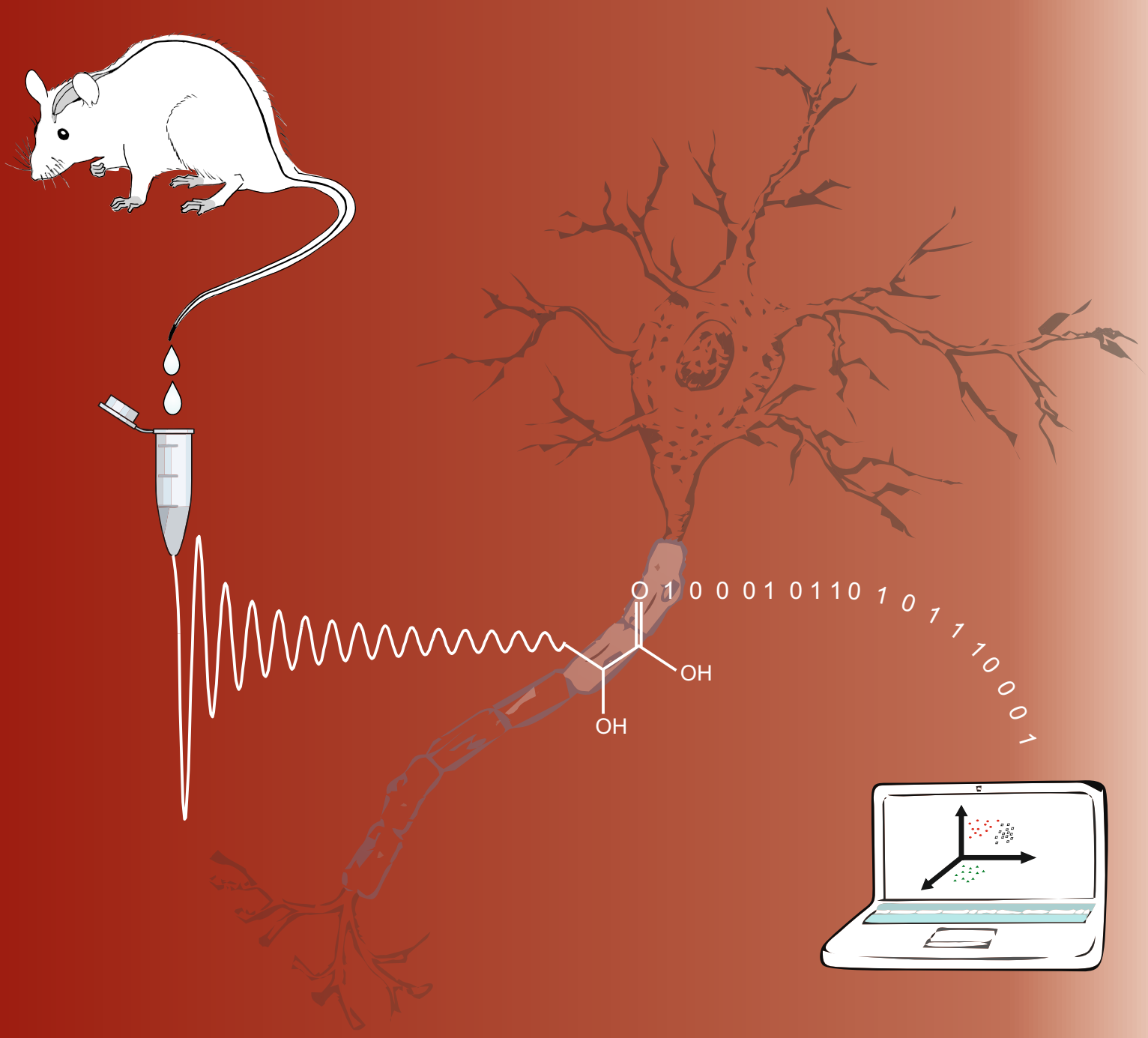
Chemometrics and NMR spectroscopy for metabolomics analysis of neurological disorders



Agnieszka Smolinska

CHEMOMETRICS AND NMR SPECTROSCOPY FOR METABOLOMICS ANALYSIS OF NEUROLOGICAL DISORDERS

Agnieszka Smolińska

ISBN: 978-94-6191-358-6

Chemometrics and NMR spectroscopy for metabolomics analysis of neurological disorders

Proefschrift

ter verkrijging van de graad van doctor
aan Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann, volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 28 augustus 2012
om 15.30 uur precies

door

Agnieszka Smolińska
geboren op 14 november 1983
te Siemianowice Sl., Polen

Promotoren: Prof. dr. L.M.C. Buydens

Prof. dr. S.S. Wijmenga

Manuscriptcommissie:

Prof. dr. A.P.M. Kentgens

Prof. dr. hab. B. Walczak (*Silesian University, Poland*)

Prof. dr. J.P.M. van Duynhoven (*Wageningen University*, *Netherlands, Unilever Vlaardingen*)

Table of contents

OUTLINE OF THE THESIS

This thesis deals with the metabolome screening of blood plasma and cerebrospinal fluid (CSF) by Nuclear Magnetic Resonance (NMR). The main focus is the biomarker discovery in Multiple Sclerosis (MScl) disease by means of NMR and pattern recognition methods. Moreover the disentanglement of CSF samples handling for metabolic measurements by NMR and variations in metabolites concentration identified by NMR in CSF of "healthy" individuals are presented. This thesis consists of 7 chapters.

In the introductory **Chapter 1**, various aspects of metabolome screening of biofluids by Nuclear Magnetic Resonance (NMR) and chemometric analysis are reviewed. Specifically, attention is paid to data acquisition, data pretreatment and preprocessing and multivariate analysis of metabolic data. The chapter gives also an overview of current developments in and status of biomarker discovery by means of NMR and pattern recognition techniques with emphasis on CSF biomarker research for MScl disease.

In **chapter 2** CSF sample handling procedures are presented, including a stability study of metabolites identified by NMR in human CSF. To mimic the probable procedure occurring in the clinic, CSF was left at room temperature for up to 120 minutes before freezing and storing it in -80^{o}C. The changes of metabolites concentration were then investigated.

Chapter 3 covers the analysis of CSF samples of "healthy" human individuals, i.e. without neurological disease. In this part, the biological variation of metabolites measurable by NMR and other analytical methods is shown. These biological variations in "healthy" individuals form a base line for detecting significant fluctuations in disease inflicted individuals. Inter-individual fluctuations in metabolite concentrations in the "healthy" individuals are compared with analytical variations and found to be much

smaller. In addition, the variation in CSF metabolites abundances in gender and group's age is demonstrated.

In **Chapter 4** a metabolic biomarker study on CSF from pre-clinical animal model of MScl, namely Experimental Autoimmune Encephalomyelitis (EAE) is investigated. The aim was to establish the metabolic profile of CSF of EAE-affected rats (representing neuroinflammation, a crucial part of early stage of MScl) and to detect the metabolic markers related to neuroinflammation. In order to find disease specific markers the NMR spectra were analyzed with two different chemometric techniques.

Chapter 5 portrays the analysis of blood plasma and CSF of EAE-affected rats. In this study both biofluids were measured by NMR spectroscopy. The obtained NMR spectra were subsequently joined by means of a mid-level data fusion scheme. Moreover, in this chapter a new approach for multi-class classification is presented.

Chapter 6 brings up a novel approach for data fusion, which is then applied to metabolomic CSF datasets from patients suffering from MScl disease. In this study two different analytical platforms, namely NMR and Gas Chromatography-Mass Spectrometry, were used to generate metabolic profiles of CSF. The approach shown in this study consists of concatenating NMR and GC-MS data in kernel space.

Lastly, **chapter 7** briefly summarizes the findings of the research described in this thesis, followed by some perspectives on future research.

CHAPTER 1

NMR AND PATTERN RECOGNITION METHODS IN METABOLOMICS. FROM DATA ACQUISITION TO BIOMARKER DISCOVERY.

A. Smolinska, L. Blanchet, L. M.C. Buydens and S. S.Wijmenga

Analytica Chimica Acta (2012), in pres, (dx.doi.org/10.1016/j.aca.2012.05.049)

ABSTRACT

Metabolomics is the discipline where endogenous and exogenous metabolites are assessed, identified and quantified in different biological samples. Metabolites are crucial components of biological system and highly informative about its functional state, due to their closeness to functional endpoints and to the organism's phenotypes. Nuclear Magnetic Resonance (NMR) spectroscopy, next to Mass Spectrometry, is one of the main metabolomics analytical platforms. The technological developments in the field of NMR spectroscopy have enabled the identification and quantitative measurement of the many metabolites in a single sample of biofluids in a non-targeted and non-destructive manner. Combination of NMR spectra of biofluids and pattern recognition methods has driven forward the application of metabolomics in the field of biomarker discovery. The importance of metabolomics in diagnostics, e.g. in identifying biomarkers or defining pathological status, has been growing exponentially as evidenced by the number of published papers. In this review, we describe the developments in data acquisition and multivariate analysis of NMR-based metabolomics data, with particular emphasis on the metabolomics of Cerebrospinal Fluid and biomarker discovery in Multiple Sclerosis.

1.1 INTRODUCTION

The terms metabonomics [1] and metabolomics [2] appeared at the end of the 90's and early 2000's, respectively [1, 3]. They describe, in broad terms, the study of the metabolome, which was first defined as the collective set of metabolites produced or present in a biosystem [3, 4]. Nowadays, metabonomics and metabolomics are often used interchangeably [4, 5], although their exact definitions are slightly different. The most often cited definition of metabonomics is the one proposed in 1999 in Xenobiotica [1]: 'Metabonomics is defined as the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification'. For Metabolomics a very similar definition is often used [4]: 'The study of the quantitative complement of metabolites in a biological system and changes in metabolite concentrations or fluxes related to genetic or environmental perturbations. Studies are typically holistic in nature though targeted studies are also encompassed in the term metabolomics'. Since, metabolomics and metabonomics terms are in practice often utilized indifferently the analytical and modeling procedures are the same and therefore in the rest of the paper we will employ the term metabolomics.

Metabolomics is a strongly developing field as evident from the exponentially growing number of papers. With a doubling time of circa three and half years, approximately 1420 papers were published in 2011 to which NMR and Mass Spectrometry equally contribute about one-third (as defined in Web of Science with keywords [metabolom* or metabonom*]. For comparison, the field of proteomics is about twice as large with about 3600 published papers in 2011. Metabolomics' approaches have developed in many areas of biomedical research, such as toxicology studies, nutritional effects, metabolic consequences of genetic modifications, inborn errors of metabolism, diabetes, cancer diagnostics, and diagnosing of neurological diseases [6-14]. The metabolic profiling was first reported in the literature in 1950 but the first progresses were slow until it became a separate scientific area[15]. To date there has been increasing emphasis on obtaining spectral "fingerprints" or metabolic profiles that can be correlated with phenotype [2, 16]. Metabolomics is a reflection of genetic factors and metabolites are often defined as the functional endpoint. The closeness of the metabolism to an organism's phenotypes

causes that it will be affected by disease and thus it is relevant to measure metabolites. Moreover, the flux of metabolites is measured in seconds in comparison to turnover in proteome which is measured in minutes to hours [17]. This shorter response time allows one to use the metabolomics as indicator of environmental perturbations. This is one of the reason why van der Greef et al. described metabolomics as a promising tool for clinical systems biology to detect early metabolic perturbations, even before the appearance of disease symptoms [18]. A single metabolite can be a substrate for a number of different enzymes, causing linkage of metabolites through complex pathways. This linkage makes difficult to assess the consequence of changes in mRNA products and proteins, but at the same time metabolites and their concentrations may report on changes in both mRNA and proteins [7, 19]. This is another reason why studying disease from the metabolic point of view is very attractive. Moreover, only 2766 metabolites (i.e. small molecules, <1500Da) are estimated to be derived from men [7] and many metabolites are species independent. Therefore, they could form the basis of translational studies, i.e. biomarkers that are found in preclinical studies and can be applied more directly during clinical studies. Metabolomics can be used either as a targeted or a non-targeted analysis of endogenous and exogenous metabolites for biomarker discovery [20]. According to the official National Institutes of Health, a biomarker is defined as 'characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention' [21]. The Food and Drug Administration defines a valid biomarker as: 'A biomarker that is measured in an analytical test system with well established performance characteristics and for which there is an established scientific framework or body of evidence that elucidates the physiologic, toxicologic, pharmacologic, or clinical significance of the test results [22]. Biomarker is defined as a laboratory measurement that is an indicator of diseased processes and also the risk of the appearance. For instance, magnetic resonance imaging measures in Alzheimer's disease or in Multiple Sclerosis (MScl) and positron emission tomographic scanning of dopamine transporters in Parkinson's disease are such markers [23]. Obviously, metabolites or metabolic profiles can be utilized as biomarkers [4, 24]. Indeed, metabolic biomarkers provide, because of their objective nature, an attractive and valuable tool for

accurate diagnose of diseases. Metabolic profiles can be particularly valuable to determine when a healthy state becomes dysfunctional in the early stage of the disease and provide new possibilities for preventing therapies.

Metabolomics' approaches have spread in many areas of biomedical research and therefore increasing number metabolic biomarkers are established [4, 8, 25]. The identification of biochemical biomarkers in biofluids for Central Nervous System (CNS) disorders has been the aim of many metabolomics studies [11, 25-33]. Presently, the clinical diagnostic of most neurological diseases is performed based on the identification of a variety of symptoms. However, it is very often difficult to identify individuals at risk or establish rapidly a definite diagnosis. This is mostly due to the complexity of CNS diseases and to the fact that the ethiopathogenesis of most such diseases is very often unclear. CNS diseases are likely to arise from deregulations of several genes, leading to complex alterations in protein and/or metabolite profiles, but also changes in environmental conditions may affect these profiles. All these aspects reflect the complexity of the CNS malfunctions[34]. Therefore, there is an increasing need to learn more about CNS diseases at the molecular systems level, to understand at this level the changes that can contribute to pathogenesis of these disorders.

In metabolic biomarker discovery, an important issue is how to extract relevant information from the data, i.e. from the metabolic profiles of biofluids. All data produced in metabolomics are highly multivariate. Therefore the use of chemometrics is required to find trends or significant information in the data, i.e. relevant metabolites. In chemometrics different multivariate methods for data exploration, visualization, classification and prediction are available. The empirical models constructed on experimental data can subsequently be used for biological interpretation and prediction.

Nowadays, proton Nuclear Magnetic Resonance (^{1}H-NMR), Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS) are well-established powerful analytical methods for generating metabolomics profiles. For the analysis of complex, biological samples like biofluids, these techniques have their advantages and disadvantages. For instance, GC-MS requires derivatization, which lengthens the sample preparation time. In general LC-MS and GC-MS need more time consuming sample preparation. On the other hand, GC-MS and LC-MS yield a higher

sensitivity than NMR and therefore may detect metabolites that are present in a concentration below the detection limit of ^{1}H-NMR. On the other hand ^{1}H-NMR requires limited sample preparation, is un-targeted, quantitative (absolute), non-destructive, reproducible and unbiased [35]. ^{1}H-NMR may detect compounds that are too volatile for GC, while metabolites without proton (phosphoric acid) are not detected by ^{1}H-NMR.

We focus in this review on the recent developments in metabolic profiling of Cerebrospinal Fluid (CSF) by ^{1}H NMR (and to some extent also blood and urine) aimed at identification of biochemical biomarkers for CNS disorders, in particular MScl. Our choice of NMR as analytical method was guided by the fact that it is a robust and reliable technique for metabolomics application and it allows for detecting a wide range of different types of metabolites simultaneously [36]. Plasma and urine are historically the two biofluids widely used and described for metabolomics application [37-41]. Many protocols for sampling and measuring of these biofluids can be found [36, 41-43]. Metabolic profiling of CSF by means of NMR has been performed to find diagnostic biomarkers for a number of neurological diseases [17, 32, 33, 44, 45]. However, only recently comprehensive protocols for metabolic (and proteomic) profiling of CSF have been established [46-48].

In this review we present comprehensive outlines of data acquisition, data preprocessing and data analysis of NMR-based metabolomics data for CSF and discuss its application to CSF biomarker discovery for MScl.

This review covers four main aspects, namely:

- analytical, i.e. samples preparation and measurements (section 2)
- data preprocessing (section 3)
- data analysis, statistical analysis (section 4)
- biomarker discovery by means of NMR and pattern recognition with focus on CSF and MScl (section 5 and 6)

The first part of the review focuses on recent developments in preparing and measuring the metabolic profiles of different biofluids, namely blood plasma, urine and CSF via NMR. In this part, the recently established protocol for measurement of the metabolomics profiles of CSF is described and discussed. In the next part, the crucial steps involved in data preprocessing are described. The third part details the multivariate data analysis. The most common pattern recognition methods used in

metabolomics are discussed. Also, current progress in data fusion is described. In the fourth part, the determination of absolute metabolite concentration and their variations in CSF of healthy controls is brought up. In this fourth part, the review centers further on providing an overview of recent developments in the application of NMR and pattern recognition methods in metabolic biomarker discovery for MScl.

1.2 NMR of BIOFLUIDS

High-resolution NMR spectroscopy is a quantitative and non-destructive technique. It is a robust and reliable analytical method with paramount reproducibility and repeatability [35]. In metabolomics studies either biofluids, cell or tissues extracts are used as main samples for metabolic fingerprints. Biofluids like urine, blood plasma or serum, CSF are the most common investigated samples in metabolomics studies. Most of these biofluids can be obtained quite easily with minimal invasion. Moreover, a high sampling frequency can be achieved [42]. In the late 1960s the developments of Fourier transform NMR spectroscopy and next in the 1970s the implementation of superconducting magnets permitted the beginning of the application of NMR spectroscopy for the metabolite profiling of biofluids. The first real applications of NMR to the analysis of biofluids, dates to early 1980s [49-51]. Further NMR technical improvements in the 1990s, namely stronger magnetic fields and introduction of cryo-cooled NMR probes, have led to an enormous boost in NMR sensitivity; the signal to noise ratio of ethylbenzene was circa 800:1 at the then highest fields of 600 MHz versus circa 8000:1 for 800 MHz nowadays. Today, the detection limit of metabolite concentration is of the order of μM. Although, the sensitivity has increased enormously and still improves, it remains a weak point compared to MS. Recent developments in spin hyperpolarisation via dynamic-nuclear polarization (DNP) or para-hydrogen-induced hyperpolarisation (PHIP) hold great promises in resolving this backlog [52, 53]. Furthermore, the increase in field strength has tremendously improved the resolution. Today, metabolomics is an exponentially growing field in which NMR together with MS each contribute about equally. A very important benefit of NMR spectroscopy for metabolic profiling is that it is quantitative and does not require time-consuming sample preparation steps, like separation or derivatization. Moreover, it does not require a-prior knowledge about compounds present in a sample and is thus ideally suited for non-targeted profiling.

1.2.1 Sampling and NMR sample preparation

Urine, blood plasma and blood serum are the most commonly used biofluids in metabolomics studies, because they contain hundreds to thousands metabolites and they both can be obtained in relatively non-invasive manner [36, 54-59]. Another biofluid, CSF, is widely used in metabolomics studies of neurological disease [28, 32, 33]. In contrast to plasma and urine its sampling is much more invasive.[31] A number of other biofluids like amniotic fluid [60, 61], bile [62, 63], seminal fluid [64-66], saliva [67-69] have been also investigated.

Biological samples should be collected under strict conditions. Usually blood is collected by venipuncture into standard vials containing either ethylene diamine tetra acetate (EDTA) or lithium heparin as anti-coagulant. One has to remember that when EDTA is used extra resonances can be observed in the NMR spectrum. This is due to formation of complexes between EDTA and ions Ca^{2+} and Mg^{2+} which are present in plasma [51].

Plasma and serum can be measured directly with minimal sample preparation. Dilution of plasma or serum is recommended, since it reduces the sample's viscosity and releases plasma-protein bound metabolites

Proteins can be removed before NMR measurements by either organic solvent precipitation or by ultra-filtration using a 10kDa molecular weight cut off filter [70]. However, a recent comparison of organic solvent precipitation with ultrafiltration has demonstrated that ultra-filtration was superior for metabolic NMR measurement [71, 72]. Before usage, the filter has to be cleaned from glycerol, which is present in many commercially available filters, by centrifugating it with water. The effect of protein removal on the ^{1}H-NMR spectrum of blood plasma is shown in Figure 1. For urine samples addition of sodium azide is required to control bacterial growth. Detailed procedures to collect, store and measure biofluids such as blood or serum, urine have been provided in the literature as guidance [36, 42].

The pH of samples has a significant influence on the chemical shifts observed in the NMR spectrum. Therefore, it is important to control the pH of the biofluid sample. In the literature pH 7.2 and 2.5 have been mostly used [33, 73, 74]. A very popular and fast method to adjust pH is addition of a phosphate buffer stock solution made up in D_2O at pH 7.0 or 7.4 [36]. Manual adjustment of pH using NaOH and HCl is another possibility [75, 76]. In

section 1.2.3, we present a detailed protocol for the metaboloic sampling and profiling of CSF by NMR.

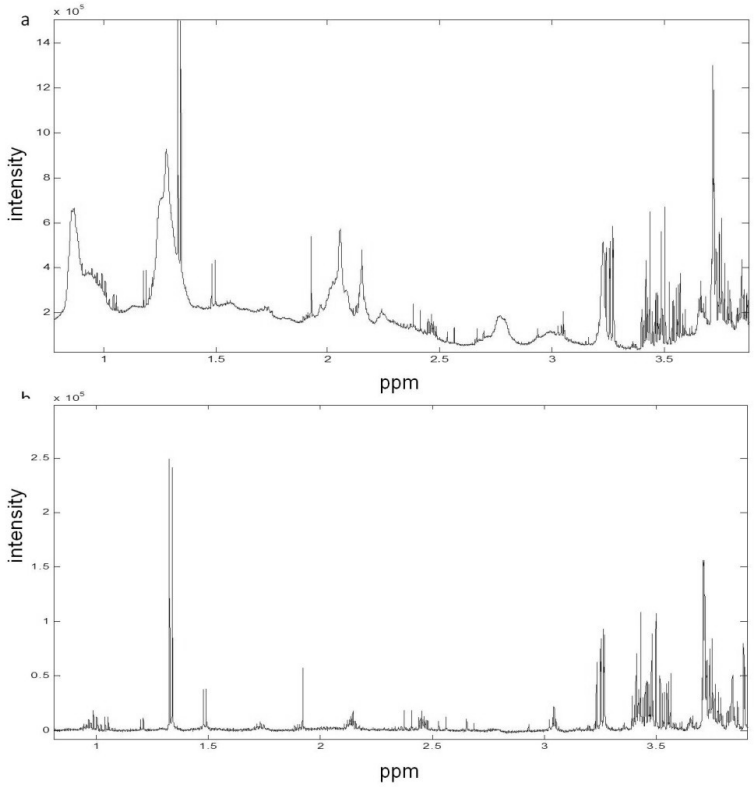


Figure 1. The 500 MHz ^{1}H-NMR spectrum of blood plasma sample: (a) before, and (b) after protein removal.

A very important advantage of NMR is absolute quantification[77]. The linear response of NMR experiment is the key benefit of NMR over other analytical methods. The signal intensities observed in NMR spectrum are directly proportional to the concentration (i.e. molar amount) of that nucleus in the sample [77, 78]. In order to obtain absolute concentration of metabolites detected by NMR, usually the addition of internal standards is used. The reference compound used for concentration reference as well as for chemical shift (δ=0.00) is usually the sodium salt of 3-trimethylsilylpropionic acid-d_4 (TSP-d_4) with deuterated methylene groups. Other references standards are 2,2-

dimethyl-2-silapentane-5-sulfonate sodium salt (DSS) or for organic solvent trimethylsilane (TMS). Another internal reference standard, 4,4-dimethyl-4-silapentane-1-ammonium trifluoroacetate has been proposed by Alum et al. as promising universal for metabolic profiling [79]. Akoka et al. have introduced a synthetic electronic reference signal, Electronic REference To access *In vivo* Concentrations (ERETIC), for quantification purpose [80]. Similar approach, QUANTification by Artificial Signal, has been recently introduced by Farrant et al.[81]. In these methods an internal standard is claimed to be not required. The standard way of quantifying compounds in NMR spectrum is integrating the different NMR resonances and comparing them to the area of the internal standard. Another possibility is spectral deconvolution, available for instance in MestReNova [82], where global spectral deconvolution is applied automatically to whole NMR spectrum. In the Chenomx NMR Suite software the spectral signatures (singlets, doublets, triplets etc.), *i.e*. the peak shapes, of a compound from an internal database of reference spectra is fitted to the experimental NMR spectrum [83]. In contrast, peak integration is very sensitive to baseline distortions. Moreover even slightly overlapping resonances cannot be reliably quantified. Peak-shape fitting, like in Chenomx, is not affected by baseline distortions and still efficient when some of the resonances and/or part of a resonance overlap with that of another compound, the peak shape can still be fitted with reasonable accuracy and the concentration of the compound reliably determined.

Usually, 5-mm diameter NMR tubes are used for NMR measurements. There are also much smaller NMR tubes with diameter of 1-3 mm [84] or a microscale SHIGEMI tube[33] in which a reduced sample size is compensated by solid glass beneath the level of sample. Recently NMR on small biofluid sample size (600nL) was investigated, where microfluidic stripline resonator was used to measure ^{1}H-NMR spectrum of human CSF [85].

The number of observed metabolites in biofluids largely depends on the magnetic field strength of NMR spectrometer. Therefore, working at the highest available magnetic field is recommended. Generally, 500 or 600 MHz NMR instruments are used in metabolomics studies, because these fields are easily accessed [16, 32, 67]. However, the use of 800 or 900 MHz has been reported [33, 47].

1.2.2 Measurements and spectral pre-treatment

The detection limit for bodyfluid NMR spectroscopy depends on many factors such as field strength, number of protons contributing to a resonance and a region of the spectrum where the resonance is observed. In general the detection limit is in the low micromolar range in the less crowded regions of the spectrum [86]. There are two experimental issues in NMR of biofluids. The first one is connected to accurate solvent suppression. Even though very effective methods like excitation sculpting [87] or WATERGATE [88] exist, the simple presaturation is the prevalent one. The second experimental issue is associated with distinction between small molecular weight metabolites (typically <1500Da) and macromolecules. Macromolecules produce broad resonances due to limited rotational diffusion and short T_2 relaxation times, causing difficulties in spectral interpretation. To overcome these problems, 1D nuclear Overhauser effect spectroscopy with presaturation (1D NOESY-presat) [89, 90] and the 1D Carr-Purcell-Meiboom-Gill (CPMG) [91] are two pulse sequences used for metabolic profiling. 1D NOESY has become the most popular sequence for NMR-based metabolic analysis. This is mostly due to high quality of water suppression with little calibration and consistency in obtained spectra. CPMG as special pulse sequence is used to remove broad proteins signals if they have not been taken away before NMR measurement.

In metabolomics studies of biofluids beside 1D ^{1}H-NMR homonuclear 2D J-resolved is also very often used [92]. Moreover 2D J-resolved spectra increase identification of biochemical substances. Other 2D NMR spectroscopy, such as correlation spectroscopy (COSY) [93] and total correlation spectroscopy (TOCSY) experiments [94] give spin-spin coupling connectivities. They provide information on which hydrogens in a molecule are close in chemical bond term. 2D NMR spectroscopy is mostly performed for better signals assignment.

The NMR data are typically processed by Fourier Transformation (FT). Before FT apodization and zero filling of Free Induction Decay (FID) is performed. The phase is then corrected to obtain absorption line shape. For spectral pre-treatment several commercial and free licensed software packages are available, such as PERCH [95], , Chenomx NMR Suite [83], MestReNova [82] and the algorithm AutoFit [96]. They all provide a

number of functions involving spectral pre-treatment, metabolite identifications and quantification.

1.2.3 Metabolic protocol of CSF

CSF is a colourless and crystal-clear fluid that surrounds the brain and spinal cord and thus protecting them from immunological and mechanical damage. The main composition of CSF is water (around 99%), proteins, nutrients needed for metabolism and electrolytes. CSF is also responsible for removing waste from surrounding tissues. CSF is constantly produced at rate of circa 500ml per day and its turnover is around 4 times per day [28].

CSF is normally sampled via lumbar puncture between the L4 and L5 vertebrae. CSF, directly after collection, should be centrifuged to remove all cellular elements. It has been shown that several metabolites concentrations were affected in porcine CSF when residual white blood cells were present [97]. Therefore, it is crucial to not only assess that CSF is free of erythrocytes but also white blood cells. After sampling CSF samples should be stored in -80^{o}C. This storing temperature assures that the metabolic composition is not affected. It has been shown that storing CSF in -20^{o}C causes activation of a number of chemical processes, which influence the metabolite concentrations [98]. Another contradictory study has demonstrated insignificant quantitative changes in human CSF after freezing in -20^{o}C and subsequent thawing [48]. It has further been demonstrated that keeping CSF samples at room temperature for a limited (hour) does not affect metabolite composition in human CSF to any great extent. Analysis of the impact of a delay in storage of CSF has revealed that out of 93 unique identified metabolites, only the concentration of erythronic acid has been elevated significantly with increased time left at room temperature [46]. Another significant aspect is connected to pH of CSF, which is naturally poorly buffered. The CSF pH rises drastically upon standing and storing in -20^{o}C [98, 99]. This rapid increase may be explained by evaporization of CO_2.

CSF in comparison to blood plasma has a relatively low protein concentration (<500mg/l). Therefore prior to NMR analysis they are not always removed. In the literature, NMR measurements on both native [29, 100-102] and deproteinized samples have

been reported [47, 73]. In our experience, the presence of protein does not affect the measured metabolite concentration to a large extent. If sufficient volume of CSF is available, the buffer solution solvated in a mixture of water and D_2O consists of TSP, mM sodium azide (NaN_3) and mM sodium phosphate dibasic dehydrate ($Na_2HPO_4 \cdot 2H_2O$) are added prior NMR measurements. If CSF sample is deproteinized and snap-frozen it is first reconstituted in D_2O and next buffer solution with TSP and NaN_3 are added. The amount of CSF that can be collected in one sampling is often not sufficient for measuring it with standard 5 mm NMR-tube. For instance, rat's brain is rather small and only a small volume of CSF can be collected (~100 uL). In that situation CSF samples can be diluted and measured in SHIGEMI tube[33] or in 1 mm NMR-tube [103]. One should keep in mind that using small volume tube does not mean an increase in sensitivity of low abundant metabolites.

An up-to-date protocol for ^1H NMR metabolomic CSF sampling and sample preparation is available in Rosenling et al. [46].

1.3 DATA PREPROCESSING

Data preprocessing is an intermediate step between raw spectra and data analysis. The main objective of data preprocessing is to transform the data in such way that the samples in the dataset are more comparable in order to ease and improve the data analysis. The crucial role of data preprocessing was pointed and discussed in many publications[104-107]. For NMR spectra preprocessing usually involves, baseline correction, alignment, binning, normalization and scaling.

1.3.1 Baseline correction

Usually, the first step of data preprocessing is the baseline removal. Baseline distortions affect not only the statistical analysis but also the quantification of the metabolites. These distortions can be corrected in many different ways; usually an automated baseline correction is applied. Currently the most popular methods are based on polynomial-fitting such as iterative polynomial fitting [108], robust estimation procedure

implemented in Chenomx NMR Suite [109], locally weighted scatterplot smoothing (Lowess) fit [110]. Asymmetric Least Squares [111] is a method that uses a different constraint for baseline correction. It tries to estimate the baseline by fitting a regression curve to a spectrum using a penalized least square approach. Baseline can also be corrected by using B-splines, B-splines with Penalization (i.e. P-splines) [112] or by applying mixture models [113].

After baseline correction spectral regions not populated with by endogenous metabolites are often removed. Therefore, the spectrum outside the window 0.2-10.0 ppm is frequently removed. Another region of the spectrum that is usually excluded belongs to solvent water. Although suppression techniques are used for water, the remaining signal dominates in the spectrum. Moreover the variance contained in water signal is not of interest and might interfere with data analysis. In metabolomics of biofluids water signal dominates the spectrum between 4.7 ppm and 5.0 ppm. In case of urine spectra the signal of urea, which is very close to the water signal, is excluded. Urea beside water is the most concentrated metabolite in urine. Moreover it is not quantitative, because protons from urea exchange with water and other exchangeable protons and thus the peak intensity varies with quality of water suppression as well as with pH. In CSF and plasma spectra no additional exclusion regions are necessary.

1.3.2 Alignment

One of the most frustrating problems with NMR profiles, from a multivariate data analysis point of view, is the presence of peak shifts between different spectra [104]. These variations obscure the discovery of patterns in the spectra [114-116]. Shifts can be due to instrumental factors, changes of the pH and temperature, changes of salt concentration, overall dilution and relative concentration of specific ions. All these parameters influence peaks shifts, although not all peaks are affected to the same extend. Recently Cloarec et al. [117] and Giskeodegard et al. [118] have shown that peaks shifts can be beneficial for discriminating groups if the locations of peaks are systematically different in studied groups. However it is not yet possible to trace back these differences into a single factor. Therefore an essential step in preprocessing is to adjust the peaks shifts, i.e. alignment or warping. NMR spectra are usually first aligned by spectral referencing. This simple,

global method for peaks alignment sets the internal reference signal of each spectrum to 0 ppm. This type of alignment removes only global shifts and is not sufficient, because in NMR mostly local shifts are observable [119]. The effect of different alignment methods for NMR spectra on classification results is presented by Giskeodegard et al. [118].

Interval correlated shifting (icoshift) was proposed for NMR spectra [120]. In this warping method spectra are divided into different length segments and aligned to the corresponding segments of a reference spectrum. Warping is performed by shifting sideways the segments so as to maximize the correlation between segments. This is done by calculating the cross-correlation between the segments by a fast Fourier transform which allows for simultaneously aligning all spectra segments of spectra. Another segmented warping method is Correlation Optimized Warping (COW) [121, 122] which also divides spectra into segments but of equal size. By linear stretching and compressing it aligns the segments with the segments of a reference spectrum. The objective is to maximize the overall correlation between two spectra. COW was originally proposed for the alignment of chromatographic data, however it has been successfully applied to NMR spectra [46, 123]. Peak alignment by beam search was made for NMR signals [124]. This method also divides spectra into segments but warps them by both shifting and stretching/compressing to maximize their respective correlation. Wu et al. have proposed fast iterative warping algorithm, i.e. fuzzy warping, for urine NMR spectral data [125]. This warping procedure tries to establish correspondence between the most intense peaks in the spectra to be aligned.

Recently a method for aligning NMR spectra, called hierarchical Cluster-based Peak Alignment (CluPA), has been proposed by Vu et al. [126]. The algorithm builds hierarchical cluster tree from peaks of reference and target spectra. It aligns two spectra using this tree. Several others warping methods have been proposed to remove complex misalignments [127-129]. It is important to mention that warping may affect peaks area and thus quantification. Therefore absolute quantification should be performed on unaligned spectra. Figure 2 shows a set of 82 NMR spectra before and after baseline correction (via ALS) and alignment (via COW).

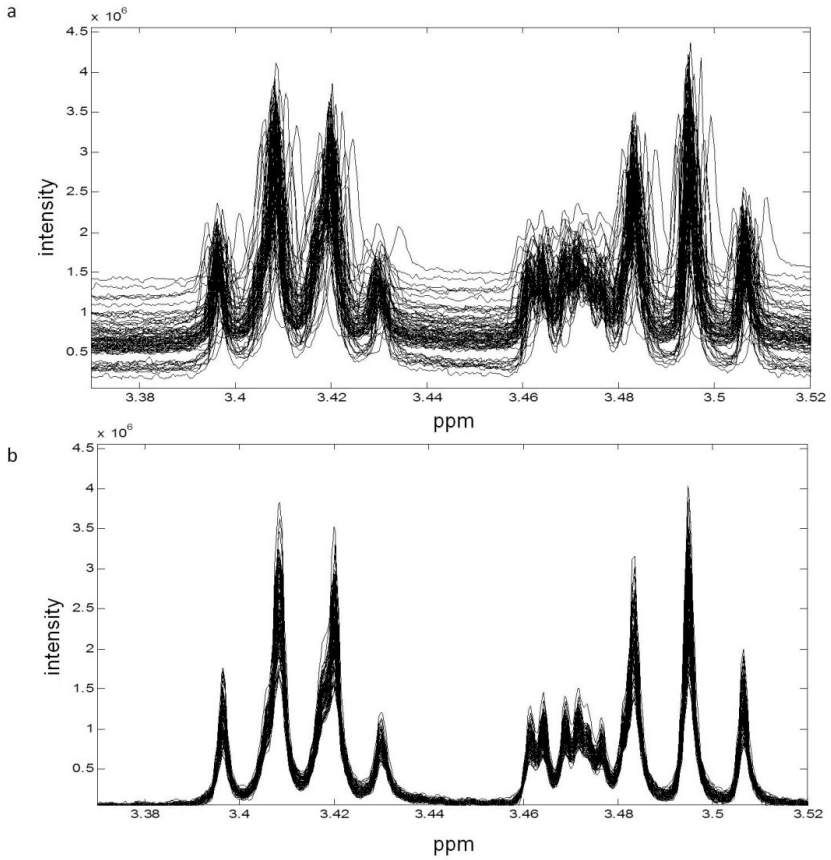


Figure 2. Example of NMR spectra: a) before baseline correction and alignment; b) after baseline and alignment.

1.3.3 Binning

For multivariate analysis of NMR metabolic data either quantitative or scored integrals of specific spectral peaks is used [130, 131]. An NMR spectrum, after excluding certain signals (e.g. water, internal reference, urea), contains approximately 22000 data point (variables). Therefore in order to reduce the data dimensionality binning (also called bucketing) is commonly used. One has to keep in mind that binning reduces the spectral resolution. In binning the spectra are divided into segments (so called bins or buckets) and the total area within each bin is calculated to represent the original spectrum.

Therefore some minor peaks shifts can be removed by spectral binning. There are many types of spectral binning. However the most common type of spectral binning is an equidistant binning of 0.04 ppm. This indicates that every spectrum is divided into evenly spaced integral regions with spectral width of 0.04 ppm [105, 106, 132]. The disadvantage of equal size binning is the lack of flexibility of the boundaries. If a peak is split between two bins, this peak frequency may significantly influence the data analysis. In order to prevent peaks being split by the boundaries of bins, methods which are based on non-equidistance spacing have been proposed, e.g. adaptive-intelligent binning (AI-binning) [133], Gaussian binning [134], adaptive binning using wavelet transform [135] and Dynamic adaptive binning[136]. These methods take into account peak positions and thus to focus on a better peak definition. Therefore it is possible to obtain binning where one bin covers only complete peaks. In Figure 3 an example of two types of spectral binning is presented, i.e. AI-binning and equidistance binning (0.04 ppm).

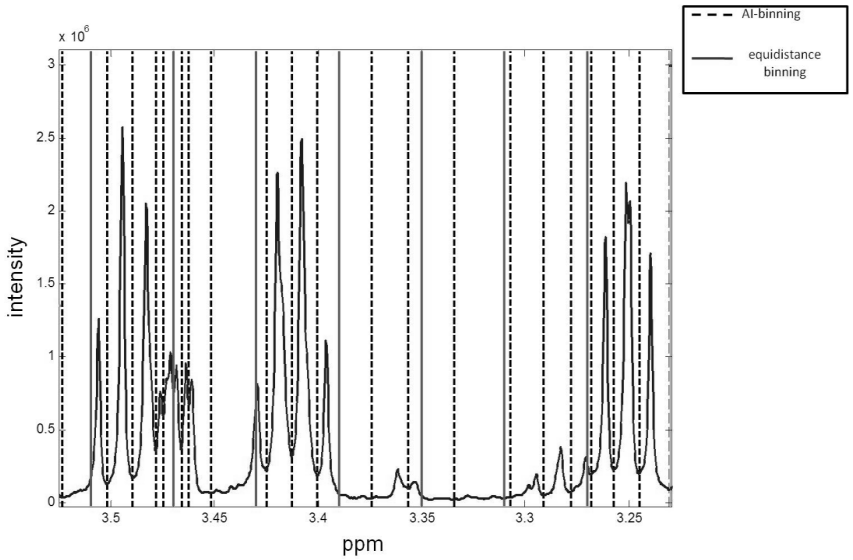


Figure 3. An example of AI-binning and equal sized binning performed on a fragment of NMR spectrum.

1.3.4 Normalization scaling and transformation

This step of preprocessing tries to account for variations of the overall concentrations of samples. The main aim is to make all the samples comparable with each other by for instance removing or minimizing total amount of material per samples or metabolite dilution. This is particularly relevant for urine studies where the variations of overall concentrations of samples are very distinctive and can vary by orders of magnitude [50]. Other inter-sample variations such as different relaxation or variations in RF pulse calibration can be corrected by normalization. Typically normalization is a multiplication of every row (i.e. every NMR spectrum) by a constant [105]. This constant can be computed in many different ways. The standard method, integral normalization, normalizes the individual spectra to constant total integrated intensity across the whole profile [137]. Integral normalization is also referred as constant sum normalization [105]. For NMR spectra of urine the normalization using the creatinine peak area as reference is common practice [138, 139]. There are other normalization methods as well, for instance probabilistic quotient normalization (PQ) [140], "histogram" normalization [141], group aggregating normalization (GAN) [142].

Metabolites can range in concentration over many orders of magnitude. Moreover the variation in metabolites level is often linked to concentration in such way that higher concentration metabolites have higher variation [104]. This causes that such metabolites have the highest influence on results of e.g. Principal Component Analysis (PCA) or Partial Least Squares (PLS). Therefore it is important to scale metabolites intensities before analysis to avoid selection of the most abundant metabolites as significant.

