to avoid ventilation with air to circumvent increasing the sample pH. Metabolites found to be affected by collection and storage factors should be interpreted with caution as potential disease markers in studies concerning samples stored in biobanks.

4.3 Paper III

"Prediction is very difficult, especially about the future." –Niels Bohr

Objective: To develop and evaluate a new feature of HMCR allowing GC-TOFMS data from new samples to be predictively resolved prior to multivariate modeling or prediction.

(a) We showed that it was possible to predictively resolve GC-TOFMS data from new samples using the mass spectral information form a representative model set.
(b) The presented findings allowed fast processing of large sample sets without compromising data quality.
(c) This was the first example of the combination of predictive data processing and multivariate predictions allowing for development of diagnostic systems based on metabolite profiles or patterns.

Comments and Discussion: To allow for metabolomics to be used as a diagnostic tool, the possibility of including and predicting new samples is crucial. In **paper III** a new feature of HMCR was introduced and tested to allow new samples to be resolved using pre-established settings for HMCR. Spectral information from previously resolved samples is used to search for the metabolites within the same defined time-windows for new samples. As a result of this the same metabolites as in the initially resolved samples will be quantified in the new samples and metabolites not present in the initially resolved sample set, will not be found in the new samples. For this reason it was of great importance to use a representative subset of samples to provide representative metabolite information.

The predictive feature with HMCR makes it possible to process in theory an unlimited number of samples. The time for processing is also decreased. Thysell et. al. have since then showed how processing of a subset of 16 samples (selected to cover the metabolic diversion amongst samples i.e. subjects) took 6 h 29 min compared to predicting 77 test samples in 10 min (<10sec/sample)[85].

This work was also a part of former PhD student's Pär Jonsson thesis "Multivariate Processing and Modeling of Hyphenated Metabolite Data" defended in 2005. He also developed HMCR with all its features during his PhD project. However, for testing the new predictive feature on human samples I was responsible for the work regarding human plasma samples all the way from sample extraction to performing the tests with predictive HMCR followed by interpretation of results.

4.4 Paper IV

"There's no strength where there's no struggle, without struggle there is no strength."
–Randy (Out of Nothing Comes Nothing)

Objective: To look for systematic differences in relation to ALS and ALS subtypes through multivariate comparisons of the human CSF metabolome between matched controls and ALS subjects.

Results:

(a) Dominant trends of decreased concentrations of metabolites were found in ALS samples in relation to control samples.
(b) FALS subjects significantly differed from their matched controls and were found to be a metabolically more homogenous group compared to subjects diagnosed with SALS.
(c) ALS subjects carrying a mutation in the gene encoding SOD1 were found to have a metabolite pattern deviating from ALS subjects without mutations in SOD1.
(d) Glutamic acid was one of the metabolites found to be systematically decreased in subjects with ALS compared to controls.

Comments and Discussion: The finding that glutamic acid was being systematically decreased in subjects with ALS compared to their matched controls is in controversy to previous studies showing increased or non altered levels in CSF. Interestingly, samples from subjects diagnosed with FALS differed significantly from their matched control samples, while this was not the case for SALS. A possible explanation to this was that the subgroup of SALS cases was found to be more heterogeneous as compared to FALS. In addition, ALS patients with a mutation in the SOD1 gene were found to have a metabolite pattern deviating from ALS subjects negative for SOD1. This might suggest a common neurodegenerative pathway for patients carrying mutations in the SOD1 gene. Although the study clearly suggests that systematic alterations in the CSF metabolome in relation to ALS and ALS subtypes exist, it is still too early to suggest specific markers or marker patterns of diagnostic value. To reach that goal verification in multiple studies will be required for finding common ALS related for extracting specific markers better suitable for classification of patients into disease sub-groups.

The large normal variation in humans is often addressed as a problem. However, by using sophisticated designs and tools for controlling and modeling such variation it may instead be seen as useful for finding stable marker patterns in human CSF more suitable for use in clinical testing.

4.5 Paper V

"Arriving at one goal is the starting point to another." –John Deweyit

Objective: To perform a detailed investigation of possible differences in metabolite patterns between carriers of different mutations in the gene encoding SOD1.

Results:

(a) ALS patients carrying a D90A SOD1mutation were found to have a different CSF metabolite pattern compared to ALS subjects carrying other SOD1 gene mutations.

Comments and Discussion: To date 151 mutations have been found in the gene encoding SOD1 in patients with ALS. However, whether or not all mutations are pathogenic is still unknown [11]. Many studies have been performed in order to unravel any common denominator for the mutants. Patients with SOD1 mutations have been reported to be clinically similar to patients without such mutations[32]. They have also been reported to have a disease onset at almost any site (known for ALS) although the dominating feature is spinal onset of a primarily LMN disorder[32, 111]. On the other hand there are no SOD1 mutation that has been associated with a predominantly UMN phenotype so far.

The question of this study has been whether patients with different mutations in SOD1 show common or separate features regarding metabolite patterns in CSF. We have discovered significant differences in the metabolome of CSF of ALS patients carrying a SOD1 mutation compared to ALS cases without mutations in the SOD1 gene. We here also reported differences between subjects carrying dissimilar mutations in the SOD1 gene. One of the more incomprehensible mutations found in ALS patients and one of the most frequent reported is the D90A SOD1 gene mutation[112]. This mutation may be inherited as a recessive trait with a characteristic of a slower progressing disease associated with longer survival times. In studies of different populations, pedigrees with ALS caused by D90A (homozygous) have members of the families carrying a D90A (heterozygous) mutation without showing symptoms of ALS. There are however some rare (and fewer) pedigrees where ALS patients have been found heterozygous for the D90A mutation.

In this study we showed that ALS cases carrying a D90A mutation are different from ALS cases carrying other SOD1mutations on a metabolite level in CSF. The finding that SOD 1 mutation cases are different from cases without a SOD1 mutation and that D90A (especially D90A homozygous) show differences in the metabolite patterns compared to other SOD1 mutation are supported by previous studies measuring neurofilament light chain in CSF indicating subjects with SOD1 gene mutations constitute a distinct subgroup within the ALS-syndrome, in particular patients with a D90 A mutation[113].