A number of scaling methods are commonly used, namely meancentering, autoscaling, pareto scaling, range scaling, vast scaling, level scaling [104, 105, 143]. Meancentering adjusts for differences between high-concentrated and low-concentrated metabolites by converting all values to vary around zero instead of around mean of metabolite level. It is not sufficient if in the data there are sub-populations with different variability (heteroscedastic data). Meancentering is usually used in combination with other scaling methods. Autoscaling scales all metabolites to unit variance and therefore the data is analyzed on the correlations basis instead of covariances. One has to be careful to use

this scaling for noisy data. Indeed autoscaling increases the influence of noisy variables. Pareto scaling is an intermediate option. It uses the square root of standard deviation as scaling factor instead of the standard deviation. This scaling method stays closer to real measurement but it is sensitive to large changes in the data. In range scaling the difference between the minimal and the maximal concentration of each metabolite is used as scaling factor. All metabolites in the data become equally important after applying ranges scaling. It is important to remember that range scaling is sensitive to outliers Because only two values are used to calculate the range. Level scaling focuses on relative response by using mean as scaling factor. It is suitable for use when large relative changes are of interest. Vast scaling can be considered as extension of autoscaling. It focuses at stable metabolites, i.e. the one having small variations.

Besides scaling methods there are several transformation approaches, like log transform [143] or the Box-Cox transformation [144] which can be used as preprocessing step. The Box-Cox transformation is a parametric preprocessing technique which decreases the effect of non-normality and heteroscedasticity. The last methodology described here is log transformation, which is nonlinear conversion of data. Large values are reduced in the data set relatively more than the small values. Log transformation removes heteroscedasticy from data if relative standard deviation is constant.

1.4 DATA ANALYSIS

1.4.1 Unsupervised analysis

Multivariate statistical methods provide an expert means of analyzing and maximizing information recovery from complex NMR data. Precise inspection of NMR data and integration of individual peaks can give valuable information on biochemical changes. After preprocessing the next stage is to analyse the data. The first step of data analysis is to explore and discover the overall structure of the data, find trends and groupings in the data. This stage of data analysis is based on blind unsupervised methods, i.e. those that do not assume any prior knowledge. These methods allow for an unbiased view of the data. There are several unsupervised methods available, among them PCA, robust-PCA, Hierarchical cluster analysis (HCA), K-means.

PCA is the workhorse [145] in multivariate analysis and is probably the most commonly used multivariate statistical analysis in metabolomics [139, 146-148]. It was invented by Pearson in 1901 [149]. PCA converts the multidimensional data space into a low-dimensional model plane. This technique expresses most of the variance within a data set using a smaller number of factors, so called Principal Components (PC's). Each PC is a linear combination of the original variables whereby each successive PC explains the maximum amount of variance, which was not accounted for by the previous PCs. Each PC is orthogonal to the other PCs and therefore exhibits different information. The variation in spectral data is described by a few PCs, compared to the number of original variables. Usually NMR spectra of 300 bins each can be summarized by a few new Principal Components (PC's). Moreover it enables to find trends, groupings, at to an extend outliers in the data.

Conversion of the original data set by PCA results in two matrices known as scores and loadings. Scores are the new coordinates for the samples. In a scores plot, each point represents a single spectrum. It provides a summary of all spectra and shows how they are related to each other. Hence, the points that are close to each other have similar profiles. On the contrary, objects that lie far away are characterized by different properties. The PC loadings describe the way in which the old variables are linearly

combined to new variables (PC's) and indicate which variables have the greatest contribution in transforming to the new variables. In the loading plot the relation among measured variables is shown. An important feature is that the directions in the score plot correspond to direction in the loading plot. Thus, any spectral clustering observed on the score plot is interpreted by examination of the loadings.

Presence of outliers affects all least squares methods, which are commonly used in multivariate data analysis. Therefore outliers detection is very essential. PCA can be used for outliers detection, however it can detect only certain type of outliers, i.e. good leverage objects (lying far away from the majority of objects). It cannot detect the orthogonal outliers, since after projecting them into PCA space they fall into cloud of data majority. Therefore robust-PCA should rather be used for outliers detection. Up till now, many robust versions of classical estimators have been proposed and their description can be found in [150]. Matlab implementation of robust-PCA can be found in [151, 152].

HCA is another unsupervised method which is widely used in modeling of metabolic data [153-155]. This method has the ability to group samples according to their similarity. HCA requires the choice of two input functions, namely the metric to be used as similarity between metabolic profiles (e.g. Euclidian, Mahalanobis or Minkowski distances) and the so-called linkage function (e.g. single, average complete or Ward's) [156, 157]. The choice of similarity metric and linkage have influence on the clustering structure. The HCA clusters the data forming a tree called dendrogram. In order to use HCA for classification, one has to decide the similarity cut-off, which divides the dendrogram into separate clusters. The main drawback of HCA is that it does not provide the information about the reason for a certain clustering. This means that HCA does not reveal which metabolites are responsible for the differences between clusters.

K-means is clustering approach has become extensively used in post-genomics, especially in analysis of transcriptomic data [156, 158-160]. Despite the popularity of k-means in many areas, in metabolomics it has not been widely applied. This might be due to the fact that there is no associated visualization or diagnostic tool. Although, recent applications involving the k-means for metabolomics data have been demonstrated [154, 161].

Lately a method called statistical correlation spectroscopy (STOCSY) [162, 163] has been developed to increase the information recovery from complex 1D-NMR metabolomics spectra. This method is based on the correlation matrix computed from 1D-NMR spectra. The correlation is calculated between all intensities in a set of NMR spectra. In this way the connections between signals from molecules that fluctuate in concentration between samples can be generated. By plotting the correlation matrix a graphical representation of samples spectra set comparable to that of a 2D correlation NMR experiment performed on one sample, namely total correlation spectroscopy (TOCSY). STOCSY is combined with supervised techniques to provide the link between relevant metabolites [117, 162]. STOCSY has been utilized for identification of drug metabolites in human urine[164]. Application of STOCSY to NMR data coupled with Liquid Chromatography arising from complex biological mixture has been demonstrated by Smith et al. [165]. STOCSY approach can be applied to 1D as well as to 2D data and to homo or heteronuclear spectral data. Heterospectroscopic-STOCSY takes into account a correlation between two different experimental data [166, 167].

Recently a combination of STOCSY and HCA has been developed, namely cluster analysis statistical spectroscopy (CLASSY) [168]. This approach has advantages over STOCSY, because correlations between the peaks from the same molecule are detected with higher accuracy.

1.4.2 Supervised analysis

Supervised techniques make use of *a priori* known structure. They use this knowledge to learn patterns and rules to predict new data. In these methods the relation between a matrix of predictors (i.e. NMR spectra) and a matrix or vector of responses (e.g. class membership or enzyme activity) is learnt. Regression is an example of supervised approach. In regression the responses are usually continuous parameters, e.g. age, blood pressure. Another major type of supervised method is classification in which responses are discrete and represented as class membership. In classification one searches for a rule that classify the objects into one of several classes. It is important to mention that regression and classification are closely related, since most regression methods can be used as classification approach. A very important aspect of

classification method is validation. Supervised methods are very powerful but it is often possible to construct a model which fits the data perfectly, giving 100% correct classification even if there is no real relation in the data. This is mostly caused by the high dimensionality of the metabolomics data (i.e. there are too many degrees of freedom with which the response can be predicted). Therefore validation of the model is a very crucial step in classification methods and will be discussed in the section entitled Validation. A wide range of methods has been used in metabolomics; here we focus on a few of them, i.e. on discriminant techniques. The advantage of these methods is to provide information about those variables that indicate differences between two or more classes. Therefore they are popular in metabolomics for biomarker discovery studies.

One of the most widely used classification method is Partial Least Squares Discriminant Analysis (PLS-DA), which is an alternative of the regression version of this technique, namely PLS [169, 170]. This technique, similarly to PCA, is a latent variable approach. It assumes that the data can be well approximated by a low dimensional subspace, i.e. by latent variables, which are assumed to be linear combinations of the original variables. A PLS-DA model can be expressed by:

Model of X: $\mathbf{X}=\mathbf{T}\mathbf{P}^T + \mathbf{E}$

Model of Y: $\mathbf{Y}=\mathbf{U}\mathbf{Q}^T + \mathbf{F}$ (1)

where *X* is an (n x m) matrix of predictors, *Y* is an (n x p) matrix of responses, *T* and U are an (n x l) matrices of projections of X (the X *score*) and Y (the Y scores), respectively, *P* and *Q* are, (m x l) and (p x l) *loading* matrices, respectively and matrices *E* and *F* are the error terms, and T indicates transposition.

The PSL-DA model can be expressed as well as:

$\mathbf{Y}=\mathbf{X}\mathbf{b}+\mathbf{r}$ (2)

Where, **X** is data matrix, **y** a vector of group memberships, **b** a vector of regression coefficients and **r** a vector of model residuals.

In Figure 4 it is shown how PCA and the PLS-DA model is obtained. As can be seen the first PC is constructed in the direction of highest variance in the data, while the first LV in the direction explaining the between-class variation of the objects.

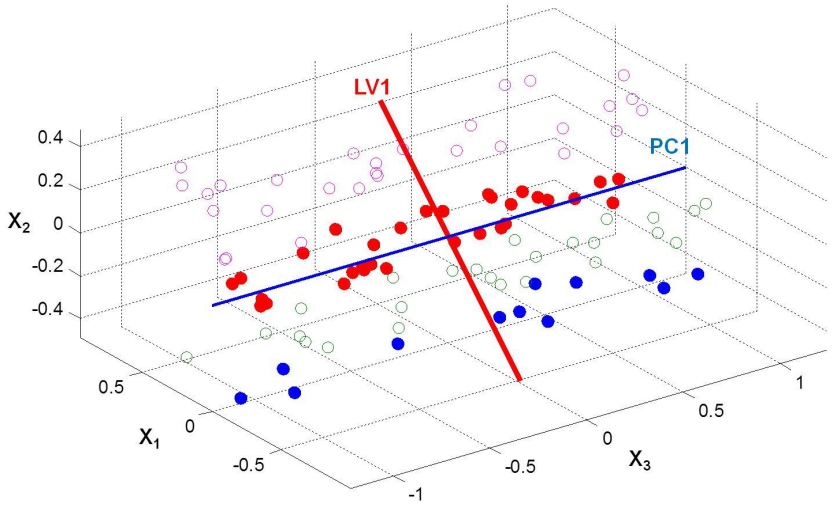


Figure 4. An illustration of PCA and PLS-DA model for simulated data containing 4 classes (pink, red, green and blue circles).

PLS-DA is a latent variable method, therefore it is possible to project the data into this new space, showing the relation between samples and variables. PLS-DA is well suited for highly correlated variables. Moreover PLS-DA can be used for two classes modeling, as well as when more than two classes are available, and then it is called PLS2-DA. In metabolomics it has been used in many applications [9, 32, 33, 58, 171].

The recent modification of PLS-DA is orthogonal PLS-DA (OPLS-DA) [172] in which the model is split into two parts. The systematic variations in X are split into two parts, i.e. one that is linearly related to response and one that is linearly uncorrelated to response (orthogonal). In that way only variation related to response are used to model it. The OPLS-DA model contains two modeled variations, namely the response-predictive

(related variation $\mathbf{T_pC^T_p}$) and the response-orthogonal ($\mathbf{T_oP^T_o}$). The OPLS-DA model can be expressed as:

Model X: $\mathbf{X=T_oP^T_o + T_pP^T_p + E}$

Model of Y: $\mathbf{Y=T_pC^T_p + F}$

In terms of prediction power OPLS-DA and PLS-DA are comparable (for the same amount of latent variable). However in terms of interpretability the OPLS-DA has advantages over standard PLS-DA, since the irrelevant variation is filtered out. This variation unrelated to response is often called structured noise and is caused by differences between subjects (e.g. different diet, age, gender). OPLS-DA similarly to PLS-DA has been successfully applied to many metabolomics data [117, 162, 173]. One has to be aware of that OPLS-DA never outperforms PLS-DA [174]. Moreover it has been shown by Kemsley and Tapp that the splitting of PLS models into y-related and y-unrelated parts can be obtained from the factorization proposed by Martens [175, 176]. It is important to mention that this is restricted to PLS1, i.e. when one y-response is available.

In the field of pattern recognition methods many other discriminant techniques are available to analyze metabolomics datasets. However metabolomic data contain highly collinear variables, usually with many more variables than samples. This causes problems with many standard pattern recognition methods, e.g. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) [156, 177-179]. Therefore in order to use them for metabolomics data special adaptations for high-dimensional data have to be undertaken. One possibility is to perform data reduction on beforehand by means of PCA. This method is then called Principal Component Discriminant Analysis (PCDA) [180, 181]. Lately a modification of the standard Canonical Variates Analysis (CVA) method to cope with collinear high-dimensional data has been developed, namely extended CVA (ECVA) [182]. This approach can be used as a supervised compression method and be used in combination with LDA as a classifier a tool for classification and discrimination of collinear data is obtained. LDA is applied to the canonical variates obtained from ECVA. This new method was recently used for extracting and compressing the information contained in metabolomics and proteomic datasets[183] .

1.4.2.1 Variable selection

In metabolomics data many variables are irrelevant or redundant for classification purpose and therefore the number of variables can be reduced with minor loss of information. It is common to apply variable selection methods to obtain a small panel of variables that are related to the response. Variable selection generally improves the model accuracy and/or reduction of the model complexity; moreover it can also be relevant for reducing the risk of overfitting. It is not straightforward to find a subset of relevant variables. An optimal variable selection approach is to perform all combinations and select the best subset. This is however rarely doable because of computational reasons. Moreover the chance of overfitting is much bigger if the number of variables is much higher than the number of samples. Many variable selection techniques exist. They work in different ways and have been developed for many different applications [184]. The simplest variable selection technique is a univariate approach, where every variable is evaluated individually. In these methods statistical values are calculated for each variable after testing (e.g. t-statistics) the differences between two classes. The disadvantage of these methods is that it is relatively easy to get high correlation by pure chance for high-dimensional datasets. To avoid such situation and to control false discovery rate multiple testing correction is needed [185, 186]. On the contrary to proteomics in the metabolomics field these univariate variable selection methods are not very popular [48, 187, 188].

Other popular techniques in metabolomics are multivariate [189-193]. Many are incorporated with the PLS algorithm (e.g. uninformative variable elimination PLS [194, 195], Cross-model Validation PLS [196, 197], backward selection method for PLS [198], iterative PLS [199], interval PLS [200], selectivity ratio plot [201]. A recent tutorial by Anderssen and Bro gives an overview of many variable selection methods with focus on regression-based calibration models [202]. Analysis of Variance-Principal Component Analysis can also be used for variable selection, as well as for classification [203]. It has been successfully applied to metabolomics data [33].

1.4.2.2 Validation of models

As was mentioned earlier validation is a very important aspect of classification. Note that validation is also essential from biological point of view. Properly validated statistical models give confidence to the findings (i.e. relevant metabolites). This is significant if the results from statistical analysis are used later, for instance in clinical application. If there are many more variables than samples it is possible to find by chance a perfect model that fits data [204]. Therefore it is obligatory to check the model for its predictive ability. There are several options for validating a model for predictive performance[196, 205, 206]. Cross-validation and double cross-validation are the preferable ones. Permutation is another way of validating the classification model. However validation by an external test sets provides a means to establish a more reliable predictive performance of the classification model. In order to obtain training and test set the data are partitioned by using different approach, e.g. the Duplex algorithm[207], Kennard-and-Stone [208, 209] or random selection. The test set is then used to validate the classification model.

Validation can be also performed using a completely new set of samples, coming from independent new experiment. This kind of validation is the ultimate one, because it allows one to test the outcomes of analysis on different population. Unfortunately such validation is rarely applied.

1.4.2.3 Non-linear methods

Biological processes are very often following a non-linear response. This is due to the complex interactions occurring in the many levels of biological organization. Some external factors (e.g. diet, medications) may produce metabolic effects that are not linearly related to group differences [210]. Therefore a number of characteristics in metabolic data argue for the consideration of non-linear pattern recognition methods. Among many non-linear techniques that have been developed in machine learning and pattern recognition area, we will discuss a particular class of non-linear models, i.e. kernel-based models (namely Support Vector Machines (SVM) and kernel-PLS).

Kernel-based models require a kernel transformation which is used beforehand to map objects into high dimensional space, called the feature space. By mapping the original space of size (n x p, where n is the number of samples and p is the number of variables)

into the feature space a kernel matrix is obtained of size (n x n). There are many kernel functions that can be used and the choice of kernel transformation is user dependent. The kernel matrix is required to be positive semi-definite and there are many kernel functions which fulfils this requirement [211]. The simplest kernel function is the dot product of the data matrix. Another commonly used kernel function is the radial basic function, which has a tuning parameter related to the width of the Gaussian. It is important to optimize this parameter, since its value has influence on the predictive ability of the classification model. An Important aspect of any kernel-based method is the loss of variable information (due to the kernel transformation). As mentioned above the kernel matrix is of size (n x n, where n is the number of samples). This causes that the information regarding the original input variables is vanished and direct interpretation of kernel-based-models is not possible. This weakness of kernel-based-methods was recently solved by Krooshof *et al.* [212] by applying a procedure, based on the non-linear biplot principle described by Gower [213].

Note that after kernel transformation any of the above discussed method can be applied. Therefore PLS-DA as well as OPLS-DA can be applied to the kernel matrix, leading to K-PLS [214, 215] and KO-PLS [216]. Another kernel-based method, SVM was introduced in the early nineties by Vapnik and coworkers [217]. SVM is a powerful, supervised method, used for binary classification. This state-of-the-art technique first maps the objects into feature space then tries to find a hyperplane which separates the data into its two classes. SVM uses a set of objects, so called support vectors, which span the range of the separating hyperplane. Kernel-based methods have been applied in many areas also in metabolomics studies [218-222].

1.4.3 Data fusion

It is common to study a single system, for instance a disease, by using different analytical platforms. Therefore in the last decade data concatenation or data fusion has become widespread in the field of metabolomics. Data fusion is an approach that combines data from different sources into a single and more complete description. Each analytical technology demonstrates different strengths and limitations regarding its

capability to distinguish between different biological conditions or to measure metabolites, depending upon factors such as sensitivity, sample preparation, analytical stability, and analytical reproducibility. The joint use of two or more analytical technologies gives then a more robust strategy for data analysis than the use of a single platform.

Different strategies for data fusion have been described by Hall et al. [223]. Three approaches for concatenating data can be distinguished, namely low-level, mid-level and high-level [224, 225]. In low-level fusion, different data sources are concatenated at the data level. All measured variables (or absolute metabolite concentrations) are put next to each other. This strategy is the simplest and the most straightforward. In the mid-level fusion, data from different sources are first treated separately for pre-processing and variable selection. Next the most optimal set of variables are concatenated into a single set. Note that datasets can be combined at the latent variables level. In high-level data fusion, different model responses (for instance prediction for each available data set) are joined to produce a final response. In high-level data fusion the individual results are usually weighted [226]. This kind of data fusion has two pitfalls, first interpretation of the model results is difficult and second it does not take into account correlation between measurements in different data sources. In low-level and midlevel data fusion the information about relevant metabolites can be easily determined. The choice of strategy for data fusion depends on the goal of the study [227]. The overview of different intra and inter-omics data integration can be found in [228].

It is also possible to fuse the datasets in a kernel space. In this approach the different sources of data are first mapped into a kernel space. The obtained kernel matrixes are then concatenated by linear combination [123, 229]. This kernel fusion falls outside the range of the classical low-, mid- and high-level fusion.

It is possible to perform data integration on pathway level. Using known metabolic pathway, data from multiple omics platforms are incorporated. This provides a more global investigation of a studied problem.

1.5 METABOLOMICS ANALYSIS of CSF by NMR

Metabolomics plays significant role in the discovery of biomarkers (or risk factors) connected with particular disorders. Many metabolomic studies are based on animal models, where inter-animal variations are limited, or directly in humans. The concentrations and fluxes of individual metabolites are the final product of interactions between gene, protein and cellular environment. Thus information delivered by metabolomics is complementary to other omics related fields. The metabolic fingerprints of biofluids obtained by NMR contain many signals related to both genetic and environmental contributions [4, 24, 230]. An important benefit of metabolic profiling for biomarker discovery is that metabolites are defined chemicals irrespective of species, genotype. Because many metabolites are not dependent on species, they can form the basis for translational study, i.e. biomarkers that are found in preclinical (animal) studies can be applied during clinical studies. A single biomarker found in metabolomics will not capture the complex process underlying a disease. Therefore, it is common practice to look at the perturbations in biological pathways and networks. This allows overall understanding of the metabolism and/or metabolic dynamics associated with the disease. Biomarker discovery is not a small task and typically requires years of validation before the clinical application phase is reached. An ideal biomarker should be sensitive, predictive, measureable in an easy accessible biofluid and cost effective [231]. The relevance of metabolomics is reflected by the fact that over 95% of diagnostic clinical assays look for small molecules, 89% of known drugs are small molecules, 50 % of all drugs are derived from pre-existing metabolites, and 30% of identified genetic disorders involve disease of small molecules metabolism [15]. A number of complex diseases result from a chronic imbalance of normal metabolism (e.g. cancer or type 2 diabetes) [232]. Therefore, metabolomics can be useful for finding biomarkers of pathological states.

1.5.1 Metabolomics profile of CSF of "healthy" control

The number of studies that use CSF for finding new metabolic biomarkers has been increasing. However, only a limited number of works focused on inspecting variations

occurring in CSF of individuals with no neurological disorders (i.e. “healthy” control). Nevertheless, it is very important to establish a comprehensive list of metabolites detectable in CSF and the corresponding variations. A proper understanding of biological fluctuation of metabolites concentrations in CSF of subjects considered as neurologically normal is crucial for trustworthy interpretation of the results in biomarker discovery studies. It is important to investigate if the variations between control and disease are caused by disease and not by inter-individual fluctuations. In the study of Jukarainen et al. ^{1}H-NMR has been used to obtain the absolute quantification of metabolites in CSF of 45 neurological controls [233]. The neurological controls chosen in the study consisted of patients examined for various neuropsychiatric symptoms, such as depression or headache but who did not give any signs of dementia and chronic neurological disease. The authors detected large inter-individual variations in metabolites concentration. In another study, Wishart et al. investigated the absolute concentration of metabolites identified in CSF by NMR and mass spectrometry in 35 individuals screened for meningitis [234]. Similar to the previous study they found considerable inter-individual variations. Wishart et al. reported that “normal” concentrations for many metabolites can vary by more than 50%. Moreover, they concluded that NMR appears to be the most suitable method for performing non-targeted metabolic profiling of CSF. Metabolic profile of CSF of controls has been investigated by Kolokolova et al. [48]. In this study, NMR was used to quantify in total 25 metabolites identified in human CSF. The absolute metabolites concentrations in CSF of controls were compared to metabolites level of patients suffering from motor neuron disease and ischemic stroke. The final findings indicated that patients with ischemic stroke had higher concentration for 19 CSF metabolites in comparison to controls.

In a recent comprehensive study by Stoop et al., [47]CSF samples of patients without neurological disease were investigated. Broad analysis of metabolite and protein concentrations, biological variation and analytical variations has been investigated in 32 human CSF samples by means of NMR and MS. Moreover, the effects of age and gender on biological variations were also addressed and found to be insignificant. Biological variations of metabolites identified by NMR and/or MS are reported to fluctuate from 15 % to 70% for the majority of metabolites. The analytical variations were

smaller than the biological variation in all cases. For NMR analysis the biological variation ranged from 8% to 53%, while the analytical error was between 2% and 9% for all metabolites. It should be pointed out that the biological variation found in Stoop et al. study was lower than in study published by Wishart et al. [232]. This difference is mostly due to the type of CSF samples used. The biological variations of CSF of "healthy controls" were also compared with fluctuations occurring in CSF of subjects with neurological diseases. Most interestingly, for a significant number of metabolites, the biological variation in diseased subjects are similar to that of normal controls, indicating that part of the CSF metabolome is more influenced by inter-individual differences and that the contribution of the diseases is only minor. Overall, it can be concluded that an understanding of the biological variation of metabolites in CSF of neurologically normal individuals is critical for a trustworthy interpretation of biomarker discovery studies for CNS disorders.

1.5.2. Metabolomics biomarkers by means of NMR

One promise of NMR is the identification of biomarkers in biofluids for early diagnosis. Metabolomics has the potential to show early biochemical changes in disease and thus gives a chance to develop biomarkers that can trigger early interventions (i.e. before symptoms are observable). Metabolic profiling of blood plasma, urine and CSF by means of NMR and chemometrics has been used effectively in clinical research for the studying a wide range of diseases. For instance plasma or serum samples have been used to evaluate lipids metabolism in obese patients [235], to detect pancreatic cancer [6] and to detect coronary heart disease and predict the severity of this disease [236]. While urine samples have been used to detect inborn metabolic diseases [14], the efficacy of immunosuppressant in renal transplant [237], for investigation of physiological perturbations during radiation sickness [56], bladder cancer [238], the effect of acute cysteamine supplementation [239] and dysfunction in peroxisomal proliferation [240]. More examples of using NMR for identifications of potential biomarkers in cancer, heart disease, diabetes, neurological disease and asthma can be found in the following references [4, 6, 8, 32, 241-244]. Each of these metabolic profiling studies was able to describe

an accurate relationship between the biofluid metabolic spectral patterns and a disease state in humans.

Finally, CSF has been used for studying a variety of neurological diseases, e.g. the detection of meningitis [245] and ventriculitis [246], Huntington disease [103]. Metabolic profiling of CSF has been conducted to distinguish first-onset schizophrenia patients from healthy controls [26]. CSF was also utilized to diagnose differentially viral, bacterial meningitis and tuberculosis in children[247]. Using an NMR-based metabolomics approach, it has been possible to investigate HIV-1-infected individuals, who are known to be susceptible to neuropsychological dysfunction[166]. In the study by Blasco et al. NMR-based metabolomic profiling of CSF in combination with PCA was used to find biomarkers of early amyotrophic lateral sclerosis [11]. Another recent study used NMR spectroscopy to identify CSF biomarkers for early diagnosis of Alzheimer's disease [27].

1.6 METABOLOMICS BIOMARKER STUDIES in MScl

MScl is a common disease of CNS which is an inflammatory, presumably autoimmune disease in which the fatty myelin sheaths which surrounds the axons of the brain and spinal cord are damaged, leading to demyelization. MScl can usually be diagnosed based on clinical examination, CSF analysis, observation from MRI and patient's history. Nevertheless, often, abnormalities found in the CSF and supported by MRI findings are not specific and sufficient enough. Many diseases can mimic conditions characteristic for MScl causing a rate of misdiagnosis of approximately 5%. An overview about different diagnosis in MScl can be found in [248-252]. Recently, new diagnostics criteria for MScl have been introduced [253]. Two factors, genetic and environmental, have influence on the risk and course of MScl. However, there is still uncertainty on aetiology and pathogenesis of MScl. MScl lesions are rarely biopsied. Therefore, there is an urgent need for new biological markers measured in CSF to better diagnose MScl. Many efforts have been made to develop such a better diagnostic for this heterogeneous disease [254].

Magnetic Resonance Techniques, ^{1}H-NMR and Magnetic Resonance Imaging and Spectroscopy (MRI/MRS), are commonly used to investigate MScl disease. ^{1}H-NMR is used to obtain metabolic profiles characteristic of distinct MScl manifestations, while

MRI/MRD is mostly employed for looking at plaques and lesions in the brain. In spite of ^{1}H in vivo Magnetic Resonance Spectroscopy (MRS) has been widely used to investigate the metabolic alterations occurring in the brain of MScl patients [255-260], it is important to notice that there are limited amount of studies where CSF of MScl is investigated by NMR. ^{1}H in vivo MRS spectra of MScl revealed increased lactate, choline, inositol and acetate level and decreased concentration of N-acetylaspartae,

CSF is mostly used to study metabolic malfunctions occurring in MScl patients. However, there are some studies where blood plasma and urine are used to find metabolic differences between MScl patients and controls [32, 261]. MScl can be studied either in pre-clinical models, e.g. the Experimental Autoimmune Encephalomyelitis (EAE) model or directly in humans. Until now there are only three studies reported where CSF of EAE animals is investigated by means of NMR [33, 183, 262]. We focus on the studies where ^{1}H-NMR of CSF was utilized.

Glucose concentrations have been measured in the CSF of MScl patients for diagnostic purpose before metabolomics was used to investigate MScl. Later it turned out that in most MScl patients the glucose concentration was normal [28]. Presently various clinical and pre-clinical metabolic studies have been carried out to characterize MScl. The first NMR studies of metabolic profile of CSF of MScl patients consisted of finding the differences between MScl patients and controls without using pattern recognition methods. Most of the NMR-based studies focus on absolute metabolite concentrations or ratios. One of the first NMR study of metabolic profiles of CSF of MScl patients was published by Lynch et al. in 1993, where a relatively big group size was investigated, i.e. 30 MScl patients and 27 controls [263]. In this study an unidentified N-methyl compound was found in MScl patients and increased acetate level and decreased formate concentration in comparison to controls. Few years later Nicoli *et al.* [30] have reported significant change in concentrations of several unidentified resonances in NMR spectra of CSF in MScl patients. Moreover, they found that some metabolites are slightly modified in MScl patients, i.e. increased lactate and fructose concentrations, decreased creatinine and phenylalanine concentrations. They reported as well that metabolic profile of CSF does not allow one to differentiate relapsing-remitting MScl and primary progressive MScl. This study [30] did not confirm the reduced formate level reported by

Lynch et al. Simone et al. [264] report increased ratios of lactate/creatine and acetate/creatine and reduced ratio of formate/creatine. Minor differences in lactate and acetate ratios were detected between RR-MScl and chronic progressive MScl. Later Simone et al. [265] have shown that the increased ratio of lactate/creatine is directly related to elevated level of lactate of demyelinating MScl plaques. In contrast to previous findings, in the study of Aasly et al. [266] the level of lactate in MScl patients was found to be down regulated.

Next a study where NMR was used to find markers of MScl in CSF was published by Lutz et al. in 2007 [45]. A highly homogenous patient group was used to characterize metabolic profile of MScl patients with and without inflammatory plaques. They demonstrated that there is a correlation between β-hydroxyisobutyrate (BHIB) [267] concentration and the presence of active inflammatory plaques. Lactate and BHIB concentrations were found to be elevated in CSF of MScl patients. In the same study, differences between clinically isolated syndromes (CIS) without any inflammatory plaques and controls were investigated. Small differences in fructose concentration were discovered. Lutz et al. [45] have shown PCA results obtained for different metabolites concentration. This is the first study, where pattern recognition methods are applied to NMR spectra of CSF of MScl patients. The PCA score plots demonstrate some groups clustering but hampered by some overlapping.

In Sinclair et al. [32], the potential of CSF NMR spectroscopy is combined with supervised analysis to find differences among MScl, idiopathic intracranial and hypertension patients. This is the first study where relative metabolite concentrations are investigated instead of absolute ones. By using PLS-DA, the sensitivity and specificity for MScl group was 83% and 53%, respectively. These results indicate that the obtained classification model is not specific enough for MScl, since almost half of the other patients were classified as MScl. Several significant changes in metabolites concentration in MScl patients were found, namely elevated levels of 2-aminobutyrate, 1,3-dimethylurrate, glutamate and acetate, and reduced levels of oxaloacetate, citrate, alanine and 3-hydroxybutyrate.

Another study applying pattern recognition, namely decision-tree-fuzzy classifier, by Aymerich et al. [245] investigated differences in metabolic profile of RR-MScl (8 patients),

PP-MScl (7 patients) and patients with other neurological disease (7 individuals. They focus on spectral regions with poor signal to noise ratio (i.e. between 6.0 and 9.7ppm). Excellent classification efficiency for cross validation was obtained. Due to the small groups size they did not present classification accuracy for independent test set. The relevant metabolites are not specified. The results presented by Aymerich et al. are promising, but it should be kept in mind that the study was performed on a small patient group.

In the last study described here, a novel approach for data fusion is applied to metabolomics datasets coming from patients suffering from MScl disease [123]. In this study, metabolic profiles of MScl patients are compared with CIS patients. The K-PLS-DA was used as classification model. The overall correct classification obtained for K-PLS-DA model of NMR data was 92.86% for the test set (one MScl sample was classified as CIS). A number of metabolites have been found as changing during progression of disease, e.g. the levels of lactate and of valine both increase, while the concentration of glutamine and citrate is reduced with disease progression. The high level of correct classification is very promising. However, the limitation is the relatively small number of studied patients (26 for MScl and 20 for CIS) and obviously, before usage in clinical practice, validation studies with larger number of patients is needed.

To summarize, the changes in several metabolites concentration in CSF were identified as related to MScl. Elevated lactate level and reduced acetate level in CSF of MScl patients were found in most studies. However, there was only one study where lactate concentration was found as being down regulated in CSF of MScl patients. Creatinine, oxaloacetate, alanine, citrate, glutamine and 3-hydroxybutyrate and phenylalanine concentrations have been reported to be decreased in CSF of MScl patients, while 2-aminobutyrate, 1,3-dimethylurrate and glutamate elevated. Interestingly, reduced alanine and citrate concentrations have been found in CSF of EAE affected animals [33].

1.7 CONCLUSIONS

In this review, a complete overview of the main steps involved in acquiring NMR spectra of biofluids, data preprocessing and modeling of NMR metabolic profiles is given. We have described the most commonly used techniques for acquiring NMR spectra of biofluids and presented an NMR-based metabolomics for CSF. Moreover different steps of data preprocessing and data analysis are described.

Metabolomics represents a major and rapidly evolving field. The development of NMR experiments allows for accurate measuring of many metabolites in different biofluids. The versatility of NMR has enabled this analytical technique to make valuable contributions to biomarker discovery in metabolomics field of biofluids. The application of NMR to the study of MScl disease is increasing. We have demonstrated several studies where several differences in the biochemical composition of CSF between MScl patients and controls were found. However, there are some discrepancies between the findings. There are several reasons why the results may differ. The most straightforward one may originate from the various sample preparations and acquisition methods. Another reason is connected to variation in CSF composition which may be dependent on the MScl patients used in the study.

Biomarker discovery by means of NMR of CSF in combination with pattern recognition techniques is likely to make an increasing contribution in uncovering disease mechanisms of complex neurological disease, like MScl. Collectively, NMR and pattern recognition methods offer great promise for the identification of clinically relevant biomarkers.

ACKNOWLEDGEMENTS

This work was performed within the framework of Dutch Top Institute Pharma, project "The CSF proteome / metabolome as primary biomarker compartment for CNS disorders" (project nr. D4-102).

REFERENCES

1. Nicholson, J. K.; Lindon, J. C.; 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.
2. Fiehn, O., Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **2002,** 48, (1-2), 155-171.
3. Oliver, S. G.; Winson, M. K.; Kell, D. B.; Baganz, F., Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **1998,** 16, (9), 373-378.
4. Dunn, W. B.; Goodacre, R.; Neyses, L.; Mamas, M., Integration of metabolomics in heart disease and diabetes research: current achievements and future outlook. *Bioanalysis* **2011,** 3, (19), 2205-2222.
5. Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E., Metabonomics:a platform for studing drug toxicity and gene function. *Nature Reviews Drug Discovery* **2002** 1, 153-161.
6. Beger, R. D.; Schnackenberg, L. K.; Holland, R. D.; Li, D. H.; Dragan, Y., Metabonomic models of human pancreatic cancer using 1D proton NMR spectra of lipids in plasma. *Metabolomics* **2006,** 2, (3), 125-134.
7. Ellis, D. I.; Dunn, W. B.; Griffin, J. L.; Allwood, J. W.; Goodacre, R., Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics* **2007,** 8, (9), 1243-1266.
8. Gowda, G. A.; Zhang, S.; Gu, H.; Asiago, V.; Shanaiah, N.; Raftery, D., Metabolomics-based methods for early disease diagnostics. *Expert Review of Molecular Diagnostics* **2008,** 8, (5), 617-33.
9. Hori, S.; Nishiumi, S.; Kobayashi, K.; Shinohara, M.; Hatakeyama, Y.; Kotani, Y.; Hatano, N.; Maniwa, Y.; Nishio, W.; Bamba, T.; Fukusaki, E.; Azuma, T.; Takenawa, T.; Nishimura, Y.; Yoshida, M., A metabolomic approach to lung cancer. *Lung Cancer* **2011,** 74, (2), 284-292.
10. Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols* **2007,** 2, (11), 2692-2703.
11. Blasco, H.; Corcia, P.; Moreau, C.; Veau, S.; Fournier, C.; Vourc'h, P.; Emond, P.; Gordon, P.; Pradat, P. F.; Praline, J.; Devos, D.; Nadal-Desbarats, L.; Andres, C. R., (1)H-NMR-Based Metabolomic Profiling of CSF in Early Amyotrophic Lateral Sclerosis. *PLoS One* **2010,** 5, (10).
12. Bogdanov, M.; Matson, W. R.; Wang, L.; Matson, T.; Saunders-Pullman, R.; Bressman, S. S.; Beal, M. F., Metabolomic profiling to develop blood biomarkers for Parkinson's disease. *Brain* **2008,** 131, 389-396.
13. Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology. *Chemical Research in Toxicology* **2008,** 21, (1), 9-27.
14. Constantinou, M. A.; Papakonstantinou, E.; Spraul, M.; Sevastiadou, S.; Costalos, C.; Koupparis, M. A.; Shulpis, K.; Tsantili-Kakoulidou, A.; Mikros, E., H-1 NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine. *Analytica Chimica Acta* **2005,** 542, (2), 169-177.
15. Goldsmith, P.; Fenton, H.; Morris-Stiff, G.; Ahmad, N.; Fisher, J.; Prasad, K. R., Metabonomics: A Useful Tool for the Future Surgeon. *Journal of Surgical Research* **2010,** 160, (1), 122-132.
16. Bernini, P.; Bertini, I.; Luchinat, C.; Nepi, S.; Saccenti, E.; Schafer, H.; Schutz, B.; Spraul, M.; Tenori, L., Individual Human Phenotypes in Metabolic Space and Time. *Journal of Proteome Research* **2009,** 8, (9), 4264-4271.
17. Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L., Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **2011,** 40, (1), 387-426.

18. van der Greef, J.; Stroobant, P.; van der Heijden, R., The role of analytical sciences medical systems biology. *Current Opinion in Chemical Biology* **2004,** 8, (5), 559-565.
19. Alm, E.; Arkin, A. P., Biological networks. *Current Opinion in Structural Biology* **2003,** 13, (2), 193-202.
20. Nicholson, J. K.; Lindon, J. C., Systems biology: Metabonomics. *Nature* **2008,** 455, (7216), 1054-6.
21. Atkinson, A. J.; Colburn, W. A.; DeGruttola, V. G.; DeMets, D. L.; Downing, G. J.; Hoth, D. F.; Oates, J. A.; Peck, C. C.; Schooley, R. T.; Spilker, B. A.; Woodcock, J.; Zeger, S. L.; Grp, B. D. W., Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **2001,** 69, (3), 89-95.
22. Industry Guidance Information on Recalls of FDA Regulated Products http://www.fda.gov/Safety/Recalls/IndustryGuidance/default.htm.
23. Katz, R., Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* **2004,** 1, (2), 189-95.
24. Griffin, J. L., The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philosophical Transactions of the Royal Society B-Biological Sciences* **2006,** 361, (1465), 147-161.
25. Holmes, E.; Tsang, T. M.; Tabrizi, S. J., The application of NMR-based metabonomics in neurological disorders. *NeuroRx* **2006,** 3, (3), 358-72.
26. Holmes, E.; Tsang, T. M.; Huang, J. T. J.; Leweke, F. M.; Koethe, D.; Gerth, C. W.; Nolden, B. M.; Gross, S.; Schreiber, D.; Nicholson, J. K.; Bahn, S., Metabolic profiling of CSF: Evidence that early intervention may impact on disease progression and outcome in schizophrenia. *Plos Medicine* **2006,** 3, (8), 1420-+.
27. Kork, F.; Holthues, J.; Hellweg, R.; Jankowski, V.; Tepel, M.; Ohring, R.; Heuser, I.; Bierbrauer, J.; Peters, O.; Schlattmann, P.; Zidek, W.; Jankowski, J., A Possible New Diagnostic Biomarker in Early Diagnosis of Alzheimer's Disease. *Current Alzheimer Research* **2009,** 6, (6), 519-524.
28. Lutz, N. W.; Cozzone, P. J., Metabolic Profiling in Multiple Sclerosis and Other Disorders by Quantitative Analysis of Cerebrospinal Fluid Using Nuclear Magnetic Resonance Spectroscopy. *Current Pharmaceutical Biotechnology* **2011,** 12, (7), 1016-1025.
29. Lutz, N. W.; Maillet, S.; Nicoli, F.; Viout, P.; Cozzone, P. J., Further assignment of resonances in 1H NMR spectra of cerebrospinal fluid (CSF). *FEBS Lett* **1998,** 425, (2), 345-51.
30. Nicoli, F.; VionDury, J.; ConfortGouny, S.; Maillet, S.; Gastaut, J. L.; Cozzone, P. J., Cerebrospinal fluid metabolic profiles in multiple sclerosis and degenerative dementias obtained by high resolution proton magnetic resonance spectroscopy. *Comptes Rendus De L Academie Des Sciences Serie Iii-Sciences De La Vie-Life Sciences* **1996,** 319, (7), 623-631.
31. Paues, J.; Strom, J. O.; Eriksson, L.; Theodorsson, A., Tuberculous meningitis with positive cell-count in lumbar puncture CSF though negative cell-count from ventricular drainage CSF. *J Infect* **2011,** 62, (5), 404-5.
32. Sinclair, A. J.; Viant, M. R.; Ball, A. K.; Burdon, M. A.; Walker, E. A.; Stewart, P. M.; Rauz, S.; Young, S. P., NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases--a diagnostic tool? *NMR Biomed* **2010,** 23, (2), 123-32.
33. Smolinska, A.; Attali, A.; Blanchet, L.; Ampt, K.; Tuinstra, T.; van Aken, H.; Suidgeest, E.; van Gool, A. J.; Luider, T.; Wijmenga, S. S.; Buydens, L. M., NMR and pattern recognition can distinguish neuroinflammation and peripheral inflammation. *J Proteome Res* **2011,** 10, (10), 4428-38.
34. Quinones, M. P.; Kaddurah-Daouk, R., Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. *Neurobiology of Disease* **2009,** 35, (2), 165-76.
35. Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B. F.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.; Holmes, E., Assessment of analytical reproducibility of H-1 NMR

spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry* **2006,** 78, (7), 2199-2208.
36. Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* **2007,** 2, (11), 2692-703.
37. Makinen, V. P.; Soininen, P.; Forsblom, C.; Parkkonen, M.; Ingman, P.; Kaski, K.; Groop, P. H.; Ala-Korpela, M., 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol Syst Biol* **2008,** 4, 167.
38. Odunsi, K.; Wollman, R. M.; Ambrosone, C. B.; Hutson, A.; McCann, S. E.; Tammela, J.; Geisler, J. P.; Miller, G.; Sellers, T.; Cliby, W.; Qian, F.; Keitz, B.; Intengan, M.; Lele, S.; Alderfer, J. L., Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *Int J Cancer* **2005,** 113, (5), 782-8.
39. Tukiainen, T.; Tynkkynen, T.; Makinen, V. P.; Jylanki, P.; Kangas, A.; Hokkanen, J.; Vehtari, A.; Grohn, O.; Hallikainen, M.; Soininen, H.; Kivipelto, M.; Groop, P. H.; Kaski, K.; Laatikainen, R.; Soininen, P.; Pirttila, T.; Ala-Korpela, M., A multi-metabolite analysis of serum by 1H NMR spectroscopy: early systemic signs of Alzheimer's disease. *Biochem Biophys Res Commun* **2008,** 375, (3), 356-61.
40. Makinen, V. P.; Forsblom, C.; Thorn, L. M.; Waden, J.; Gordin, D.; Heikkila, O.; Hietala, K.; Kyllonen, L.; Kyto, J.; Rosengard-Barlund, M.; Saraheimo, M.; Tolonen, N.; Parkkonen, M.; Kaski, K.; Ala-Korpela, M.; Groop, P. H., Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes. *Diabetes* **2008,** 57, (9), 2480-7.
41. Ala-Korpela, M., Critical evaluation of 1H NMR metabonomics of serum as a methodology for disease risk assessment and diagnostics. *Clin Chem Lab Med* **2008,** 46, (1), 27-42.
42. Keun, H. C.; Athersuch, T. J., Nuclear magnetic resonance (NMR)-based metabolomics. *Methods Mol Biol* **2011,** 708, 321-34.
43. Lindon, J. C.; Nicholson, J. K., Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *Trends in Analytical CHemistry* **2008,** 27, (3), 194-204.
44. Chiasserini, D.; Di Filippo, M.; Candeliere, A.; Susta, F.; Orvietani, P. L.; Calabresi, P.; Binaglia, L.; Sarchielli, P., CSF proteome analysis in multiple sclerosis patients by two-dimensional electrophoresis. *European Journal of Neurology* **2008,** 15, (9), 998-1001.
45. Lutz, N. W.; Viola, A.; Malikova, I.; Confort-Gouny, S.; Audoin, B.; Ranjeva, J. P.; Pelletier, J.; Cozzone, P. J., Inflammatory multiple-sclerosis plaques generate characteristic metabolic profiles in cerebrospinal fluid. *PLoS One* **2007,** 2, (7), e595.
46. Rosenling, T.; Stoop, M. P.; Smolinska, A.; Muilwijk, B.; Coulier, L.; Shi, S. N.; Dane, A.; Christin, C.; Suits, F.; Horvatovich, P. L.; Wijmenga, S. S.; Buydens, L. M. C.; Vreeken, R.; Hankemeier, T.; van Gool, A. J.; Luider, T. M.; Bischoff, R., The Impact of Delayed Storage on the Measured Proteome and Metabolome of Human Cerebrospinal Fluid. *Clinical Chemistry* **2011,** 57, (12), 1703-1711.
47. Stoop, M. P.; Coulier, L.; Rosenling, T.; Shi, S.; Smolinska, A. M.; Buydens, L.; Ampt, K.; Stingl, C.; Dane, A.; Muilwijk, B.; Luitwieler, R. L.; Sillevis Smitt, P. A.; Hintzen, R. Q.; Bischoff, R.; Wijmenga, S. S.; Hankemeier, T.; van Gool, A. J.; Luider, T. M., Quantitative proteomics and metabolomics analysis of normal human cerebrospinal fluid samples. *Mol Cell Proteomics* **2010,** 9, (9), 2063-75.
48. Kolokolova, T. N.; Savel'ev, O. Y.; Sergeev, N. M.; Shpigun, O. A.; Sokolov, K. V.; Skvortsova, V. I., Nuclear magnetic resonance spectroscopy in solving the analytical problems of medicine: Analysis of cerebrospinal fluid. *Journal of Analytical Chemistry (Translation of Zhurnal Analiticheskoi Khimii)* **2010,** 65, (10), 1073-1081.
49. Nicholson, J. K.; O'Flynn, M. P.; Sadler, P. J.; Macleod, A. F.; Juul, S. M.; Sonksen, P. H., Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochem J* **1984,** 217, (2), 365-75.

50. Nicholson, J. K.; Wilson, I. D., High-Resolution Proton Magnetic-Resonance Spectroscopy of Biological-Fluids. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1989,** 21, 449-501.
51. Nicholson, J. K.; Buckingham, M. J.; Sadler, P. J., High resolution 1H n.m.r. studies of vertebrate blood and plasma. *The Biochemical journal* **1983,** 211, (3), 605-15.
52. Hamans, B. C.; Andreychenko, A.; Heerschap, A.; Wijmenga, S. S.; Tessari, M., NMR at earth's magnetic field using para-hydrogen induced polarization. *Journal of Magnetic Resonance* **2011,** 212, (1), 224-228.
53. Adams, R. W.; Aguilar, J. A.; Atkinson, K. D.; Cowley, M. J.; Elliott, P. I. P.; Duckett, S. B.; Green, G. G. R.; Khazal, I. G.; Lopez-Serrano, J.; Williamson, D. C., Reversible Interactions with para-Hydrogen Enhance NMR Sensitivity by Polarization Transfer. *Science* **2009,** 323, (5922), 1708-1711.
54. Potts, B. C. M.; Deese, A. J.; Stevens, G. J.; Reily, M. D.; Robertson, D. G.; Theiss, J., NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse. *Journal of Pharmaceutical and Biomedical Analysis* **2001,** 26, (3), 463-476.
55. Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Everett, J. R., Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance* **2000,** 12, (5), 289-320.
56. Khan, A. R.; Rana, P.; Tyagi, R.; Kumar, I. P.; Devi, M. M.; Javed, S.; Tripathi, R. P.; Khushu, S., NMR spectroscopy based metabolic profiling of urine and serum for investigation of physiological perturbations during radiation sickness. *Metabolomics* **2011,** 7, (4), 583-592.
57. Tyagi, R.; Rana, P.; Khan, A. R.; Bhatnagar, D.; Devi, M. M.; Chaturvedi, S.; Tripathi, R. P.; Khushu, S., Study of acute biochemical effects of thallium toxicity in mouse urine by NMR spectroscopy. *J Appl Toxicol* **2011,** 31, (7), 663-70.
58. Rocha, C. M.; Carrola, J.; Barros, A. S.; Gil, A. M.; Goodfellow, B. J.; Carreira, I. M.; Bernardo, J.; Gomes, A.; Sousa, V.; Carvalho, L.; Duarte, I. F., Metabolic Signatures of Lung Cancer in Biofluids: NMR-Based Metabonomics of Blood Plasma. *Journal of Proteome Research* **2011,** 10, (9), 4314-4324.
59. Diaz, S. O.; Pinto, J.; Graca, G.; Duarte, I. F.; Barros, A. S.; Galhano, E.; Pita, C.; Almeida, M. D.; Goodfellow, B. J.; Carreira, I. M.; Gil, A. M., Metabolic Biomarkers of Prenatal Disorders: An Exploratory NMR Metabonomics Study of Second Trimester Maternal Urine and Blood Plasma. *Journal of Proteome Research* **2011,** 10, (8), 3732-3742.
60. Graca, G.; Duarte, I. F.; Goodfellow, B. J.; Barros, A. S.; Carreira, I. M.; Couceiro, A. B.; Spraul, M.; Gil, A. M., Potential of NMR Spectroscopy for the study of human amniotic fluid. *Analytical Chemistry* **2007,** 79, (21), 8367-8375.
61. Graca, G.; Duarte, I. F.; Goodfellow, B. J.; Carreira, I. M.; Couceiro, A. B.; Domingues, M. D.; Spraul, M.; Tseng, L. H.; Gil, A. M., Metabolite profiling of human amniotic fluid by hyphenated nuclear magnetic resonance spectroscopy. *Analytical Chemistry* **2008,** 80, (15), 6085-6092.
62. Gowda, G. A. N.; Shanaiah, N.; Cooper, A.; Maluccio, M.; Raftery, D., Bile Acids Conjugation in Human Bile Is Not Random: New Insights from (1)H-NMR Spectroscopy at 800 MHz. *Lipids* **2009,** 44, (6), 527-535.
63. Gowda, G. A.; Shanaiah, N.; Cooper, A.; Maluccio, M.; Raftery, D., Visualization of Bile Homeostasis Using (1)H-NMR Spectroscopy as a Route for Assessing Liver Cancer. *Lipids* **2009,** 44, (1), 27-35.
64. DeFeo, E. M.; Wu, C. L.; McDougal, W. S.; Cheng, L. L., A decade in prostate cancer: from NMR to metabolomics. *Nature Reviews Urology* **2011,** 8, (6), 301-311.
65. Jordan, K. W.; Cheng, L. L., NMR-based metabolomics approach to target biomarkers for human prostate cancer. *Expert Review of Proteomics* **2007,** 4, (3), 389-400.

66. Maher, A. D.; Cloarec, O.; Patki, P.; Craggs, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Dynamic Biochemical Information Recovery in Spontaneous Human Seminal Fluid Reactions via (1)H NMR Kinetic Statistical Total Correlation Spectroscopy. *Analytical Chemistry* **2009,** 81, (1), 288-295.
67. Bertram, H. C.; Eggers, N.; Eller, N., Potential of Human Saliva for Nuclear Magnetic Resonance-Based Metabolomics and for Health-Related Biomarker Identification. *Analytical Chemistry* **2009,** 81, (21), 9188-9193.
68. Wei, J. E.; Xie, G. X.; Zhou, Z. T.; Shi, P.; Qiu, Y. P.; Zheng, X. J.; Chen, T. L.; Su, M. M.; Zhao, A. H.; Jia, W., Salivary metabolite signatures of oral cancer and leukoplakia. *International Journal of Cancer* **2011,** 129, (9), 2207-2217.
69. Bertram, H. C.; Eggers, N.; Eller, N., Potential of human saliva for nuclear magnetic resonance-based metabolomics and for health-related biomarker identification. *Analytical chemistry* **2009,** 81, (21), 9188-93.
70. Issaq, H. J.; Van, Q. N.; Waybright, T. J.; Muschik, G. M.; Veenstra, T. D., Analytical and statistical approaches to metabolomics research. *J Sep Sci* **2009,** 32, (13), 2183-99.
71. Tiziani, S.; Emwas, A. H.; Lodi, A.; Ludwig, C.; Bunce, C. M.; Viant, M. R.; Gunther, U. L., Optimized metabolite extraction from blood serum for 1H nuclear magnetic resonance spectroscopy. *Anal Biochem* **2008,** 377, (1), 16-23.
72. Daykin, C. A.; Foxall, P. J.; Connor, S. C.; Lindon, J. C.; Nicholson, J. K., The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Anal Biochem* **2002,** 304, (2), 220-30.
73. Wevers, R. A.; Engelke, U.; Wendel, U.; Dejong, J. G. N.; Gabreels, F. J. M.; Heerschap, A., Standardized Method for High-Resolution H-1-Nmr of Cerebrospinal-Fluid. *Clinical Chemistry* **1995,** 41, (5), 744-751.
74. Lehnert, W.; Hunkler, D., Possibilities of Selective Screening for Inborn-Errors of Metabolism Using High-Resolution H-1-Ft-Nmr Spectrometry. *European Journal of Pediatrics* **1986,** 145, (4), 260-266.
75. Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J., Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical Chemistry* **2007,** 79, (18), 6995-7004.
76. Viant, M. R.; Ludwig, C.; Rhodes, S.; Guenther, U. L.; Allaway, D., Validation of a urine metabolome fingerprint in dog for phenotypic classification. *Metabolomics* **2007,** 3, (4), 453-463.
77. Pauli, G. F.; Jaki, B. U.; Lankin, D. C., Quantitative 1H NMR: development and potential of a method for natural products analysis. *Journal of natural products* **2005,** 68, (1), 133-49.
78. Lane, S.; Boughtflower, B.; Mutton, I.; Paterson, C.; Farrant, D.; Taylor, N.; Blaxill, Z.; Carmody, C.; Borman, P., Toward single-calibrant quantification in HPLC. A comparison of three detection strategies: evaporative light scattering, chemiluminescent nitrogen, and proton NMR. *Analytical chemistry* **2005,** 77, (14), 4354-65.
79. Alum, M. F.; Shaw, P. A.; Sweatman, B. C.; Ubhi, B. K.; Haselden, J. N.; Connor, S. C., 4,4-dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA), a promising universal internal standard for NMR-based metabolic profiling studies of biofluids, including blood plasma and serum. *Metabolomics* **2008,** 4, (2), 122-127.
80. Akoka, S.; Barantin, L.; Trierweiler, M., Concentration measurement by proton NMR using the ERETIC method. *Analytical Chemistry* **1999,** 71, (13), 2554-2557.
81. Farrant, R. D.; Hollerton, J. C.; Lynn, S. M.; Provera, S.; Sidebottom, P. J.; Upton, R. J., NMR quantification using an artificial signal. *Magnetic Resonance in Chemistry* **2010,** 48, (10), 753-762.
82. Research, M. http://mestrelab.com/. (January 2012),

83. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M., Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry* **2006,** 78, (13), 4430-42.
84. Verwaest, K. A.; Vu, T. N.; Laukens, K.; Clemens, L. E.; Nguyen, H. P.; Van Gasse, B.; Martins, J. C.; Van der Linden, A.; Dommisse, R., (1)H NMR based metabolomics of CSF and blood serum: A metabolic profile for a transgenic rat model of Huntington disease. *Biochimica Et Biophysica Acta-Molecular Basis of Disease* **2011,** 1812, (11), 1371-1379.
85. Bart, J.; Kolkman, A. J.; Oosthoek-de Vries, A. J.; Koch, K.; Nieuwland, P. J.; Janssen, H.; van Bentum, P. J. M.; Ampt, K. A. M.; Rutjes, F. P. J. T.; Wijmenga, S. S.; Gardeniers, H.; Kentgens, A. P. M., A Microfluidic High-Resolution NMR Flow Probe. *Journal of the American Chemical Society* **2009,** 131, (14), 5014-+.
86. Wevers, R. A.; Engelke, U.; Heerschap, A., High-resolution 1H-NMR spectroscopy of blood plasma for metabolic studies. *Clin Chem* **1994,** 40, (7 Pt 1), 1245-50.
87. Hwang, T. L.; Shaka, A. J., 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.
88. Piotto, M.; Saudek, V.; Sklenar, V., Gradient-Tailored Excitation for Single-Quantum Nmr-Spectroscopy of Aqueous-Solutions. *Journal of Biomolecular Nmr* **1992,** 2, (6), 661-665.
89. Neuhaus, D.; Ismail, I. M.; Chung, C. W., ''FLIPSY'' - A new solvent-suppression sequence for nonexchanging solutes offering improved integral accuracy relative to 1D NOESY. *Journal of Magnetic Resonance Series A* **1996,** 118, (2), 256-263.
90. McKay, R. T., How the 1D-NOESY Suppresses Solvent Signal in Metabonomics NMR Spectroscopy: An Examination of the Pulse Sequence Components and Evolution. *Concepts in Magnetic Resonance Part A* **2011,** 38A, (5), 197-220.
91. Meiboom, S.; Gill, D., Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Review of Scientific Instruments* **1958,** 29, (8), 688-691.
92. Aue, W. P.; Karhan, J.; Ernst, R. R., Homonuclear Broad-Band Decoupling and 2-Dimensional J-Resolved Nmr-Spectroscopy. *Journal of Chemical Physics* **1976,** 64, (10), 4226-4227.
93. Aue, W. P.; Bartholdi, E.; Ernst, R. R., 2-Dimensional Spectroscopy - Application to Nuclear Magnetic-Resonance. *Journal of Chemical Physics* **1976,** 64, (5), 2229-2246.
94. Braunschweiler, L.; Ernst, R. R., Coherence Transfer by Isotropic Mixing - Application to Proton Correlation Spectroscopy. *Journal of Magnetic Resonance* **1983,** 53, (3), 521-528.
95. Software, P. N. http://perchnmrsoftware.com.
96. Mercier, P.; Lewis, M. J.; Chang, D.; Baker, D.; Wishart, D. S., Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. *Journal of Biomolecular NMR* **2011,** 49, (3-4), 307-323.
97. Rosenling, T.; Slim, C. L.; Christin, C.; Coulier, L.; Shi, S.; Stoop, M. P.; Bosman, J.; Suits, F.; Horvatovich, P. L.; Stockhofe-Zurwieden, N.; Vreeken, R.; Hankemeier, T.; Gool, A. J.; Luider, T. M.; Bischoff, R., The Effect of Preanalytical Factors on Stability of the Proteome and Selected Metabolites in Cerebrospinal Fluid (CSF). *Journal of Proteome Research* **2009,** 8, (12), 5511-5522.
98. Wuolikainen, A.; Hedenstrom, M.; Moritz, T.; Marklund, S. L.; Antti, H.; 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-U8.
99. Cunniffe, J. G.; WhitbyStrevens, S.; Wilcox, M. H., Effect of pH changes in cerebrospinal fluid specimens on bacterial survival and antigen test results. *Journal of Clinical Pathology* **1996,** 49, (3), 249-253.
100. Maillet, S.; Vion-Dury, J.; Confort-Gouny, S.; Nicoli, F.; Lutz, N. W.; Viout, P.; Cozzone, P. J., Experimental protocol for clinical analysis of cerebrospinal fluid by high resolution proton magnetic resonance spectroscopy. *Brain Res Brain Res Protoc* **1998,** 3, (2), 123-34.

101. Sweatman, B. C.; Farrant, R. D.; Holmes, E.; Ghauri, F. Y.; Nicholson, J. K.; Lindon, J. C., 600 MHz 1H-NMR spectroscopy of human cerebrospinal fluid: effects of sample manipulation and assignment of resonances. *J Pharm Biomed Anal* **1993,** 11, (8), 651-64.
102. Commodari, F.; Arnold, D. L.; Sanctuary, B. C.; Shoubridge, E. A., 1H NMR characterization of normal human cerebrospinal fluid and the detection of methylmalonic acid in a vitamin B12 deficient patient. *NMR Biomed* **1991,** 4, (4), 192-200.
103. Verwaest, K. A.; Vu, T. N.; Laukens, K.; Clemens, L. E.; Nguyen, H. P.; Van Gasse, B.; Martins, J. C.; Van Der Linden, A.; Dommisse, R., (1)H NMR based metabolomics of CSF and blood serum: a metabolic profile for a transgenic rat model of Huntington disease. *Biochimica et Biophysica Acta* **2011,** 1812, (11), 1371-9.
104. Ebbels, T. M.; Lindon, J. C.; Coen, M., Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. *Methods Mol Biol* **2011,** 708, 365-88.
105. Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C., Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem* **2006,** 78, (7), 2262-7.
106. De Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Rietzschel, E.-R.; De Buyzere, M. L.; Langlois, M. R.; Bekaert, S.; Martins, J. C.; Van Criekinge, W., Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum 1H-NMR spectra. *Anal. Bioanal. Chem.* **2010,** 398, (4), 1781-1790.
107. Zhang, S.; Zheng, C.; Lanza, I. R.; Nair, K. S.; Raftery, D.; Vitek, O., Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolic profiling. *Anal Chem* **2009,** 81, (15), 6080-8.
108. Gan, F.; Ruan, G. H.; Mo, J. Y., Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems* **2006,** 82, (1-2), 59-65.
109. Chang, D.; Banack, C. D.; Shah, S. L., Robust baseline correction algorithm for signal dense NMR spectra. *Journal of Magnetic Resonance* **2007,** 187, (2), 288-292.
110. Xi, Y.; Rocke, D. M., Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC bioinformatics* **2008,** 9, 324.
111. Eilers, P. H. C., A perfect smoother. *Analytical Chemistry* **2003,** 75, (14), 3631-3636.
112. Eilers, P. H. C.; Marx, B. D., Flexible smoothing with B-splines and penalties. *Statistical Science* **1996,** 11, (2), 89-102.
113. de Rooi, J. J.; Eilers, P. H. C., Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory System* **2011**.
114. Defernez, M.; Colquhoun, I. J., Factors affecting the robustness of metabolite fingerprinting using H-1 NMR spectra. *Phytochemistry* **2003,** 62, (6), 1009-1017.
115. Witjes, H.; van den Brink, M.; Melssen, W. J.; Buydens, L. M. C., Automatic correction of peak shifts in Raman spectra before PLS regression. *Chemometrics and Intelligent Laboratory Systems* **2000,** 52, (1), 105-116.
116. Vogels, J. T. W. E.; Tas, A. C.; Venekamp, J.; VanderGreef, J., Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *Journal of Chemometrics* **1996,** 10, (5-6), 425-438.
117. Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E., Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in H-1 NMR spectroscopic metabonomic studies. *Analytical Chemistry* **2005,** 77, (2), 517-526.
118. Giskeodegard, G. F.; Bloemberg, T. G.; Postma, G.; Sitter, B.; Tessem, M. B.; Gribbestad, I. S.; Bathen, T. F.; Buydens, L. M., Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Anal Chim Acta* **2010,** 683, (1), 1-11.

119. Giskeodegard, G. F.; Bloemberg, T. G.; Postma, G.; Sitter, B.; Tessem, M. B.; Gribbestad, I. S.; Bathen, T. F.; Buydens, L. M. C., Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Analytica Chimica Acta* **2010,** 683, (1), 1-11.
120. Savorani, F.; Tomasi, G.; Engelsen, S. B., icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* **2010,** 202, (2), 190-202.
121. Nielsen, N. P. V.; Carstensen, J. M.; Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **1998,** 805, (1-2), 17-35.
122. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **2004,** 18, (5), 231-241.
123. Smolinska, A.; Blanchet, L.; Coulier, L.; Ampt, K. A. M.; Luider, T.; Hintzen, R. Q.; Wijmenga, S. S.; Buydens, L. M. C., Interpretation and Visualization of Non-Linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis. *PLoS ONE* **2012,** 7, (6), e38163.
124. Lee, G. C.; Woodruff, D. L., Beam search for peak alignment of NMR signals. *Analytica Chimica Acta* **2004,** 513, (2), 413-416.
125. Wu, W.; Daszykowski, M.; Walczak, B.; Sweatman, B. C.; Connor, S. C.; Haseldeo, J. N.; Crowther, D. J.; Gill, R. W.; Lutz, M. W., Peak alignment of urine NMR spectra using fuzzy warping. *Journal of Chemical Information and Modeling* **2006,** 46, (2), 863-875.
126. Vu, T. N.; Valkenborg, D.; Smets, K.; Verwaest, K. A.; Dommisse, R.; Lemiere, F.; Verschoren, A.; Goethals, B.; Laukens, K., An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* **2011,** 12.
127. Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R., Improved parametric time warping for proteomics. *Chemometrics and Intelligent Laboratory Systems* **2010,** 104, (1), 65-74.
128. Forshed, J.; Schuppe-Koistinen, I.; Jacobsson, S. P., Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta* **2003,** 487, (2), 189-199.
129. Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K., Recursive Segment-Wise Peak Alignment of Biological (1)H NMR Spectra for Improved Metabolic Biomarker Recovery. *Analytical Chemistry* **2009,** 81, (1), 56-66.
130. Gartland, K. P.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K., Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine. *Mol Pharmacol* **1991,** 39, (5), 629-42.
131. Gartland, K. P.; Sanins, S. M.; Nicholson, J. K.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C., Pattern recognition analysis of high resolution 1H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data. *NMR Biomed* **1990,** 3, (4), 166-72.
132. Izquierdo-Garcia, J. L.; Villa, P.; Kyriazis, A.; del Puerto-Nevado, L.; Perez-Rial, S.; Rodriguez, I.; Hernandez, N.; Ruiz-Cabello, J., Descriptive review of current NMR-based metabolomic data analysis packages. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2011,** 59, (3), 263-270.
133. de Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Tsiporkova, E.; Rietzschel, E. R.; De Buyzere, M. L.; Gillebert, T. C.; Bekaert, S.; Martins, J. C.; Van Criekinge, W., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry* **2008,** 80, (10), 3783-3790.
134. Anderson, P. E.; Reo, N. V.; DelRaso, N. J.; Doom, T. E.; Raymer, M. L., Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics* **2008,** 4, (3), 261-272.

135. Davis, R. A.; Charlton, A. J.; Godward, J.; Jones, S. A.; Harrison, M.; Wilson, J. C., Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems* **2007,** 85, (1), 144-154.
136. Anderson, P. E.; Mahle, D. A.; Doom, T. E.; Reo, N. V.; DelRaso, N. J.; Raymer, M. L., Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics* **2011,** 7, (2), 179-190.
137. Spraul, M.; Neidig, P.; Klauck, U.; Kessler, P.; Holmes, E.; Nicholson, J. K.; Sweatman, B. C.; Salman, S. R.; Farrant, R. D.; Rahr, E.; Beddell, C. R.; Lindon, J. C., Automatic Reduction of Nmr Spectroscopic Data for Statistical and Pattern-Recognition Classification of Samples. *Journal of Pharmaceutical and Biomedical Analysis* **1994,** 12, (10), 1215-1225.
138. Fauler, G.; Leis, H. J.; Huber, E.; Schellauf, C.; Kerbl, R.; Urban, C.; Gleispach, H., Determination of homovanillic acid and vanillylmandelic acid in neuroblastoma screening by stable isotope dilution GC-MS. *J Mass Spectrom* **1997,** 32, (5), 507-14.
139. Holmes, E.; Foxall, P. J.; Nicholson, J. K.; Neild, G. H.; Brown, S. M.; Beddell, C. R.; Sweatman, B. C.; Rahr, E.; Lindon, J. C.; Spraul, M.; et al., Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal Biochem* **1994,** 220, (2), 284-96.
140. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **2006,** 78, (13), 4281-90.
141. Torgrip, R. J. O.; Aberg, K. M.; Alm, E.; Schuppe-Koistinen, I.; Lindberg, J., A note on normalization of biofluid 1D H-1-NMR data. *Metabolomics* **2008,** 4, (2), 114-121.
142. Dong, J. Y.; Cheng, K. K.; Xu, J. J.; Chen, Z.; Griffin, J. L., Group aggregating normalization method for the preprocessing of NMR-based metabolomic data. *Chemometrics and Intelligent Laboratory Systems* **2011,** 108, (2), 123-132.
143. van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Bmc Genomics* **2006,** 7.
144. Sakia, R. M., The Box-Cox Transformation Technique - a Review. *Statistician* **1992,** 41, (2), 169-178.
145. Jackson, T. E., *A Users guide to Principal Components.* Wiley: New York, 1991.
146. Beckwith-Hall, B. M.; Nicholson, J. K.; Nicholls, A. W.; Foxall, P. J.; Lindon, J. C.; Connor, S. C.; Abdi, M.; Connelly, J.; Holmes, E., Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chem Res Toxicol* **1998,** 11, (4), 260-72.
147. el-Deredy, W., Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review. *NMR Biomed* **1997,** 10, (3), 99-124.
148. Holmes, E.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Connelly, J. C.; Haselden, J. N.; Damment, S. J.; Spraul, M.; Neidig, P.; Nicholson, J. K., Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem Res Toxicol* **2000,** 13, (6), 471-8.
149. Pearson, K., On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine A: Physics of Condensed Matter: Defects and Mechanical Properties* **1901,** 2, (6), 559-572.
150. Daszykowski, M.; Kaczmarek, K.; Stanimirova, I.; Vander Heyden, Y.; Walczak, B., Robust SIMCA-bounding influence of outliers. *Chemometrics and Intelligent Laboratory Systems* **2007,** 87, (1), 95-103.
151. Daszykowski, M.; Serneels, S.; Kaczmarek, K.; Van Espen, P.; Croux, C.; Walczak, B., TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemometrics and Intelligent Laboratory Systems* **2007,** 85, (2), 269-277.

152. Verboven, S.; Hubert, M., LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems* **2005,** 75, (2), 127-136.
153. Kim, H. K.; Saifullah; Khan, S.; Wilson, E. G.; Kricun, S. D. P.; Meissner, A.; Goraler, S.; Deelder, A. M.; Choi, Y. H.; Verpoorte, R., Metabolic classification of South American Ilex species by NMR-based metabolomics. *Phytochemistry* **2010,** 71, (7), 773-784.
154. Cuperlovic-Culf, M.; Belacel, N.; Cuif, A. S.; Chute, I. C.; Ouellette, R. J.; Burton, I. W.; Karakach, T. K.; Walter, J. A., NMR metabolic analysis of samples using fuzzy K-means clustering. *Magnetic Resonance in Chemistry* **2009,** 47, S96-S104.
155. Mahle, D. A.; Anderson, P. E.; DelRaso, N. J.; Raymer, M. L.; Neuforth, A. E.; Reo, N. V., A generalized model for metabolomic analyses: application to dose and time dependent toxicity. *Metabolomics* **2011,** 7, (2), 206-216.
156. Webb, A., *Statistical Pattern Recognition*. John Wiley & Sons Ltd.: 2002.
157. Duda, R. O.; Hart, P. E.; Stork, D. G., *Pattern Classification*. Jogn Wiley & Sons Inc.: New York, 2000.
158. von Luxburg, U., A tutorial on spectral clustering. *Statistics and Computing* **2007,** 17, (4), 395-416.
159. Raman, B.; McKeown, C. K.; Rodriguez, M., Jr.; Brown, S. D.; Mielenz, J. R., Transcriptomic analysis of Clostridium thermocellum ATCC 27405 cellulose fermentation. *BMC Microbiol* **2011,** 11, 134.
160. Moulos, P.; Papadodima, O.; Chatziioannou, A.; Loutrari, H.; Roussos, C.; Kolisis, F. N., A transcriptomic computational analysis of mastic oil-treated Lewis lung carcinomas reveals molecular mechanisms targeting tumor cell growth and survival. *BMC Med Genomics* **2009,** 2, 68.
161. Hageman, J. A.; van den Berg, R. A.; Westerhuis, J. A.; Hoefsloot, H. C. J.; Smilde, A. K., Bagged K-means clustering of metabolome data. *Critical Reviews in Analytical Chemistry* **2006,** 36, (3-4), 211-220.
162. Holmes, E.; Cloarec, O.; Nicholson, J. K., Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: Application to HgCl2 toxicity. *Journal of Proteome Research* **2006,** 5, (6), 1313-1320.
163. Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J., Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Analytical Chemistry* **2005,** 77, (5), 1282-1289.
164. Holmes, E.; Loo, R. L.; Cloarec, O.; Coen, M.; Tang, H. R.; Maibaum, E.; Bruce, S.; Chan, Q.; Elliott, P.; Stamler, J.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K., Detection of urinary drug metabolite (Xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy. *Analytical Chemistry* **2007,** 79, (7), 2629-2640.
165. Smith, L. M.; Maher, A. D.; Cloarec, O.; Rantalainen, M.; Tang, H. R.; Elliott, P.; Stamler, J.; Lindon, J. C.; Holmes, E.; Nicholson, J. K., Statistical correlation and projection methods for improved information recovery from diffusion-edited NMR spectra of biological samples. *Analytical Chemistry* **2007,** 79, (15), 5682-5689.
166. Maher, A. D.; Cysique, L. A.; Brew, B. J.; Rae, C. D., Statistical Integration of (1)H NMR and MRS Data from Different Biofluids and Tissues Enhances Recovery of Biological Information from Individuals with HIV-1 infection. *Journal of Proteome Research* **2011,** 10, (4), 1737-1745.
167. Coen, M.; Hong, Y. S.; Cloarec, O.; Rhode, C. M.; Reily, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Heteronuclear H-1-P-31 statistical total correlation NMR spectroscopy of intact liver for metabolic biomarker assignment: Application to galactosamine-induced hepatotoxicity. *Analytical Chemistry* **2007,** 79, (23), 8956-8966.
168. Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M. D.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K., Cluster Analysis Statistical Spectroscopy Using Nuclear

Magnetic Resonance Generated Metabolic Data Sets from Perturbed Biological Systems. *Analytical Chemistry* **2009,** 81, (16), 6581-6589.
169. Wold, S.; Sjostrom, M.; Eriksson, L., PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001,** 58, (2), 109-130.
170. Hoskuldsson, A., PLS regression methods. *Journal of Chemometrics* **1988,** 2, 211-228.
171. Gu, H. W.; Pan, Z. Z.; Xi, B. W.; Asiago, V.; Musselman, B.; Raftery, D., Principal component directed partial least squares analysis for combining nuclear magnetic resonance and mass spectrometry data in metabolomics: Application to the detection of breast cancer. *Analytica Chimica Acta* **2011,** 686, (1-2), 57-63.
172. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **2002,** 16, (3), 119-128.
173. Coen, M.; Hong, Y. S.; Clayton, T. A.; Rohde, C. M.; Pearce, J. T.; Reily, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., The mechanism of galactosamine toxicity revisited; A metabonomic study. *Journal of Proteome Research* **2007,** 6, (7), 2711-2719.
174. Tapp, H. S.; Kemsley, E. K., Notes on the practical utility of OPLS. *Trac-Trends in Analytical Chemistry* **2009,** 28, (11), 1322-1327.
175. Kemsley, E. K.; Tapp, H. S., OPLS filtered data can be obtained directly from non-orthogonalized PLS1. *Journal of Chemometrics* **2009,** 23, (5-6), 263-264.
176. Martens, H.; Naes, T., *Multivariate Calibration (2nd edn.)*. Wiley: Chichester, 1989.
177. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; Jong, S. D.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimerics: Part B.* Elsevier: Amsterdam, 1998.
178. Fisher, R. A., Use of multiple measurements in taxonomic problems. *Ann.Eugen.* **1936,** 7, 179-188.
179. Krzanowski, W. J., *Principles of Multivariate Analysis (Revised edn)*. New York, 2000.
180. Xu, C. J.; Hoefsloot, H. C. J.; Smilde, A. K., To aggregate or not to aggregate high-dimensional classifiers. *Bmc Bioinformatics* **2011,** 12.
181. Schoonen, W. G. E. J.; Kloks, C. P. A. M.; Ploemen, J. P. H. T. M.; Smit, M. J.; Zandberg, P.; Horbach, G. J.; Mellema, J. R.; Thijssen-VanZuylen, C.; Tas, A. C.; van Nesselrooij, J. H. J.; Vogels, J. T. W. E., Uniform procedure of H-1 NMR analysis of rat urine and toxicometabonomics part II: comparison of NMR profiles for classification of hepatotoxicity. *Toxicological Sciences* **2007,** 98, (1), 286-297.
182. Norgaard, L.; Bro, R.; Westad, F.; Engelsen, S. B., A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics* **2006,** 20, (8-10), 425-435.
183. Blanchet, L.; Smolinska, A.; Attali, A.; Stoop, M. P.; Ampt, K. A.; van Aken, H.; Suidgeest, E.; Tuinstra, T.; Wijmenga, S. S.; Luider, T.; Buydens, L. M., Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics* **2011,** 12, (1), 254.
184. Abrahamsson, C.; Johansson, J.; Sparen, A.; Lindgren, F., Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemometrics and Intelligent Laboratory Systems* **2003,** 69, (1-2), 3-12.
185. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **1995,** 57, (1), 289-300.
186. Han, B.; Kang, H. M.; Eskin, E., Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers. *Plos Genetics* **2009,** 5, (4).
187. Broadhurst, D. I.; Kell, D. B., Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006,** 2, (4), 171-196.

188. Chadeau-Hyam, M.; Ebbels, T. M. D.; Brown, I. J.; Chan, Q.; Stemler, J.; Huang, C. C.; Daviglus, M. L.; Ueshima, H.; Zhao, L. C.; Holmes, E.; Nicholson, J. K.; Elliott, P.; De Iorio, M., Metabolic Profiling and the Metabolome-Wide Association Study: Significance Level For Biomarker Identification. *Journal of Proteome Research* **2010,** 9, (9), 4620-4627.
189. Wehrens, R.; Franceschi, P.; Vrhovsek, U.; Mattivi, F., Stability-based biomarker selection. *Analytica Chimica Acta* **2011,** 705, (1-2), 15-23.
190. Whitley, D., A Genetic Algorithm Tutorial. *Statistics and Computing* **1994,** 4, (2), 65-85.
191. Lloyd, G. R.; Wongravee, K.; Silwood, C. J. L.; Grootveld, M.; Brereton, R. G., Self Organising Maps for variable selection: Application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product. *Chemometrics and Intelligent Laboratory Systems* **2009,** 98, (2), 149-161.
192. Rajalahti, T.; Kvalheim, O. M., Multivariate data analysis in pharmaceutics: A tutorial review. *International Journal of Pharmaceutics* **2011,** 417, (1-2), 280-290.
193. Cao, D. S.; Wang, B.; Zeng, M. M.; Liang, Y. Z.; Xu, Q. S.; Zhang, L. X.; Li, H. D.; Hu, Q. N., A new strategy of exploring metabolomics data using Monte Carlo tree. *Analyst* **2011,** 136, (5), 947-954.
194. Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C., Elimination of uninformative variables for multivariate calibration. *Anal Chem* **1996,** 68, (21), 3851-8.
195. Daszykowski, M.; Wu, W.; Nicholls, A. W.; Ball, R. J.; Czekaj, T.; Walczak, B., Identifying potential biomarkers in LC-MS data. *Journal of Chemometrics* **2007,** 21, (7-9), 292-302.
196. Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H., Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems* **2006,** 84, (1-2), 69-74.
197. Gidskehaug, L.; Anderssen, E.; Alsberg, B. K., Cross model validation and optimisation of bilinear regression models. *Chemometrics and Intelligent Laboratory Systems* **2008,** 93, (1), 1-10.
198. Fernandez Pierna, J. A.; Abbas, O.; Baeten, V.; Dardenne, P., A Backward Variable Selection method for PLS regression (BVSPLS). *Anal Chim Acta* **2009,** 642, (1-2), 89-93.
199. Osborne, S. D.; Jordan, R. B.; Kunnemeyer, R., Method of wavelength selection for partial least squares. *Analyst* **1997,** 122, (12), 1531-1537.
200. Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B., Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy* **2000,** 54, (3), 413-419.
201. Rajalahti, T.; Arneberg, R.; Berven, F. S.; Myhr, K. M.; Ulvik, R. J.; Kvalheim, O. M., Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems* **2009,** 95, (1), 35-48.
202. Andersen, C. M.; Bro, R., Variable selection in regression-a tutorial. *Journal of Chemometrics* **2010,** 24, (11-12), 728-737.
203. de Haan, J. R.; Wehrens, R.; Bauerschmidt, S.; Piek, E.; van Schaik, R. C.; Buydens, L. M. C., Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* **2007,** 23, (2), 184-190.
204. Brereton, R. G., Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *Trac-Trends in Analytical Chemistry* **2006,** 25, (11), 1103-1111.
205. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., Assessment of PLSDA cross validation. *Metabolomics* **2008,** 4, (1), 81-89.
206. Faber, N. M.; Rajko, R., How to avoid over-fitting in multivariate calibration - The conventional validation approach and an alternative. *Analytica Chimica Acta* **2007,** 595, (1-2), 98-106.
207. Snee, R. D., Validation of regression models: Methods and examples. *Technometrics* **1977,** 19, (4), 415-428.

208. Kennard, R. W., Computer Aided Design of Experiments. *Technometrics* **1968,** 10, (2), 423-&.
209. Kennard, R. W.; Stone, L. A., Computer Aided Design of Experiments. *Technometrics* **1969,** 11, (1), 137-&.
210. Lindon, J. C.; Nicholson , J. K.; Holmes, E., *The Handbook of Metabonomics and Metabolomics*. Elsevier: Amsterdam, 2007.
211. Shawe-Taylor, J.; Cristianini, N., *Kernel Methods for Pattern Analysis*. Cambridge University Press: Canbridge, 2004.
212. Krooshof, P. W. T.; Ustun, B.; Postma, G. J.; Buydens, L. M. C., Visualization and Recovery of the (Bio)chemical Interesting Variables in Data Analysis with Support Vector Machine Classification. *Analytical Chemistry* **2010,** 82, (16), 7000-7007.
213. Gower, J. C.; Harding, S. A., Nonlinear Biplots. *Biometrika* **1988,** 75, (3), 445-455.
214. Walczak, B.; Massart, D. L., The radial basis functions - Partial least squares approach as a flexible non-linear regression technique. *Analytica Chimica Acta* **1996,** 331, (3), 177-185.
215. Walczak, B.; Massart, D. L., Application of Radial Basis Functions - Partial Least Squares to non-linear pattern recognition problems: Diagnosis of process faults. *Analytica Chimica Acta* **1996,** 331, (3), 187-193.
216. Fonville, J. M.; Bylesjo, M.; Coen, M.; Nicholson, J. K.; Holmes, E.; Lindon, J. C.; Rantalainen, M., Non-linear modeling of (1)H NMR metabonomic data using kernel-based orthogonal projections to latent structures optimized by simulated annealing. *Analytica Chimica Acta* **2011,** 705, (1-2), 72-80.
217. Vapnik, V., *Statistical Learning Theory*. John Willey & Sons: New York, 1998.
218. Lin, X. H.; Wang, Q. C.; Yin, P. Y.; Tang, L.; Tan, Y. X.; Li, H.; Yan, K.; Xu, G. W., A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics* **2011,** 7, (4), 549-558.
219. Lin, X.; Zhang, Y.; Ye, G.; Li, X.; Yin, P.; Ruan, Q.; Xu, G., Classification and differential metabolite discovery of liver diseases based on plasma metabolic profiling and support vector machines. *Journal of Separation Science* **2011,** 34, (21), 3029-36.
220. Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S., Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics* **2011,** 27, (11), 1537-45.
221. Yetukuri, L.; Tikka, J.; Hollmen, J.; Oresic, M., Functional prediction of unidentified lipids using supervised classifiers. *Metabolomics* **2010,** 6, (1), 18-26.
222. Henneges, C.; Bullinger, D.; Fux, R.; Friese, N.; Seeger, H.; Neubauer, H.; Laufer, S.; Gleiter, C. H.; Schwab, M.; Zell, A.; Kammerer, B., Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. *BMC Cancer* **2009,** 9.
223. Hall, D. L.; Llinas, J., An introduction to multisensor data fusion. *Proceedings of the Ieee* **1997,** 85, (1), 6-23.
224. Roussel, S.; Bellon-Maurel, V.; Roger, J. M.; Grenier, P., Fusion of aroma, FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. *Chemometrics and Intelligent Laboratory Systems* **2003,** 65, (2), 209-219.
225. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van-der Vat, B. J. C.; Jellema, R. H., Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry* **2005,** 77, (20), 6729-6736.
226. Hall, D. L.; McMullen, S. A. H., *Mathematical techniques in multisensor data fusion*. Boston, 2004.
227. Hall, D. L.; Garga, A. K., Pitfalls in Data Fusion (and How to Avoid Them). *In Proceedings of the 2nd International Conference on Information Fusion – FUSION'99* **1999,** 1, 429-436.
228. Richards, S. E.; Dumas, M. E.; Fonville, J. M.; Ebbels, T. M. D.; Holmes, E.; Nicholson, J. K., Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework. *Chemometrics and Intelligent Laboratory Systems* **2010,** 104, (1), 121-131.