A number of amino acids were found to be generally decreased in the CSF metabolome in subjects carrying a D90A mutation compared to subjects carrying other SOD1 mutations, FALS and SALS. A number of unidentified compounds were also found altered and work to unravel the identity of these compounds will be needed to allow for a better understanding of potential mechanisms involved.

This study was performed on a limited number of ALS patients carrying a SOD1 mutation. The cases carrying a heterozygous D90A mutation showed a rather slow rate of progression. To provide a broader understanding about the D90A mutation in the SOD1 gene, inclusion of cases showing a faster disease progression would be of great interest. For this purpose follow up studies in a larger cohort will be needed.

5. Conclusion

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."
-Douglas Adams

Characterizing the human CSF metabolome in search for diagnostic biomarkers for ALS is a difficult task, which has yet to be fully addressed.

Working towards this goal we have developed a working methodology for screening and comparing groups of human CSF samples based on a comprehensive metabolic fingerprint. This methodology is based on a combination of GC-TOFMS analysis for metabolite detection and quantification, HMCR for data processing and multivariate data analysis for multiple sample comparisons. Furthermore, a predictive metabolomics approach has been developed and further modified and extended to work specifically for CSF. This approach allows screening of large numbers of CSF samples with maintained high data quality for quantification and identification of metabolites. It can also be seen as the first step towards developing truly predictive systems for diagnosis based on characteristic patterns of metabolites in CSF.

Studying the effects of collection and storage of CSF on metabolite stability for metabolomics analyses suggested that CSF sampled and stored under different conditions also expressed different metabolomic profiles. This was especially evident for the temperature of storage, where samples kept in -20 °C showed a clearly altered metabolic profile in a number of important metabolites (e.g. glyceric acid, glutamic acid, pyroglutamic acid and citric acid) as well as an increase in pH. These results highlighted the importance of standardized protocols for sample handling and storage in metabolomics, but also emphasized the importance of considering confounding factors effect on the metabolite profiles when screening samples stored in biobanks.

In this work it has also become clear that the selection of a representative control group is a crucial factor in human metabolomics studies. This may however be difficult to practically achieve for less accessible biofluids such as CSF. Another important factor herein is that the aim of the study must be well defined so that controls with the right properties for the objective are selected. In this work with the aim of finding markers or marker patterns specific for ALS, CSF samples from patients with differential diagnoses were included as a part of the control group, together with healthy controls, in order to find markers separating diseases with similar symptoms. The controls were also matched according to age, sex and the time the samples had been stored. This was done to achieve the best possibilities for making reliable and unbiased comparisons between sample groups as well as between matched control and ALS samples.

Using the developed methodology, together with a careful study design, including selection of ALS samples and matched controls, we could detect significant systematic patterns in the data related to ALS and ALS subtypes.

A general pattern related to ALS was seen as a decrease in the majority of the detected metabolites. A similar pattern has earlier been seen in a metabolomics study in blood plasma for ALS versus healthy controls[30].

Interestingly, we detected a larger metabolic heterogeneity among SALS cases compared to FALS, which were clearly more defined as a group. This was also reflected in models of SALS and FALS against their respective matched controls, where no significant difference from control was found for SALS while the FALS samples significantly differed from their matched controls. One possible explanation to this could be that ALS in fact consists of a group of diseases or that it is one disease with different characteristic metabolic profiles originating from different combinations of symptoms or being dependent on the rate of progression of the disease. Another possibility is that multiple mechanisms may participate in the pathological process.

It was also possible to differentiate between ALS cases carrying a D90A mutation and ALS cases carrying other SOD1 mutations. The reported findings have support in a previous studies measuring neurofilament light chain in CSF. This study concludes that subjects with SOD1 gene mutations make up a distinct ALS subgroup, especially patients with a verified D90 A mutation[113].

In summary we believe that we have a well working strategy for targeting the CSF metabolome in the hunt for diagnostic biomarkers or biomarker patterns for ALS or even more likely for specific ALS subtypes. In addition we already have a methodology in place for developing a diagnostic system if or when we manage to detect and validate a biomarker pattern for ALS. However, although the findings in this work are interesting and some maybe even promising, there is still a lot of work to be done before there is a diagnostic method for ALS based on metabolomics data in place. Even though we might be one step closer to diagnosing ALS we are still only in the infancy of exploring the complexity and wealth of the metabolome in relation to ALS.

6. Research ethics

”Great words won´t cover ugly actions, good frames won´t save bad paintings.” -Refused (New Noise)

Medical research demands great efforts in the area of research ethics. Studies conducted on human subjects require an even higher level of knowledge of the people working with such studies. The Declaration of Helsinki is a well known, established set of ethical principles concerning experimentation on humans that has been developed by the World Medical Association (WMA). The studies included in this thesis have been performed in accordance with the Declaration of Helsinki and have been approved by the medical ethical research board at Umeå University, Sweden (94-135, 98-240, 03-398, 09-160M).

Good research ethics may be thought of as following rules and directions. This should however only be seen as a part of conducting ethical research. For research to be considered ethical there are still far more aspects to cover. For example, the collected samples and generated data must be stored safely. A plan for the outcomes of studies and how results should be used and presented is also necessary to consider. Performing ethical studies may be seen as being one step ahead of the research.

The CSF samples collected and used in the studies included in this thesis were coded at the timepoint of sampling and stored in the biobank (#472) at (Umeå University Hospital) in locked -80 °C freezers. Selected samples were moved in boxes cooled by CO_2(s) to a -80 °C freezer located at the Department of Chemistry, secured by a temperature supervised alarm system. Data security was achieved using anti-virus software and backup of raw data to external hardware and DVD-discs was done continuously. Work in progress was secured by external backup against a distant server.

All studies were performed on CSF from human subjects. If possible the ratio between male/female samples selected for analysis was 50:50. In the study of collection and storage of CSF (**paper II**), only male subjects gave their informed consent and were included in the study. The study was designed to allow an extension in terms of new subjects (potentially female subjects).

Replicates of samples were either designed or randomly selected and deviating samples were excluded based on deviations in chromatograms due to technical problems and/or samples detected as outliers in multivariate space. All sample exclusions have been reported. The results have been validated using multiple samples and/or datasets. The combination of methods used throughout this thesis provides us with tools to overview the work in a way to assure a high quality.