229. Yu, C.; Tranchevent, L. C.; De Moor, B.; Moreau, Y., *Kernel-based Data Fusion for Machine Learning. Methods and applications in Bioinformatics and Text mining.* Springer: Berlin, 2011.
230. Barton, R. H., A decade of advances in metabonomics. *Expert Opinion on Drug Metabolism and Toxicology* **2011,** 7, (2), 129-136.
231. van Gool, A. J.; Henry, B.; Sprengers, E. D., From biomarker strategies to biomarker activities and back. *Drug Discovery Today* **2010,** 15, (3-4), 121-6.
232. Wishart, D. S., Proteomics and the human metabolome project. *Expert Review of Proteomics* **2007,** 4, (3), 333-335.
233. Jukarainen, N. M.; Korhonen, S. P.; Laakso, M. P.; Korolainen, M. A.; Niemitz, M.; Soininen, P. P.; Tuppurainen, K.; Vepsalainen, J.; Pirttila, T.; Laatikainen, R., Quantification of H-1 NMR spectra of human cerebrospinal fluid: a protocol based on constrained total-line-shape analysis. *Metabolomics* **2008,** 4, (2), 150-160.
234. Wishart, D. S.; Lewis, M. J.; Morrissey, J. A.; Flegel, M. D.; Jeroncic, K.; Xiong, Y. P.; Cheng, D.; Eisner, R.; Gautam, B.; Tzur, D.; Sawhney, S.; Bamforth, F.; Greiner, R.; Li, L., The human cerebrospinal fluid metabolome. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **2008,** 871, (2), 164-173.
235. Griffin, J. L.; Nicholls, A. W., Metabolomics as a functional genomic tool for understanding lipid dysfunction in diabetes, obesity and related disorders. *Pharmacogenomics* **2006,** 7, (7), 1095-1107.
236. Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W. L.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J., Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics (vol 8, pg 1439, 2002). *Nature Medicine* **2003,** 9, (4), 477-477.
237. Holmes, E.; Foxall, P. J.; Nicholson, J. K., Proton NMR analysis of plasma from renal failure patients: evaluation of sample preparation and spectral-editing methods. *Journal of Pharmaceutical and Biomedical Analysis* **1990,** 8, (8-12), 955-8.
238. Hyndman, M. E.; Mullins, J. K.; Bivalacqua, T. J., Metabolomics and bladder cancer. *Urol Oncol* **2011,** 29, (5), 558-61.
239. Liu, G. M.; Wang, Y.; Wang, Z. S.; Cai, J. Y.; Lv, X. Z.; Zhou, A. G., Metabolomic studies on the biochemical profile of urine from rats with acute cysteamine supplementation. *Metabolomics* **2011,** 7, (4), 536-541.
240. Ringeissen, S.; Connor, S. C.; Thakkar, H.; Sweatman, B. C.; Hodson, M. P.; Hutton, K. A.; Kenny, S. P.; McGill, P.; Nunez, D. J.; Haselden, J. N.; Waterfield, C. J., Identification of potential non-invasive biomarkers of peroxisome proliferation in the rat. *Toxicology* **2004,** 194, (3), 246-247.
241. Fonville, J. M.; Maher, A. D.; Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Evaluation of Full-Resolution J-Resolved (1)H NMR Projections of Biofluids for Metabonomics Information Retrieval and Biomarker Identification. *Analytical Chemistry* **2010,** 82, (5), 1811-1821.
242. Serkova, N. J.; Standiford, T. J.; Stringer, K. A., The Emerging Field of Quantitative Blood Metabolomics for Biomarker Discovery in Critical Illnesses. *American Journal of Respiratory and Critical Care Medicine* **2011,** 184, (6), 647-655.
243. Carraro, S.; Rezzi, S.; Reniero, F.; Heberger, K.; Giordano, G.; Zanconato, S.; Guillou, C.; Baraldi, E., Metabolomics applied to exhaled breath condensate in childhood asthma. *American Journal of Respiratory and Critical Care Medicine* **2007,** 175, (10), 986-990.
244. Jiang, N.; Yan, X.; Zhou, W.; Zhang, Q.; Chen, H.; Zhang, Y.; Zhang, X., NMR-based metabonomic investigations into the metabolic profile of the senescence-accelerated mouse. *Journal of Proteome Research* **2008,** 7, (9), 3678-86.

245. Aymerich, F. X.; Alonso, J.; Cabanas, M. E.; Comabella, M.; Sobrevilla, P.; Rovira, A., Decision tree based fuzzy classifier of (1)H magnetic resonance spectra from cerebrospinal fluid samples. *Fuzzy Sets and Systems* **2011,** 170, (1), 43-63.
246. Coen, M.; O'Sullivan, M.; Bubb, W. A.; Kuchel, P. W.; Sorrell, T., Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clinical Infectious Diseases* **2005,** 41, (11), 1582-90.
247. Subramanian, A.; Gupta, A.; Saxena, S.; Gupta, A.; Kumar, R.; Nigam, A.; Kumar, R.; Mandal, S. K.; Roy, R., Proton MR CSF analysis and a new software as predictors for the differentiation of meningitis in children. *NMR in Biomedicine* **2005,** 18, (4), 213-225.
248. Fadil, H.; Kelley, R. E.; Gonzalez-Toledo, E., Differential diagnosis of multiple sclerosis. *International Review of Neurobiology* **2007,** 79, 393-422.
249. Rolak, L. A.; Fleming, J. O., The differential diagnosis of multiple sclerosis. *Neurologist* **2007,** 13, (2), 57-72.
250. Gasperini, C., Differential diagnosis in multiple sclerosis. *Neurological Sciences* **2001,** 22 Suppl 2, S93-7.
251. Alpini, D.; Caputo, D.; Pugnetti, L.; Giuliano, D. A.; Cesarani, A., Vertigo and multiple sclerosis: aspects of differential diagnosis. *Neurological Sciences* **2001,** 22 Suppl 2, S84-7.
252. Scolding, N., The differential diagnosis of multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry* **2001,** 71 Suppl 2, ii9-15.
253. Schaffler, N.; Kopke, S.; Winkler, L.; Schippling, S.; Inglese, M.; Fischer, K.; Heesen, C., Accuracy of diagnostic tests in multiple sclerosis--a systematic review. *Acta Neurologica Scandinavica* **2011,** 124, (3), 151-64.
254. Tumani, H.; Hartung, H. P.; Hemmer, B.; Teunissen, C.; Deisenhammer, F.; Giovannoni, G.; Zettl, U. K.; Grp, B. S., Cerebrospinal fluid biomarkers in multiple sclerosis. *Neurobiology of Disease* **2009,** 35, (2), 117-127.
255. Mader, I.; Roser, W.; Kappos, L.; Hagberg, G.; Seelig, J.; Radue, E. W.; Steinbrich, W., Serial proton MR spectroscopy of contrast-enhancing multiple sclerosis plaques: absolute metabolic values over 2 years during a clinical pharmacological study. *AJNR: American Journal of Neuroradiology* **2000,** 21, (7), 1220-7.
256. Wattjes, M. P.; Harzheim, M.; Lutterbey, G. G.; Bogdanow, M.; Schild, H. H.; Traber, F., High field MR imaging and 1H-MR spectroscopy in clinically isolated syndromes suggestive of multiple sclerosis: correlation between metabolic alterations and diagnostic MR imaging criteria. *Journal of Neurology* **2008,** 255, (1), 56-63.
257. Wattjes, M. P.; Harzheim, M.; Lutterbey, G. G.; Klotz, L.; Schild, H. H.; Traber, F., Axonal damage but no increased glial cell activity in the normal-appearing white matter of patients with clinically isolated syndromes suggestive of multiple sclerosis using high-field magnetic resonance spectroscopy. *AJNR: American Journal of Neuroradiology* **2007,** 28, (8), 1517-22.
258. Zaaraoui, W.; Rico, A.; Audoin, B.; Reuter, F.; Malikova, I.; Soulier, E.; Viout, P.; Le Fur, Y.; Confort-Gouny, S.; Cozzone, P. J.; Pelletier, J.; Ranjeva, J. P., Unfolding the long-term pathophysiological processes following an acute inflammatory demyelinating lesion of multiple sclerosis. *Magnetic Resonance Imaging* **2010,** 28, (4), 477-486.
259. Blinkenberg, M.; Mathiesen, H. K.; Tscherning, T.; Jonsson, A.; Svarer, C.; Holm, S.; Sellebjerg, F.; Paulson, O. B.; Hanson, L. G.; Sorensen, P. S., Cerebral metabolism, magnetic resonance spectroscopy and cognitive dysfunction in early multiple sclerosis: an exploratory study. *Neurological Research* **2012,** 34, (1), 52-8.

260. Davie, C. A.; Hawkins, C. P.; Barker, G. J.; Brennan, A.; Tofts, P. S.; Miller, D. H.; McDonald, W. I., Serial proton magnetic resonance spectroscopy in acute multiple sclerosis lesions. *Brain* **1994,** 117 (Pt 1), 49-58.
261. t'Hart, B. A.; Vogels, J. T.; Spijksma, G.; Brok, H. P.; Polman, C.; van der Greef, J., 1H-NMR spectroscopy combined with pattern recognition analysis reveals characteristic chemical pattern in urines of MS patients and non-human primates with MS-like disease. *Journal of the Neurological Sciences* **2003** 212, (1-2), 21-30.
262. Smolinska, A.; Posma, J. M.; Blanchet, L.; Ampt, K. A.; Attali, A.; Tuinstra, T.; Luider, T.; Doskocz, M.; Michiels, P. J.; Girard, F. C.; Buydens, L. M.; Wijmenga, S. S., Simultaneous analysis of plasma and CSF by NMR and hierarchical models fusion. *Analytical and bioanalytical chemistry* **2012,** 403, (4), 947-59.
263. Lynch, J.; Peeling, J.; Auty, A.; Sutherland, G. R., Nuclear magnetic resonance study of cerebrospinal fluid from patients with multiple sclerosis. *Can J Neurol Sci* **1993,** 20, (3), 194-8.
264. Simone, I. L.; Federico, F.; Trojano, M.; Tortorella, C.; Liguori, M.; Giannini, P.; Picciola, E.; Natile, G.; Livrea, P., High resolution proton MR spectroscopy of cerebrospinal fluid in MS patients. Comparison with biochemical changes in demyelinating plaques. *Journal of the Neurological Sciences* **1996,** 144, (1-2), 182-190.
265. Simone, I. L.; Tortorella, C.; Federico, F.; Liguori, M.; Lucivero, V.; Giannini, P.; Carrara, D.; Bellacosa, A.; Livrea, P., Axonal damage in multiple sclerosis plaques: a combined magnetic resonance imaging and H-1-magnetic resonance spectroscopy study. *Journal of the Neurological Sciences* **2001,** 182, (2), 143-150.
266. Aasly, J.; Garseth, M.; Sonnewald, U.; Zwart, J. A.; White, L. R.; Unsgard, G., Cerebrospinal fluid lactate and glutamine are reduced in multiple sclerosis. *Acta Neurologica Scandinavica* **1997,** 95, (1), 9-12.
267. Lutz, N. W.; A., V.; Malikova, I.; Confort-Gouny, S.; Ranjeva, J. P.; Cozzone, P. J., A branched-chain organic acid linked to multiple sclerosis:First identification by NMR spectroscopy of CSF. *Biochemical and Biophysical Research Communications* **2007,** 354, (1), 16-164.

CHAPTER 2

THE IMPACT OF DELAYED STORAGE ON THE MEASURED PROTEOME AND METABOLOME OF HUMAN CEREBROSPINAL FLUID (CSF)

T. Rosenling, M. P. Stoop, **A. Smolinska**, B. Muilwijk, L. Coulier, S. Shi, A. Dane, C. Christin, F. Suits, P. L. Horvatovich, S. S. Wijmenga, L. M.C. Buydens, R. Vreeken, T. Hankemeier, A. J. van Gool, T. M. Luider and R. Bischoff

Clinical Chemistry (2011), 57(12), pp. 1703-11

ABSTRACT

BACKGROUND: Cerebrospinal fluid is in close contact with diseased areas in neurological disorders and is therefore an important source of material in the search for molecular biomarkers. CSF is withdrawn from patients in a clinical setting where sample handling might not always be adequate in view of proteomics and metabolomics studies. To study the effect of a time delay between sampling and freezing, we have performed a combined proteomics and metabolomics study.

METHODS: CSF was left for 0, 30 and 120 min at room temperature directly after sample collection and centrifugation/removal of the cell pellet. CSF samples were analyzed at five separate laboratories using five different analytical platforms. The techniques used for proteome analysis were nanoLC Orbitrap-MS and chipLC QTOF-MS after tryptic digestion. Metabolome analysis was performed by NMR, GC-MS, and LC-MS. Targeted analyses of Cystatin C and Albumin were performed by LC-MS/MS in the selected reaction monitoring mode.

RESULTS: Our results show that storage of CSF at room temperature after centrifugation does not lead to significant changes in the measured proteome and metabolome except for two peptides and one metabolite 2,3,4-trihydrobutanoic acid among 5780 identified peptides and 93 identified metabolites. A sensitive protein stability marker, Cystatin C, was not affected.

CONCLUSIONS: The measured proteome/metabolome profile of centrifuged, human CSF with all cells removed is stable at room temperature for up to two hours. This gives the laboratory personnel at the collection site sufficient time to aliquot samples before freezing and storage at -80 °C.

2.1 INTRODUCTION

Conditions during the journey of a biological sample from the clinical collection site to the analytical research laboratory might not always be adequate for subsequent proteomics and metabolomics analyses, especially in cases where the sample collection was not originally performed with these large-scale analyses in mind. In order to detect reliable molecular biomarkers it is imperative to handle biological fluids according to standardized procedures and to evaluate the effect of pre-analytical parameters on the final result to avoid artifacts [1, 2]. Earlier studies on urine, plasma and cerebrospinal fluid (CSF) have shown that sample handling can affect the stability of proteins as well as metabolites [3-11]. Sample handling according to standardized procedures is also important when trying to compare results between different laboratories [12-14]. In the search for molecular biomarkers related to disorders of the central nervous system, CSF is the most promising bio-fluid because of its close contact to the affected tissue [13, 15-20].
In this study we analyzed a set of human CSF samples in order to assess protein and metabolite stability at room temperature after a low-speed centrifugation step to remove cells. To cover a wide range of proteins and metabolites, the results from a number of analytical platforms comprising LC-MS, GC-MS and NMR were combined.

2.2 MATERIALS and METHODS

2.2.1 Sample set

Six human CSF samples were obtained from the Department of Neurology at the Erasmus University Medical Center (Rotterdam, The Netherlands). The CSF samples were collected as part of routine clinical examination of patients with various symptoms (Table 1). All samples were withdrawn via lumbar puncture between the 3rd and 4th lumbar vertebrae using a Spinocan needle (0.90 × 88 mm). The Medical Ethical Committee of the Erasmus University Medical Center (Rotterdam, The Netherlands) approved the study protocol and all patients gave their informed consent. Samples were centrifuged (10 min at 956 g) within five minutes after collection to remove cells. Aliquots were directly snap-frozen in liquid nitrogen or left at room temperature for 30 and 120 min before snap freezing and storage at -80 °C. Routine CSF diagnostics including total protein and albumin concentration measurements as well as intrathecal cell count were performed and absence of hemoglobin and apolipoprotein B100 was assured to eliminate the possibility that samples were contaminated with blood. Sample H1 was analyzed by the chipLC QTOF-MS and the nanoLC Orbitrap-MS/MS platforms only. Samples H2 - H6 were analyzed by all platforms. Protein digestion for proteomic analysis was performed as previously described [21]. Before analysis on each platform the samples were exposed to two freeze-thaw cycles.

Table 1. Description of CSF samples used for stability studies.[1]

Sample	Age	Gender	Diagnosis	Protein conc. (mg/L)	Albumine conc. (mg/L) Clinic	Albumine conc. (mg/L) SRM[2]	# Cells/ µL (after centrifugation)
H1	49	M	Migraine	415	193	192.3	0
H2	56	M	Idiopathic intracranial hypertension	472	247	237.5	0
H3	69	F	Headache	395	236	221.0	0
H4	48	M	Idiopathic intracranial hypertension	436	241	225.6	0
H5	29	F	Clinical isolated syndrome (Neuromyelitis optica)	387	226	213.9	0
H6	38	F	Epilepsy	381	184	194.2	0

[1] Sample H1 was only analyzed with respect to proteomics

[2] Average over 3 time points (supplementary Table S4).

2.2.2 ChipLC QTOF-MS proteomic analysis

Half a microliter trypsin-digested CSF was randomly injected in quintuplicate with 0.5 µL digested QC samples (pooled CSF spiked with cytochrome C; Fluka, part # 30396, final concentration: 375 fmol/µL) and blanks injected between every 10th sample for LC-MS analysis on an Agilent chipLC QTOF-MS system as reported previously [21]. Enrichment and separation was done using an LC chip (G4240-63001 SPQ110, Agilent Technologies [separation column: 150 mm × 75 µm Zorbax 300SB-C18, 5 µm; trap column: 160 nL Zorbax 300SB-C18, 5 µm]). The LC separations were carried out as described earlier using the following gradient: 80 min linear gradient from 3 to 40% B;

10 min linear gradient from 40 to 50% B; 10 min linear gradient from 50 to 3% B [21]. MS analysis was performed under the following conditions; mass range: 200-2000 m/z in profile mode, acquisition rate: 1 spectrum/s, fragmentor voltage: 175 V, skimmer voltage: 65 V, OCT 1 RF Vpp: 750 V. The spray voltage was ~1800 V and the drying gas (N_2) was 6 L/min at a temperature of 325 ºC. Mass correction was done for each spectrum using internal standards (methyl stearate m/z: 299.294457 and HP-1221 m/z: 1221.990637) evaporating from a wetted wick inside the spray chamber. Reproducibility was monitored on selected cytochrome C peaks in the QC samples. Mass difference between theoretical and measured values was within +/- 4 ppm. The selected peaks showed a peak area RSD of less than +/- 20% and a retention time (RT) RSD of less than 2%.

Data was processed using a pipeline developed in C++ as previously described [21, 22]. MzData.XML data were converted to ASCII format over a mass range of 200 to 1600 m/z (no multiply charged peptide ions were detected outside this range), a retention time range of 3 to 80 min (peptide elution range) and an intensity threshold of 300 counts. A double cross validated Nearest Shrunken Centroid (NSC) algorithm was applied to the complete peak matrix; the NSC comparison gives a cross validation error between 0 and 1 depending on the shrinkage value, where 1 implies that class assignment is incorrect, 0.5 that class assignment is random and 0 that class assignment is correct [23]. NSC selected features were compared by univariate statistical analysis (Student's t-test with Bonferroni correction for multiple comparisons) and ANOVA (Microsoft Excel 2007 and SPSS 16.0). Features were considered significantly different based on a *p*-value below 0.05 (T0 vs. T120 and T0 vs. T30) in at least five out of six samples (T0 vs. T120 and T0 vs. T30). Each discriminatory feature was analyzed by targeted tandem MS for identification. Principal component analysis (PCA) [24] was applied to the complete peak matrix (10 000 peaks) as well as to the NSC-selected features (MatLab, R2009a). For visualization, box and whisker plots were created in Origin 7.0.

2.2.3 Nano LC ORBITRAP-MS/MS SHOTGUN proteomics analysis

Trypsin-digested CSF samples were injected in random order and analyzed by MS/MS (shotgun approach) on an Ultimate 3000 nano LC system (Dionex, the Netherlands) online coupled to a hybrid linear ion trap/Orbitrap mass spectrometer (LTQ Orbitrap XL; Thermo Fisher Scientific, Bremen, Germany) as previously described [21].

Data files were analyzed and pre-processed using the Progenesis LC-MS software package (Nonlinear Dynamics, United Kingdom). Retention times were aligned and the intensities of the ions were normalized. To assess inter-patient variability all identified peaks were analyzed by PCA. All identified peaks were also analyzed by the NSC algorithm for classification [23]. Peptides were analyzed for differential abundance between the groups by ANOVA. *P*-values below 0.01 were considered significant.

All MS/MS spectra were searched against the UniProt/SwissProt database (version 57.6, taxonomy: *Homo sapiens*, 20070 sequences) using Mascot (version 2.2.06). Search parameters were; parent ion tolerance: 2 ppm, amino acid modifications: carbamidomethylation of cysteine (fixed) and oxidation of methionine (variable).

2.2.4 Nano LC-MS/MS Analyses in the selected reaction monitoring (SRM) mode

Trypsin-digested CSF samples were spiked with known concentrations of stable isotope-labeled peptide standards corresponding to sequences 427-434 (FQNALLVR) of human serum albumin and 52-62 (ALDFAVGEYNK) of human cystatin C for quantitation by Selected Reaction Monitoring (SRM) (supplementary Table S1 and S2).

Chromatographic separation of spiked CSF digests, was performed on an Ultimate 3000 nano LC system (Dionex). One microliter of spiked CSF digest was loaded onto a C18 trap column (PepMap C18, 300 µm ID x 5mm, 5 µm particle size and 100 Å pore size; Dionex, the Netherlands) and washed for 5 min at a flow rate of 20 µL/min 0.1% TFA in H_2O. Next, the trap column was switched in line with the analytical column (PepMap C18, 75 µm ID x 150 mm, 3 µm particle size and 100 Å pore size; Dionex). Peptides were eluted at a flow rate of 300 nL/min with the following gradient: 0-45% solvent B in 30 min, solvent A (H_2O/acetonitrile (ACN) 98/2 (v/v), 0.1% formic acid (FA)) and solvent B (H_2O/ACN 20/80 (v/v), 0.1% FA). The separation of the peptides was monitored at 214 nm.

SRM detection was performed by means of a triple quadrupole tandem mass spectrometer (4000 QTRAP; AB SCIEX, Concord, Ontario, Canada) in the positive ion mode. As technical control for the measurements, a single spiked CSF digest was measured after every 6th run. A technical control for the enzymatic digestion (one sample digested at three separate times) was included in this quantitative analysis. Data Analysis was performed using the SRM data analysis program Skyline (version 0.7) [25], using the ratio of the analyte peptide to the known concentration of the spiked isotopically labelled internal peptide standard to calculate the concentrations of the original peptides. For the cystatin C peptide the average of both fragment ion ratios was used for the determination of the protein concentration. A paired, two-sided t-test was used to test for differences in peptide concentrations between time points.

2.2.5 GC-MS metabolomics analysis

CSF samples were treated with an oximation reagent followed by silylation prior to GC-MS analysis [21, 26]. Each sample was injected twice in random order and analyzed on an Agilent 6890 gas chromatograph coupled to an Agilent 5973 quadrupole mass spectrometer as described earlier [21].

Peaks were characterized by retention time and m/z ratio and identified by comparison with a spectral data base (TNO) [21]. All detected metabolites were analyzed by PCA. A two tailed Student's t-test was applied to all known metabolites (T0 vs. T30 and T0 vs. T120). Metabolites with a *p*-value below 0.05 were considered significantly affected by storage time.

2.2.6 NMR metabolomics analysis

Samples were randomized prior to sample preparation and analysis. Fifty microliters of CSF were diluted in 200 µL of D_2O (99.99 % D). Twenty-five µL of 8.8 mM TSP-d_4 (3-(Trimethylsilyl)propionic acid-d_4 sodium salt, 99 % D) stock solution in D_2O were added to 250 µL CSF to a final concentration of 0.8 mM TSP as internal standard and as chemical shift reference (δ0.00). The TSP-d_4 stock solution was prepared from dry TSP-d_4. The pH was adjusted (7.0 – 7.1) by adding phosphate buffer (9.7 µL of a 1 M stock solution) to a final concentration of 35 mM [27]. Finally the sample (284.7 µL) was

transferred to a SHIGEMI microcell NMR tube for measurements. Each sample was analyzed once.

1D ^{1}H NMR spectra were acquired on an 800 MHz Inova (Varian) system equipped with a 5 mm triple-resonance, Z-gradient HCN cold-probe. Suppression of water was achieved using WATERGATE (delay: 85 μs) [28]. For each spectrum 256 scans of 18000 data points were accumulated with a spectral width of 9000 Hz. The acquisition time for each scan was 2 s. An 8 s relaxation delay was employed between scans. Prior to spectral analysis, all acquired Free Induction Decays (FIDs) were zero-filled to 32000 data points, multiplied with a 0.3 Hz line broadening function, Fourier transformed and manually phased. Calibration of the chemical shift scale was done on the external reference standard TSP-d_4 by using ACD/SpecManager software (Advanced Chemistry Development Inc., Toronto, Canada). Spectra were transformed to MatLab, version 7.6 (R2008b) (Mathworks, Natick, MA) for further analysis.

NMR spectral data was preprocessed by baseline correction using the Asymmetric Least Squares method [29] and aligned with the Correlation Optimized Warping (COW) method [30]. Each spectrum was divided (along the chemical shift axis) into equally sized bins (0.04 ppm) and each data point was averaged over each bin. The areas of the bins were summed to provide an integral so that the intensities of the peaks in such defined spectral regions were extracted. Each NMR spectrum was reduced to 210 variables, calculated by integrating regions of equal width (0.04 ppm) corresponding to the regions of δ0.7-9. To remove effects of variation in water resonance suppression, spectral regions between δ4.4-5.4 were removed. All spectra thus reduced were normalized to unit area.

The data was further processed by supervised vast scaling, in order to determine group-specific scaling factors [31]. To visualize possible systematic variation, grouping, trends and outliers, PCA was applied to the entire data set. To remove biological (patient-to-patient) variation, data was further mean-centered per patient and vast-scaled (see supplementary Fig. S1).

2.2.7 LC-MS/MS amino acid analysis

CSF samples were prepared in triplicate as previously described (21). One microliter of each reaction mixture was injected in duplicate on an ACQUITY UPLC™ system (Waters Chromatography B.V., Etten-Leur, The Netherlands) coupled to a Quattro Premier Xe tandem quadrupole mass spectrometer (Waters Corporation) operated under the MassLynx data acquisition software (version 4.1; Waters). Quantification and pre-analysis of the data was done using LC-QuanLynx (Waters) and Microsoft Excel 2003, respectively. The complete set of amino acids in all samples was examined by PCA. The data was analyzed by a two-tailed Student's t-test (T0 vs. T30 and T0 vs. T120) and amino acids with *p*-values below 0.05 were considered discriminatory.

2.3 RESULTS

2.3.1 Proteomics analysis

The Orbitrap-MS/MS shotgun analysis resulted in a list of 55421 peaks out of which 5780 peptides were identified. All identified peptides from the Orbitrap-MS/MS data and the 10000 most intense QTOF-MS peaks (complete peak matrix) were used for unsupervised multivariate statistical analysis (PCA). No trend with respect to delay before storage was visible (Fig. 1A and 2A). PCA showed that biological variation was more prominent than the effect of delay time, since data points clustered according to the individual patients rather than according to time points (Fig. 1B-D and 2B-D).

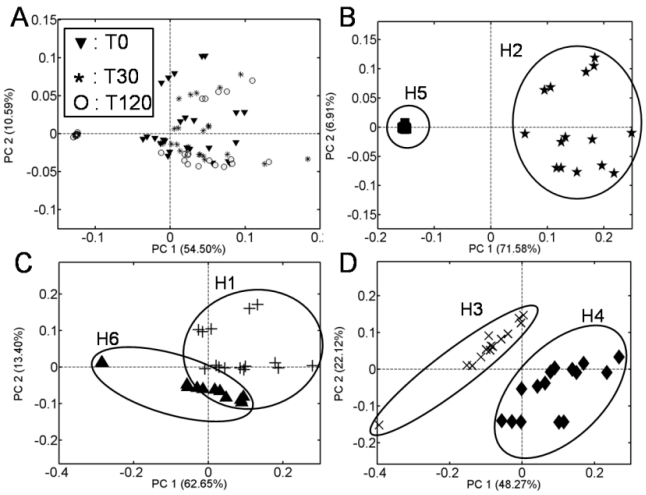


Figure 1. Multivariate statistical analysis (PCA) of the 10.000 most intense peaks selected from the chipLC QTOF-MS proteomic data (quintuplicate sample analysis). There is no separation based on time between sampling and freezing (T0 [▼] / T30 [*] / T120 [O]), while data from individual samples cluster together indicating that the inter-individual differences are larger than those related to time. (A) All samples. (B) Samples H2 (*) and H5 (■). (C) Samples H1 (+) and H6 (▲). (D) Samples H3 (x) and H4 (♦).

Comparison of the Orbitrap-MS/MS data by ANOVA according to delay time resulted in only 56 peaks with a *p*-value below 0.01, which is well below 554, the number of peaks

that would receive this *p*-value when comparing samples containing no differences (null hypothesis). NSC analysis pointed also to only random differences between the time groups with a cross validation error of 0.5. This lead to the conclusion that there was no significant discrimination between the samples stored at -80 °C immediately after centrifugation and samples left at room temperature for 30 or 120 min prior to being frozen and stored based on the observed peptides.

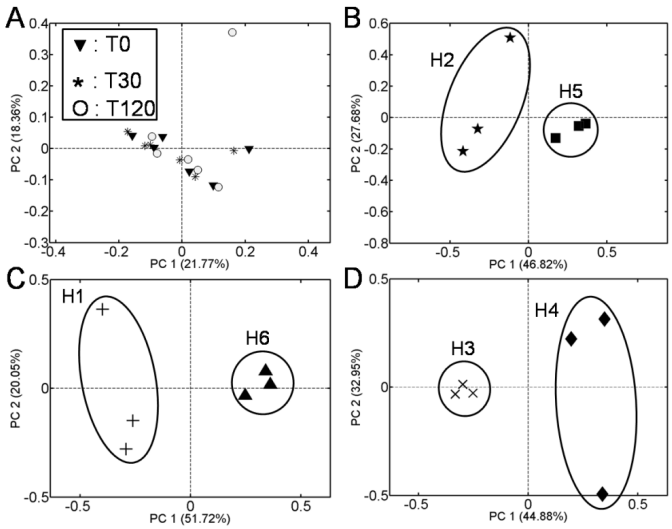


Figure 2. Multivariate statistical analysis (PCA) of 5780 identified peaks from nanoLC Orbitrap-MS/MS proteomic data (single sample analyses). There is no separation based on time between sampling and freezing (T0 [▼] / T30 [*] / T120 [O]), while data from individual samples cluster together indicating that the inter-individual differences are larger than those related to time. (A) All samples, (B) Samples H2 [*]) and H5 (■). (C) Samples H1 (+) and H6 (▲). (D) Samples H3 (x) and H4 (♦)

NSC analysis of the QTOF-MS data confirmed that differences between T0 and T120 were random, with a double cross validation error of 0.5. Comparison of T0 versus T30 by NSC reached a minimal average cross validation error of 0.34. PCA on the NSC-

selected peaks (T0 vs. T30) from the QTOF data showed no clear discrimination but a weak tendency of clustering according to time groups (Fig. 3A).

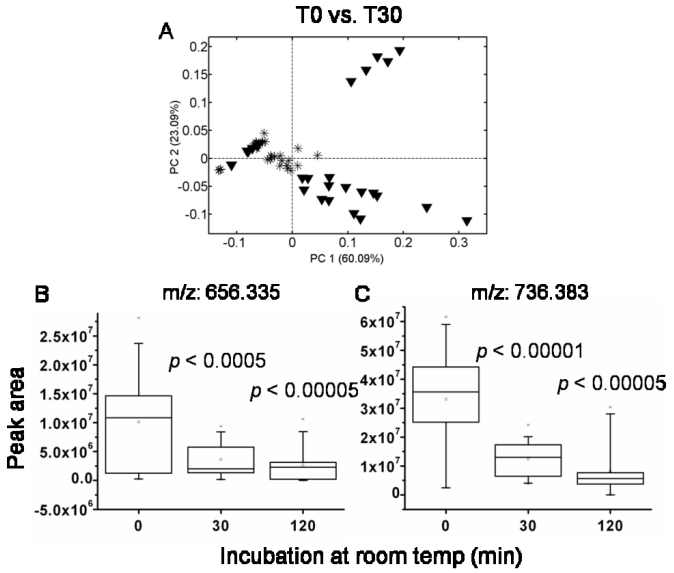


Figure 3. Multivariate statistical analysis by PCA based on NSC-selected peaks derived from chipLC QTOF-MS proteomics data (T0 [▼] vs. T30 [*]). (B and C) and univariate statistical analysis of two peaks that decreased significantly with respect to delay time between CSF sampling and freezing at room temperature. Data are represented as box and whisker plots with significant p-values marked ($p < 5 \times 10^{-5}$) (T0 vs. T30 and T0 vs. T120). (A) PCA of NSC-selected peaks for T0 vs. T30. (B) Peak at m/z: 656.335 that decreased significantly after 30 and 120 min at room temperature. (C) Peak at m/z 736.383 that decreased significantly after 30 and 120 min at room temperature. The statistical analysis was based on two-tailed Students t-tests with Bonferroni correction of the combined data from five repetitive analyses of six human CSF samples (H1-H6).

The concentration of two proteins in CSF, albumin and cystatin C, were determined by targeted mass spectrometric analysis in the SRM mode. These proteins are exemplary of the vast majority of CSF proteins, which we found to remain unchanged after 120 min at room temperature. Albumin was chosen as it represents the largest part of CSF total

protein, a parameter that is often used in CSF-based clinical diagnostics, and Cystatin C was chosen as a protein that is sensitive to storage conditions[1, 2] . Concentrations of cystatin C and albumin were calculated based on the measured ratios of the corresponding spiked isotopically labeled internal peptide standards to their biological counterparts, confirming that variation between different time points was not statistically significant (supplementary Tables S3 and S4). The measured concentrations of both proteins were both found to agree with reported CSF concentrations [32, 33].The albumin concentrations measured by SRM were also in agreement with albumin concentrations measured by standard clinical chemistry techniques (Table 1). Additionally, trypsin cleavage efficiency was assessed by monitoring the release of a tag from the lysine end of the cystatin C peptide during the regular digestion procedure. After overnight digestion the entire peak in the LC-MS data corresponding to the peptide including the tag had completely disappeared, indicating that complete digestion had taken place. The observed relative standard deviation for the cystatin C measurements was below 10% and those for albumin were less than 4%. Technical variability with sample pre-treatment was below 4% and without sample pre-treatment less than 2% (Table 2).

Table 2. Relative standard deviation (RSD) of SRM measurements of cystatin C and albumin. Relative standard deviations of protein concentrations in individual patients are slightly higher than technical controls (measuring a single sample multiple times). Average RSD for cystatin C in the 6 patient samples ± 2SD = 6.25 ± 4.99. Average RSD for albumin in the 6 patient samples ± 2SD = 2.40 ± 0.66.

Sample	No. of samples	RSD cystatin C (%)	RSD albumin (%)
H 1	3	5.20	2.19
H 2	3	6.62	1.78
H 3	3	7.72	3.13
H 4	3	7.55	3.05
H 5	3	9.15	2.59
H 6	3	2.72	1.51
Technical variation sample pre-treatment	3	3.74	2.34
Technical variation mass spectrometry measurement	4	0.82	1.74

2.3.2 Metabolomics analysis

GC-MS analysis quantified 88 metabolites, of which 67 were assigned to known compounds based on spectral libraries. This analysis was complemented by targeted LC-MS of 19 natural amino acids. NMR analysis identified and quantified 51 metabolites. Thirteen of the metabolites were detected by all three methods, 24 were detected by GC-MS and NMR, 14 by GC-MS and LC-MS and 16 by NMR and LC-MS. In total 93 unique identified metabolites were quantified. PCA of the data from the different analytical platforms showed that clustering occurs primarily according to the individual patients rather than to the time points when all data are considered (Fig. 4A-C). Meancentering the NMR data per patient and vast scaling showed further that there is no variation in the metabolome according to delay time (supplementary Fig. S1). The

absolute metabolite concentrations identified by NMR can be found in supplementary material. Statistical analysis revealed that the concentration of 2,3,4-trihydrobutanoic acid (erythronic acid, threonic acid), detected by GC-MS (Fig. 4D) increased in all samples with increasing delay time at room temperature. In sample H2 the increase of this metabolite was extremely high after 120 minutes. Non-parametric ANOVA (Kruskal-Wallis) showed that discrimination between T0 and T120 was significant ($p < 5 \times 10^{-3}$).

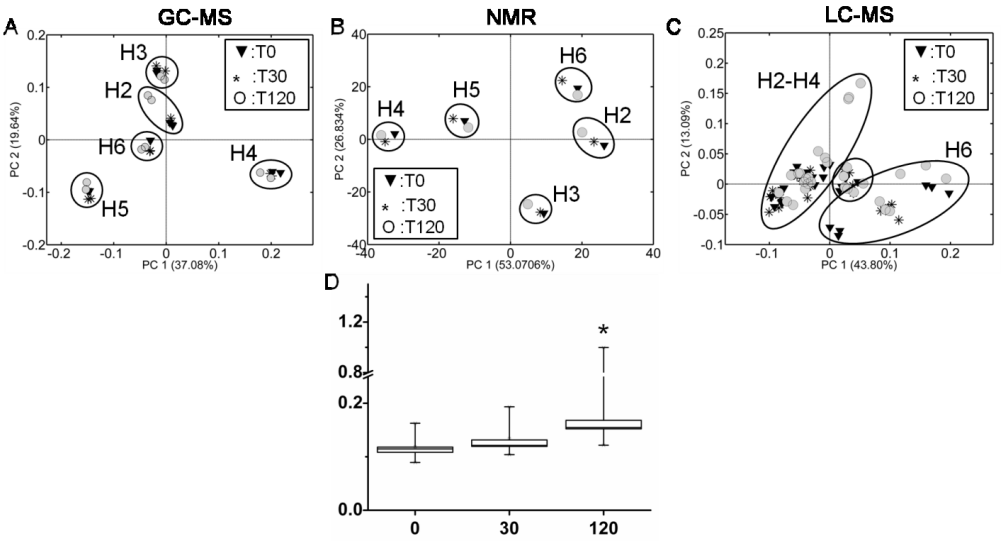


Figure 4. Statistical analysis of metabolomics data derived from human CSF (patients H2-H6). Multivariate statistical analysis by PCA based on all detected metabolites and univariate statistical analysis (Kruskal-Wallis non-parametric ANOVA) of 2,3,4-trihydroxybutanoic acid (T0 vs. T120; (*) $p < 5 \times 10^{-3}$). (A) GC-MS (90 metabolites, duplicate sample analysis). (B) NMR (51 metabolites, single analysis), (C) LC-MS targeting 19 natural amino acids (sextuplicate analysis) and (D) 2,3,4-trihydroxybutanoic acid ((*) $p < 5 \times 10^{-3}$).

2.3 DISCUSSION

We present a study of the stability of the measured proteome and metabolome of human CSF when leaving samples at room temperature for up to 2 h between lumbar puncture and storage at -80 °C to mimic delayed storage in clinical routine practice.

Unsupervised multivariate statistical analysis (PCA) showed that patient-to-patient variation is most prominent overriding variation that is due to delay time. Following variable selection based on pre-classification of the samples according to delay time, we found that only two peptides and one metabolite changed significantly over time from amongst approximately 6000 detected peptides and 93 identified and quantified metabolites. Our results demonstrate that human CSF prepared according to the described procedure is suitable for proteomics and metabolomics analysis even when left at room temperature for 2 hours provided that all cells have been removed by centrifugation. Quantitation of albumin and cystatin C by targeted mass spectrometry in the SRM mode using stable isotope labeled internal standard peptides confirmed that there is no statistically significant difference over two hours of delay time.

Another study on the stability of the proteome in CSF at room temperature pointed in the same direction, with the detection of only two polypeptides that changed after storage [33]. These samples were, however, contaminated with blood since both polypeptides were derived from hemoglobin. Another study showed that blood contamination decreases the stability of the CSF proteome [11] corroborating our earlier results [21]. Metabolomics revealed increased levels of 2,3,4-trihydroxybutanoic acid after storage at room temperature. This increase may be caused by oxidative degradation of ascorbic acid [32, 34, 35] as ascorbic acid levels were slightly decreased with increased time at room temperature (statistically not significant). Interestingly, 2,3,4-trihydroxybutanoic acid decreased in CSF containing white blood cells [21], which might be due to further metabolism of the acid by enzymes released from white blood cells. The concentration of 2,3,4-trihydroxybutanoic acid was too low for NMR detection.

The biological variability of metabolites and proteins is another important factor when designing biomarker studies. A study from our team showed that biological variation of

some proteins and peptides in CSF can exceed 100% which limits their potential as biomarker candidates [36].

In conclusion, we assessed the stability of CSF with respect to delay time using five different analytical platforms. Overall we observed only minor changes in either peptides (two out of approximately 6000) or metabolites (one out of 93). Earlier studies showed that blood or white blood cell contamination reduces CSF stability considerably, emphasizing the importance of the initial centrifugation step. As we did not add antioxidants, we cannot draw conclusions about oxygen-sensitive metabolites such as the catecholamines. The observed increase in 2,3,4-trihydroxybutanoic acid over time indicates, however, that oxygen-sensitive metabolites require additional protective measures during sample preparation and storage.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr. Rogier Hintzen from the Department of Neurology, Erasmus University Medical Center, Rotterdam, The Netherlands for providing the CSF samples. The study was performed within the framework of the Top Institute Pharma project number D4-102. The work was also supported by the project BioRange 2.2.3 from the Netherlands Proteomics and The Netherlands Bioinformatics Center.

REFERENCES

1. Hansson, S. F.; Simonsen, A. H.; Zetterberg, H.; Andersen, O.; Haghighi, S.; Fagerberg, I.; Andreasson, U.; Westman-Brinkmalm, A.; Wallin, A.; Ruetschi, U.; Blennow, K., Cystatin C in cerebrospinal fluid and multiple sclerosis. *Annals of neurology* **2007,** 62, (2), 193-6; discussion 205.
2. Irani, D. N.; Anderson, C.; Gundry, R.; Cotter, R.; Moore, S.; Kerr, D. A.; McArthur, J. C.; Sacktor, N.; Pardo, C. A.; Jones, M.; Calabresi, P. A.; Nath, A., Cleavage of cystatin C in the cerebrospinal fluid of patients with multiple sclerosis. *Annals of neurology* **2006,** 59, (2), 237-47.
3. Anesi, A.; Rondanelli, M.; d'Eril, G. M., Stability of neuroactive amino acids in cerebrospinal fluid under various conditions of processing and storage. *Clinical chemistry* **1998,** 44, (11), 2359-60.
4. Kaiser, E.; Schonknecht, P.; Thomann, P. A.; Hunt, A.; Schroder, J., Influence of delayed CSF storage on concentrations of phospho-tau protein (181), total tau protein and beta-amyloid (1-42). *Neuroscience letters* **2007,** 417, (2), 193-5.
5. Kraut, A.; Marcellin, M.; Adrait, A.; Kuhn, L.; Louwagie, M.; Kieffer-Jaquinod, S.; Lebert, D.; Masselon, C. D.; Dupuis, A.; Bruley, C.; Jaquinod, M.; Garin, J.; Gallagher-Gambarelli, M., Peptide storage: are you getting the best return on your investment? Defining optimal storage conditions for proteomics samples. *Journal of proteome research* **2009,** 8, (7), 3778-85.
6. Levine, J.; Panchalingam, K.; McClure, R. J.; Gershon, S.; Pettegrew, J. W., Stability of CSF metabolites measured by proton NMR. *J Neural Transm* **2000,** 107, (7), 843-8.
7. Schaub, S.; Wilkins, J.; Weiler, T.; Sangster, K.; Rush, D.; Nickerson, P., Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int* **2004,** 65, (1), 323-32.
8. Teahan, O.; Gamble, S.; Holmes, E.; Waxman, J.; Nicholson, J. K.; Bevan, C.; Keun, H. C., Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical chemistry* **2006,** 78, (13), 4307-18.
9. West-Nielsen, M.; Hogdall, E. V.; Marchiori, E.; Hogdall, C. K.; Schou, C.; Heegaard, N. H., Sample handling for mass spectrometric proteomic investigations of human sera. *Analytical chemistry* **2005,** 77, (16), 5114-23.
10. Wuolikainen, A.; Hedenstrom, M.; Moritz, T.; Marklund, S. L.; Antti, H.; Andersen, P. M., Optimization of procedures for collecting and storing of CSF for studying the metabolome in ALS. *Amyotroph Lateral Scler* **2009,** 10, (4), 229-36.
11. You, J. S.; Gelfanova, V.; Knierman, M. D.; Witzmann, F. A.; Wang, M.; Hale, J. E., The impact of blood contamination on the proteome of cerebrospinal fluid. *Proteomics* **2005,** 5, (1), 290-6.
12. Diamandis, E. P., Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* **2004,** 3, (4), 367-78.
13. Giovannoni, G., Multiple sclerosis cerebrospinal fluid biomarkers. *Disease markers* **2006,** 22, (4), 187-96.
14. Villanueva, J.; Philip, J.; Chaparro, C. A.; Li, Y.; Toledo-Crow, R.; DeNoyer, L.; Fleisher, M.; Robbins, R. J.; Tempst, P., Correcting common errors in identifying cancer-specific serum peptide signatures. *Journal of proteome research* **2005,** 4, (4), 1060-72.
15. Dekker, L. J.; Burgers, P. C.; Kros, J. M.; Smitt, P. A.; Luider, T. M., Peptide profiling of cerebrospinal fluid by mass spectrometry. *Expert review of proteomics* **2006,** 3, (3), 297-309.
16. Lutz, N. W.; Viola, A.; Malikova, I.; Confort-Gouny, S.; Ranjeva, J. P.; Pelletier, J.; Cozzone, P. J., A branched-chain organic acid linked to multiple sclerosis: first identification by NMR spectroscopy of CSF. *Biochem Biophys Res Commun* **2007,** 354, (1), 160-4.
17. Myint, K. T.; Aoshima, K.; Tanaka, S.; Nakamura, T.; Oda, Y., Quantitative profiling of polar cationic metabolites in human cerebrospinal fluid by reversed-phase nanoliquid chromatography/mass spectrometry. *Analytical chemistry* **2009,** 81, (3), 1121-9.
18. Noben, J. P.; Dumont, D.; Kwasnikowska, N.; Verhaert, P.; Somers, V.; Hupperts, R.; Stinissen, P.; Robben, J., Lumbar cerebrospinal fluid proteome in multiple sclerosis: characterization by ultrafiltration, liquid chromatography, and mass spectrometry. *Journal of proteome research* **2006,** 5, (7), 1647-57.
19. Stoop, M. P.; Dekker, L. J.; Titulaer, M. K.; Lamers, R. J.; Burgers, P. C.; Sillevis Smitt, P. A.; van Gool, A. J.; Luider, T. M.; Hintzen, R. Q., Quantitative matrix-assisted laser desorption ionization-fourier

transform ion cyclotron resonance (MALDI-FT-ICR) peptide profiling and identification of multiple-sclerosis-related proteins. *Journal of proteome research* **2009,** 8, (3), 1404-14.
20. Zhang, J.; Goodlett, D. R.; Montine, T. J., Proteomic biomarker discovery in cerebrospinal fluid for neurodegenerative diseases. *J Alzheimers Dis* **2005,** 8, (4), 377-86.
21. Rosenling, T.; Slim, C. L.; Christin, C.; Coulier, L.; Shi, S.; Stoop, M. P.; Bosman, J.; Suits, F.; Horvatovich, P. L.; Stockhofe-Zurwieden, N.; Vreeken, R.; Hankemeier, T.; van Gool, A. J.; Luider, T. M.; Bischoff, R., The effect of preanalytical factors on stability of the proteome and selected metabolites in cerebrospinal fluid (CSF). *Journal of proteome research* **2009,** 8, (12), 5511-22.
22. Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P., Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Analytical chemistry* **2008,** 80, (9), 3095-104.
23. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G., Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** 99, (10), 6567-72.
24. Wold, S.; Esbensen, K.; Geladi, P., Principal Component Analysis. *Chemometr Intell Lab* **1987,** 2, (1-3), 37-52.
25. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** 26, (7), 966-8.
26. Koek, M. M.; Muilwijk, B.; van der Werf, M. J.; Hankemeier, T., Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical chemistry* **2006,** 78, (4), 1272-81.
27. Cunniffe, J. G.; Whitby-Strevens, S.; Wilcox, M. H., Effect of pH changes in cerebrospinal fluid specimens on bacterial survival and antigen test results. *J Clin Pathol* **1996,** 49, (3), 249-53.
28. Piotto, M.; Saudek, V.; Sklenar, V., Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J Biomol NMR* **1992,** 2, (6), 661-5.
29. Eilers, P. H., A perfect smoother. *Analytical chemistry* **2003,** 75, (14), 3631-6.
30. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemometr* **2004,** 18, (5), 231-241.
31. van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006,** 7, 142.
32. Deutsch, J. C., Ascorbic acid oxidation by hydrogen peroxide. *Anal Biochem* **1998,** 255, (1), 1-7.
33. Berven, F. S.; Kroksveen, A. C.; Berle, M.; Rajalahti, T.; Flikka, K.; Arneberg, R.; Myhr, K. M.; Vedeler, C.; Kvalheim, O. M.; Ulvik, R. J., Pre-analytical influence on the low molecular weight cerebrospinal fluid proteome. *Proteomics Clin Appl* **2007,** 1, (7), 699-711.
34. Mystkowski, E. M.; Lasocka, D., Factors preventing oxidation of ascorbic acid in blood serum. *Biochem J* **1939,** 33, 1460-4.
35. Kuellmer, V., *Vitamins: Ascorbic acid*. Wiley Interscience: 1999.
36. Stoop, M. P.; Coulier, L.; Rosenling, T.; Shi, S.; Smolinska, A. M.; Buydens, L.; Ampt, K.; Stingl, C.; Dane, A.; Muilwijk, B.; Luitwieler, R. L.; Sillevis Smitt, P. A.; Hintzen, R. Q.; Bischoff, R.; Wijmenga, S. S.; Hankemeier, T.; van Gool, A. J.; Luider, T. M., Quantitative proteomics and metabolomics analysis of normal human cerebrospinal fluid samples. *Mol Cell Proteomics* **2010,** 9, (9), 2063-75.

CHAPTER 3

QUANTITATIVE PROTEOMICS AND METABOLOMICS ANALYSIS OF NORMAL CEREBROSPINAL FLUID SAMPLES

M. P. Stoop, L. Coulier, T. Rosenling, S. Shi, **A. M. Smolinska**, L. Buydens, K. Ampt, Ch. Stingl, A. Dane, B. Muilwijk, R. L. Luitwieler, P. A. E. S. Smitt, R. Q. Hintzen, R. Bischoff, S. S. Wijmenga, T. Hankemeier, A. J. van Gool, and T. M. LuiderMolecular & Cellular Proteomics (2010), 9 (9), pp. 2063-75

ABSTRACT

The analysis of cerebrospinal fluid (CSF) is employed in biomarker discovery studies for various neurodegenerative central nervous system disorders. However, little is known about variation of CSF proteins and metabolites between patients without neurological disorders, a baseline for a large number of CSF compounds appears to be lacking. To analyze the variation in CSF protein and metabolite abundances in a number of well-defined individual samples of patients undergoing routine, non-neurological, surgical procedures we determined the variation of various proteins and metabolites by multiple analytical platforms.

A total of 126 common proteins were assessed for biological variations between individuals by ESI-Orbitrap. A large spread in inter-individual variation was observed (RSDs ranged from 18% to 148%), for both high abundant and low abundant proteins. Technical variation was between 15% and 30% for all 126 proteins. Metabolomics analysis was performed by means of GC-MS and NMR and amino acids were specifically analyzed by LC-MS/MS, resulting in the detection of more than 100 metabolites. Interestingly, the variation in the metabolome appears to be much more limited compared to the proteome, as the observed RSDs ranged from 12% to 70%. Technical variation was below 20% for almost all metabolites.

Consequently, an understanding of the biological variation of proteins and metabolites in CSF of neurologically normal individuals appears to be essential for reliable interpretation of biomarker discovery studies for central nervous system disorders, because such results may be influenced by natural inter-individual variations. Therefore proteins and metabolites with high variation between individuals ought to be assessed with caution as candidate biomarkers because at least part of the difference observed between the diseased individuals and the controls will not be caused by the disease, but rather by the natural biological variation between individuals.

3.1 INTRODUCTION

The analysis of cerebrospinal fluid (CSF) is indispensable in the diagnosis and understanding of various neurodegenerative central nervous system (CNS) disorders [1-3]. CSF is a fluid that has different functions, such as the protection of the brain to forces from outside, transport of biological substances and excretion of toxic and waste substances. It is in close contact with the extracellular fluid of the brain. Therefore, the composition of CSF can reflect biological processes of the brain [4]. By characterization of the proteome and metabolome of CSF better insight in, for example, the pathogenesis of CNS disorders may be achieved; as for many of these disorders the aetiology is still unclear.

CSF is produced in the ventricles of the brain and in the subarachnoidal spaces. Humans normally produce around 500 mL of CSF each day, and the total volume of CSF at a given time is approximately 150 mL. CSF reflects the composition of blood plasma although the concentrations of most proteins and metabolites in CSF are lower. However, individual proteins and metabolites can act differently. Active transport from blood and secretion from the brain contribute to the specific composition of CSF. This composition can be disturbed in neurological disorders [5, 6]. Since CNS specific proteins and metabolites are typically low in abundance compared to blood, this change in composition is more likely to be found in CSF, because in blood the higher abundant plasma proteins can completely mask the signal of the lower abundant proteins. Also, if the disease markers do not cross the blood-brain-barrier (BBB), then CSF is the only viable biofluid source. CSF might therefore be an excellent source for biomarker discovery for CNS disorders, following the hypothesis that neurological diseases induce alterations in CSF protein and metabolite levels.

Analysis of metabolites in CSF has been common practice in clinical chemistry for decades to analyse biomarkers for inborn errors of metabolism. The approaches used are either metabolite profiling of CSF using NMR [7], or targeted analysis of one or a few metabolites using specific analytical methods [8]. Metabolomics includes the analysis of metabolites in biofluids by NMR or MS-based approaches, i.e. LC-MS or GC-MS. Several metabolite profiling studies were carried out on CSF using NMR, some of which were published only recently [9, 10]. Surprisingly, very few metabolomics studies using MS-

based methods have been performed on CSF yet [11, 12]. One of the reasons is the fact that the human CSF metabolome has not been characterized very well yet. Many CSF metabolites remain unidentified and for those that have been identified there is not much known about normal concentration ranges. A systematic categorization of the CSF metabolome is necessary and expected to be beneficial for future biomarker discoveries. Recently Wishart *et al.* made a good start in exploring the human CSF metabolome. Computer-aided literature survey resulted in 308 detectable metabolites in human CSF [13].

The CSF proteome has been characterized to a much larger extent than the CSF metabolome and is currently topic of investigations in several research groups worldwide. Recently, studies have been published with numerous identities and quantities of CSF proteins. Pan and co-workers were able to identify 2.594 proteins in well-characterized pooled human CSF samples using strict proteomics criteria with a combination of LTQ-FT and MALDI TOF/TOF equipment [14]. Also they were able to quantify several proteins using a targeted LC MALDI TOF/TOF approach [15]. Hu and co-workers have studied the intra- and inter-individual variation in human CSF, and found large variations in protein concentrations in six patients by means of 2D-gel electrophoresis [16], focussing mainly on the variations within individuals at two different time-points. Although only a limited number of proteins was analyzed, the variation between the time-points was profound, exceeding 200% for seven proteins.

Unique CSF biomarkers may contribute to a deeper understanding of the mechanisms of CNS disorders. However, for this assumption to come true, there are still challenges ahead. Even though CSF is not as complex as blood (almost missing the cellular part and the clotting system present in blood), it is expected to consist of thousands of organic- and non-organic salts, sugars, lipids and proteins. A large part of the CSF consists of a few high abundant metabolites and proteins, which hamper, if no precautions are undertaken, the identification and quantification of metabolites and proteins that occur in lower amounts. The analysis of the CSF metabolome is complicated due to the diverse chemical nature of metabolites and the lower concentration of metabolites compared to blood. Analytical method development is still required as it is not possible to identify the entire range of CSF metabolites with one

single analytical method. Though in proteome research efforts have been made to quantify proteins, metabolomics studies up to now do not provide quantitative information or only give information for the most abundant metabolites.

Another challenge is the sample amount obtained by lumbar puncture to collect CSF. Lumbar puncture is an invasive method that is not performed as frequently as blood sampling. However, often after the analysis of various clinical parameters only a limited amount of CSF sample is available for biomarker discovery. Metabolomics studies are hampered by limited CSF sample amount. Therefore analytical methods are required that are suitable to handle relative small sample volumes.

The main objective of this study was firstly to analyze the variation in CSF protein and metabolite abundances in a number of well-defined individual samples by multiple analytical platforms. Secondly, the goal was to integrate metabolomics and proteomics and to present biological variations in metabolite and protein abundances and compare these with technical variations with the currently employed analytical methods. The results will facilitate and increase the application of CSF for future biomarker discovery studies in the field of neurodegenerative diseases and neuro-oncology.

3.2 EXPERIMENTAL PROCEDURE

3.2.1 CSF sampling

CSF samples were obtained by lumbar puncture in the Erasmus University Medical Centre (Rotterdam, the Netherlands). An experienced medical doctor selected ten samples, which were taken from patients receiving spinal anaesthesia prior to non-neurological surgery. These subjects had no neurological diseases, were not using any medication and were considered to have neurologically normal CSF. Immediately after sampling, the CSF samples were centrifuged (10 minutes at 3.000 rpm) to discard cellular elements. The samples were subsequently used for routine CSF diagnostics. This included quantification of total protein concentration by routine clinical chemistry measurements and quantification of the cell count (< 5 white blood cells per mL). The remaining volume of the samples was aliquoted and stored at –80°C immediately after centrifugation. As a standard procedure the samples were checked for blood contamination, and any sample in which a hemoglobin or apolipoprotein B100 peptide was identified with a significant score by nanoLC-Orbitrap MS was excluded from the study.

For pooling of the samples (n=10), the originally obtained samples were thawed on ice and 0.75 mL from each of the samples was joined, resulting in a 7.5 mL pooled CSF sample. This pooled CSF sample was vortexed for 30 seconds and then subdivided into 75 portions of 100 µL in sterile cryogenic vials (Nalgene Nunc Int., Rochester, NY, USA). The portions were immediately frozen at –80 °C. The characteristics of the pooled sample are described in Table 1. This pooled sample was used to assess the technical variation in the proteomics experiments by measuring it five times. For the measurements of the individual patients only nine CSF samples were used because there was insufficient volume of one sample.

The 28 CSF samples from the validation sample set were also taken by an experienced anesthesiologist from patients receiving spinal anaesthesia prior to non-neurological surgery, but these samples were taken at another hospital (Sint Franciscus Gasthuis (Rotterdam, the Netherlands)). These subjects had no neurological diseases, were not using any medication and were considered to have neurologically normal CSF.

Table 1. Details on the pooled sample, including gender, average age and average protein concentration.

Gender	**Male 8; Female 2**
Mean age (years)	51 (SD = 14)
Total protein concentration (g/L)	0.4 (SD = 0.1)
Glucose concentration (mmol/L)	3.3 (SD = 0.3)

The CSF samples used in the experimental sample set were selected by an experienced neurologist and taken from patients undergoing tests for clinical diagnosis. These samples, taken from multiple sclerosis and headache patients were subjected to the same, strict post-sampling procedure as the samples mentioned previously. In these samples no significant difference in protein concentration between the two groups was observed, so there was no leakage in the blood-CSF barrier. All CSF samples used in this study were taken in the morning at approximately 10 AM. The Medical Ethical Committees of the Erasmus University Medical Centre in Rotterdam, The Netherlands, and the Sint Franciscus Gasthuis in Rotterdam, The Netherlands, approved the study protocol and all study participants gave written consent. The average age and protein concentration of the samples in all three sample sets is listed in Table 2 and age, gender and protein concentration of the individual samples is listed in the Supplementary Material.

Table 2. Details on the three sample sets: the original sample set (n=9), the validation sample set (n=28), and the experimental sample sets (n=36/42). Age and protein concentration values are averages (standard deviation in brackets). The gender, age and protein concentration of all individual patients is listed in the Supplementary Material.

	Original sample set	**Validation sample set**	**Experimental sample set proteomics**	**Experimental sample set metabolomics**
Gender	7M / 2F	13M / 15F	13M / 23F	23M / 19F
Age (years)	51.0 (14.8)	44.5 (14.5)	41.7 (10.8)	43.5 (12.8)
Protein concentration (g/L)	0.39 (0.12)	0.37 (0.11)	0.38 (0.11)	0.38 (0.13)

3.2.2 Proteomics

3.2.2.1 Sample preparation for nanoLC-Orbitrap MS and MALDI-FT-ICR MS

For measurement of proteins in CSF, samples were enzymatically digested with trypsin to obtain peptides. An amount of 50 µL Rapigest (Waters, Milford, USA) in 50 mM ammonium bicarbonate and 1 µL 100 mM DTT was added to 50 µL CSF. The mixture was heated at 60°C for 30 minutes, upon which it was cooled down to room temperature in approximately 20 minutes. Iodoacetamide (5 µL of 0.3 M solution) was added and this mixture was left for 30 minutes in dark at room temperature. Trypsin was added (10 µL, 0.1 mg/mL) and all samples, processed in one batch, were incubated overnight at 37°C. To stop digestion, 2 µL of a 50% TFA/50% water solution was added. The sample was then incubated for 45 minutes at 37°C.

3.2.2.2 NanoLC-Orbitrap MS analysis

These measurements were carried out on a Ultimate 3000 nanoLC system (Dionex, Germering, Germany) online coupled to a hybrid linear ion trap / Orbitrap MS (LTQ Orbitrap XL; Thermo Fisher Scientific, Bremen, Germany). Five µL digest were loaded on to a C18 trap column (C18 PepMap, 300µm ID x 5mm, 5µm particle size, 100 Å pore size; Dionex, Amsterdam, The Netherlands) and desalted for 10 minutes using a flow rate of 20 µL /min 0.1% TFA. Then the trap column was switched online with the analytical column (PepMap C18, 75 µm ID x 150 mm, 3 µm particle and 100 Å pore size; Dionex, Amsterdam, The Netherlands) and peptides were eluted with following binary gradient of solvent A and B: 0% - 25% solvent B in 120 min and 25% - 50% solvent B in further 60 minutes, where solvent A consist of 2% acetonitrile and 0.1% formic in water and solvent B consists of 80% acetonitrile and 0.08% formic acid in water. Column flow rate was set to 300 nL/min. For MS detection a data dependent acquisition method was used: high resolution survey scan from 400 – 1800 Th. was performed in the Orbitrap (value of target of automatic gain control AGC 10^6, resolution 30,000 at 400 m/z; lock mass was set to 445.120025 u (protonated $(Si(CH_3)_2O)_6$) [17]). Based on this survey scan the 5 most intensive ions were consecutively isolated (AGC target set to 10^4 ions) and fragmented by collision-activated dissociation (CAD) applying 35% normalized collision energy in the linear ion trap. After precursors were selected for MS/MS, they were

excluded for further MS/MS spectra for 3 minutes. Proteins were identified using the Bioworks 3.2 (peak picking by Extract_msn, default settings) software package (Thermo Fisher Scientific, Bremen, Germany), and SEQUEST (Thermo Fisher Scientific, Bremen, Germany), taking the HUPO criteria, with XC scores of 1.8, 2.2 and 3.75 for single, double and triple charged ions, respectively, into account. The used database was the SwissProt-database (version 56.0, human taxonomy (20069 entries)). Carboxymethylation of cysteine (+57.021 u) as fixed and oxidation of methionine (+15.996 u) as variable modifications and tryptic cleavage were considered. The number of allowed missed cleavages was 2, the mass tolerance for precursor ions was 10 ppm and for fragment ions 0.5 Da. The cut-off for mass differences with the theoretical mass of the identified peptides was set at 2 ppm.

The Orbitrap data was subsequently analysed using the Progenesis LC-MS software package (version 2.5, Nonlinear Dynamics, Newcastle-upon-Tyne, United Kingdom), in which the LC runs were aligned and the biological variation between the samples was calculated to assess variation between individuals in this data set. A S/N > 4 and the presence of at least 3 isotope peaks per peptide were used as a minimum threshold for quantitation. Variation was assessed by comparing the area-under-the-curve of all peptides of a protein. The mean area-under-the-curve, corrected for the total ion current, of all peptides of a protein was compared between the individuals, and the relative standard deviation (RSD) of this value was considered to be the inter-individual variation (listed as RSD (in percentages) in the supplementary material). Technical variation was assessed by performing the same comparison on the five replicas of the pooled sample.

3.2.2.3 MALDI-FT-ICR MS analysis

The CSF samples were handled according to the same protocol we reported previously [18], in which the samples were tryptically digested and desalted using C18 material. Using a 2,5-dihydroxybenzoic acid matrix the samples were all measured manually on an APEX IV Qe 9.6 Tesla MALDI-FT-ICR mass spectrometer (Bruker Daltonics, Billarica, USA), using a multishot accumulation as recommended by Mize *et al.*, Moyer *et al.,* and O'Connor *et al.* [19-21]. External mass calibration was applied using a quadratic equation. Quantitative MALDI-FT-ICR has previously been applied to quantify HIV-1

protease inhibitors in cell lysates [22] as well as peptides in CSF [18], indicating that quantitative MALDI-FT-ICR methods are readily applicable, which is due to the fact that variation in peak height in MALDI-FT-ICR mass spectrometry is much more reproducible than in, for example, MALDI-TOF mass spectrometry. The sum of the height of 14 omnipresent albumin peaks of each sample was then compared to albumin concentrations obtained by routine clinical chemistry measurements. Standard deviations of the peak height of the albumin peaks were between 9-16% for all 14 albumin peaks.

3.2.2.4 Biological variation in an experimental setting

To compare the results on the variation of protein abundances found in the neurologically normal individual CSF samples to an experimental setting, an identical experiment was performed on a larger set of samples. A total of 36 CSF samples, taken from patients with either multiple sclerosis or headaches, was used. It must be noted that these samples, especially those of the multiple sclerosis patients, originate from people suffering from neurological problems. Hence the variation in protein abundance, like for example immunoglobulin levels, which are known to be elevated in neuro-inflammatory diseases such as multiple sclerosis [23-25], is potentially far more extensive than in the nine well-defined individuals measured previously.

3.2.3 Metabolomics

3.2.3.1 GC-MS analysis

Human CSF samples from the original sample set (60 µL) were deproteinized by adding 250 µL methanol and subsequently centrifuged for 10 min at 10000 rpm. Human CSF samples from the validation sample set (100 µL) were deproteinized by adding 400 µL methanol. The supernatant was dried under N_2 followed by derivatization with methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) in pyridine similar to Koek *et al.* [26]. During the different steps in the sample work-up, i.e. prior to deproteinization, derivatization and injection, different (deuterated) internal standards were added at a level of approx. 20 ng/µL. The final volume was 45 µL for the original sample set and 135 µL for the validation sample set and 1 µL aliquots of the derivatized samples were injected in

splitless mode on a HP5-MS 30 m x 0.25 mm x 0.25 µm capillary column (Agilent Technologies, Palo Alto, USA) using a temperature gradient from 70^{o}C to 320^{o}C at a rate of 5^{o}C/min. GC-MS analysis was performed using an Agilent 6890 gas chromatograph coupled to an Agilent 5973 mass selective detector (Agilent Technologies, Palo Alto, USA). Detection was carried out using MS detection in electron impact mode and full scan monitoring mode (m/z 15-800). The electron impact for the generation of ions was 70 eV.

Sample work-up was carried out in duplicate for the original sample set. For the validation sample set samples were injected in duplicate. For both sample sets a pooled human CSF sample was analyzed in sextuplicate to determine the analytical error in the analysis of metabolites by GC-MS. Data-pre-processing was carried out by composing target lists of peaks detected in the samples based on retention time and mass spectra and these peaks were integrated for all samples. All peak areas were subsequently normalized using internal standards. The resulting target lists were used for further statistical analysis. Identities were assigned based on the presence of identical mass spectra in an in-house database.

3.2.3.2 LC-MS/MS analysis

To 10 µL of human CSF sample, 10 µL of an internal standard solution containing $^{13}C^{15}N$-amino acids was added followed by addition of 100 µL of MeOH. The mixture was vortexed for 10 s and centrifuged at 10.000 rpm for 10 min at 10 ºC. The supernatant was dried under N_2. The residues were dissolved in 80 µL borate buffer (pH 8.5) and after 10 s vortexing 20 µL of AQC reagent (Waters, Etten-Leur, The Netherlands) was added and the mixture was vortexed immediately. The samples were heated 10 min at 55ºC. After cooling down, a 1µL sample of the reaction mixture was injected into the UPLC-MS/MS system.

An ACQUITY UPLC^{TM} system with autosampler (Waters, Milford, USA) was coupled online with a Quattro Premier XE Tandem quadrupole mass spectrometer (Waters, Milford, USA) and was used in positive-ion electrospray mode. The instrument was operated under Masslynx data acquisition software (version 4.1; Waters). The samples were analyzed by UPLC-MS/MS using a AccQ-TagTM Ultra 100 mm x 2.1 mm (1.7µm

particle size) column (Waters, Milford, USA). A binary gradient system of water – eluent A (10:1, v/v) (AccQ Tag, Waters) and 100% eluent B (AccQ Tag, Waters), was used. Elution of the analytes was achieved by ramping the percentage of eluent B from 0.1 to 90.0 in approx. 9.5 minutes using a combination of both linear and convex profiles. The flow-rate was 0.7 mL/min. the column temperature was maintained at 60ºC and the temperature of the autosampler tray was set to 10ºC. After each injection the injection needle was washed with 200 µL strong wash solvent (95% ACN), and 600µL weak wash solvent (5% ACN).

The Quattro Premier XE was used in the positive-ion electrospray mode and all analytes were monitored in Selective Reaction Monitoring (SRM) using nominal mass resolution (FWHM 0.7 amu). Next to the derivatisation reagent all amino acids were selectively monitored via the transition from the protonated molecule of the AccQ-Tag derivative to the common fragment at m/z 171. Collision energy and collision gas (Ar) pressure were 22eV and 2.5 mbar, respectively. The complete chromatogram was divided into 6 time windows, restricting the number of SRM transitions to follow and allowing quantitative information to be gathered in each segment. Acquired data was evaluated using Quantlynx (Waters, Milford, USA). All samples were analyzed in duplicate.

Data pre-processing was carried out by calculating the concentration of 18 amino acids in all samples by peak integration, followed by normalization using relevant internal standards and quantification using external calibration curves. The analytical variation was determined from the duplicate analysis of the samples using weighted regression [27, 28]

3.2.3.3 NMR analysis

CSF samples from the original sample set (280 µL) were centrifuged (2000g, 15 minutes) using a filter with a cut-off of 10 kDa (Centrisart I 13239-E) to remove proteins. Next, 25 µL of 8.8 mM TSP-d_4 stock solution in D_2O was added to 250 µL filtrated CSF to a final concentration of 0.8 mM TSP. The pH of the filtrated CSF was adjusted to around 7 (7.0 – 7.1) by adding phosphate buffer (9.7 µL 1M, to a final concentration of 35 mM). The final CSF NMR sample (284.7 µL) was then transferred to a Shigemi microcell NMR tube for NMR measurements (called non-diluted CSF samples). For the

validation sample set, samples (100 µL) were first diluted with 270 µL D_2O before protein removal.

As a duplicate, for establishing the analytical variation, 100 µL of CSF of the individuals was diluted into 180 µL D_20 and subsequently worked up as described above (called further on diluted CSF samples).

The 1D ^1H NMR spectra of diluted and non-diluted CSF samples were acquired on an 800 MHz Inova or 600 MHz (Varian Inc., Palo Alto, USA) system equipped with either a 5 mm triple-resonance, XYZ-gradient HCN room-temperature probe or a 5 mm triple-resonance, Z-gradient HCN cold-probe, respectively. Suppression of water was achieved by using WATERGATE (delay: 85 µs) [29] or presaturation. For each 1D ^1H NMR spectrum 512 scans of 18K data points were accumulated with a spectral width of 9000 Hz. The acquisition time for each scan was 2 s. Between scans an 8 s relaxation delay was employed. Prior to spectral analysis, all acquired Free Induction Decays (FIDs) were zero-filled to 64K data points, multiplied with a 0.3 Hz line broadening function, Fourier transformed and manually phase - and baseline corrected by using ACD/SpecManager software. Spectra were subsequently transformed to the Chenomx NMR Suite Professional software package version 5.1 for further analysis [30]. Metabolite identification and quantification were done by using the 800 MHz library of metabolite NMR spectra from the Chenomx NMR Suite 5.1 (pH 6-8) for the original sample set (Chenomx NMR Suite 6.1 for the validation sample set). The metabolite spectra in the library are predicted based on a database of pure compound spectra acquired using particular pulse sequence and acquisition parameters, e.g. the tn-noesy-presaturation pulse sequence with 4s acquisition time and 1s of recycle delay. The Chenomx NMR Suite software fits the spectral signatures (singlets, doublets, triplets etc.), i.e. the peak shapes, of a compound from an internal database of reference spectra to the experimental NMR spectrum. The resonance assignments derived from the Chenomx NMR Suite software were further checked against literature spectra. For quantification, Chenomx NMR Suite 5.1 uses the concentration of the known reference signal as calibration (in this case TSP-d_4).

The analytical variation on the individual metabolite concentrations was determined from the NMR analysis of the dilute and non-dilute CSF samples, completely independently, and the quintuplicate measurement of the diluted CSF sample of one individual.

3.2.3.4 Biological variation in an experimental setting

To assess the results on the variation of metabolite abundances found in original and validation sample set, i.e. neurologically normal CSF, an identical experiment was performed using GC-MS on a set of 42 human CSF samples (100 µL), taken from patients with multiple sclerosis and other (inflammatory) neurological diseases.

3.3 RESULTS

3.3.1 Proteomics

None of the CSF samples was contaminated with plasma, as according to the criteria set hemoglobin and apolipoprotein B100 were not identified in any of the samples. All sequenced peptides and identified proteins are listed in the supplementary material (including the number of unique peptides per protein and the sequence coverage for all proteins identified with two or more peptides).

Using MALDI-FT-ICR mass spectrometry we analysed the height of albumin peptide peaks of the nine samples of the original sample set and their correlation to albumin concentration levels in CSF as measured by routine clinical chemistry diagnostics. The sum of the height of 14 omnipresent albumin peaks showed positive correlation to the values measured by clinical chemistry (R^2 = 0.919). These values (median: 0.219 g/L, range 0.097-0.403 g/L) clearly show a large variation between individuals, which was also apparent from the differences in height of the peaks measured by MALDI-FT-ICR (Figure 1). The area under the curve of all peptides identified to be part of albumin in the ESI-Orbitrap experiments was also plotted against the albumin concentration, showing good correlation (R^2 = 0.971). Relative standard deviations (RSD) were comparable for all three methods (43.7% for clinical chemistry, 66.7% for MALDI-FT-ICR and 39.1% for ESI-Orbitrap).

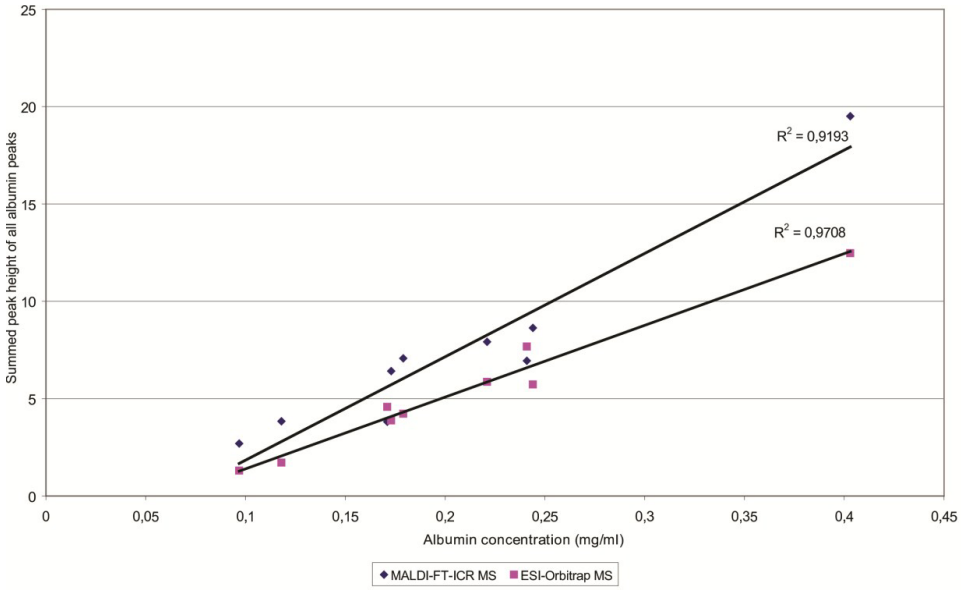


Figure 1. Correlation between the measured albumin concentration by clinical chemistry diagnostics and the sum of the height of 14 omnipresent albumin peaks as measured by MALDI-FT-ICR (R^2 = 0.919) and by ESI-Orbitrap (R^2 = 0.971).

A total of 126 proteins, all identified by multiple peptides and present in all nine normal CSF samples, was analysed in the nine individual CSF samples by ESI-Orbitrap to assess the variance in protein abundances in CSF, based on the averages of peak heights of all the peptides of a single protein. The RSD ranged from 18% to 148% (median: 43%) in peak height. The far greater part of the examined proteins (119 of 126, i.e. 94.4%) showed lower than 100% RSD in average peptide peak height per protein between the nine individual CSF samples. These results were subsequently tested in two larger sample sets, a validation set (28 samples) and an experimental sample set (36 samples), in which similar profiles for the variation in protein abundance, based on the averages of the area-under-the-curve of the peptides in the ESI-Orbitrap, was

observed (Figure 2). In this figure the variation (in RSD (%)) for all proteins is plotted for all three data sets, and the protein-numbers used here are the same as in the Supplementary Material. The slightly lower RSD's found for the validation sample set are at least partially due to the fact that a longer nanoLC column (50 cm) was used and consequently more peptides were measured per protein. Using these data the correlation of the individual protein variations between the datasets was calculated. This resulted in an R^2 of 0.94 for the correlation between the original dataset and the validation dataset, and an R^2 of 0.66 for the correlation between the original dataset and the experimental dataset. This is a strong indication that the same proteins have a high inter-individual variation in the healthy CSF patients in both the original and the validation sample set, but that this is quite different in the experimental sample set of patients with known neurological disorders. In the original sample set the 126 proteins were observed in all 9 normal control samples, but in the experimental sample set 11 proteins were not observed in all 36 samples. This may indicate a greater variance in the experimental samples, while in the validation sample set the variation is slightly lower than in the original sample set. In the experimental sample set, the RSD in peptide abundance per protein ranged from 30% to 182% (median: 91%). The greater variation in the experimental sample set is at least partially due to the sample choice for this set of samples. As referenced earlier, in multiple sclerosis it is known that immunoglobulins are elevated and since this sample set contained both multiple sclerosis CSF samples and samples from patients with headaches it is not surprising to note that many of the proteins with the highest RSD between individuals are all immunoglobulin types and proteins related to inflammatory response, which were indeed elevated in the multiple sclerosis samples. In essence, in the normal controls we observed the biological variation between the individuals, whereas in the experimental sample set both the biological variation as well as the disease-related variation was observed.

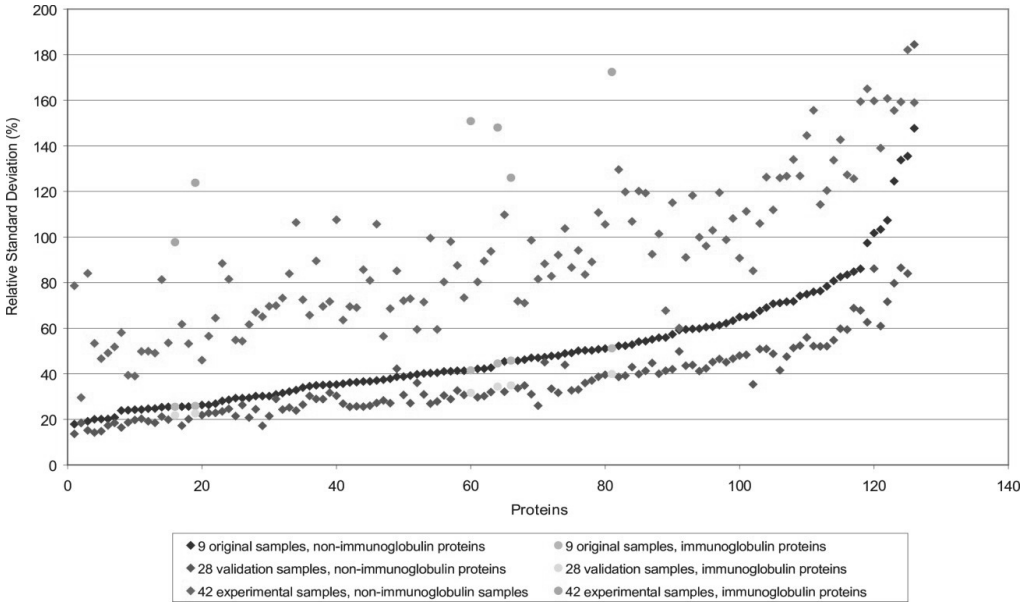


Figure 2. The proteins sorted by the variation in the original samples paired to the variation in the validation samples and the experimental samples. A trend is clearly visible, but due to the nature of the experimental samples (multiple sclerosis and headaches), the immunoglobulins do not correspond with the overall trend, which is to be expected considering the well-known inflammatory component of multiple sclerosis. Numbers on x-axis correlate to protein numbers mentioned in Supplementary Material.

Although all three sample sets are distinct, a number of similarities can be observed. In the sample sets, there is a clear division that can be seen between proteins whose abundances vary highly among individuals and proteins that show a much more limited variation between individuals. Among the proteins that showed limited variation between individual CSF samples were serotransferrin (25% RSD in the original sample set, 18% RSD in the validation sample set and 50% RSD in the experimental sample set), tetranectin (21%, 18% and 52%, respectively), and gelsolin (24%, 19% and 58%, respectively). Proteins with high variation between individuals in all three sample sets included cadherin-13 (82% RSD in the original sample set, 60% RSD in the validation sample set and 143% RSD in the experimental sample set), contactin-2 (124%, 80%

and 156%, respectively), and haptoglobin (135%, 84% and 182%, respectively). The full list of variations between the individuals for the 126 proteins can be found in the Supplementary Material.