Study results have been communicated at national and international congresses in addition to the published works. The research has also been described to the community in the journal Reflex (NHR), #3, 2007 and a local newspaper.

7. References

[1] Haverkamp LJ, Appel V, Appel SH. Natural-History of Amyotrophic-Lateral-Sclerosis in a Database Population - Validation of a Scoring System and a Model for Survival Prediction. *Brain*. **1995**;118:707-719.

[2] Andersen PM, Borasio GD, Dengler R, Hardiman O, Kollewe K, Leigh PN, et al. Good practice in the management of amyotrophic lateral sclerosis: Clinical guidelines. An evidence-based review with good practice points. EALSC Working Group. *Amyotrophic Lateral Sclerosis*. **2007**;8(4):195-213.

[3] Forsgren L, Almay BGL, Holmgren G, Wall S. Epidemiology of Motor Neuron Disease in Northern Sweden. *Acta Neurologica Scandinavica.* **1983**;68(1):20-29.

[4] Eisen A. Amyotrophic lateral sclerosis: A 40-year personal perspective. *Journal of Clinical Neuroscience*. **2009**;16(4):505-512.

[5] Swash M, Desai J, Misra VP. What is primary lateral sclerosis? 16th World Federation of Neurology Congress of Neurology; Buenos Aires, Argentina: Elsevier Science Bv; **1999**. p. 5-10.

[6] Swash M, Desai J. Motor neuron disease: Classification and nomenclature. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*. **2000**;1(2):105-112.

[7] Li TM, Alberman E, Swash M. Comparison of Sporadic and Familial Disease Amongst 580 Cases of Motor Neuron Disease. *Journal of Neurology Neurosurgery and Psychiatry*. **1988**;51(6):778-84.

[8] van der Graaff M.M, de Jong J, Baas F, de Visser M. Upper motor neuron and extra-motor neuron involvement in amyotrophic lateral sclerosis: A clinical and brain imaging review. *Neuromuscular Disorders*. **2009**;19(1):53-58.

[9] de Aguilar JLG, Echaniz-Laguna A, Fergani A, Rene F, Meininger V, Loeffler JP, et al. Amyotrophic lateral sclerosis: all roads lead to Rome. *Journal of Neurochemistry*. **2007**;101(5):1153-1160.

[10] Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, et al. Mutations in Cu/Zn Superoxide-Dismutase Gene Are Associated with Familial Amyotrophic-Lateral-Sclerosis. *Nature.* **1993**;362(6415):59-62.

[11] Felbecker AC, Valdmanis, P, Sperfeld, A, Waibel, S, Winter, S, Birve, A, Steinbach, P, Kassubek, J, Rouleau, GA, Ludolph, AC, Andersen, PM. Four familial ALS pedigrees discordant for two SOD1 gene mutations: Are all SOD1 gene mutations pathogenic? *Submitted: J Neuro Neurosurg Psychiatry*. **2009**.

[12] Li TM, Day SJ, Alberman E, Swash M. Differential-Diagnosis of Motoneuron Disease from Other Neurological Conditions. *Lancet*. **1986**;2(8509):731-733.

[13] Brooks BR. El-Escorial World Federation of Neurology Criteria for the Diagnosis of Amyotrophic-Lateral-Sclerosis. *Journal of the Neurological Sciences*. **1994**;124:96-107.

[14] Brooks BR, Miller RG, Swash M, Munsat TL. El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders.* **2000**;1(5):293-299.

[15] Traynor BJ, Codd MB, Corr B, Forde C, Frost E, Hardiman O. Amyotrophic lateral sclerosis mimic syndromes - A population-based study. *Archives of Neurology*. **2000**;57(1):109-113.

[16] Belsh JM, Schiffman PL. The amyotrophic lateral sclerosis (ALS) patient perspective on misdiagnosis and its repercussions. 6th International Symposium on MND/ALS; 1995 Oct; Dublin, Ireland: Elsevier Science Bv; **1995**. p. 110-116.

[17] Davenport RJ, Swingler RJ, Chancellor AM, Warlow CP. Avoiding false positive diagnoses of motor neuron disease: Lessons from the Scottish Motor Neuron Disease Register. *Journal of Neurology Neurosurgery and Psychiatry*. **1996**;60(2):147-151.

[18] Chio A. Update on ISIS survey: Europe, North America and South America. 2nd Consensus Conference on Debating the Possibility of Earlier Diagnosis of Patients with Amyotrophic Lateral Sclerosis; 1999 Jan 30; Versailles, France: Martin Dunitz Ltd; **1999**. p. S9-S11.

[19] Chio A. ISIS Survey: an international study on the diagnostic process and its implications in amyotrophic lateral sclerosis. 9th International Symposium on ALS/MND; 1998 Nov 17; Munich, Germany: Dr Dietrich Steinkopff Verlag; **1998**. p. 1-5.

[20] Brooks B. Earlier is better: the benefits of early diagnosis. *Neurology*. **1999**;8 Suppl 5:S53-7.

[21] Miller RGM, JD, Lyon M, Moore DH. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). Cochrane Database Systematic Review. **2007**(1):CD001447.

[22] Lacomblez L, Bensimon G, Leigh PN, Guillet P, Meininger V. Dose-ranging study of riluzole in amyotrophic lateral sclerosis. *Lancet*. **1996**;347(9013):1425-1431.

[23] Zhong ZH, Deane R, Ali Z, Parisi M, Shapovalov Y, O'Banion MK, et al. ALS-causing SOD1 mutants generate vascular changes prior to motor neuron degeneration. *Nature Neuroscience*. **2008**;11(4):420-422.

[24] Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*. *Clin Pharmacol Ther*. **2001**;69(3):89-95.

[25] Turner MR, Kiernan MC, Leigh PN, Talbot K. Biomarkers in amyotrophic lateral sclerosis. *Lancet Neurology*. **2009**;8(1):94-109.

[26] Sussmuth SD, Brettschneider J, Ludolph AC, Tumani H. Biochemical markers in CSF of ALS patients. *Current Medicinal Chemistry*. **2008**;15(18):1788-801.

[27] Kalra S, Arnold DL, Cashman NR. Biological markers in the diagnosis and treatment of ALS. *Journal of the Neurological Sciences*. **1999**;165:S27-S32.