Gender and age related inter-individual variations were assessed in the validation sample set. In this sample set the number of males and females was roughly equal (13 males and 15 females) and we defined 3 different age groups (below 35 years of age, between 35 and 50 years of age and above 50 years of age). Although both parameters (gender and age) appear to influence the variation, their influence appears limited (complete list of individual protein variations specified for age-group and gender is listed in the Supplementary Material). The inter-individual variation in protein abundance appears slightly larger in females than in males, especially in proteins that have a high inter-individual variation, but this is not exclusively the case as there are also proteins of which the variation is higher in males. The same is true for the inter-individual variation of the proteins when comparing age groups. The variation seems slightly larger in the oldest age group, again most clearly for the proteins with the highest inter-individual variation. But, as with age, this is not exclusively so, as there are also proteins that have a higher inter-individual variation in the group of patients below age 35 (full list of age and gender variation in Supplementary Material (Figures S1 and S2)). T-test show 1.6% and 3.2% of the proteins to be significantly different based on variation of their RSD between males and females and between the age groups, respectively ($p<0.01$, Supplementary Material).

Total abundance of most of the measured proteins was slightly higher in males than in females, and with regards to age the abundance of most of the measured proteins was highest in the oldest patient group of the validation sample set.

3.3.2 Metabolomics

Three different analytical methods were applied to analyse the individual as well as the pooled CSF samples. The methods included untargeted GC-MS and NMR methods and a targeted LC-MS/MS method specifically for amino acids.

First a small set of CSF samples, i.e. original sample set consisting of 9 samples for GC-MS, 8 for LC-MS/MS and 5 for NMR, was analyzed by the three methods followed by a

larger set of samples, i.e. validation sample set consisting of 28 samples for GC-MS and 27 samples for NMR and LC-MS/MS. The original sample set was used to have a quick screen of what type of metabolites could be detected and to have a rough idea of biological variation and analytical error. However, for reliable data with respect to biological variation as well as possible gender and age effect and comparison of the three methods, the validation sample set was used. Analysis of original samples CSF with GC-MS resulted in a list of 108 metabolites of which 93 could be identified (see Supplementary Material). The unknown metabolites covered both metabolites that were observed in other biofluids, i.e. plasma and/or urine, as well as metabolites that seemed to be specific for CSF. The metabolites detected by GC-MS cover many different compound classes, i.e. amino acids, organic acids, nucleosides, fatty acids, mono- and disaccharides. Of the 93 identified metabolites, some were only present in trace amounts and were therefore not used for further analysis. Interestingly, all identified metabolites were observed in all samples. The analytical variation for each metabolite was determined from the repeated (n=6) analysis of the pooled CSF sample. Results show that the analytical variation (< 20%) was less than the biological variation for all metabolites (15 to 85%) (see Supplementary Material). The concentration or relative peak area for each metabolite in the pooled human CSF sample is given in Supplementary Material. As expected, the average concentrations and relative peak areas found for the 9 individual CSF samples were very similar to that of the pooled human CSF sample.

Next, the validation sample set was analyzed by GC-MS using a somewhat different sample work-up including different sample volumes leading to a less concentrated extracts. As a result some of the low abundant compounds could not be detected. In total 68 metabolites could be detected and most of them were also detected in the original sample set. The biological variation of these 68 metabolites ranged from 7 to 214%, while the analytical error ranged from 1 to 36% and in all cases the analytical error was equal or less than the biological variation (see Table 3 and Supplementary Material). The biological variation for the metabolites observed with GC-MS in the validation sample set shows a normal distribution, as can be deduced from Figure 3.

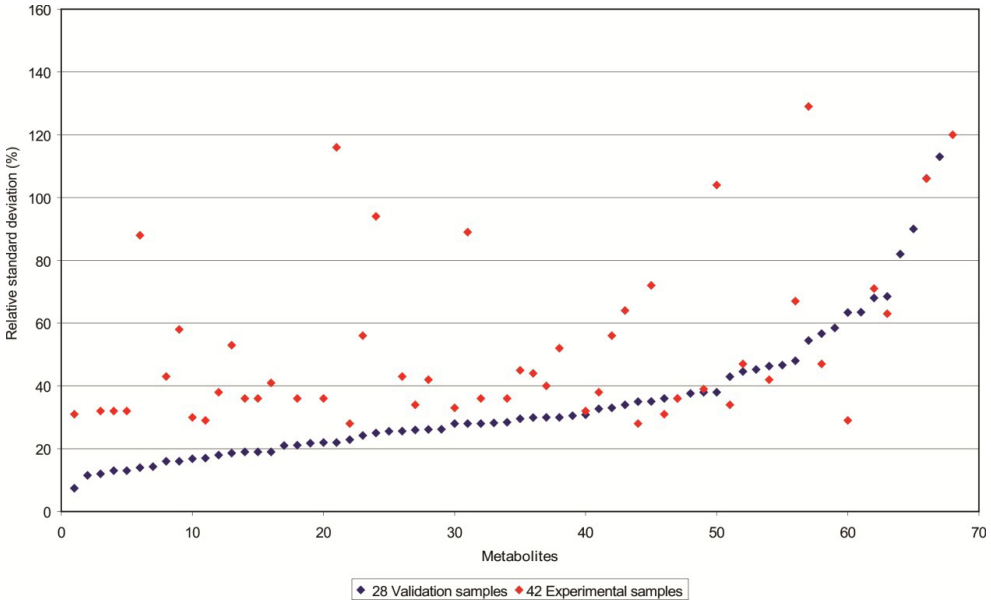


Figure 3. Metabolites detected by GC-MS in both the validation (blue) as well as the experimental (red) sample set, sorted by variation in the validation samples paired to the variation in the experimental samples. A trend is visible, but due to the nature of the experimental samples (multiple sclerosis and other neurological disorders), a number of metabolites do not correspond with the overall trend and show significantly higher biological variation. Numbers on x-axis correlate to metabolite numbers mentioned in Supplementary Material.

Multivariate data analysis using PCA (principal component analysis) showed that overall biological variation was dominant over age and gender effects (see Figures S3 and S4 in the Supplementary Material).

Table 3. Metabolites detected by GC-MS, NMR and LC-MS/MS in the validation sample set and their biological variation.

Metabolite	GC-MS RSD (%) n=28	LC-MS RSD (%) n=27	NMR RSD (%) n=27
1,5-anhydro-D-Glucitol	34		
1-methylhistidine			49
1-monopalmitoylglycerol	30		
1-monostearoylglycerol	19		
2,3-dihydroxybutanoic acid	28		
2,4-dihydroxybutanoic acid	26		
2-aminobutyric acid	22		28
2-hydroxybutanoic acid	35		29
2-hydroxyisovaleric acid	24		30
2-piperidinon	113		
3,4-dihydroxybutanoic acid	54		
3-hydroxybutanoic acid	106		15
3-hydroxyisovaleric acid	31		15
3-hydroxypropanoic acid	14		
3-methylhistidine			143
Acetic acid			52
Acetoacetic acid	38		26
Acetone			20
Aconitic acid			28
Alanine	28	32	27
Arabinose	19		
Arginine		25	19
Ascorbic acid	25		
Asparagine		22	
C16:0 Fatty acid	23		
C18:0 Fatty acid	7		
C18:1 fatty acid	58		
Cholesterol	30		
Choline			24
Citric acid	18		15
Citrulline		34	
Creatine			15
Creatinine	63		17
Dimethylamine			25
Erythronic acid	28		
Formic acid			19
Fructose	38		24
Fucose	17		
Galactitol			23
Gluconic acid	26		
Glucose	12		12
Glutamic acid	30		
Glutamine		14	18

Glyceric acid	14		
Glycerol	16		
Glycerol-galactopyranoside	22		
Glycine	35	30	24
Glycolic acid	13		
Histidine		16	15
Inositol	68		
Inositol related compound	26		
Iso-citric acid	21		
Iso-leucine	36	30	25
Lactic acid	13		14
Leucine	22	27	26
Lysine	43	22	16
Mannitol	26		
Mannose	17		
Meso-erythrytol	16		
Methanol			21
Methionine	57	34	31
Myo-inositol	19		25
Ornithine	46		
Phenylalanine	19	23	30
Phosphoric acid	214		
Phosphorylethanolamine	47		
Proline	45	49	
Pseudo uridine	21		
Pyroglutamic acid	31		
Pyruvic acid	28		23
Quinic acid	90		
Ribitol	30		
Ribonic acid	26		
Ribose	12		
Serine	36	17	
sn-Glycerol-3-Phosphate	48		
Succinic acid			23
Sucrose	64		
Threonic acid	33		
Threonine	38	27	20
Trimethylamine-N-oxide			18
Tryptophan		24	
Tyrosine		30	27
Urea	28		
Uric acid	69		
Valine	33	28	28
Xanthine			86
Xylonic acid	45		
Xylose	82		

For the experimental sample set of 42 human CSF samples from patients having neurological diseases a similar profile for the variation in metabolite level could be observed as for the validation sample set from neurologically normal individuals (see Figure 3 and Supplementary Material). A number of metabolites show a significantly higher RSD for the experimental samples, which is most probably due to the heterogeneity of the experimental CSF samples, as discussed in the proteomics section. Metabolites that show significantly higher RSDs for the experimental samples are 3,4-dihydroxybutanoic acid, fructose, ascorbic acid, glyceric acid, pyruvic acid and 2-aminobutyric acid. However there is no clear relation between these metabolites and the neurological disease in the experimental sample set.

For NMR only five individual CSF samples were analysed in the original sample set due to limited available sample volumes. Analysis of the CSF samples by NMR resulted in a list of 51 metabolites of which 41 could be quantified (see Supplementary Material). All metabolites observed with NMR were detected in all samples. The biological variation ranged from 8 to 53% while the analytical error was between 3 and 9 % for all metabolites (Supplementary Material). The concentrations found for the pooled CSF sample were very similar to that of the averages of the individual samples.

The 27 CSF samples in the validation sample set were analyzed by NMR using somewhat different conditions, i.e. more diluted and 600 MHz instead of 800 MHz. However, the same set of metabolites could be quantified in these samples, except for urea. In this sample set the biological variation ranged from 12 to 143% which is somewhat higher compared to the original sample set (see Table 3 and Supplementary material). The analytical error is in the same range as observed for the original sample set, i.e. 2 to 9%. PCA on the NMR data did not show any significant age or gender effect (see Figures S5 and S6 in Supplementary Material).

Of the 41 metabolites quantified by NMR in the validation sample set, 21 were also detected by GC-MS. Some of the more volatile metabolites, like acetone and methanol, can only be analyzed by NMR, showing that despite the overlap, NMR and GC-MS are complementary techniques. Furthermore, NMR can detect a number of metabolites that are difficult to analyze by GC-MS, because they cannot be derivatized, like choline, or they can give unstable derivatives, like arginine. On the other hand, a range of

metabolites was only observed by GC-MS and not by NMR. In most cases these metabolites either have no proton signal, e.g. uric acid and phosphoric acid, or the concentration is below the detection limit of NMR, e.g. dihydroxybutanoic acids and proline.

The absolute concentrations of amino acids in eight individual CSF samples of the original sample set were determined by a targeted LC-MS/MS. One of the individual CSF samples was omitted due to technical failure. The analytical error is less than the biological variation for all amino acids. The biological variation ranges from 28 to 52% while the analytical error is less than 12% (see Supplementary Material). Despite the differences between samples, all amino acids were present in every individual sample. Again, it can be seen that the concentrations found for the pooled CSF sample were very similar to that of the averages of the individual samples. Glutamic acid and aspartic acid could not be quantified in a reliable way in the CSF samples.

Analysis of the same amino acids in the 27 samples of the validation sample set resulted in similar biological variation, i.e. 14 to 49%, and analytical error, i.e. 1 to 9% (see Table 3 and Supplementary Material). PCA showed no significant age or gender effects (see Figures S7 and S8 in Supplementary Material).

Most of the amino acids analyzed by LC-MS/MS were also detected either by GC-MS or NMR. However, one of the advantages of the targeted LC-MS/MS method is the low sample volume required for analysis, i.e. 10 µl vs. 60-100 µl for NMR and GC-MS, respectively.

Comparison of the RSD of metabolites that could be analyzed with more than one of the analytical methods, as shown in Table 3, shows that on average the biological variation of a metabolite is similar for different methods. Deviations occur mainly for low abundant metabolites, like 3-hydroxybutanoic acid, and metabolites that show relative high analytical errors for certain methods, e.g. creatinine with GC-MS.

3.3 DISCUSSION

In this study, we investigated metabolite and protein identities, and their abundances and inter-individual variations in abundance in CSF by analyzing a unique and well-defined set of CSF samples and a corresponding pooled CSF sample. Here we have strictly defined criteria to exclude blood contaminated CSF. These criteria warrant that at a certain threshold no contamination is observed, however contamination not exceeding this threshold can still exist and cannot be ruled out.

Combination of three different analytical techniques for metabolites used in this study resulted in a list of about 89 identified metabolites in CSF that can routinely be analyzed, which is about a third of the metabolites in CSF present in the human metabolome database [30]. It is expected that many of the metabolites that are not detected by NMR and GC-MS are low abundant metabolites, i.e. neurotransmitters, steroids, eicosanoids, for which more specific, targeted methods are required [31-33]. However, these methods often require significant amounts of CSF and should therefore only be used in metabolomics studies when there is evidence that these metabolites are of importance and/or when enough sample volume is available. Furthermore, some metabolites are (almost) absent in normal controls and are only detectable in diseased persons [7].

All endogenous metabolites detected with the three analytical methods in this study were observed in all individual CSF samples. This implies that the qualitative metabolite composition of CSF in normal controls is relatively similar between individuals. This is generally also observed for plasma of healthy persons in contrary to urine, which is more influenced by dietary intake.

With NMR and LC-MS/MS it was possible to determine the absolute concentration of metabolites. This in contrast to GC-MS for which metabolites can only be quantified when either internal standards or calibration curves for each metabolite are used, which is practically not feasible and therefore this method, like many other non-targeted methods, is used to measure relative differences in metabolite concentrations between groups or individuals. The absolute concentrations determined by LC-MS/MS and NMR agree well with values reported in literature (13, 34-36). For example, comparison of the concentrations of metabolites detected by NMR with values determined by Wishart et al. (13) and literature values referred to in this paper show that in most cases the values fall

within the same concentration range (see Supplementary Material). Furthermore, the concentrations of metabolites determined by more than one method in the validation sample set are also in good agreement.

Although the study of Hu *et al.* [16] mainly focussed on the variation of specific protein abundances within individuals, they concluded that inter-individual variation is far more extensive than intra-individual variation. Yet in that study two different stages of Alzheimer's disease were included, which could potentially influence the levels of protein variation between individuals. Here we examined well-defined CSF samples taken from patients without neurological disorders, and also found profound differences in protein abundances between individuals. Characterization of variation of CSF levels of amyloid beta [34] and apolipoprotein E [35] in patients with Alzheimer's disease have been published, but this is the first attempt to characterize a large number of proteins in CSF of patients without neurological afflictions. Some proteins, such as serotransferrin and fibulin-1 appear to be more constant than others with regards to abundance levels in CSF, as these showed only limited variations in both the sample set of nine non-neurological individuals and the validation set as well as in the experimental sample set. Other proteins, such as contactin-2 and cadherin-13, showed large variations in abundance levels in all three data sets, while proteins related to inflammatory response showed the largest variation in the experimental sample set (Figure 2). This is, in all likelihood, due to the well-known neuroinflammatory component of multiple sclerosis [36-38], because the abundance of neuroinflammatory proteins was far higher in the multiple sclerosis samples. This was most obvious in the primary progressive multiple sclerosis patients, so the increased variance of immunoglobulin proteins in this data set is likely due to this group of samples as this type of multiple sclerosis is characterized by continuous inflammation in the central nervous system, which would result in higher concentrations of inflammation-related proteins.

Additionally, it must be noted that the proteins with high inter-individual variation were not only low abundant proteins, but also high abundant proteins such as haptoglobin, indicating that these high variations were not caused by measuring at the limit of the detection capabilities of the machines. Proteins specific to the central nervous system appear to be more variable between individuals than proteins that originate from blood.

The majority of the central nervous system specific proteins have high inter-individual variations, like for example neurotrimin and neuroserpin, whereas a large number of the proteins from blood show low inter-individual variation, such as serotransferrin and ceruloplasmin. However, there are also blood specific proteins that display high inter-individual variation, like haptoglobin.

Matching biological samples for age, gender and protein concentration is an essential step in biomarker research. However, in the validation sample set we observed that proteins with high inter-individual variance were influenced by gender and age only in a limited way. A good example of this is apolipoprotein E, a protein that is reported to be present at lower concentrations in CSF of patients with Alzheimer's disease, regardless of age and gender variability [35]. Here we found a similar RSD (23-29%) for all variables (gender and age-group) tested, suggesting a limited influence of both age and gender on the inter-individual protein abundance of apolipoprotein E.

Although all metabolites detected were present in all individual samples the concentration of metabolites differed strongly between individuals. For all metabolites the analytical variation was significantly less than the biological variation. The biological variation was not caused in a significant way by age or gender. The biological variation in this study is about 15% for 70% for the majority of the metabolites which is lower than was observed by Wishart *et al.* [13], but very similar to the variation reported in the recent study of Crews *et al.* [39]. The main difference between the two studies is the type of CSF sample that was used, i.e. persons without neurological disorders vs. patients screened for meningitis. Therefore, it can be concluded that the biological variation for normal controls is, as expected, less than for neurologically diseased individuals.

Lactic acid and glucose, two high abundant metabolites that can be detected by both NMR and GC-MS, show relatively low biological variation, i.e. < 20%, as well as some medium abundant metabolites like citric acid and glutamine (Table 3). In general most medium abundant compounds show RSD < 30%. Most compounds that show high biological variation are relatively low abundant, e.g. proline, 1- and 3-methylhistidine, xanthine. Even more interesting is the fact that some low abundant metabolites observed with GC-MS show low biological variation, like 3-hydroxypropionic acid, arabinose, fucose, glyceric acid, glycerol, glycolic acid, ribose, while others show very

high biological variation, e.g. 3-hydroxybutanoic acid, 2-piperidon, inositol, phosphoric acid, sucrose, uric acid and xylose (see Table 3). Both types of metabolites include organic acids and carbohydrates and there is yet no clear biological reason why some metabolites in this study show much higher biological variation than others. The biological variation of the experimental sample set of 42 human CSF samples showed a similar trend as the nine CSF samples, although for a number of metabolites the biological variation was significantly higher (Figure 3). The latter can of course be attributed to the different diseases of the subjects that might lead to general differences in metabolite levels. Metabolites that showed a significant higher biological variation in the experimental samples could not be directly related to neurological disorders. Further analysis of the data of the experimental samples is necessary in order to find relations between metabolites and the different types of neurological disorders present in the samples set, including the different stages of multiple sclerosis. More interestingly, for a significant number of metabolites, the biological variation in diseased subjects are similar to that of normal controls, indicating that part of the CSF metabolome is more influenced by person to person differences and that the contribution of the diseases is only minor.

The biological variation of metabolites that were detected by more than one analytical method, showed in general good agreement (see Table 3). Amino acids determined by LC-MS/MS and NMR showed in general <10% difference in RSD and the same was true, with some exceptions, for metabolites analyzed by both GC-MS and NMR.

The work discussed above showed that metabolomics, i.e. non-targeted analysis of as many metabolites as possible, of CSF is possible with a combination of analytical techniques currently available. Depending on the amount of CSF available and existing knowledge with respect to the biological question that has to be answered, a combination of non-targeted and targeted analytical methods is preferred to cover different classes of metabolites ranging from high to low abundant metabolites. The metabolites that were detected in CSF seemed to be quite similar between normal controls although the concentration of metabolites can differ between individuals up to 60% depending on the specific metabolite.

3.4 CONCLUDING REMARKS

From the previous discussion we conclude that for most proteins the biological variation in two sets of normal control CSF samples, i.e. patients without any significant neurological disorders, appears to be limited, e.g. serotransferrin with RSD 25% (original sample set) and 18% (validation sample set), which includes a technical variation of approximately 20%. The majority of the identified proteins show lower than 60% RSD. However, for 28% of the identified proteins the RSD is above 60% and for a limited number of proteins (5% of total) the inter-individual variation is extensive (RSD > 100%, original sample set). The results of extensive inter-individual variation for 5% of the identified proteins is not limited to low abundant proteins but also several high abundant proteins shows extensive biological variation, e.g. haptoglobin with RSD of 135% in the original sample set (Table 4).

Table 4. Proteins with high and low biological variation in normal control CSF samples (original sample set).

Low variation between individuals			**High variation between individuals**		
Accession number, protein	Biological variation (%)	Technical variation (%)	Accession number, protein	Biological variation (%)	Technical variation (%)
P01011, Alpha-1-antichymotrypsin	26	17	Q9BYH1, Seizure 6-like protein	102	21
P07339, Cathepsin D	26	17	Q8TCZ2, Voltage-dependent calcium channel subunit alpha-2/delta-1	103	19
P23142, Fibulin-1	27	16	P54764, Ephrin type-A receptor 4	107	19
P02774, Vitamin D-binding protein	29	21	Q02246, Contactin-2	124	19
P17900, Ganglioside GM2 activator	29	23	P00738, Haptoglobin	135	28
P02749, Beta-2-glycoprotein	30	18	Q86YZ3, Hornerin	148	28

Metabolomics analysis on the same CSF samples showed that the biological variation for most CSF metabolites is limited especially compared to the proteomics results. Only a few metabolites were observed with a biological RSD > 70%. However, within the group of metabolites that could be quantified with the different analytical methods, substantial differences in RSDs could be observed between individual metabolites. For example, glucose, a high abundant metabolite, showed a RSD of only 12% while for acetic acid a RSD of 52% was observed.

These results show that for CSF biomarker discovery research, it is essential to have an understanding of the biological variation between normal controls, because observation of differential abundance between controls and diseased individuals must necessarily be weighed against known inter-individual variations in normal controls. Proteins and metabolites showing high RSD in healthy CSF ought to be assessed with caution as candidate biomarker, since a large part of the observed difference will not be due to the disease under investigation, but to the natural biological variation between individuals.

ACKNOWLEDGEMENTS

This study was performed within the framework of Top Institute Pharma project number D4-102.

REFERENCES

1. Frankfort, S. V.; Tulner, L. R.; van Campen, J. P.; Verbeek, M. M.; Jansen, R. W.; Beijnen, J. H., Amyloid beta protein and tau in cerebrospinal fluid and plasma as biomarkers for dementia: a review of recent literature. *Current clinical pharmacology* **2008,** 3, (2), 123-31.
2. Helbok, R.; Broessner, G.; Pfausler, B.; Schmutzhard, E., Chronic meningitis. *Journal of neurology* **2009,** 256, (2), 168-75.
3. Lewczuk, P.; Hornegger, J.; Zimmermann, R.; Otto, M.; Wiltfang, J.; Kornhuber, J., Neurochemical dementia diagnostics: assays in CSF and blood. *European archives of psychiatry and clinical neuroscience* **2008,** 258 Suppl 5, 44-9.
4. Romeo, M. J.; Espina, V.; Lowenthal, M.; Espina, B. H.; Petricoin, E. F., 3rd; Liotta, L. A., CSF proteome: a protein repository for potential biomarker identification. *Expert review of proteomics* **2005,** 2, (1), 57-70.
5. Johnston, I.; Teo, C., Disorders of CSF hydrodynamics. *Childs Nerv Syst* **2000,** 16, (10-11), 776-99.
6. Taniguchi, M.; Okayama, Y.; Hashimoto, Y.; Kitaura, M.; Jimbo, D.; Wakutani, Y.; Wada-Isoe, K.; Nakashima, K.; Akatsu, H.; Furukawa, K.; Arai, H.; Urakami, K., Sugar chains of cerebrospinal fluid transferrin as a new biological marker of Alzheimer's disease. *Dementia and geriatric cognitive disorders* **2008,** 26, (2), 117-22.
7. Moolenaar, S., von Engelke, U., Hoenderop, S., Morava, E., van der Graaf, M., Wevers, R., *Handbook of ^{1}H-NMR spectroscopy in inborn errors of metabolism*. Heilbronn: SPS verlagsgesellschaft: 2002.
8. Baran, R.; Reindl, W.; Northen, T. R., Mass spectrometry based metabolomics and enzymatic assays for functional genomics. *Current opinion in microbiology* **2009**.
9. Lindon, J. C., Nicholson, J. K., Everett, J. R., NMR spectroscopy of biofluids. In *Annual reports on NMR spectroscopy*, Webb, G. A., Ed. Academic Press: London, 1999; Vol. 38, pp 1-88.
10. Lutz, N. W.; Viola, A.; Malikova, I.; Confort-Gouny, S.; Audoin, B.; Ranjeva, J. P.; Pelletier, J.; Cozzone, P. J., Inflammatory multiple-sclerosis plaques generate characteristic metabolic profiles in cerebrospinal fluid. *PLoS ONE* **2007,** 2, (7), e595.
11. Kawashima, H.; Oguchi, M.; Ioi, H.; Amaha, M.; Yamanaka, G.; Kashiwagi, Y.; Takekuma, K.; Yamazaki, Y.; Hoshika, A.; Watanabe, Y., Primary biomarkers in cerebral spinal fluid obtained from patients with influenza-associated encephalopathy analyzed by metabolomics. *The International journal of neuroscience* **2006,** 116, (8), 927-36.
12. Myint, K. T.; Aoshima, K.; Tanaka, S.; Nakamura, T.; Oda, Y., Quantitative profiling of polar cationic metabolites in human cerebrospinal fluid by reversed-phase nanoliquid chromatography/mass spectrometry. *Analytical chemistry* **2009,** 81, (3), 1121-9.
13. Wishart, D. S.; Lewis, M. J.; Morrissey, J. A.; Flegel, M. D.; Jeroncic, K.; Xiong, Y.; Cheng, D.; Eisner, R.; Gautam, B.; Tzur, D.; Sawhney, S.; Bamforth, F.; Greiner, R.; Li, L., The human cerebrospinal fluid metabolome. *Journal of chromatography* **2008,** 871, (2), 164-73.
14. Pan, S.; Zhu, D.; Quinn, J. F.; Peskind, E. R.; Montine, T. J.; Lin, B.; Goodlett, D. R.; Taylor, G.; Eng, J.; Zhang, J., A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *Proteomics* **2007,** 7, (3), 469-73.
15. Pan, S.; Rush, J.; Peskind, E. R.; Galasko, D.; Chung, K.; Quinn, J.; Jankovic, J.; Leverenz, J. B.; Zabetian, C.; Pan, C.; Wang, Y.; Oh, J. H.; Gao, J.; Zhang, J.; Montine, T.; Zhang, J., Application of targeted quantitative proteomics analysis in human cerebrospinal fluid using a liquid chromatography matrix-assisted laser desorption/ionization time-of-flight tandem mass spectrometer (LC MALDI TOF/TOF) platform. *Journal of proteome research* **2008,** 7, (2), 720-30.
16. Hu, Y.; Malone, J. P.; Fagan, A. M.; Townsend, R. R.; Holtzman, D. M., Comparative proteomic analysis of intra- and interindividual variation in human cerebrospinal fluid. *Mol Cell Proteomics* **2005,** 4, (12), 2000-9.
17. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **2005,** 4, (12), 2010-21.
18. Stoop, M. P.; Dekker, L. J.; Titulaer, M. K.; Lamers, R. J.; Burgers, P. C.; Sillevis Smitt, P. A.; van Gool, A. J.; Luider, T. M.; Hintzen, R. Q., Quantitative matrix-assisted laser desorption ionization-fourier

transform ion cyclotron resonance (MALDI-FT-ICR) peptide profiling and identification of multiple-sclerosis-related proteins. *Journal of proteome research* **2009,** 8, (3), 1404-14.
19. Mize, T. H.; Amster, I. J., Broad-band ion accumulation with an internal source MALDI-FTICR-MS. *Analytical chemistry* **2000,** 72, (24), 5886-91.
20. Moyer, S. C.; Budnik, B. A.; Pittman, J. L.; Costello, C. E.; O'Connor, P. B., Attomole peptide analysis by high-pressure matrix-assisted laser desorption/ionization Fourier transform mass spectrometry. *Analytical chemistry* **2003,** 75, (23), 6449-54.
21. O'Connor, P. B.; Costello, C. E., Application of multishot acquisition in Fourier transform mass spectrometry. *Analytical chemistry* **2000,** 72, (20), 5125-30.
22. van Kampen, J. J.; Burgers, P. C.; de Groot, R.; Osterhaus, A. D.; Reedijk, M. L.; Verschuren, E. J.; Gruters, R. A.; Luider, T. M., Quantitative analysis of HIV-1 protease inhibitors in cell lysates using MALDI-FTICR mass spectrometry. *Analytical chemistry* **2008,** 80, (10), 3751-6.
23. Frequin, S. T.; Barkhof, F.; Lamers, K. J.; Hommes, O. R.; Borm, G. F., CSF myelin basic protein, IgG and IgM levels in 101 MS patients before and after treatment with high-dose intravenous methylprednisolone. *Acta neurologica Scandinavica* **1992,** 86, (3), 291-7.
24. Presslauer, S.; Milosavljevic, D.; Brucke, T.; Bayer, P.; Hubl, W., Elevated levels of kappa free light chains in CSF support the diagnosis of multiple sclerosis. *Journal of neurology* **2008,** 255, (10), 1508-14.
25. Williams, K. C.; Ulvestad, E.; Hickey, W. F., Immunology of multiple sclerosis. *Clinical neuroscience (New York, N.Y* **1994,** 2, (3-4), 229-45.
26. Koek, M. M.; Muilwijk, B.; van der Werf, M. J.; Hankemeier, T., Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical chemistry* **2006,** 78, (4), 1272-81.
27. Massart, L. M., Vandenginste, B. G. M., Buydens, L. M. C., De Jong, S., Lewi, P. J., Smeyers-Verbeke, J. , . *Handbook of Chemometrics and Qualimetrics, Part A*. Elsevier: Amsterdam, 1997.
28. http://cran.r-project.org/web/packages/chemCal/index.html. http://cran.r-project.org/web/packages/chemCal/index.html
29. Sweatman, B. C.; Farrant, R. D.; Holmes, E.; Ghauri, F. Y.; Nicholson, J. K.; Lindon, J. C., 600 MHz 1H-NMR spectroscopy of human cerebrospinal fluid: effects of sample manipulation and assignment of resonances. *Journal of pharmaceutical and biomedical analysis* **1993,** 11, (8), 651-64.
30. Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L., HMDB: the Human Metabolome Database. *Nucleic acids research* **2007,** 35, (Database issue), D521-6.
31. Kim, Y. S.; Zhang, H.; Kim, H. Y., Profiling neurosteroids in cerebrospinal fluids and plasma by gas chromatography/electron capture negative chemical ionization mass spectrometry. *Analytical biochemistry* **2000,** 277, (2), 187-95.
32. Obata, T.; Nagakura, T.; Maeda, H.; Yamashita, K.; Maekawa, K., Simultaneous assay of prostaglandins and thromboxane in the cerebrospinal fluid by gas chromatography-mass spectrometry-selected ion monitoring. *J Chromatogr B Biomed Sci Appl* **1999,** 731, (1), 73-81.
33. Rodriguez, S.; Vio, K.; Wagner, C.; Barria, M.; Navarrete, E. H.; Ramirez, V. D.; Perez-Figares, J. M.; Rodriguez, E. M., Changes in the cerebrospinal-fluid monoamines in rats with an immunoneutralization of the subcommissural organ-Reissner's fiber complex by maternal delivery of antibodies. *Experimental brain research. Experimentelle Hirnforschung* **1999,** 128, (3), 278-90.
34. Bateman, R. J.; Wen, G.; Morris, J. C.; Holtzman, D. M., Fluctuations of CSF amyloid-beta levels: implications for a diagnostic and therapeutic biomarker. *Neurology* **2007,** 68, (9), 666-9.
35. Lefranc, D.; Vermersch, P.; Dallongeville, J.; Daems-Monpeurt, C.; Petit, H.; Delacourte, A., Relevance of the quantification of apolipoprotein E in the cerebrospinal fluid in Alzheimer's disease. *Neuroscience letters* **1996,** 212, (2), 91-4.
36. Compston, A.; Coles, A., Multiple sclerosis. *Lancet* **2002,** 359, (9313), 1221-31.
37. Frohman, E. M.; Eagar, T.; Monson, N.; Stuve, O.; Karandikar, N., Immunologic mechanisms of multiple sclerosis. *Neuroimaging clinics of North America* **2008,** 18, (4), 577-88, ix.
38. Uccelli, A.; Pedemonte, E.; Narciso, E.; Mancardi, G., Biological markers of the inflammatory phase of multiple sclerosis. *Neurol Sci* **2003,** 24 Suppl 5, S271-4.

39. Crews, B.; Wikoff, W. R.; Patti, G. J.; Woo, H. K.; Kalisiak, E.; Heideker, J.; Siuzdak, G., 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), 8538-44.

CHAPTER 4

NMR AND PATTERN RECOGNITION CAN DISTINGUISH NEUROINFLAMMATION AND PERIPHERAL INFLAMMATION

A. Smolinska, A. Attali, L. Blanchet, K. Ampt, T. Tuinstra, H. van Aken, E. Suidgeest, A. J. van Gool, T. Luider, S. S. Wijmenga, and L. M.C. Buydens

Journal of Proteome Research (2011), 10 (10), pp 4428–4438

ABSTRACT

Multiple Sclerosis (MScl) is a neurodegenerative disease of the CNS, associated with chronic neuroinflammation. Cerebrospinal fluid (CSF), being in closest interaction with CNS, was used to profile neuroinflammation in order to discover disease-specific markers. We used the commonly accepted animal model for the neuroinflammatory aspect of MScl: the Experimental Autoimmune/Allergic Encephalomyelitis (EAE). An combination of advanced ^{1}H-NMR spectroscopy and pattern recognition methods, was used to establish the metabolic profile of CSF of EAE-affected rats (representing neuroinflammation) and of two control groups (healthy and peripherally inflamed) to detect specific markers for early neuroinflammation. We found that the CSF metabolic profile for neuroinflammation is distinct from healthy and peripheral inflammation and characterized by changes in concentrations of metabolites such as creatine, arginine and lysine. Using these disease specific markers we were able to detect early stage neuroinflammation, with high accuracy in a second independent set of animals. This confirms the predictive value of these markers. These findings from the EAE model may help to develop a molecular diagnosis for the early stage MScl in humans.

4.1 INTRODUCTION

Multiple Sclerosis (MScl) is a chronic progressive inflammatory, presumably autoimmune disease of the human central nervous system (CNS) in which the fatty myelin sheaths which surround the axons of the neurons of the brain and spinal cord are damaged, leading to demyelination.[1] MScl is one of the most common neurological diseases and usually starts in early adulthood and progresses to serious neurological disability. It has an enormous impact on the health care system and economy of different countries. Currently, MScl afflicts approximately one million people worldwide and the total cost of MScl has been estimated at 12.5 billion € per year.[2, 3]

The animal model of MScl, the Experimental Autoimmune/Allergic Encephalomyelitis (EAE) model,[4-9] has become an important tool for understanding the human disease. EAE is a cell-mediated experimental autoimmune disorder of the CNS and shares its clinical expression and pathological picture with that of MScl. For instance, the Lewis rats in which neuroinflammation is induced by means of Myelin Basic Protein (MBP), as in our study here, display a typical disease curve resembling that of the beginning of MScl Relapsing-Remitting type (RR-MScl). RR-MScl is the most common type of the early stage of MScl. EAE is therefore a useful model for the neuroinflammatory aspect of MScl. Note that EAE does not mimic neurodegeneration and widespread demyelination seen in MScl.[10] In EAE demyelination might be present in the ventral root exit and dorsal root entry zone of the spinal cord or even absent. Also, EAE is monophasic, whereas MScl has a random relapsing-remitting or chronic progressive pattern.[11] It is worth mentioning that EAE has led to development of the hypothesized immunological basis of MScl pathology.[12] Moreover, the EAE has been instrumental in discovering and developing three of the six currently approved therapies for MScl that diminish its symptoms: Copaxone, Mitoxantrone, and Natalizumab.[13, 14] For instance, Mitoxantrone, a known cancer drug, was tested in EAE and found to have positive effects, which led to its use in relapsing-remitting MScl, slowing its progression into secondary progressive MScl.[15]

Unambiguous current clinical diagnosis of MScl remains difficult, particularly in its early stage, due to the complexity of its pathology and the similarities of these pathologies to that of other neurological diseases/inflammations. Diagnosis in an early stage is

important as early intervention appears beneficial to slow down the long-term progression of the disease.[16] A molecular biomarker derived for neuroinflammation in the animal model EAE may form a first step towards such an early diagnosis.

The main objective of this study is to find in the cerebrospinal fluid (CSF) of EAE induced Lewis rats metabolic markers of neuroinflammation and to differentiate neuroinflammation from peripheral inflammation. For this a controlled animal study was set-up, with a healthy control group, a group injected with inflammatory 'booster' (Complete Freund Adjuvant emulsion, CFA) to induce peripheral inflammation and a group injected with CFA and in addition MBP to induce neuroinflammation. Disease progression was monitored and CSF samples were collected at two different time points. To analyze the metabolic profile of the CSF we used an untargeted and unbiased biomarker discovery approach in which high-field 1D ^{1}H-NMR is combined with pattern recognition methods. To our knowledge, this is the first study applying ^{1}H-NMR to analyze rat CSF in the EAE model. A fact that very few rat CSF metabolite studies have been done by NMR is likely due to the limited amount of CSF in rodents. This is not a standard procedure to use such little sample volume in NMR. High quality data could be obtained using only 10 µL of CSF, thanks to the use of advanced NMR including high-field and cryo-probe technology.

We found that the CSF metabolic profile for neuroinflammation is distinct from that of healthy and peripheral inflammation. The metabolites, identified as specific for neuroinflammation, were investigated in a second independent set of animals for the prediction of the early stage of neuroinflammation. The validation of the findings with an external experiment is often acclaimed but rarely practiced in a single study. These disease specific markers, detected in the animal model, may lead to a better understanding of the metabolism underlying neuroinflammation and help to develop a molecular diagnosis for the early stage of MScl in humans.

4.2 MATERIALS AND METHODS

4.2.1 Experimental design of EAE models

The Experimental Autoimmune/Allergic Encephalomyelitis (EAE) is one of the most intensively examined and best characterized animal models of autoimmune disease [9]. EAE shares similarities with MScl. Although the EAE model does not mimic all aspects of MScl, this rodent disease model is an excellent experimental system for understanding aspects of the MScl disease.

The first set of male Lewis rats (Harlan Laboratories B.V., the Netherlands) was inoculated on Day 0 as previously described.[17] Briefly, a 100 µL saline based emulsion containing 50 uL Complete Freund Adjuvant H37 RA (CFA, Difco Laboratories, Detroit, MI), 500 µL Mycobacterium tuberculosis type H37RA (Difco) and 20 µg guinea pig myelin basic protein (MBP) was injected subcutaneously in the pad of the left hind paw of isoflurane anaesthetized animals. CFA was injected for boosting of the immune system, while MBP was injected for induction of neuroinflammation. Next to these MBP challenged rats, referred to as the EAE group (or neuroinflamed group), two control groups were included: a group of animals receiving the same emulsion without MBP (CFA group = peripheral inflammation) and a healthy group undergoing anesthesia only (healthy control). Each group consisted of 30 animals. In each group half of the animals was sacrificed to collect CSF on Day 10 (Day of onset in EAE group) and the other half on Day 14 (peak of disease in EAE group). The typical progression of disease is shown in Figure 1a, while the design of the first EAE experiment is summarized in Table 1.

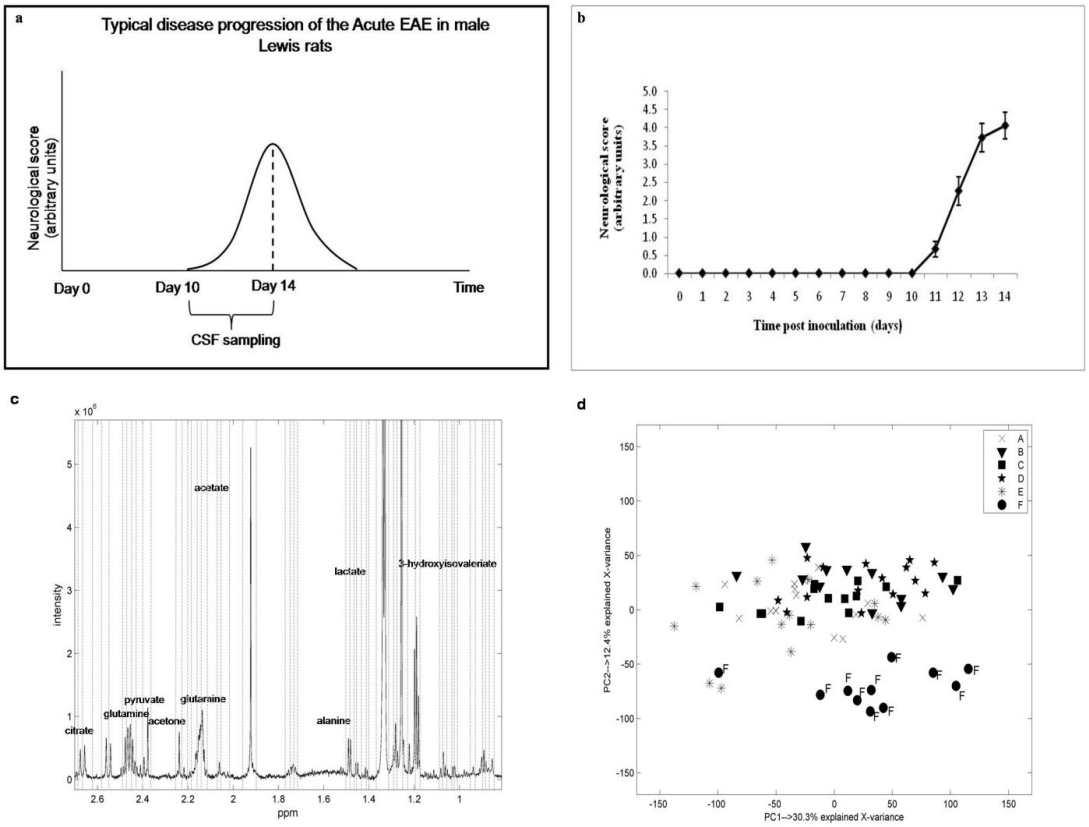


Figure 1. (a) EAE disease progression. (**b)** The average neurological score screened for group "N14". The vertical bars indicate standard error of the mean; (**c)** ^{1}H-800 MHz NMR spectrum of rat CSF. Vertical lines indicate the bins; **(d)** PCA score plot of the ^{1}H-NMR spectra of rat CSF.