[28] Kolarcik C, Bowser R. Plasma and cerebrospinal fluid-based protein biomarkers for motor neuron disease. *Molecular Diagnosis & Therapy*. **2006**;10(5):281-292.
[29] Pradat PF, Dib M. Biomarkers in Amyotrophic Lateral Sclerosis Facts and Future Horizons. *Molecular Diagnosis & Therapy*. **2009**;13(2):115-125.
[30] Rozen S, Cudkowicz ME, Bogdanov M, Matson WR, Kristal BS, Beecher C, et al. Metabolomic analysis and signatures in motor neuron disease. *Metabolomics*. **2005**;1(2):101-108.
[31] Zoccolella S, Simone IL, Lamberti P, Samarelli V, Tortelli R, Serlenga L, et al. Elevated plasma homocysteine levels in patients with amyotrophic lateral sclerosis. *Neurology*. **2008**;70(3):222-225.
[32] Andersen P. Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Current Neurology and Neuroscience Reports*. **2006**;6(1):37-46.
[33] Aggarwal A, Nicholson G. Detection of preclinical motor neurone loss in SOD1 mutation carriers using motor unit number estimation. *Journal of Neurology Neurosurgery and Psychiatry*. **2002**;73(2):199-201.
[34] Maekawa S, Leigh NP, King A, Jones E, Steele JC, Bodi I, et al. TDP-43 is consistently co-localized with ubiquitinated inclusions in sporadic and Guam amyotrophic lateral sclerosis but not in familial amyotrophic lateral sclerosis with and without SOD1 mutations. **2009**.
[35] Kuhle J, Lindberg RLP, Regeniter A, Mehling M, Steck AJ, Kappos L, et al. Increased levels of inflammatory chemokines in amyotrophic lateral sclerosis. *European Journal of Neurology*. **2009**;16(6):771-774.
[36] Laaksovirta H, Soinila S, Hukkanen V, Roeytta M, Soilu-Hanninen M. Serum level of CNTF is elevated in patients with amyotrophic lateral sclerosis and correlates with site of disease onset. *European Journal of Neurology*. **2008**;15(4):355-359.
[37] Suhy J, Miller RG, Rule R, Schuff N, Licht J, Dronsky V, et al. Early detection and longitudinal changes in amyotrophic lateral sclerosis by H-1 MRSI. *Neurology*. **2002**;58(5):773-779.
[38] Niessen HG, Debska-Vielhaber G, Sander K, Angenstein F, Ludolph AC, Hilfert L, et al. Metabolic progression markers of neurodegeneration in the transgenic G93A-SOD1 mouse model of amyotrophic lateral sclerosis. *European Journal of Neuroscience*. **2007**;25(6):1669-1677.
[39] Yin H, Lim CCT, Ma L, Gao YG, Cai YQ, Li DJ, et al. Combined MR spectroscopic imaging and diffusion tensor MRI visualizes corticospinal tract degeneration in amyotrophic lateral sclerosis. *Journal of Neurology*. **2004**;251(10):1249-1254.
[40] Dunckley T, Coon KD, Stephan DA. Discovery and development of biomarkers of neurological disease. *Drug Discovery Today*. **2005**;10(5):326-334.
[41] Ellis DI, Goodacre R. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst*. **2006**;131(8):875-885.
[42] Wishart DS, Lewis MJ, Morrissey JA, Flegel MD, Jeroncic K, Xiong YP, et al. The human cerebrospinal fluid metabolome. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*. **2008**;871(2):164-173.

[43] Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*. **1998**;16(9):373-378.
[44] Nicholson JK, Holmes E, Wilson ID. Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology*. **2005** May;3(5):431-438.
[45] Claudine M, Jane H, Rafael L, Augustin S. The complex links between dietary phytochemicals and human health deciphered by metabolomics. **2009**:In press.
[46] Sauer U, Heinemann M, Zamboni N. Genetics - Getting closer to the whole picture. *Science*. **2007**;316(5824):550-551.
[47] Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*. **1999**;29(11):1181-1189.
[48] Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nature Biotechnology*. **2000**;18(11):1157-1161.
[49] Pauling L, Robinson AB, Teranish.R, Cary P. Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. *Proceedings of the National Academy of Sciences of the United States of America*. **1971**;68(10):2374-.
[50] Dalglies CE, Horning EC, Horning MG, Knox KL, Yarger K. A Gas-Liquid-Chromatographic Procedure for Separating a Wide Range of Metabolites Occurring in Urine or Tissue Extracts. *Biochemical Journal.* **1966**;101(3):792-.
[51] Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu JD, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*. **2009**;457(7231):910-U176.
[52] Nordström A, Lewensohn R. Metabolomics: Moving to the Clinic. *Journal of Neuroimmune Pharmacology*. **2009**.
[53] Goodacre R. Metabolomics - the way forward. *Metabolomics*. **2005**;1(1):1-2.
[54] Jonsson P, Stenlund H, Moritz T, Trygg J, Sjostrom M, Verheij ER, et al. A strategy for modelling dynamic responses in metabolic samples characterized by GC/MS. *Metabolomics*. **2006**;2(3):135-143.
[55] Azmi J, Griffin JL, Antti H, Shore RF, Johansson E, Nicholson JK, et al. Metabolic trajectory characterisation of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data. *Analyst*. **2002**;127(2):271-276.
[56] Wiklund S, Karlsson M, Antti H, Johnels D, Sjostrom M, Wingsle G, et al. A new metabonomic strategy for analysing the growth process of the poplar tree. *Plant Biotechnology Journal.* **2005**;3(3):353-362.
[57] Pohjanen E, Thysell E, Jonsson P, Eklund C, Silfver A, Carlsson IB, et al. A multivariate screening strategy for investigating metabolic effects of strenuous physical exercise in human serum. *Journal of Proteome Research*. **2007**;6(6):2113-2120.
[58] Stenlund H, Madsen R, Vivi A, Calderisi M, Lundstedt T, Tassini M, et al. Monitoring kidney-transplant patients using metabolomics and dynamic modeling. *Chemometrics and Intelligent Laboratory Systems*. **2009**;98(1):45-50.

[59] Assfalg M, Bertini I, Colangiuli D, Luchinat C, Schafer H, Schutz B, et al. Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences of the United States of America*. **2008**;105(5):1420-1424.
[60] Gibney MJ, Walsh M, Brennan L, Roche HM, German B, van Ommen B. Metabolomics in human nutrition: opportunities and challenges. *American Journal of Clinical Nutrition*. **2005**;82(3):497-503.
[61] Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the human metabolome database. *Nucleic Acids Research*. **2007**;35:D521-D6.
[62] Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*. **2007**;3(3):211-221.
[63] Trygg J. Parsimonious Multivariate Models, Umeå, **2001**.
[64] Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. New York: John Wiley & Sons, Inc. **1978**.
[65] Lundstedt T, Seifert E, Abramo L, Thelin B, Nystrom A, Pettersen J, et al. Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*. **1998**;42(1-2):3-40.
[66] Eriksson LJ, Erik, Kettaneh-Wold N, Wikström C, Wold S. Design of Experiments; Principles and Applications. Umeå, **2000**.
[67] Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*. **2006**;20(8-10):341-351.
[68] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*. **2002**;16(3):119-128.
[69] Wold S, Esbensen K, Geladi P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*. **1987**;2(1-3):37-52.
[70] Wold S, Hellberg STL, Sjöström M, Wold H. PLS Model Building: Theory and applications. PLS modeling with latent variables in two or more dimensions.: Frankfurt am Main **1987**.
[71] Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Multi- and Megavariate Data Analysis; Principles and Applications. Umeå, **2001**.
[72] Wold S. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*. **1995**;30(1):109-115.
[73] Martens H, Næs T. Multivariate Calibration: John Wiley & Sons **1989**.
[74] Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. **1987**:185-196.
[75] Fisher RA. The arrangements of field experiments. *Journal of the Ministry of Agriculture of Great Britain*. **1926**;33:503-513.
[76] Wold S, Sjostrom M, Carlson R, Lundstedt T, Hellberg S, Skagerberg B, et al. Multivariate Design. *Analytica Chimica Acta*. **1986**;191:17-32.
[77] Linusson A, Wold S, Norden B. Statistical molecular design of peptoid libraries. *Molecular Diversity*. **1998**;4(2):103-114.
[78] Turnbull DK, Shepherd DB. Post-dural puncture headache: pathogenesis, prevention and treatment. *British Journal of Anaesthesia*. **2003**;91(5):718-729.

[79] Andersen PM, Nilsson P, Alahurula V, Keranen ML, Tarvainen I, Haltia T, et al. Amyotrophic-Lateral-Sclerosis Associated with Homozygosity for an Asp90ala Mutation in Cuzn-Superoxide Dismutase. *Nature Genetics*. **1995**;10(1):61-66.
[80] A J, Trygg J, Gullberg J, Johansson AI, Jonsson P, Antti H, et al. Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry*. **2005**; 15;77(24):8086-8094.
[81] Pears MR, Salek RM, Palmer DN, Kay GW, Mortishire-Smith RJ, Griffin JL. Metabolomic investigation of CLN6 neuronal ceroid lipofuscinosis in affected South Hampshire sheep. 7th International Conference on Brain Energy Metabolism; 2006 Aug 15-18; Lausanne, SWITZERLAND: Wiley-Liss; **2006**. p. 3494-3504.
[82] Gullberg J, Jonsson P, Nordstrom A, Sjostrom M, Moritz T. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical Biochemistry*. **2004**;331(2):283-295.
[83] Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, et al. High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Analytical Chemistry*. **2005**; 1;77(17):5635-5642.
[84] Jonsson P, Moritz T, Trygg J, Sjöström M, Antti H. Validated deconvolution of metabolic gas chromatography-mass spectrometry data. *Submitted*. **2009**.
[85] Thysell E, Chorell E, Svensson MB, Mortiz T, Jonsson P, Antti H. Efficient processing of human metabolism data by predictive metabolomics. *Manuscript*. **2009**.
[86] Hwang TL, Shaka AJ. Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic Resonance Series A*. **1995**;112(2):275-279.
[87] Hoffmann GF, Meieraugenstein W, Stockler S, Surtees R, Rating D, Nyhan WL. Physiology and Pathophysiology of Organic-Acids in Cerebrospinal-Fluid. *Journal of Inherited Metabolic Disease*. **1993**;16(4):648-669.
[88] MqQuarrie I, Shohl AT. A Coloritmetric Method for the Determination of the pH of Cerebrospinal Fluid. *The Journal of Biological Chemistry*. **1925**;lxvi:367-374.
[89] Aylward G, Findlay T. SI Chemical Data. 4th ed. Milton: John Wiley & Sons Australia **1998**.
[90] Thorén L. Vätskebalans. Stockholm: Almqvist & Wiksell Förlag AB **1983**.
[91] Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. **1901**;2:559-572.
[92] Demoor BLR, Golub GH. The Restricted Singular Value Decomposition - Properties and Applications. *Siam Journal on Matrix Analysis and Applications*. **1991**;12(3):401-425.
[93] Wold H. Nonlinear estimation by iterative least squares procedures. New York: Ed. Wiley **1966**.
[94] Eastment HT, Krzanowski WJ. Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*. **1982**;24(1):73-77.
[95] Andersson M. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*. **2009**.

[96] Rothstein JD, Tsai G, Kuncl RW, Clawson L, Cornblath DR, Drachman DB, et al. Abnormal excitatory amino acid metabolism in amyotrophic lateral sclerosis. *Annals of Neurolog.y* **1990**:18-25.

[97] Spreux-Varoquaux O, Bensimon G, Lacomblez L, Salachas F, Pradat PF, Le Forestier N, et al. Glutamate levels in cerebrospinal fluid in amyotrophic lateral sclerosis: a reappraisal using a new HPLC method with coulometric detection in a large cohort of patients. *Journal of the Neurological Sciences*. **2002**;193(2):73-78.

[98] Perry TL, Krieger C, Hansen S, Eisen A. Amyotrophic-Lateral-Sclerosis - Amino-Acid Levels in Plasma and Cerebrospinal-Fluid. *Annals of Neurology*. **1990**;28(1):12-17.

[99] Wiklund S, Johansson E, Sjostrom L, Mellerowicz EJ, Edlund U, Shockcor JP, et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*. **2008**;80(1):115-122.

[100] Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems.* **2009**;95(1):35-48.