Animals were kept under normal housing conditions with water and food and libitum, weighing between 175 and 225 grams at the start of the experiment. Animals were group housed per 3 and cages were randomized across treatments and disease duration.

A second set of male Lewis rats was used to perform another EAE experiment, one year after the first one. The animals were inoculated with CFA and/or MBP as described above. The experimental design, the amount of animals (i.e. 15 animals per group) and settings of this model were similar to the first one. The experimental design and the

number of animals per group are summarized in Table 1S in the supplementary material.

Table 1. Experimental Design of EAE model; "n" indicates number of rats (samples) for each group.

Treatment Day 0	Group description	Day 10	Day 14
Anesthesia only	Healthy	C10 n=15	C14 n=15
CFA	Peripheral inflammation	P10 n=15	P14 n=15
CFA+MBP	Neuroinflammation + peripheral inflammation	N10* n=15	N14** n=15

*1 sample was discarded due to blood contamination

**3 samples were discarded due to blood contamination

4.2.3 Data acquisition and analysis

4.2.3.1 Neurological scores

Disease symptoms and weights of all animals from both EAE models were recorded daily. The following scores for motor dysfunctions were used: 0, healthy animal with normal curling reflex at the tail; 1, paralysis of the tip of the tail; 2, loss of muscle tone at the base of the tail; 3, low posture of hind limbs; 4, instability at hips; 5, partial hind limb paralysis; 6, complete hind limb paralysis; 7, paralysis include midriff; 8, quadriplegia; 9, moribund; 10, death due to EAE (supplementary material). All experimental procedures were approved by Abbott's Institutional Animal Care and Use Committee.

4.2.3.2 CSF sampling, sample preparation and data acquisition

On Day 10 and 14, animals were euthanized with CO_2/O_2. Terminal CSF samples were obtained by direct insertion of an insulin syringe needle (Myjector, 29G x 1/2") via the

arachnoid membrane into the Cisterna Magna. For this purpose a skin incision was made followed by a horizontal incision in the musculus trapezius pars descendens to reveal the arachnoid membrane. A maximum volume of 60 µL of CSF was collected per animal. Each sample was centrifuged within 20 min after sampling, for 10 min at 2000g at 4°C. After centrifugation the supernatants were stored at -80° C for further analysis. Previous experiments have shown that collecting up to 60 µL using this technique under these conditions provides hemoglobin-free CSF samples as measured by ESI-Orbitrap (unpublished data). As an additional check fresh samples, supernatant and pellet size were visually scored for hemolysis and samples were discarded if positive.

From the set of 90 samples from the first EAE experiment (Table 1) 4 were contaminated with blood therefore they were excluded from the measurements. A set of 86 CSF samples were prepared and measured as described below.

10µL of rat CSF were thawed at room temperature and 240 µL D_2O (99.96 at.%D) were added to the biofluid in order to obtain sufficient amount of sample for the NMR measurement. TSP-d4 (Sodium 3-(trimethylsilyl)propionate-2,2,3,3-d4) (99 at.%D) was used as internal standard for chemical shift reference (δ 0.00 ppm). For the latter, 25µL of 8.8 mM TSP-d4 stock solution in D2O was added to 250 µL of rat CSF to a final concentration of 0.8 mM TSP. The TSP-d4 stock was prepared by weighing in dry TSP-d4. The pH of the CSF was adjusted to around 7 (7.0 – 7.1) by adding phosphate buffer (9.7 µL 1M, to a final concentration of 35 mM). The final CSF NMR sample (284.7 µL) was then transferred to a SHIGEMI microcell tube for measurements.

The 1D ^{1}H NMR spectra of rat CSF samples were acquired on an 800 MHz Inova (Varian) system equipped with a 5 mm triple-resonance, Z-gradient HCN cold-probe. Suppression of water was achieved by using WATERGATE (delay: 85 µs).[18] For each 1D ^{1}H NMR spectrum 512 scans were accumulated with a spectral width of 9000 Hz resulting in a total of 18K points. The acquisition time for each scan was 2s. Between scans, a 8s relaxation delay was employed. Prior to spectral analysis, all acquired Free Induction Decays (FIDs) were zero-filled to 32K data points, multiplied with a 0.3 Hz line broadening function, Fourier transformed, manually phased and the TSP internal reference peak was set to 0 ppm - by using ACD/SpecManager software version 11.0.[19]

All 86 rat CSF spectra were acquired and preprocessed as described above. However

due to high line broadening of the internal standard (TSP) four spectra were not included in spectral analysis. In total 82 spectra were subsequently transferred to Matlab, version 7.6 (R2008b) (Mathworks, Natick, MA) for further analysis.
The preparation and data acquisition of CSF samples from the second EAE experiment is described in the supplementary material.

4.2.3.3 Preprocessing of NMR spectra

The NMR spectral data was preprocessed, which typically involves baseline correction, alignment, binning, normalization and scaling. Baseline correction of NMR spectra was performed by applying Asymmetric Least Squares method.[20] Fluctuation in experimental conditions like sample temperature, pH, ionic strength can lead to chemical shift variations, therefore NMR spectra were aligned by using improved parametric time warping (I-PTW).[21] A further problem is the high dimensionality of the data (circa 10000 variables). It is common to apply binning to this kind of data, which reduces the number of variables. To perform proper spectral bucketing we used adaptive intelligent binning.[22] The chemical shift range δ 0.75 – 4.15 was used for the binning procedure because it contained relevant information. Next, spectral resonances corresponding to one metabolite were summed and regrouped in one bin. This procedure was applied to the resonances where no overlapping was present and it led to 153 bins in total. To make spectra comparable between different samples, the final step of preprocessing consisted of integral normalization and supervised vast scaling applied to the binned data.[23]

4.2.3.4 Metabolites identification and quantification

Metabolite identification was carried out by using the 800 MHz library of metabolite NMR spectra from the Chenomx NMR Suite 5.1 (pH 6-8). The library of metabolite spectra is obtained based on a database of pure compound spectra acquired using a particular pulse sequence and acquisition parameters, the tn-noesy-presaturation pulse sequence with 4s acquisition time and 1s of recycle delay.[24] The Chenomx NMR Suite software fits the spectral signatures (singlets, doublets, triplets etc), i.e. the peak shapes, of a compound from an internal database of reference spectra to the experimental NMR spectrum.

For quantification, that is determination of the concentrations of individual metabolites, Chenomx NMR Suite 5.1 uses the concentration of the known reference signal as calibration (in this case TSP-d4). Note that Chenomx approach is that peak shapes are fitted to the experimental ones. In contrast, peak integration, which is often employed for quantification, is very sensitive to baseline distortions and even slightly overlapping resonances cannot reliably be integrated. Peak-shape fitting, like employed in Chenomx, is instead not much affected by base-line distortions. Moreover, even when some of the resonances and/or part of a resonance signature (triplet etc.) of a compound overlap with that of another compound, the peak shape can still be fitted with reasonable accuracy and the concentration of the compound reliably determined.

4.2.3.5 Data analysis strategy

Explorative analysis, by means of Principal Component Analysis (PCA),[25] was used first in order to extract and display the systematic variation in the data.

The strategy for multivariate supervised analysis, using Discriminant Analysis by Projection on Latent Structures (PLS-DA) [26] was designed in accordance with the experimental design summarized in Table 1. The data contain three main groups (healthy control “C10” and “C14”, peripheral inflammation “P10” and “P14” and mixture of neuroinflammation and peripheral inflammation “N10” and “N14”) in two time points (Day 10 and Day 14). The two time points (Day 10 and Day 14) were analyzed separately and we always differentiated between two groups. We have used PLS-DA, since it gives more interpretable results and it enabled us to observe the influence of single effects (i.e. peripheral inflammation, neuroinflammation or disease-progression) on the CSF metabolic profile.

To corroborate the results we used another supervised method, ANalysis Of VAriance-Principal Component Analysis (ANOVA-PCA). The analysis was performed on each time point separately as well as on the two time points combined. Variables (metabolites) differentially profiled across groups of interest were selected based on regression coefficients in PLS-DA and on ANOVA-PCA, Hotelling T^2 statistic with p-value inferior than 0.05).[27]

4.2.3.6 Model construction and validation

PLS-DA is a variation of PLS.[28] PLS-DA uses the group information to maximize the separation between groups of observations. It is currently widely used in metabolomics because of its ability to cope with high correlations between variables. In PLS-DA a linear model is constructed according to equation 1:

$$\mathbf{y=Xb + r} \quad (1)$$

Where, **X** is a dataset matrix, y a vector of group memberships, **b** a vector of regression coefficient and **r** a vector of model residuals. The regression coefficients reflect the relative importance of the variables in the PLS-DA model. We divided the data into a training and a test set using the Duplex algorithm[29] in such a way that the number of samples in the training set was equal for every considered group. The amount of samples in the test set was equal to at least 25% of the total number of samples but not more than 30%. To prevent model overfitting we applied the cross model validation (CMV) procedure, introduced by Anderssen et al. and Gidskehaug et al. [30, 31] in which double cross validation procedures are included for model optimization, the variable selection based on jack-knifing and final model performance assessment. All MATLAB routines for performing variable selection can be found in this reference.[31]

Finally, all PLS-DA models were validated with the independent test set. For every considered model the specificity and sensitivity of the test set were calculated. The final PLS-DA model was applied to the whole dataset if the accuracy of the independent test set was satisfactory (i.e. above 90% of correct classification).

In order to predict the class labels ($\hat{\mathbf{Y}}$) of validation set the scaled spectra ($\mathbf{X}_{new}$) have to be multiply by regression coefficient (**b**) obtained from PLS-DA model (i.e. $\hat{\mathbf{Y}}=\mathbf{X}_{new}\mathbf{b}$). Please keep in mind that scaling performed on validation set applied all necessary parameters (i.e. mean and standard deviation) from training set.

4.2.3.7ANOVA-PCA

Previously we showed that this approach allows identifying the relevant variables to distinguish groups.[27] In this study the main factors are the metabolites effect (M) and the

treatments effect (T) (groups) plus a random dimension linked to the individual variations. The ANOVA model is given in equation 2:

$$X_{ijk}=\mu + +M_i + T_j + MT_{ij} + e_{ijk} \quad (2)$$

Where μ is a general mean, e is the error term and MT_{ij} indicates the interactions between main effects. The interaction between metabolites and treatments, i.e. groups, (MT_{ij}) is of highest interest because it reflects the influence of the different treatments on the intensity of metabolites when all others effects are averaged out. Therefore PCA is performed on the matrix of interaction effects to identify metabolites that contribute to this interaction. If the treatment (j) is relevant for the metabolite (i), the interaction MT_{ij} is significantly different from zero.

4.2.3.8 Heat map and correlation network map

In order to represent the relevant metabolites concentrations in groups “C14”, “P10” and “N14” we generated a heat map from NMR metabolomics data. The heat map is a graphical representation of NMR data in two-dimensional map, where the metabolites concentrations are illustrated as colors. Every concentration value was standardized according to the reference group (healthy group “C14”), i.e. by subtracting mean and dividing by the standard deviation of healthy controls. In that way the metabolites concentrations are expressed in values of standard deviation from the control group.

The concentration of a certain metabolite does not fluctuate independently, but may be correlated and change with the others metabolites. To examine the level of observed changes among metabolites we calculated the Spearman’s correlation between relative metabolites concentrations in groups “C14” and “N14”, as well in groups “P10” and “N14”. Spearman correlation calculates the correlation of the ranks for metabolites concentrations, given in equation 3.

$$C_r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (3)$$

Where, C_r is a Spearman’s correlation, d_i is a difference between ranks of each sample on the two metabolites, n is the number of samples.

Extreme variations in concentration values have less influence on the Spearman correlation than on the Pearson's correlation. The calculated correlations are then transferred to Graphviz 1.01 for network map visualization (**http://www.graphviz.org**). In the network map two metabolites are connected with a link if their correlation coefficient is relevant, i.e. exceeds a given threshold. The correlation between metabolites concentrations was significant if its value was superior than 0.7 and p-value was inferior than 0.05.

4.3 RESULTS

4.3.1 Overview of the available data

86 NMR spectra were measured, while 82 spectra were included in further analysis. An example of a ^{1}H-NMR spectrum of rat CSF is shown in Figure 1c in the chemical shift region between 0.8 and 2.7 ppm. This region contains several signals from carboxylic acids and amino acids protons. Some of the metabolites are named in Figure 1c. For the analysis the NMR spectra were divided into 153 bins, which contain resonances of 33 identified metabolites and some unidentified signals. The identified metabolites are listed in the supplementary material. The information concerning the second EAE model can be found in the supplementary material.

The average neurological scores and standard deviations for group “N14” are shown in Figure 1b, the information of each individual animal is listed in the supplementary material. As can be observed the animals showed motor dysfunction at day 11. The average day of onset was 11.56 (+/- 0.2). This means that on average, animals were disabled for a period of 2.4 days (+/- 0.2) at the sampling day. In group “N14” only one animal showed disease duration shorter than 2 days (onset on day 13). The maximum score reached was on average 4.5 (+/- 0.2). Six animals showed a maximum score above this average. The average peak day (as defined by the first day of maximum score per individual animal) was 13.3 (+/-0.2). An idea would be to use the neurological scores for regression purpose for groups “N10” and “N14”. However all animals in group “N10” had neurological scores equal zero, while animals in group “N14” quite spread values from 0 till 6. This would cause an artificial compactness of samples from group “N10”, while group “N14” very spread. This would lead to biased PLS results. Therefore the neurological scores were not used.

4.3.2 Explorative analysis by PCA

Before performing pattern recognition, the spectra of 82 rat CSF samples were checked for outliers. In total 3 spectra were detected as outliers and were removed from the final analysis. Figure 1d shows the PCA score plot of the total dataset (79 vast scaled spectra). This plot shows that the samples belonging to group “N14” are clearly separated from the others samples along PC2. It also reveals that a large source of

variance in the data does not correspond to the available groups. No clear grouping is present since most of the groups overlap. With PCA only, we are not able to distinguish all analyzed groups. Therefore we needed to use more dedicated methods.

4.3.3 Supervised multivariate analysis by PLS-DA and ANOVA-PCA

4.3.3.1 PLS-DA of ^{1}H-NMR CSF data

PLS-DA models of the ^{1}H-NMR spectra were performed to extract information on the metabolic effects of the different group treatments as presented in Table 1. In Table 2S in the supplementary material the specificity, sensitivity, the correct classification rate for the test set and the total number of variables in the PLS-DA models are presented. Metabolites that contributed significantly to the group separation are listed in Table 3S in the supplementary materials. The results of these PLS-DA models urged us to revise the assumed effects for some groups of Table 1. The summary of the particular effects investigated by the different PLS-DA models is presented in Table 4S in the supplementary materials. As the first part of supervised analysis the comparison between “C10” and “C14” has been performed in order to find the metabolites that represent metabolic evolution over days. Later, this information was used to check if any of the discriminating metabolites are related to metabolic variation over time. In the remaining of this paragraph a detailed analysis of the results of the different models is given.

The score plots of final PLS-DA models are presented in Figures 2a-2e. Note that score plots are used only for illustration purpose not determine the classification of PLS-DA model.

Effect of neuroinflammation and peripheral inflammation

The PLS-DA model for groups “C10” versus “N10” allows one to study the combined effect of neuroinflammation and peripheral inflammation. This model has a high prediction ability of 100% for the independent test set. The area under the curve (93.4%) indicates that group “C10” and group “N10” are well separated and the PLS-DA has a high model performance. The CSF metabolic profile of the EAE-affected group can thus without a doubt be differentiated from the metabolic profile of the healthy group.

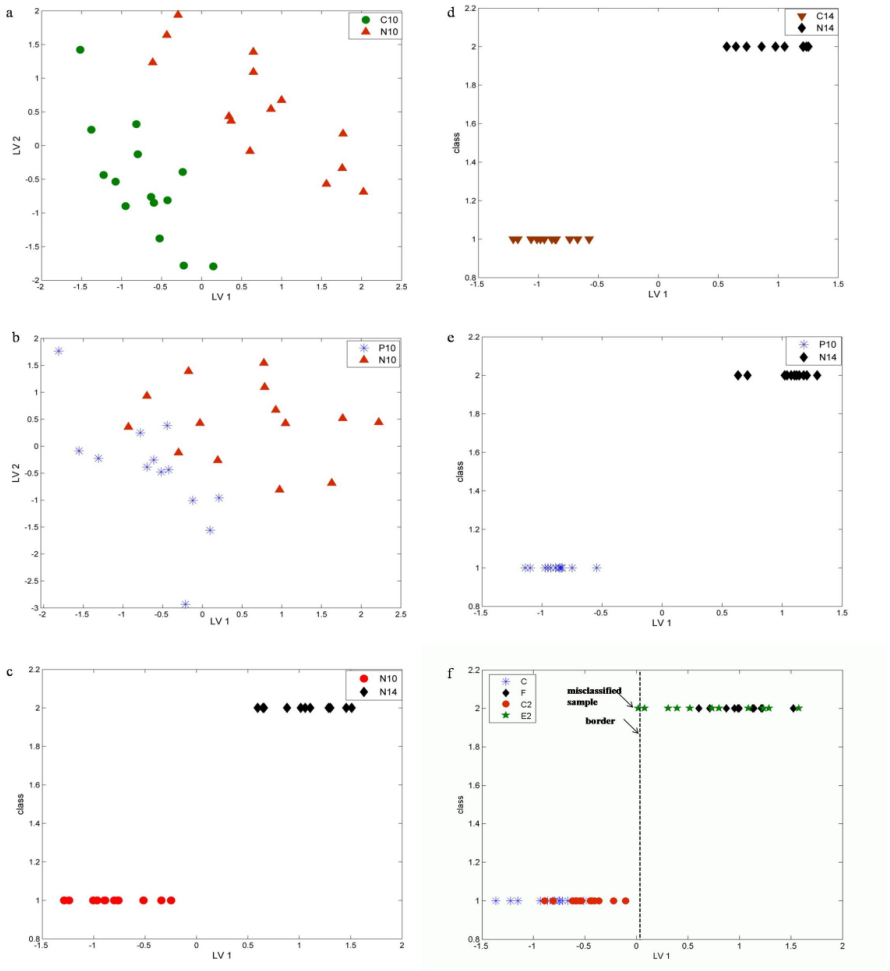


Figure 2. PLS-DA score plots derived from ^{1}H-NMR spectra of rat CSF belonging to: **(a)** groups "C10" and "N10". The amount of Y explained variance for two latent variables was equal to 87.9%. **(b)** groups "P10" and "N10". The amount of explained variance in Y for two latent variables was equal 58.3%. **(c)** groups "N10" and "N14". The amount of explained variance in Y for one latent variable was equal to 89.6%. **(d)** groups "C14" and "N14". The amount of explained variance in Y for one latent variable was equal to 95.2%. **(e)** groups "P10" and "N14". The amount of explained variance in Y for one latent variable was equal to 97.1%. **(f)** The projection of the independent samples from the second EAE model (group "P10-2" and group "N10-2") on the PLS-DA score plot derived from group "P10" versus "N14" of the first EAE model.

Effect of neuroinflammation

The PLS-DA model for groups “P10” versus “N10” allows one to discriminate the effects of peripheral inflammation and neuroinflammation on the CSF metabolic profile. This model shows poor prediction (accuracy for the independent test set is 62.5%) and the sensitivity for group “N10” is very low (predicted as belonging to group “P10”). The score plot (Figure 2b) illustrates two subgroups in group “N10”, one overlapping with group “P10” and a second cluster further away. This result suggests that animals in group “N10” are heterogeneous regarding to the disease response. Probably, within group “N10” some animals did not show any neuroinflammation yet, indeed at Day 10, no neurological deficits have been observed in this particular experiment. Therefore for some animals the peripheral inflammation is still the dominant effect since the neuroinflammation has not developed yet. A hierarchical clustering was performed on group “N10” to confirm the heterogeneity of the group (data not shown). Three samples were identified as outliers for group “N10”. This suggests that the response of the animals to the disease in the EAE model was not uniform and some animals’ response is shifted along time. The neurological scores recorded for group “N10” on average were equal to 0. This means that by examination of the rat’s motor system no external neurological or pathological physical signs were observed. The heterogeneity of group “N14” was further investigated with ROC curve analysis. The area under curve (60%) shows that group “P10” and group “N10” are not well separated.

The PLS-DA model for groups “N10” versus “N14” allows one to study the progression of disease. The classification model has perfect accuracy (overall classification of 100% for independent test set).

Investigation of the second time point (Day 14) by PLS-DA allows one to identify the metabolites differently profiled during the peak of disease. The PLS-DA model of “C14” versus “N14” has perfect prediction ability (overall classification is 100% for independent test set).

Effect of peripheral inflammation during the peak of the disease

The overall correct classification rate for independent test set of this model is only 65%. The area under the ROC curve is 59%. These results suggest that the metabolic profiles of group “C14” and group “P14” are quite similar. Based on these data we can make the

statement that at Day 14 the effect of an immunopotentiator, i.e. CFA, was not measurable based on the metabolite profile. Our interpretation of these results is that peripheral inflammation has vanished by Day 14 and group "P14" is actually healthy. This interpretation is coherent with the absence of external symptoms (i.e. fever and swelling pad). Since group "P14" is not

inflamed anymore the PLS-DA models of group "P14" vs. "N14" and "N14" vs. ("C14&P14") give similar results as the PLS-DA model of the healthy group "C14" vs. group "N14". In both situations the overall correct classification for independent test set is 100%. The PLS-DA model of "P10" vs. "P14" is described in the supplementary material.

P10 vs N14 : Peripheral inflammation effect versus neuroinflammation effect

Based on the experimental design the effect in group "N14" is a combination of peripheral inflammation (CFA effect) and neuroinflammation (MBP effect). However, considering the conclusion that at Day 14 the peripheral inflammation has vanished, group "N14" should show only neuroinflammation. Because group "N10" was ~~is~~ heterogeneous in disease onset, we used groups "P10" and "N14" to distinguish peripheral inflammation from neuroinflammation. We performed this comparison, knowing that groups "P10" and "N14" represent different conditions, i.e. Day 10 for group "P10" and Day 14 for group "N14". This comparison should yield the metabolites influenced by peripheral inflammation or neuroinflammation.

The "P10" vs."N14" PLS-DA model should carry the information about the metabolites differentially profiled in peripherally inflamed and neuroinflamed subjects. The PLS-DA model shows perfect prediction ability with an overall correct classification of 100% for independent test set. Based on our results these two groups indeed have different metabolic profiles.

4.3.3.2 ANOVA-PCA of ^{1}H-NMR CSF data

To confirm the results from the PLS-DA models we used ANOVA-PCA as a corroborative technique to identify metabolites that vary significantly between groups. The factors modeled in this study are the factor treatment (CFA, CFA+MBP or no treatment), and the different metabolites (the Metabolite factor). The resulting biplots of

the ANOVA-PCA are displayed in Figures S2 (a-d), while the group specific metabolites are summarized in Table S5 in the supplementary materials. In the next part of this paragraph an overview of the results derived from ANOVA-PCA are discussed.
Except for a few metabolites, the same set of metabolites for peripheral as well as neuroinflammation was selected by ANOVA-PCA. ANOVA-PCA showed that the metabolic profiles of healthy group, peripheral inflamed and neuroinflamed animals can be distinguished. Moreover it showed that groups “C14” and “P14” have similar metabolic profiles.

The most discriminating metabolites and their absolute concentrations are presented in Table 2; their short biological interpretation is further discussed in the section ‘Discussion and Conclusion'.
As a final step, to make the PLS-DA model more robust and generic (i.e. machine independent), PLS-DA model was constructed for the absolute concentration of the metabolites. This is, however, a time consuming step and therefore we performed it only for the “P10” and “N14” groups. The PLS-DA model for group “P10” versus group “N14” shows perfect prediction ability with an accuracy of 100% on the independent test set. The score plot of this model is shown in the supplementary material (Figure 3S). This model is further validated with an independent set of animals, derived from a second EAE experiment.

4.3.4 Heat map and correlation network map

In order to represent the individual differences in metabolites concentrations of healthy controls (“C14”), peripherally inflamed (“P10”) and neuroinflamed (“N14”) animals a heat map of the metabolomics data was constructed (Figure 3a). The metabolite concentrations were standardized with respect to the healthy group “C14”. The metabolites included in the heat map correspond to the significant ones shown in Table 2. The heat map shows that the metabolite concentrations, characteristic for neuroinflammation, change significantly in group “N14”. For instance arginine concentration is elevated in group “N14” in comparison to healthy and peripherally

inflamed controls. On the contrary alanine is reduced. The intra-individual variations are still observable.

To assess the correlation between metabolites, we calculated Spearman's correlation between relative metabolites concentrations in groups "C14" and "N14", as well as in groups "P10" and "N14". Those correlations are visualized as a network map (Figure 3b-3c) to provide an overview of similarities between different metabolites in the considered groups. In Figure 3b a correlation network map between metabolites in healthy group "C14" and neuroinflamed group "N14" is presented. There are just a few shared correlations (indicated with a dashed line), i.e. significant correlations seen in both groups. These correlations are thus not influenced by neuroinflammation. Figure 3c shows the correlation network map between metabolites in peripherally inflamed group "P10" and neuroinflamed group "N14". There are only a few correlations present in group "P10" (solid, black line) and shared correlations between groups "P10" and "N14" are absent. What is particularly noteworthy is the fact that many correlations, present in the healthy control, vanished due to neuroinflammation, but also some new correlations appeared.

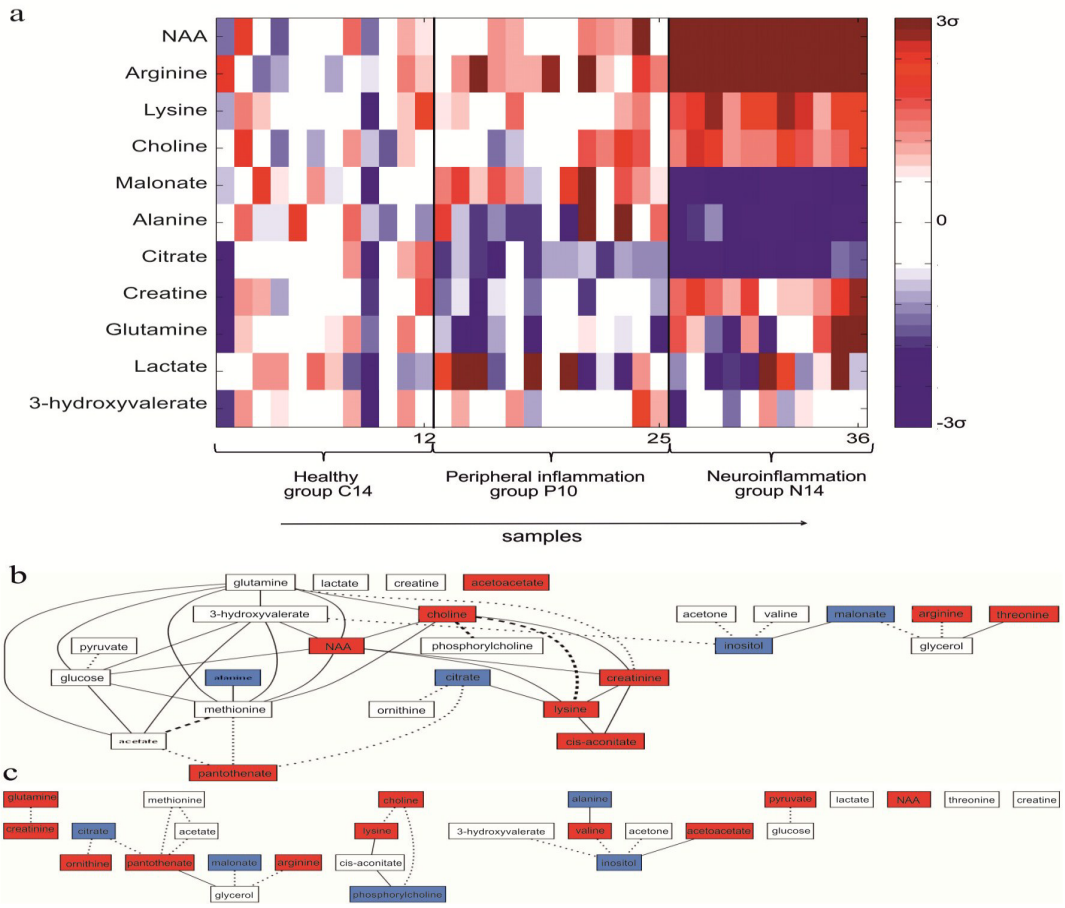


Figure 3. (a) Heat map of selected metabolites for groups “C14”, “P10” and “N14”. Each row represents a metabolite, and each column corresponds to an animal from healthy group “C14” or peripherally inflamed group “P10”, or neuroinflamed group “N14”. The relative metabolite concentrations were standardized with respect to healthy group. Red and blue colors represent elevation or reduction of a given metabolite concentration, respectively. NAA stands for N-acetylaspartate, while σ indicates the standard deviation of the healthy control. **(b)** Correlation network map between metabolites identified with high certainty**:** in healthy group “C14” and neuroinflamed group “N14”; **(c)** in peripherally inflamed group “P10” and neuroinflamed group “N14”. A solid black line indicates a significant correlation (i.e. correlation > 0.7 and p-value < 0.05) occurring in the first group (healthy group “C14” or peripherally inflamed group “P10”). A dotted line signifies the significant correlation unique for neuroinflamed group “N14”. A dashed

line represents a significant correlation seen in each group. The red and blue boxes represent elevated or reduced metabolite concentrations in neuroinflamed group “N14”, respectively. NAA stands for N-acetylaspartate.

4.3.5 Prediction of neuroinflammation in an independent set of animals

To confirm the validity of the discovered disease specific markers for neuroinflammation, we used an independent set of samples coming from another EAE experiment, in which we want to predict neuroinflammation in an early stage. The absolute concentration of our markers was determined for the group with peripheral inflammation at Day 10 (here called group “P10-2”) and the group with peripheral and early neuroinflammation at Day 10 (here called group “N10-2”). All peripherally inflamed animals (group “P10-2”) are correctly classified. Animals with peripheral inflammation and early neuroinflammation (group “N10-2”) are also correctly classified, except for one animal. The overall correct classification for the independent set of samples is 95.8%, with specificity for group “P10-2” and group “N10-2” equal to 100% and 91%, respectively. The projection of these independent samples on the PLS-DA score plot derived from group “P10” versus “N10” is shown in Figure 2f. It is important to note that the misclassified sample is located close to the borderline between the two groups. This result suggests that proposed markers have a similar behavior in the second study.

4.4 DISCUSSION AND CONCLUSIONS

Supervised analysis of CSF ^{1}H-NMR data revealed significant changes in biochemical composition of the CSF metabolic profile amongst the analyzed groups. PLS-DA and ANOVA-PCA have been used to model the metabolic profiles from rat CSF. The small differences between PLS-DA and ANOVA-PCA results may arise from the principles of the two methods. PLS-DA aims to select a subset of metabolites which gives the highest separations between groups. This may imply that not all discriminatory metabolites are selected. In the case of ANOVA-PCA the amount of selected variables mostly depend on α level Hotelling T^2. In addition the two different scaling approaches applied before PLS-DA and ANOVA-PCA may have an influence on the selection of metabolites.

Interpretation of PLS-DA models and ANOVA-PCA yielded a set of relevant metabolites, which are shown in Table 2. In this section we further discuss their biological interpretation

Table 2. Average absolute concentration (C_μ) and standard deviation (σ) of metabolites important for peripheral inflammation or neuroinflammation selected based on PLS-DA and ANOVA-PCA.

	Groups							
	healthy		P10		N10		N14	
Metabolites	C_μ [µM]	σ [µM]	C_μ [µM]	σ [µM]	C_μ [µM]	σ [µM]	C_μ [µM]	σ [µM]
NAA*	13.5	2.5	14.8	1.7	16.5	2.3	21.0	3.6
arginine	39.0	10.5	37.5	8.4	46.2	11.8	62.3	17.8
lysine	72.5	9.6	68.5	11.6	89.0	15.0	135.6	29.5
choline	5.6	1.4	5.9	1.2	8.0	1.1	10.6	1.8
malonate	120.4	17.2	126.6	17.2	122.2	33.2	40.5	10.9
alanine	51.9	5.1	55.5	7.3	51.8	8.8	34.3	4.6
citrate	200.2	21.8	143.0	21.2	145.9	37.5	105.4	23.5
creatine	70.6	9.3	69.1	11.2	80.1	9.1	91.3	8.7
glutamine	606.4	93.4	531.9	96.4	544.7	71.6	654.3	127.2
lactate	2660	551.9	2757.3	387.2	3723.0	671.8	2869.3	282.0
3-hydroxy*	494.4	17.2	430.7	25.8	439.9	33.5	510.8	15.4

* 3-hydroxy stands for 3hydroxyisovalerate; NAA stands for N-acetylaspartate

Metabolites related to neuroinflammation: Choline, N-acetylaspartate, creatine, lysine, arginine, alanine and malonate

The concentration of choline was elevated in groups “N10” and “N14”. However, the choline concentration stayed invariable between two control groups “C10” and “P10”. In a previous study, where MScl patients and EAE model in marmoset were studied by ^{1}H-NMR spectroscopy, higher choline concentrations have been reported as a marker for demyelination in urine.[32] This metabolite is required for synthesis of neurotransmitter acetylcholine, and phosphatidylcholine. The increase in choline concentration could be due to demyelination and/or cell membrane breakdown.

Another metabolite found as being highly correlated with EAE-affected groups “N10” and “N14” and which enabled to differentiate these groups from the others was N-acetylaspartate (NAA). This compound is known as a marker of neuronal damage. Its concentration was elevated in the “N14” group. In addition this metabolite distinguished neuroinflammation from peripheral inflammation in the PLS-DA model. NAA is a free amino acid present in neuronal cell and it has been culpably involved in many processes of the nervous system: for instance it may be involved in the myelin production, regulation of neuronal protein synthesis or the metabolism of several neurotransmitters such as aspartate or N-acetyl-aspartyl-glutamate.

Two other metabolites, creatine and malonate, were found to vary between the two EAE-effected groups. Creatine level was found to be up regulated in group “N14”. Creatine is considered as one of the principal brain metabolite. Its changes in concentration are seen in many other neuro-degenerative disorders and are caused by gliosis [33] or scarring of neuron (demyelination). Elevated level of creatine in patients with MScl was found in a previous study and has been associated as marker of gliosis.[34, 35 36] Up regulated creatine level could be due to a change in the cellular composition, either increased inflammatory cells or glial cells. Malonate level was reduced in neuroinflamed group “N14”. Another metabolite found to be reduced in neuroinflamed group “N14” is alanine. Reduced level of alanine in patients with MScl in comparison to patients with cerebrovascular disease was found in a previous study, where Sinclair et al used NMR spectroscopy to evaluate the ability of metabolomics analysis to differentiate neurological disease.[37] Alanine is used as a source for pyruvate for energy metabolism or to synthesis of macromolecules within neural and immune cells.[38] Reduction of alanine concentration may be connected to energy metabolism, since it might be used by invading cells.

Lysine and arginine were found to have a high correlation with EAE-affected groups “N10” and “N14”. Lysine and arginine levels were up regulated in group “N14” compared to “N10” and in “N10” compared to “P10”. This indicates their relation to neuroinflammation. Lysine and arginine are metabolites which may differentiate peripheral inflammation from neuroinflammation. Arginine is used to synthesize nitric oxide (NO). Elevated levels of NO oxidation products in the CNS have been shown in

bacterial meningitis in cerebral lupus erythematosus. Recently it was demonstrated that increased levels of NO oxidation plays a part in the generation of MScl symptoms.[39] Inhibition of NO synthesis may suppress[40] or emphasis EAE.[41] However, the precise role of NO in EAE and MScl still remains elusive and unclear. Qureshi and cowworkers have studied a role of neurotransmitters amino acids in CSF of MScl patients. [42] They have reported increased level of lysine in CSF in MScl patiens.

Metabolites related to peripheral inflammation: citrate, glutamine, lactate and 3-hydroxyisovalerate

Citrate is a key metabolite to differentiate the healthy group "C10" from the peripherally inflamed group, "P10" but also the peripherally inflamed group from the EAE affected group "N14". Reduced citrate is in line with a previous study of Sinclair and coworkers.[37] The difference in citrate level is larger between the peripherally inflamed and neuroinflamed group than between the healthy control and the peripherally inflamed group. The citrate level may indicate the degree of inflammation. Citrate is released to a larger extent from astrocytes than from neurons. This metabolite is an intermediate in tricarboxylic acid cycle (TCA). In the study of Smith and coworkers[43] the metabolic activity of proteins from myelin and non-myelin fractions of spinal cords of Lewis rats with EAE was investigated using [1-^{14}C]leucine as a protein precursor. In this study they showed that the decreased activity of the TCA cycle exists. However, the implication of the citrate alternation is unclear although it was already noted in Alzheimer's and in MScl disease.[37, 44]

Glutamine is an amino acid, which plays an important role in brain metabolism. This metabolite is involved in energy metabolism. It was shown that glutamine is a necessary nutrient for cell proliferation, serving as a specific fuel for inflammatory cells and enterocytes and, when present in appropriate concentrations, enhancing cell function. During inflammatory states, glutamine consumption may outstrip endogenous production and a relative glutamine deficiency state may exist.[45]

A higher level of lactate was found in group "N10" in comparison to group "N14". Predominant lactate peaks have already been reported in inflammatory CNS diseases. The amount of CSF lactate depends largely on production from CNS glycolysis.

Increased lactate production by immune cells is observable in the presence of inflammation. Although elevations in CSF lactate may occur because of many different processes, for instance hypoxia of inflamed tissues, reduced blood flow from cerebral edema, and granulocyte and bacterial metabolism.[46] In addition elevated lactate levels have been identified in vitreous.

The reduced level of 3-hydroxyisovalerate was established for group "N10" and group "P10" when compared with the healthy group. This suggests that this metabolite is involved in peripheral inflammation. However this result deserves more attention since a previous publication demonstrated an increase in 3-hydroxyisovalerate level in some MScl patients.[47]

Conclusions

We investigated the effect of neuroinflammation and peripheral inflammation on the metabolic state of CSF in the rat EAE model, a mimic of the neuroinflammatory aspect of the early stage of MScl. In the animal study untargeted and unbiased biomarker discovery approach consisting of high-field 1D ^{1}H NMR combined with multivariate data analysis was employed.

CSF is demonstrated as a valuable biofluid for the investigation of neurological disorders in the CNS. We found that ^{1}H-NMR is a powerful technique capable of providing information for the identification and quantification of a large number of metabolites in CSF.

The use of two statistical techniques (PLS-DA and ANOVA-PCA) contributes significantly to the reliability of the results. The two methods are corroborative, because the overall results obtained by PLS-DA and ANOVA-PCA were found to be coherent.

The CSF metabolic profile for neuroinflammation is distinct from that of healthy and peripheral inflammation and characterized by changes in concentrations of metabolites such as creatine, arginine and lysine. Peripheral inflammation was only seen at Day 10 and absent at Day 14. A further interesting observation was that the correlation network map is much more complex for the healthy group than for the groups affected by peripheral inflammation or neuroinflammation. Disappearance of correlation between metabolites in peripherally inflamed and neuroinflamed animals might be related to change of penetrability of the blood-brain-barrier (BBB). Under standard physiological

circumstances BBB controls the homeostasis of the interstitial cerebral fluid.[48] It is known that during EAE changes in the BBB function occur. It causes disruption in the BBB and affects the saturable transport system of substances involved in disease process.[49] Injection of CFA can itself lead to increased BBB permeability to small molecules and even certain serum proteins.[50] These disturbances might then cause changes in metabolites flux across BBB and relations between them. For instance the correlation might become more non-linear or weaker and therefore be absent in disease stage under selected threshold. In addition in EAE there is a strong increase in infiltration of the BBB by monocytes and activated lymphocytes which is bound to change the metabolite profile of the fluids. Not only a disrupted BBB leads to "leakage", but activated immune cells crossing the BBB and entering the interstitial fluid and the CSF produce metabolites that change the overall profile of the fluids they are in.

Interestingly some markers of neuroinflammation have been connected to demyelination and neuronal damage. In EAE model induced with MBP demyelination is missing or limited to the ventral root exit and dorsal root entry zone of the spinal cord. In the case of Lewis rats primary demyelination is restricted to occasional perivenous myelin sheaths. In general, demyelination is more distinct when addition of other CNS antigens to the MBP results in pronounced demyelination. In addition, some demyelination was observed in guinea pig EAE model incorporated with MBP inoculums utilized for sensitization.[51] However the presence of demyelination in the EAE model induced with MBP is a matter of debate. Therefore we believe that it needs additional investigations. In current study the presence of demyelination could be investigated by for instance repeating the EAE model and then performing histology studies or electron microscopy of spinal cord. Another possibility would be to use Magnetic magnetic resonance imaging to study the pathology of rats' brain and spinal cord. These steps could demonstrate the presence or absence of demyelination.

By using an independent set of animals, i.e. coming from another EAE experiment, we demonstrated that this model and ipso facto the disease specific markers have ability to predict neuroinflammation in its early stage with high accuracy. Thus, these animal-model based markers may be used to diagnose the early stage of neuroinflammation.

Further developments will include the investigation and translation of our results to a clinical context, i.e. how these results can be used to predict MScl.

ACKNOWLEDGEMENTS

This study was performed within the framework of Top Institute Pharma project number D4-102.