[101] Levine J, Panchalingam K, McClure RJ, Gershon S, Pettegrew JW. Stability of CSF metabolites measured by proton NMR. *Journal of Neural Transmission.* **2000**;107(7):843-848.

[102] Chorell E, Moritz T, Branth S, Antti H, Svensson MB. Predictive Metabolomics Evaluation of Nutrition-Modulated Metabolic Stress Responses in Human Blood Serum During the Early Recovery Phase of Strenuous Physical Exercise. *Journal of Proteome Research*. **2009** Jun;8(6):2966-2977.

[103] Bruce SJ, Jonsson P, Antti H, Cloarec O, Trygg J, Marklund SL, et al. Evaluation of a protocol for metabolic profiling studies on human blood plasma by combined ultra-performance liquid chromatography/mass spectrometry: From extraction to data analysis. *Analytical Biochemistry*. **2008**;372(2):237-249.

[104] Bylesjo M, Nilsson R, Srivastava V, Gronlund A, Johansson AI, Jansson S, et al. Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen. *Journal of Proteome Research*. **2009**;8(1):199-210.

[105] Wibom C, Surowiec I, Mörén L, Bergström P, Johansson M, Antti H, et al. Metabolomic patterns in malignant glioma and changes during radiotheraphy -a clinical microdialysis study. *Manuscript.* **2009**.

[106] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*. **2004**;22(5):245-252.

[107] Zhang Y, Jiye A, Wang GJ, Huang Q, Yan B, Zha WB, et al. Organic solvent extraction and metabonomic profiling of the metabolites in erythrocytes. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*. **2009**;877(18-19):1751-7.

[108] Crews B, Wikoff WR, Patti GJ, Woo H-K, Kalisiak E, Heideker J, et al. Variability Analysis of Human Plasma and Cerebral Spinal Fluid Reveals Statistical Significance of Changes in Mass Spectrometry-Based Metabolomics Data. *Analytical Chemistry*. **2009**;81(20):8538-8544.

[109] Dunn WB, Broadhurst D, Ellis DI, Brown M, Halsall A, O'Hagan S, et al. A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *International Journal of Epidemiology*. **2008** Apr;37:23-30.
[110] Graca G, Duarte IF, Goodfellow BJ, Barros AS, Carreira IM, Couceiro AB, et al. Potential of NMR Spectroscopy for the study of human amniotic fluid. *Analytical Chemistry*. **2007**;79(21):8367-8375.
[111] Andersen PM, Sims KB, Xin WW, Kiely R, O'Neill G, Ravits J, et al. Sixteen novel mutations in the Cu/Zn superoxide dismutase gene in amyotrophic lateral sclerosis: a decade of discoveries, defects and disputes. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*. **2003**;4(2):62-73.
[112] Andersen PM, Forsgren L, Binzer M, Nilsson P, AlaHurula V, Keranen ML, et al. Autosomal recessive adult-onset amyotrophic lateral sclerosis associated with homozygosity for Asp90Ala CuZn-superoxide dismutase mutation - A clinical and genealogical study of 36 patients. *Brain.* **1996**;119:1153-1172.
[113] Zetterberg H, Jacobsson J, Rosengren L, Blennow K, Andersen PM. Cerebrospinal fluid neurofilament light levels in amyotrophic lateral sclerosis: impact of SOD1 genotype. *European Journal of Neurology*. **2007**;14(12):1329-1333.
[114] Fave G, Beckmann ME, Draper JH, Mathers JC. Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics? *Genes and Nutrition*. **2009**;4(2):135-141.

8. Acknowledgements

"I've got stars in my great big sky
I shall gaze upon, without leaving small ones behind
because they're harder to find" -The Starting Line (Something Left to Give)

Jag vill börja med att rikta ett stort tack till de patienter som ingått i studierna som denna avhandling bygger på. Utan er medverkan hade inga av resultaten i denna avhandling varit möjliga. Jag vill dessutom passa på att rikta ett tack till NHR och Hjärnfonden för stort stöd till forskningen samt Kempe stiftelsen, Anna Cederbergs stiftelse och Klinisk Neurovetenskap, Umeå universitet.

Den personen som har betytt mest för både min egen utveckling under dessa år såväl som för forskningen är definitivt min huvudhandledare **Peter M Andersen**. Tusen tack Peter för allt, du har med största inlevelse och finess guidat mig fram inom såväl forskningsfältet som genom djungeln av blanketter och regler. Du är en stor inspirationskälla och det är fantastiskt att jobba med dig. Jag är väldigt lyckligt lottad som har haft dig som handledare dessa år. De finns inte många som kan berätta och förklara saker på ett sätt som inspirerar tankeverksamheten som du.

Henrik Antti för att du är en klippa på att förstå vad jag menar när jag inte kan formulera det i ord. Dessutom tar du dig alltid dig tid att förklara och diskutera samma sak för sjuhundrade gången (ibland på samma dag) när jag inte kan släppa ett ämne. Ett stort tack även för att du bidrar med en fantastisk forskningsmiljö, god stämning och ser till att koordinera gruppen så att problemlösning sker i team. Det hade varit otroligt svårt att vara ensam doktorand inom ALS metabolomik utan denna struktur som grund.

Ett stort tack vill jag även rikta till **Thomas Moritz** för all hjälp under åren. Med dig som handledare har jag aldrig behövt känna mig utelämnad när de har strulat med instrument och analyser. Det är inte många proffessorer som det är större chans att hitta på labbet än på kontoret. Du har alltid en idé om hur man kan lösa problem och hur man kan analysera spännande metaboliter. Jag önskar bara att jag hade haft mer tid att testa allt.

Stefan Marklund för att du alltid tar dig tid att gå igenom saker från grunden med mig. Du är en fantastisk korrekturläsare med öga för detaljer som alltid resulterar i konstruktiv kritik. Det har varit spännande år och du har lärt mig så ofantligt mycket. Tack även för utbildning i vinkunskap (även om jag hade noll rätt på vinprovningen) och sabrering.

Krister Lungren, Annika Johansson, Jonas Gullberg och Ingabritt Carlsson som spenderat många timmar med att dels lära mig metod för metabolomics, förklarat GC-TOFMS, visat mig allt från derivatisering, normalisering och identifiering till växthus och hur man gör ett riktigt spex. Tack dessutom för all er hjälp i tid och otid med krånglande instrument.

Ett stort tack till mina medarbetere (och i somliga fall forna medarbetare) Sabine Björk (du borde få minst tre sidor acknowledgments), Jörgen Andersson, Johan Trygg, Pär Jonsson, Elin Thysell, Mattias Hedenström, Elin Chorell, Hans Stenlund, Susanne Wiklund, Ing-Marie Olsson, Anton Lindström, Michael Granström, Sofie Edvinsson, Sara Nilsson, Lina Mörén, Michael Sjöström, Tommy Löfstedt, Rasmus Madsen och Max Bylesjö.

Jonathan Gilthorpe for his splendid work helping me to proof-read this thesis (and correcting my most horrible Swenglish, all abuse of the English language is definitely added after you read my thesis). Henke, Iza, Lina, Tommy, Hans, Elin, Maria, Oskar, pappa och alla andra som läst och berättat vart jag tappat tråden samt hjälpt mig att fånga upp den.

Izabella and Steve for help with instruments and teaching me chromatography and mass spectrometry. Jiye A for team-work in 2005 (especially that time my design resulted in work in the middle of the night AND the morning after, of course on a Sunday).

Erik Johansson för ditt engegemang och genuina vilja att hjälpa till att tänka när svåra situationer av dataanalys uppstår. Du är en klippa när det kommer till att analysera vilka fallgropar som finns och bör undvikas i den statistiska djungeln. Jag kommer speciellt ihåg när du kom helt dyngsur efter en cykeltur i spöregn när jag inte lyckats förklara problematik via mail. I slutskedet har du även fått hoppa in som samordnare av problemlösning med korrupta filer. Oändligt mycket tack till dig Erik (och alla som stått bakom din hjälp på Umetrics).

Jag vill även tacka alla underbara människor på kemi, UPSC och sjukhuset som är på tok för många för att nämnas vid namn. Jag hoppas ni vet vilka ni är och att hur betydelsefulla ni varit för utfallet av dessa studier. Dock bör jag speciellt nämna Kerstin Gruffman som har stenkoll på mig och mina administrativa bryderier.

Ett tack till Suss, Tommy, Elin och Elin mina roomies och ex-roooomies för spännande disskusioner och konstruktiv kritik på jobb och musiksmak.

Hasse och Linus för att ni tar våra lunchdiskussioner till en helt ny (och oftast låg) nivå. Man ska aldrig underskatta utvecklingen som en gnällig fikarast (jag syftar på en kafferast när gnällfokus låg på ett speciellt copy/paste-tangentbord och du introducerade mig för ditt RDA-scpript). Tack dessutom Hans för att du alltid tar dig tid att formulera personligt anpassade förklaringar när jag är förvirrad. Trots att du är den värsta snackpåsen jag känner så är du samtidigt den bästa pedagogen! Jag uppskattar även din förmåga att se när jag behöver lakrits och uppmuntran.

Sabine, Caroline och Anna: Tack för alla spännande forskningsdiskussioner och allra mest för trevligt sällskap och god mat. Hoppas vi får många fler sådana stunder framöver, främst kanske på ALS/MND mötet i Berlin.

Oskar min personliga uppslagsbok ”Medicine for Dummies” under avhandlingsskrivandet. Johan Bucht med sin superkamera (som rymde hela UPLC-MS maskinen) för hjälp med foto till diverse presentationer under årens lopp. Benny på kemiförrådet som alltid hittar det matrial, kemikalier och liknande jag behöver ASAP (!) i djungeln av företag samt aldrig sinar på historier om grill-torsdagar och bastubravader.

Clas, Torbjörn, Marie, Gunnar och Johan (Gottfries) som peppade mig att bli doktorand inom metabolomics. Stina Saitton, Lotta Holm och Tomas Gustavsson utan vilkas hjälp jag aldrig blivit en kemist som kunde hantera saker som brinner i luft (frågan är om jag kan det fortfarande, reagensen vi använder inom metabolomics är rätt så harmlösa i jämförelse).

David, Idol-Åsa, Maria, Stina, Kricke, Fredrik och Kung-Emil för att jag tog mig igenom kemistprogrammets första absolut dödstråkiga del (analys, kvantmekanik (parentes i parentesen, nej ser fortfarande inte nyttan av det) och termodynamik). Jag syftar dock inte på era eminenta studiecirklar under kursen i envariablel-analys I.

Simon för att inget är för dumt för att diskuteras, Lottis för att trots att du är långt borta alltid kommer ihåg mig. Värdens bästa särbo, favorit-Fredrik som fick mig att lägga ner tv-tittande för flera år sedan till fördel för ”The Real World Fredde”. Helena och Camilla, det är svårt att hitta tjejer som är lika bra som ni och som dessutom gillar att pyssla, äta vegmat och lyssna på hardcore en lördagkväll. Elvispens upphovskvinna Maria och sötaste Theia för många trevliga party-of-two, välplanerade upptåg (planering är din grej), sällskap i tvättstugan (tvätta kan visst vara skoj), middagar bestående av pizza med extra smältost samt titt på stjänklar himmel. Krille-krokodil, Arvidsjaur-Emma, Kiruna-Lisa med familj, Svenkan med familj, Trosa-Erik, Piddan, Kung-Uken, PK (läs vidare du orkar), Chrillaz (Kick-hisass-Seabass), sötaste Tessan, Aron (vems tur är det att ringa och be om ursäkt?), O my O, Ödlan (skejt, skejt, skejt, nu, nu, nu), Stor-Robban, turk-Ciss-it, Stefan (mäster-kock), Robban (Xrob), Davve, D2 (high five för att jobba dygnet runt), Janne (min fot är läkt, snowboard ist för longboard kanske?), Vakt-Lars, Ronja, Camilla, Holma (du är världens sötaste damp-unge), Elin, Lina, Dave Lee, Berra, Isbjörn, Dove, Karin, Martin och alla ni andra som gör mitt liv så mycket roligare!

Pyret my partner in crime, för att du är precis som du är och har varit min vän i vått och torrt så länge jag kan minnas. MacGyverlösningar, sista-minuten entréer, livsnjuteri och tjejmiddagar som är svåra att beskriva är våran grej. Ingen känner mig som du och ingen är som du. Jag ser fram emot kommande tjejhäng i utökad skara.

Min släkt och extrasläkt som är för många för att nämnas vid namn: tack för all uppmuntran och stöd. Utan dina historier farfar hade jag helt lagt ner ämnet linjär algebra för väldigt (väldigt, väldigt) länge sedan.

Min helt otroligt fantastiska familj. Utan dig mamma hade jag svultit ihjäl många gånger och utan dig pappa hade jag bott i en u-landslägenhet och förmodligen fått gå till jobbet mest jämt (nya cykeln funkar än). Tack även till min käre lillebror PK för allt från middagssällskap till att jag får låna dina muskler (eftersom jag inte har några).

Sist men inte minst vill jag tacka Marie-Louise Rönnmark som är en oändlig glädjespridare och förebild när det kommer till att visa att man kan hinna med fler saker på en dag än vad många tror är möjligt på en livstid.

9. Populärvetenskaplig sammanfattning

”Det finns alltid en tredje utväg, de gäller bara att hitta den.” -Selma Lagerlöf

I denna avhandling har vi mätt nivåer av metaboliter (små kemiska molekyler) i cerebrospinalvätska (den vätskan som omger hjärnan). Syftet har varit att hitta en eller flera metaboliter som får förändrad koncentration då man drabbas av den dödliga sjukdomen amyotrofisk lateralskleros (ALS). Varje år drabbas runt 300 personer i Sverige av ALS. Ett av de mer kända fallen av sjukdomen var Rapports nyhetsankare Ulla-Carin Lindquist som fick diagnosen ALS våren 2003.

I ALS dör de nervceller i hjärnan och ryggmärgen som styr musklerna (de sk. alfa-motorneuronen). När nervcellerna dör avtar signalerna till musklerna vilket resulterar i att musklerna förtvinar. En fortlöpande försämring av muskelstyrka är ett av symptomen på ALS, men sjukdomssymptom har setts variera mycket från fall till fall. Det som är gemensamt är att sjukdomen alltid leder till döden, oftast inom loppet av 3 år. I dagsläget finns det ingen effektiv diagnosmetod eller behandling av ALS. Rilutek® är hittills det enda läkemedlet som har visats ha en viss bromsande effekt på sjukdomsförloppet. Studier pekar mot att ju tidigare Rilutek® sätts in, desto bättre effekt har läkemedlet för att öka livslängden. Det finns dock inget säkert test för att ställa diagnosen ALS. En rad liknande sjukdomar måste därför uteslutas innan diagnosen ALS kan ges. Detta medför oftast långa utredningar och risken finns att fel diagnos ställs om någon sjukdom med liknande symptom missas under utredningen. Det är i dagsläget därför av hög prioritet att hitta och utveckla en bättre, säkrare och snabbare diagnosmetod för ALS.

Arbetet som denna avhandling bygger på har främst syftat till att leta efter kemiska markörer som på sikt ska kunna utvecklas till en diagnosmetod för ALS. Genom att mäta många metaboliter är förhoppningen att öka chanserna att hitta en eller flera metaboliter som kan påvisa ALS. Genom att tolka mönster av förändrade metaboliter kan förhoppningsvis även en ökad förståelse fås för vad som händer i kroppen om man drabbas av ALS, dvs. få ledtrådar om vad som orsakar sjukdomen.

Med dagens analytiska tekniker kan man mäta hundratals lågmolekylära metaboliter samtidigt i ett prov. Detta genererar enorma mängder data som måste omvandlas till tolkningsbar information. Kemometri är ett koncept som går ut på att designa och strukturera forskningsförsök så att informationen lättare kan tolkas. Kemometri använder sig av multivariata analysmetoder som kan hantera de datatyper som genereras när många metaboliter mäts samtidigt. Dessa metoder gör det möjligt att överblicka informationen i datat, undersöka vilka individer som är lika repektive olika varandra. Metoderna gör att det även är möjligt att hitta metaboliter som är förändrade mellan sjuka och friska patientgrupper.

I denna avhandling har kemometriska metoder används för att söka efter metaboliter som skiljer sig åt för olika typer av ALS i relation till andra sjukdomar. Förändrade metabolitmönster hittades i cerebrospinalvätska för patienter med ALS i jämförelse mot kontroller. En trend som sågs var att ALS patienter hade lägre halter av många metaboliter jämfört med kontrollgruppen. Glutamat (aminosyra som agerar signalmolekyl i hjärnan) var en av de metaboliter som sågs systematiskt minskad hos ALS patienterna. Halter av glutamat har tidigare rapporterats både som ökade och oförändrade hos ALS patienter. En av andledningarna till dessa tvetydiga resultat har vi i en studie visat kan bero på ostabilitet under förvaring av prover.

En viss grupp av ALS patienter har genförändingar (mutationer) i en gen som kodar för ett enzym (SOD1) i kroppen. Vi har sett skillnader i metabolitmönster i cerebrospinalvätskan från ALS patienter med olika typer av mutationer i denna gen (hittills känner man till 151 olika mutationer i SOD1). ALS är en sjukdom som kan ge många olika typer av förlopp (snabbare eller långsammare) och symptom (beroende på vilka motor neuron som är mest drabbade av sjukdomen). Det är därför intressant att undersöka om patienter med ALS uppvisar likheter eller skillnader i cerebrospinalvätskans metabolit mönster. Tidigare forskningsresultat har rapporterats då man sett att patienter med mutationer i SOD1 skulle kunna vara en sub-typ av ALS, då speciellt en av de funna mutationerna ska kunna vara annorlunda.

Genom att förstå sjukdomen bättre och kunna klassificiera eventuella fall i subgrupper skulle forskningen kring ALS kunna göras mer inriktad. Det skulle kunna vara en av nyklarna till att hitta fungerande diagnostiska verktyg samt i längden kunna leda till att man kan hitta verksamma läkemedel. Det kan vara så att det inte är en diagnostisk markör eller signatur (dvs. flera metaboliter i mönster) vi letar efter utan flera olika (eftersom vi fortfarande inte vet om ALS är en eller flera sjukdomar). Vi vet heller inte om det kommer behövas olika läkemedel för olika typer av ALS (eftersom vi inte vet om det är en eller flera mekanismer som orsakar ALS). Genom detta arbete har vi förhoppningsvis kommit lite närmare en lösning till gåtan om sjukdomen ALS.