REFERENCES

1. Pilz, G.; Wipfler, P.; Ladurner, G.; Kraus, J., Modern multiple sclerosis treatment - what is approved, what is on the horizon. *Drug Discov. Today* **2008** 13, (23-24), 1013-1025.
2. Sobocki, P.; Pugliatti, M.; Lauer, K.; Kobelt, G., Estimation of the cost of MS in Europe: extrapolations from a multinational cost study. *Mult. Scler.* **2007,** 13, (8), 1054-64.
3. Kantarci, O.; Wingerchuk, D., Epidemiology and natural history of multiple sclerosis: new insights. *Curr. Opin. Neurol.* **2006,** 19, (3), 248-254.
4. Rivers, T. M.; Sprunt, D. H.; Berry, G. P., Observations on Attempts to Produce Acute Disseminated Encephalomyelitis in Monkeys. *J. Exp. Med.* **1933,** 58, (1), 39-53.
5. Schwentker, F. F.; Rivers, T. M., The Antibody Response of Rabbits to Injections of Emulsions and Extracts of Homologous Brain. *J. Exp. Med.* **1934,** 60, (5), 559-74.
6. Kabat, E. A.; Wolf, A.; Bezer, A. E., Rapid Production of Acute Disseminated Encephalomyelitis in Rhesus Monkeys by Injection of Brain Tissue With Adjuvants. *Science* **1946,** 104, (2703), 362-3.
7. Morgan, I. M., Allergic Encephalomyelitis in Monkeys in Response to Injection of Normal Monkey Nervous Tissue. *J. Exp. Med.* **1947,** 85, (1), 131-40.
8. Steinman, L.; Zamvil, S. S., How to successfully apply animal studies in experimental allergic encephalomyelitis to research on multiple sclerosis. *Ann. Neurol.* **2006,** 60, (1), 12-21.
9. Baxter, A. G., The origin and application of experimental autoimmune encephalomyelitis. *Nat. Rev. Immunol.* **2007,** 7, (11), 904-912.
10. Bradl, M.; Linington, C., Animal models of demyelination. *Brain Pathol.* **1996,** 6, (3), 303-11.
11. Lublin, F. D., Relapsing experimental allergic encephalomyelitis. An autoimmune model of multiple sclerosis. *Springer Semin. Immunopathol.* **1985,** 8, (3), 197-208.
12. Wolf, A.; Kabat, E. A.; Bezer, A. E., The Pathology of Acute Disseminated Encephalomyelitis Produced Experimentally in the Rhesus Monkey and Its Resemblance to Human Demyelinating Disease. *J. Neuropathol. Exp. Neurol.* **1947,** 6, (4), 333-357.
13. Teitelbaum, D.; Meshorer, A.; Arnon, R.; Sela, M., Suppresion of experimental allergic encephalomyelitis by a synthetic polypeptide. . *Eur. J. Immunol.* **1971,** 1, 242-248.
14. Yednock, T. A.; Cannon, C.; Fritz, L. C.; Sanchezmadrid, F.; Steinman, L.; Karin, N., Prevention of Experimental Autoimmune Encephalomyelitis by Antibodies against Alpha-4-Beta-1 Integrin. *Nature* **1992,** 356, (6364), 63-66.
15. Ridge, S. C.; Sloboda, A. E.; Mcreynolds, R. A.; Levine, S.; Oronsky, A. L.; Kerwar, S. S., Suppression of Experimental Allergic Encephalomyelitis by Mitoxantrone. *Clin. Immunol. Immunopathol.* **1985,** 35, (1), 35-42.
16. Tintore, M., Early MS treatment. *Int. MS J.* **2007,** 14, 5-10.
17. Hendricks, J. J. A.; Alblas, J.; van der Pol, S. M. A.; van Tol, E. A. F.; Dijkstra, C. D.; de Vries, H. E., Flavonoids influence monocytic GTPase activity and are protective in experimental allergic encephalitis. *J. Exp. Med.* **2004,** 200, (12), 1667-1672.
18. Piotto, M.; Saudek, V.; Sklena, V., Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **1992,** 2, (6), 661-665.
19. ACD/1D HNMR Manager, v., Advanced Chemistry Development, Inc, Toronto On, Canada. *www.acdlabs.com* **2003**.
20. Eilers, P. H. C., A perfect smoother. *Anal. Chem.* **2003,** 75, (14), 3631-3636.
21. Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R., Improved parametric time warping for proteomics. *Chemom. Intell. Lab. Syst.* **2010**.
22. de Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Tsiporkova, E.; Rietzschel, E. R.; De Buyzere, M. L.; Gillebert, T. C.; Bekaert, S.; Martins, J. C.; Van Criekinge, W., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem.* **2008,** 80, (10), 3783-3790.
23. Keun, H. C.; Ebbels, T. M. D.; Antt, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson , J. K., Improved analysis of multivariate data by variable stability scaling:aplication to NMR-based metabolic profiling. *ANAL CHIM ACTA* **2003,** 490, (1-2), 265-276.
24. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M., Targeted profiling: Quantitative analysis of H-1 NMR metabolomics data. *Anal. Chem.* **2006,** 78, (13), 4430-4442.

25. Trygg, J.; Holmes, E.; Lundstedt, T., Chemometrics in metabonomics. *J. Proteome Res.* **2007,** 6, (2), 469-479.
26. Giskeodegard, G. F.; Grinde, M. T.; Sitter, B.; Axelson, D. E.; Lundgren, S.; Fjosne, H. E.; Dahl, S.; Gribbestad, I. S.; Bathen, T. F., Multivariate Modeling and Prediction of Breast Cancer Prognostic Factors Using MR Metabolomics. *J. Proteome Res.* **2010,** 9, (2), 972-979.
27. de Haan, J. R.; Wehrens, R.; Bauerschmidt, S.; Piek, E.; van Schaik, R. C.; Buydens, L. M. C., Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* **2007,** 23, (2), 184-190.
28. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., Assessment of PLSDA cross validation. *Metabolomics* **2008,** 4, (1), 81-89.
29. Snee, R. D., Validation of regression models: Methods and examples. *Technometrics* **1977,** 19, (4), 415-428.
30. Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H., Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* **2006,** 84, (1-2), 69-74.
31. Gidskehaug, L.; Anderssen, E.; Alsberg, B. K., Cross model validation and optimisation of bilinear regression models. *Chemom. Intell. Lab. Syst.* **2008,** 93, (1), 1-10.
32. t'Hart, B. A.; Vogels, J. T.; Spijksma, G.; Brok, H. P.; Polman, C.; van der Greef, J., 1H-NMR spectroscopy combined with pattern recognition analysis reveals characteristic chemical pattern in urines of MS patients and non-human primates with MS-like disease. *J. Neurol. Sci.* **2003** 212, (1-2), 21-30.
33. Hattingen, E.; Magerkurth, J.; Pilatus, U.; Hubers, A.; Wahl, M.; Ziemann, U., Combined (1)H and (31)P spectroscopy provides new insights into the pathobiochemistry of brain damage in multiple sclerosis. *NMR Biomed* **2010**.
34. Qian, J.; Herrera, J. J.; Narayana, P. A., Neuronal and axonal degeneration in experimental spinal cord injury: in vivo proton magnetic resonance spectroscopy and histology. *J. Neurotrauma* **2010,** 27, (3), 599-610.
35. Sajja, B. R.; Wolinsky, J. S.; Narayana, P. A., Proton magnetic resonance spectroscopy in multiple sclerosis. *Neuroimaging Clin. N. Am.* **2009,** 19, (1), 45-58.
36. Sarchielli, P.; Presciutti, O.; Tarducci, R.; Gobbi, G.; Alberti, A.; Pelliccioli, G. P.; Chiarini, P.; Gallai, V., Localized (1)H magnetic resonance spectroscopy in mainly cortical gray matter of patients with multiple sclerosis. *J Neurol* **2002,** 249, (7), 902-10.
37. Sinclair, A. J.; Viant, M. R.; Ball, A. K.; Burdon, M. A.; Walker, E. A.; Stewart, P. M.; Rauz, S.; Young, S. P., NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases--a diagnostic tool? *NMR Biomed.* **2010,** 23, (2), 123-32.
38. Noga, M. J.; Dane, A.; Shi, S.; Attali, A.; van Aken, H.; Suidgeest, E.; Tuinstra, T.; Muilwijk, B.; Coulier, L.; Luider, T. M.; Reijmers, T. H.; Vreeken, R. J.; Hankemeier, T., Metabolomics of cerebrospinal fluid reveals changes in central nervous system metabolism in a rat model of multiple sclerosis. *Metabolomics* **2011**.
39. Danilov, A. I.; Andersson, M.; Bavand, N.; Wiklund, N. P.; Olsson, T.; Brundin, L., Nitric oxide metabolite determinations reveal continuous inflammation in multiple sclerosis. *J. Neuroimmunol.* **2003,** 136, (1-2), 112-118.
40. Ding, M. Z.; Zhang, M.; Wong, J. L.; Rogers, N. E.; Ignarro, L. J.; Voskuhl, R. R., Cutting edge: Antisense knockdown of inducible nitric oxide synthase inhibits induction of experimental autoimmune encephalomyelitis in SJL/J mice. *J. Immunol.* **1998,** 160, (6), 2560-2564.
41. Zielasek, J.; Jung, S.; Gold, R.; Liew, F. Y.; Toyka, K. V.; Hartung, H. P., Administration of Nitric-Oxide Synthase Inhibitors in Experimental Autoimmune Neuritis and Experimental Autoimmune Encephalomyelitis. *J. Neuroimmunol.* **1995,** 58, (1), 81-88.
42. Qureshi, G. A.; Baig, S. M., Role of Neurotransmitter Amino-Acids in Multiple-Sclerosis in Exacerbation, Remission and Chronic Progressive Course. *Biog. Amines* **1993,** 10, (1), 39-48.
43. Smith, M. E.; Rauch, H. C., Metabolic-Activity of Cns Proteins in Rats and Monkeys with Experimental Allergic Encephalomyelitis (Eae). *J. Neurochem.* **1974,** 23, (4), 775-783.
44. Ghauri, F. Y. K.; Nicholson, J. K.; Sweatman, B. C.; Wood, J.; Beddell, C. R.; Lindon, J. C.; Cairns, N. J., Nmr-Spectroscopy of Human Postmortem Cerebrospinal-Fluid - Distinction of Alzheimers-Disease from Control Using Pattern-Recognition and Statistics. *NMR Biomed.* **1993,** 6, (2), 163-167.
45. Wilmore, D. W.; Shabert, J. K., Role of glutamine in immunologic responses. *Nutrition* **1998,** 14, (7-8), 618-626.

46. Watson, M. A.; Scott, M. G., Clinical utility of biochemical analysis of cerebrospinal fluid. *Clin. Chem.* **1995,** 41, (3), 343-60.
47. Lutz, N. W.; A., V.; Malikova, I.; Confort-Gouny, S.; Ranjeva, J. P.; Cozzone, P. J., A branched-chain organic acid linked to multiple sclerosis:First identification by NMR spectroscopy of CSF. *Biochem. Biophys. Res. Commun.* **2007,** 354, (1), 16-164.
48. Kunz, J.; Krause, D.; Gehrmann, J.; Dermietzel, R., Changes in the Expression Pattern of Blood-Brain Barrier-Associated Pericytic Aminopeptidase-N (Pap-N) in the Course of Acute Experimental Autoimmune Encephalomyelitis. *J. Neuroimmunol.* **1995,** 59, (1-2), 41-55.
49. Pan, W.; Banks, W. A.; Kennedy, M. K.; Gutierrez, E. G.; Kastin, A. J., Differential permeability of the BBB in acute EAE: enhanced transport of TNT-alpha. *Am. J. Physiol.* **1996,** 271, (4 Pt 1), E636-42.
50. Pan, W.; Banks, W. A.; Kennedy, M. K.; Gutierrez, E. G.; Kastin, A. J., Peripheral injections of Freund's adjuvant in mice provoke leakage of serum proteins through the blood-brain barrier without inducing reactive gliosis
Brain Res. Bull. **1999,** 832, (1-2), 84-96.
51. Linington, C.; Bradl, M.; Lassmann, H.; Brunner, C.; Vass, K., Augmentation of Demyelination in Rat Acute Allergic Encephalomyelitis by Circulating Mouse Monoclonal-Antibodies Directed against a Myelin Oligodendrocyte Glycoprotein. *Am. J. Pathol.* **1988,** 130, (3), 443-454.

CHAPTER 5

SIMULTANEOUS ANALYSIS OF PLASMA AND CSF BY NMR AND HIERARCHICAL MODEL FUSION

A. Smolinska, J. Posma, L. Blanchet, K. A.M. Ampt, A. Attali, T. Tuinstra, T. Luider, M. Doskocz, P. J. Michiels, F. C. Girard, L. M.C. Buydens, and S. S. Wijmenga

Analytical and Bioanalytical Chemistry (2012), 403(4), pp. 947-59

ABSTRACT

Cerebrospinal fluid (CSF) is the biofluid in closest interaction with the central nervous system. Therefore it holds promise as a reporter on neurological disease such as Multiple Sclerosis (MScl). To characterize the metabolomics signature of the neuroinflammation aspects of this disease we studied an animal model of MScl: the Experimental Autoimmune/Allergic Encephalomyelitis (EAE). CSF also exchanges metabolites with blood via the blood-brain-barrier. Therefore, malfunctions occurring in the CNS may be reflected in the biochemical composition of blood plasma. The combination of blood plasma and CSF provides more complete information about the disease. Both biofluids can be studied using NMR spectroscopy. It is then necessary to perform a combined analysis of the two different datasets. In consequence mid-level data fusion was applied to blood plasma and CSF datasets. Firstly, the relevant information is extracted per biofluid dataset using linear support vector machine recursive feature elimination. The selected variables per dataset are concatenated in order to be analyzed jointly by Partial Least Squares Discriminant Analysis (PLS-DA). The combined metabolomics information from plasma and CSF allows for a more efficient and reliable discrimination of the onset of the EAE. Secondly, we introduce Hierarchical Models Fusion, in which previously developed PLS-DA models are hierarchically combined. We show that this approach allows one to distinguish neuroinflamed rats (even on the day of onset) from either healthy or peripherally inflamed rats. Moreover, the progression of EAE can be investigated thanks to the model separating the onset and the peak of the disease.

5.1 INTRODUCTION

Multiple Sclerosis (MScl) is an inflammatory, presumably autoimmune disease of Central Nervous System (CNS) in which the fatty myelin sheaths which surround the axons of the brain and spinal cord are damaged, leading to demyelization[1]. MScl is one of the most common neurological disease affecting young adults and it has enormous effect on the health system and economy of different countries. The cause of MScl is still elusive. However it is believed that it is a combination of genetic and environmental factors with a possible infectious origin. Signs of MScl can be observed not only in CNS but also in the peripheral nervous system (PNS) [2].

Diagnosis of MScl still remains challenging, especially in its early stage. Currently the diagnosis of MScl is mostly based on clinical evidence complemented with laboratory investigations, like the presence of lesions in the brain and/or spinal cord (visualized by magnetic resonance imaging (MRI)). However, lesions have been found in other neurological diseases, like Guillain-Barré syndrome [3], as well as in non-neurological diseases such as systematic vasculitis [4] or sarcoidosis [5]. Furthermore, brain lesions have been found in healthy individuals [6]. Therefore, brain lesions are not specific enough for proper, early diagnosis. In order to improve diagnosis of MScl it is necessary to combine the information from Cerebrospinal fluid analysis, MRI results and all clinical symptoms.

In order to fingerprint MScl at the molecular level the biological samples have to be analyzed. Since CSF is the biofluid in direct contact with the brain and spinal cord, it is the most suitable choice for fingerprinting MScl. The investigation of biochemical composition of CSF may indicate the abnormal status of the brain. The CSF is absorbed into the blood via a semi-permeable membrane, the blood-brain barrier (BBB). Therefore, effects of CNS diseases can potentially also be seen in the biochemical composition of blood plasma. Obviously, cross-over effects from the plasma to the CSF may also cause changes in biochemical composition of the CSF. In MScl, the BBB is often damaged causing “leakage” [7]. This suggests that plasma may contain predictive information about the disease. Therefore we propose to study the metabolic profiles of both CSF and plasma. These types of samples are relatively difficult to obtain in humans and interesting information is very often obscured by other factors, like genetic,

environmental and dietetic backgrounds. Thus we opt for the possibility of using samples from designed and controlled experiments in rodents.

The animal model of MScl, the Experimental Autoimmune/Allergic Encephalomyelitis (EAE) model has become an important tool to study the neuroinflammatory aspect of MScl [8, 9]. EAE is a cell-mediated experimental autoimmune disorder of the CNS and shares its clinical expression and pathological picture with that of MScl. EAE is used as a pre-clinical model of a single episode of MScl. Similar to MScl, in EAE a strong increase in infiltration of the BBB occurs, which leads to increased exchange between CSF and plasma.

In this study we have extracted CSF and plasma samples at two time points of the disease progression, namely at the onset and the peak; these samples were obtained for healthy, immune booster (a group of animals injected with Complete Freund Adjuvant emulsion, CFA) and EAE (resembling MScl) Lewis rats. The metabolic profile of the CSF and plasma was measured using the untargeted and unbiased technique of high-field 1D proton Nuclear Magnetic Resonance (^{1}H-NMR). This method allows one to analyze both biofluids with very similar measurements protocol. The ^{1}H NMR data of each biofluid can be analyzed separately or the two complementary NMR data sets can be combined (fused) in the analysis. In this work CSF and blood plasma NMR spectra were used in a mid-level data fusion. The metabolite information extracted for each biofluid can be directly translated into relative concentrations for each biofluid and compared.

To obtain such information from the individual or combined data sets several analysis challenges have to be solved. First, the disease must be distinguished from the healthy condition but also from other diseases such as peripheral inflammation. This means that we need to construct a multi-class classifier. Second, even if well controlled the experiment still carries additional variances unrelated to the study, i.e. biological and experimental variations. Thirdly, the number of variables recorded by NMR is much larger than the number of samples which implies specific statistical problems. Moreover most of these variables are probably unrelated to the studied problem or redundant. To solve these problems we propose the following architecture for the data analysis. Linear support vector machine recursive feature elimination (SVM-RFE) is used as variable selection technique for both data sets. The selected variables are fused and analyzed

using either one multi-class Partial Least Squares Discriminant Analysis (PLS2-DA) model or multiple 2 classes PLS-DA models. The latter using a novel approach where a hierarchical structure permits to introduce prior knowledge. We introduce this method as Hierarchical Models Fusion (HMF). We show that by using HMF, EAE affected rats can be distinguished from either healthy or peripherally inflamed rats at the day of onset (when no physical symptoms of neuroinflammation are present) with a 100% correct classification rate. In addition the progression of EAE can be described. In summary, HMF allows one to characterize all studied groups simultaneously without applying multiclass classifier.

5.2 MATERIALS AND METHODS

5.2.1 Experimental design of EAE models

The Experimental Autoimmune/Allergic Encephalomyelitis (EAE) is the common animal model for studying the neuroinflammation aspect of the autoimmune disease of Multiple Sclerosis (MScl). The experimental setup was as previously described by Smolinska et al. and/or Hendricks [10, 11]. Here, we briefly summarize the main points. Three sets of male Lewis rats (Harlan Laboratories B.V., the Netherlands) were inoculated on Day 0. First, a set of 30 animals was injected with guinea pig myelin basic protein (MBP), Complete Freund Adjuvant H37 RA (CFA, Difco Laboratories, Detroit, MI) and Mycobacterium tuberculosis type H37RA (Difco). Another group of 30 animals was injected with CFA only. Next to these MBP and CFA challenged rats, referred to as the EAE group and peripherally inflamed group, respectively, a healthy group undergoing anesthesia only (healthy control) was included. In each group, half of the animals were sacrificed to collect both CSF and plasma on Day 10 (onset of disease in EAE group) and the other half on Day 14 (peak of disease in EAE group). The typical progression of the disease is shown in Figure 1S in the supplementary material, while the details of the design of the EAE experiment are summarized in Table 1.

Table 1. Experimental design of EAE model; "n" indicates a number of rats; p – indicates a number of common samples between CSF and plasma.

Group	Inflammation type	Day 10	Day 14
Healthy	None	C10 n = 15 p= 14	C14 n = 15 p=14
CFA	Peripheral	P10 n = 15 p= 14	P14 n = 15 p=15
CFA+MBP	Peripheral& neuroinflammation	N10 n = 15* p=14	N14 n = 15** p=11

* 1 sample from CSF was discarded due to blood contamination and one from blood plasma due to sampling

** 3 samples from CSF was discarded due to blood contamination and two from blood plasma due to sampling and preparation

5.2.2 CSF and plasma sampling, sample preparation and data acquisition

On Day 10 and 14, animals were euthanized with CO_2/O_2 and blood and CSF were collected. Sampling, sample preparation and data acquisition of CSF NMR spectra was ~~as~~ previously described [11]. The blood was sampled intravenously with a Heparin-treated syringe. Next, every blood sample was centrifuged for 10 minutes at 4°C with a relative centrifugal force of 2000g in order to extract plasma. After centrifugation the supernatants were stored at -80°C for further analysis.

For the NMR measurements, an aliquot of 50 µL of the stored frozen plasma was left at room temperature for thawing. Next, the plasma sample was diluted into 200 µL of water and then proteins were removed by centrifugation for 30 minutes at 2000g (filter 10 kDa Centrisart I 13239-E, Sigma Aldrich, St. Louis, MO, USA) [12]. After protein removal, the supernatant was lyophilized. Prior to NMR measurements the lyophilized plasma samples were re-dissolved in 50µL of water, after which 550 µL of buffer solution was added to obtain sufficient volume for NMR measurement. The buffer solution consisted of 2.85 mM TSP-d_4 (Sodium 3-(trimethylsilyl)propionate-2,2,3,3-d_4) (99 atom %D), 6.92 mM Sodium-azide (NaN_3), 42.08 mM Sodium Phosphate dibasic dehydrate ($Na_2HPO_4 \cdot 2H_2O$) and 7.30 mM HCl solvated in a H_2O:D_2O (99.96 atom %D) mixture (7.93:1). The final TSP concentration in each plasma sample was equal to 2.61 mM.

The ^{1}H-NMR spectra of the 86 plasma samples were acquired on an AVANCE III (Bruker BioSpin, Bruker Inc., Billerica (MA), USA) 500 MHz system equipped 5mm cryoprobes, CPTCI (1H-13C/15N/2H + Z-gradients) (Bruker BioSpin, Bruker Inc., Billerica (MA), USA). Water suppression was achieved using pre-saturation. For each 1D ^{1}H NMR spectrum 256 scans were accumulated with a spectral width of 10273 Hz resulting in a total of 18K points. The acquisition time for each scan was 3.2s. Between scans, a 4s relaxation delay was employed. Prior to spectral analysis, all acquired Free Induction Decays (FIDs) were zero-filled to 32K data points, multiplied with a 0.3 Hz line broadening function, Fourier transformed, manually phased and the TSP internal reference peak was set to 0 ppm - by using ACD/SpecManager software version 12.0 [13].

All 86 rat CSF spectra were acquired and preprocessed as described in [11]. However, due to high line broadening of the internal standard (TSP) four spectra from CSF and plasma were not included in the spectral analysis. Ultimately, 82 CSF spectra and 86

plasma spectra were transferred to Matlab (version 7.9, Mathworks Inc., Natick (MA), USA) for further analysis, of which 82 samples showed overlap between both CSF and plasma spectra (see Table 1).

5.2.3 Preprocessing of CSF and plasma NMR spectra

The ^{1}H NMR spectral data was preprocessed in Matlab, which typically involved baseline correction, alignment, binning, normalization and scaling. For 4 samples (out of the 86) only plasma ^{1}H NMR spectra were available. Therefore these four spectra were not used in the pre-processing and analysis process. Baseline correction of NMR spectra was performed by applying Asymmetric Least Square method [14]. Fluctuation in chemical shift variations were removed by applying improved parametric time warping (I-PTW)[15]. Each CSF and plasma spectrum was normalized to a total area-under-curve (AUC) of 1, to correct for potential differences in sample concentration. In order to reduce the high dimensionality of the data, binning was performed by means of adaptive intelligent binning [16]. This procedure led to 409 bins for CSF and 478 for plasma, which can be considered as relative metabolites concentrations. The absolute quantification of metabolites in CSF and plasma samples was not performed and used. Data analysis was performed on binned data, i.e. on relative metabolites concentrations. The final step of preprocessing consisted of autoscaling.

5.2.4 Data analysis

Explorative analysis by means of robust-Principal Component Analysis (R-PCA) was first used to control the presence of outliers in both datasets [17]. The strategy for supervised data analysis consisted of data division into a training set (75% of samples per class) and an independent test set (25% of samples per class) by using the Duplex algorithm [18], variable selection by Support Vector Machine Recursive Feature Elimination (SVM-RFE) for linear kernels [19] performed on each dataset (CSF and plasma) separately and discriminant analysis by PLS-DA [20] employed on both individual and fused datasets. For data fusion so called mid-level data fusion architecture was employed [21]. In this approach the two data sources are first pre-processed and analyzed separately to extract relevant information and next they are fused and analyzed as unique dataset.

We used this method, since it was shown that it eliminates variables redundancy. Particular steps of this type of data fusion are described in sections 4.1 and 4.2. In this fusion approach every data source is treated separately for pre-processing, scaling and variable selection. Next, the most optimal set of variables is combined into a single set and analyzed with PLS-DA. In the last step of data analysis the approach for cumulative fusion by means of Hierarchical Models Fusion (HMF) was carried out. This method, proposed for the first time in this paper, is described in detail in the section 4.2. The results of this method are compared to PLS2-DA, a variation of PLS-DA which allows more than 2 groups to be analyzed simultaneously (see 4.2).

5.2.4.1 SVM-RFE

SVM-RFE was originally proposed by Guyon et al. [19] and applied to a microarray dataset in a cancer study. The method is based on the binary classification method SVM. This technique first maps objects into a feature space using kernel transformation and then tries to find a hyperplane which separates the data into two classes [22]. From all separating hyperplanes, SVM looks for the one that gives the biggest separation between the borderline training samples of the two classes. The borderline training samples are called support vectors. All support vectors have an alpha value, indicating how supporting this object is for the position of the hyperplane. Non-supporting objects have alpha value equal 0, while alpha equal to 1 indicates the highest support. RFE is a backward elimination algorithm, which ranks features based on the weights of linear SVM. The algorithm starts with a full training set to train a linear SVM. Next, the variables are ranked by sorting in descending order the square of the SVM's weights $\{w_j^2\}$

$$w_j^2=(\sum_{i=\varphi}\alpha_i y_i x_{ij})^2 \qquad (1)$$

where $\boldsymbol{\varphi}$ contains the indexes of support vectors, α_i are alpha values and y_i are the class labels. A variable with smallest weight w_j is then removed. Indeed the smaller the weight of a variable is, the less it contributes to the size of the margin between classes. The remaining variables are used to train another linear SVM and all the process is repeated

until all variables have been eliminated. In our paper, one variable is removed in each iteration.

We have used a Leave-One-Out (LOO) Cross-Validation (CV) approach to select the optimal set of variables per data sets. In this procedure one sample from the training set is left out and a variables ranking is obtained based on the remaining objects. The procedure is repeated until every object is left once. The final ranking was obtained by sorting the variables based on the amount of times it was selected in the LOOCV procedure. The variables selected median+1 times made up the optimal set. The complete scheme for LOOCV can be found in the supplementary material.

5.2.4.2 Classification of individual and fused plasma and CSF datasets

After selecting an optimal set of variables, the features of both data sets are concatenated and autoscaled. Subsequently, the variables of the fused sets were ranked by SVM-RFE. Classification of fused sets was made by PLS-DA, a well-known method used in many omics fields [20, 23]. PLS-DA uses the group information to maximize the separation between groups of observations. It is currently widely used in metabolomics because of its ability to cope with high correlations between variables. In PLS-DA a linear model is constructed according to equation 2:

$$\mathbf{y = Xb + r} \qquad (2)$$

Where, **X** is a dataset matrix, y a vector of group memberships, **b** a vector of regression coefficient (i.e. weights of individual variable) and **r** a vector of model residuals. The regression coefficients reflect the relative importance of the variables in the PLS-DA model. PLS2-DA is a variation of PLS-DA, where the response “y” is not a vector but a matrix, which allows more than 2 groups to be analyzed simultaneously.

The optimal complexity (i.e. number of latent variable (LV)) for all individual as well as fused models was determined by LOOCV performed on training sets. The optimal number of LV was selected based on the minimal error of the root mean square error of cross-validation (RMSECV). For all individual as well as fused models the optimal model complexity was determined to be one LV. All PLS-DA models were validated with

an independent test set. A PLS-DA model is considered statistically valid if it shows good prediction ability. After validation, a final model is then reconstructed using all available samples. The model can be visualized in a score plot. The importance of all variables on the predictive model can be investigated by means of the regression coefficients [24].

After the individual two-class (binary) models are optimized they can be used for HMF.

5.2.4.3 Hierarchical Models Fusion

In this paper, we propose a new approach Hierarchical Models Fusion, HMF, which uses hierarchically multiple simple 2-class classification models to represent individually certain parts of the inter-class variation. This approach uses, as any supervised method, a priori knowledge of the classes (for instance, type of inflammation) and establishes commonalities between them.

The use of simple two-class models makes the results easier to interpret. The method proposed here aims to describe the relevant differences gradually instead of explaining all variation from all classes at once. This gradual process becomes possible by applying statistically optimized binary models to the data at each step and then to combine the outcomes. Since it fuses the outcome of all earlier optimized models it describes and shows all the relevant differences in the data. By using this approach it is possible to visualize separation between studied classes and relation between objects without applying multiclass classification models (like PLS2-DA or Linear Discriminant Analysis (LDA)).

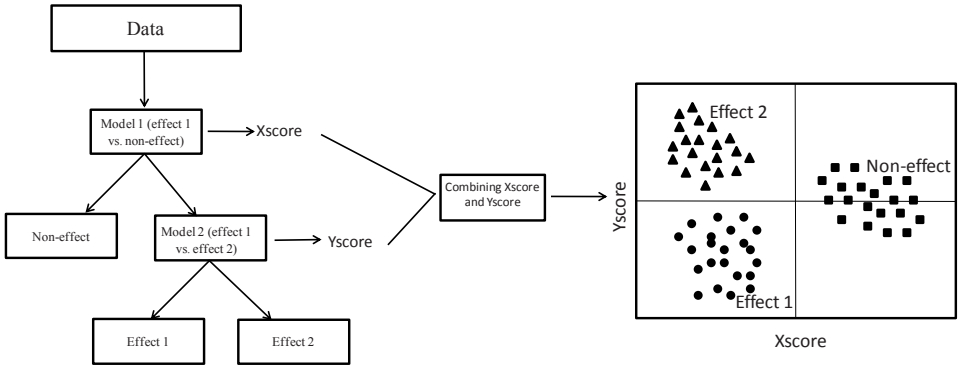


Figure 1. A graphical representation of Hierarchical Models Fusion.

To demonstrate the HMF approach let's consider a dataset with three classes: non-effect (i.e. healthy), effect 1 (e.g. peripheral inflammation) and effect 2 (e.g. neuroinflammation). First individual binary PLS-DA models of interest have to be optimized (i.e. model 1 and 2 from Figure 1). A graphical representation of HMF is shown in Figure 1. These models can then be hierarchically applied to the data in accordance to a priori knowledge (here experimental design). For example having data containing three classes, i.e. effect 1, effect 2 and non-effect, it is possible to apply HMF to separate all three classes. In the first step, model 1 (effect 1 versus non-effect) is used to obtain a new score for all samples in data matrix **X**. This new score (here called Xscore) separates non-effect objects from objects belonging to group effect 1 and group effect 2. In the next step, another model (model 2: effect 1 versus effect 2) can be utilized on matrix X to assess and distinguish these two effects. In that way a second score is obtained for all samples in data matrix **X** (here called Yscore). At each step, a new score is obtained by multiplying data matrix **X** with PLS-DA weights**.** These two new scores (Xscore and Yscore) can then be combined and used to visualize the relationship between studied groups. When the new scores are orthogonal they can be represented as usual (i.e. with perpendicular axis).

Because HMF is based on a hierarchical structure, the complexity of the studied problem is reduced by describing every difference on a different level (i.e. step). It decomposes the difficulty of multi-class separation into simpler, solvable two-class problems. Indeed, the representation of HMF as a decision tree (see Figure 1) is similar to the Classification and Regression Trees (CART) [25]. However, in HMF at each step (node) not a single variable but a PLS-DA model is used to separate objects. In order to represent the usefulness of HMF for analyzing multiple classes, simulated data were created. The results are shown in the supplementary material.

Obviously, the presented method can be used not just for visualization of relations between samples but also for prediction of new samples (e.g. coming from an extra experiment). Moreover, information about variables significant for discrimination is associated with PLS-DA weights. Therefore biological interpretation is feasible as well.

Any results obtained by predictive methods must be validated before drawing any conclusions. In HMF the validation is twofold. First all the individual PLS-DA models are

validated using independent test sets. Second, the complete HMF structure is also validated. Moreover to reduce the possibility of random classification we performed permutation test for HMF.

5.2.5 Metabolite identification

Metabolite identification of the most relevant set of variables was carried out by using the 800 MHz library (for CSF) and 500 MHz library (for plasma) of metabolite NMR spectra from the Chenomx NMR Suite 7.0 (Chenomx Inc., Edmonton (AD), Canada). The libraries of metabolite spectra were obtained based on a database of pure compound spectra acquired using a particular pulse sequence and acquisition parameters, namely, the noesy-presaturation pulse sequence with 4s acquisition time and 1s of recycle delay [26]. The Chenomx NMR Suite software fits the spectral signatures (singlets, doublets, triplets etc), i.e. the peak shapes, of a compound from an internal database of reference spectra to the experimental NMR spectrum.

5.3 RESULTS

5.3.1 Explorative analysis of the CSF - and plasma datasets

The 82 NMR data of plasma and the 82 NMR data of CSF were each pre-processed as described in Materials and Methods. Examples of plasma and CSF spectra are shown in Figure 2. As can be noted the intensities of many metabolites (normalized to TSP signal, only for visualization purpose), like alanine, arginine, in the plasma spectrum are higher in plasma than in CSF. Most metabolites present in CSF can be observed in plasma. A few volatile metabolites are not visible in plasma, because of lyophilisation. Some metabolites are detected only in plasma, for instance glutamate or phenylalanine. This is mostly due to low concentration of these metabolites in CSF. The NMR spectra of CSF were divided into 409 bins, which contain resonances of 33 identified metabolites and some unidentified signals. In the case of plasma, the NMR spectra were divided into 478 bins, which correspond to resonances of 50 identified metabolites and some unidentified signals.

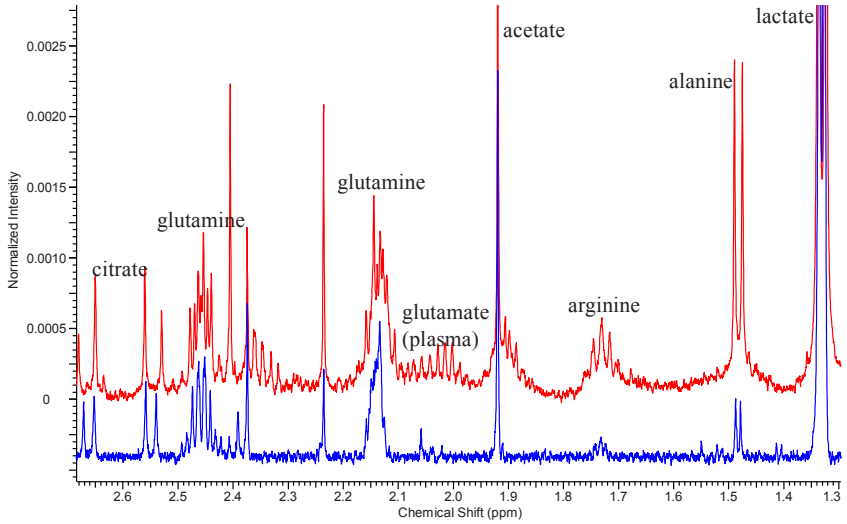


Figure 2. Section of the 800 MHz ^{1}H NMR spectrum of CSF (blue) and ^{1}H 500 MHz NMR spectrum of plasma (red).

After pre-processing, explorative analysis was performed by means of R-PCA and PCA. Initially, R-PCA was applied to the autoscaled spectra of 82 rat CSF and plasma samples, to check for outliers. No outliers were detected. Figures 3a-3b show the PCA score plots of the plasma and CSF NMR spectra, respectively. These figures show that the samples belonging to group "N14" are clearly separated from the others samples along PC2 for plasma data and along PC4 for CSF data. In both situations PC1, which is the main source of variance, does not show any group information. This indicates that a large source of variance in the data does not correspond to the available groups. . No clear grouping is present since most of the groups overlap. It is important to mention that further PC's did not show groupings either.

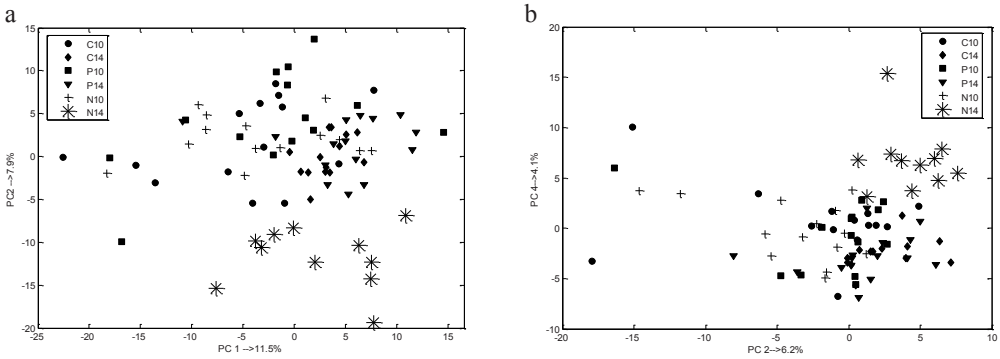


Figure 3. PCA score plot of: (a) plasma NMR spectra; (b) CSF NMR spectra.

5.3.2. Supervised analysis

The most straightforward approach for separating the 6 groups present in CSF and plasma data simultaneously is to apply a multi-class method, for instance PLS2-DA. The two datasets can be analyzed separately by PLS2-DA. Alternatively, the CSF and plasma data can be fused and PLS2-DA can be applied to the fused data. However, PLS2-DA has to describe all group-related variations at the same time. This might lead, on the one hand, to worse results in comparison to multiple binary PLS-DA models and on the other hand, to difficulties in the biological interpretation. One can apply binary PLS-DA models to handle individual biological platforms (CSF and plasma) and the (mid-level) fused data sets. This implies that for a full description many binary PLS-DA

models have to be constructed and optimized. Therefore, we propose and present a new approach, namely HMF. In HMF, a limited number of multiple binary PLS-DA models are employed to still fully describe the fused CSF and plasma data. The fusion was achieved using the approach described in Materials and Methods (section 4.1 and 4.2). Binary PLS-DA models were applied to the fused datasets to extract information on the metabolic effects of the different group treatments shown in Table 1 and to establish the variables significance. All optimal binary PLS-DA models were constructed using only 1 LV. Next these optimized binary PLS-DA models are used in HMF. Below, we first present the results of PLS2-DA, subsequently the binary PLS-DA models, and finally those of HMF. The outcomes of HMF are compared with PLS2-DA of the fused datasets.

PLS2-DA – complete EAE model for plasma data, CSF data and fused sets

We applied PLS2-DA to separate simultaneously all 6 groups of the CSF dataset and of the plasma dataset. The variables included in the PLS2-DA model are selected by linear SVM-RFE. The number of LV's in the PLS2-DA model was optimized by cross validation. The correct classification rate for an independent test set was equal to 57.1% for plasma data and 56% for CSF data (the correct classification per class is included in Table 1S in the supplementary material). It is interesting to note that better results are obtained for the binary PLS-DA models (at least for the binary PLS-DA models considered) than for PLS2-DA.However, it is important to mention that PLS2-DA has a more difficult problem to solve (i.e. separate 6 classes at once) than PLS-DA.

A similar situation is encountered for fused datasets. The correct classification for an independent test set is equal to 65% for PLS2-DA (correct classification per class can be found in the supplementary material, Table 1S). This result is much worse than multiple PLS-DA models. PLS2-DA performance (64 % classification) is in turn still much better than a random classifier (correct classification 17 %) but still insufficient for a proper diagnostic. However, one should notice that some groups are completely classified correctly (100 % e.g. "C10"), while some are totally misclassified.

PLS-DA models for plasma data, CSF data and for mid-level fused sets: the onset of neuroinflammation

We present the results of PLS-DA obtained for the group “P10” versus “N10”, as this represents the interesting case of early onset of neuroinflammation. Binary PLS-DA models were derived for the separate CSF and plasma data sets as well as for mid-level fused data sets. The predictive models for the problem “P10” vs. “N10” are displayed as PLS-DA score plots in Figures 4a, 4b, and 4c. These 1 LV score plots are presented as the density distribution of the entire group. The PLS-DA model of CSF is constructed based on 87 variables. The PLS-DA model of CSF alone has no prediction ability, as follows from the 50% correct classification for an independent test set. Accordingly the groups “P10” and “N10” (for CSF) are not separated in Figure 4a. For plasma, the PLS-DA model separates the classes somewhat better, as follows from the classification for independent test sets of 75%. However, there is still quite some overlap and the groups of points are still mixed, as can be seen on the horizontal axis of Figure 4b.

Since the individual analysis did not bring satisfactory results, we decided to fuse the selected variables from plasma and CSF data. The SVM-RFE carried out on the fused sets, led to 11 variables (out of the 112 firstly selected variables). The resulting PLS-DA model of fused datasets is shown in Figure 4c. The correct classification for independent test set is equal to 100%, demonstrating the statistical adequacy of this model. As can be seen from Figure 4c, there is a clear separation. Figure 4d shows the regression coefficients of this PLS-DA model. Interestingly, the fusion model consists of 6 CSF and 5 plasma variables. This suggests that both biofluids contribute significantly to the group separation.

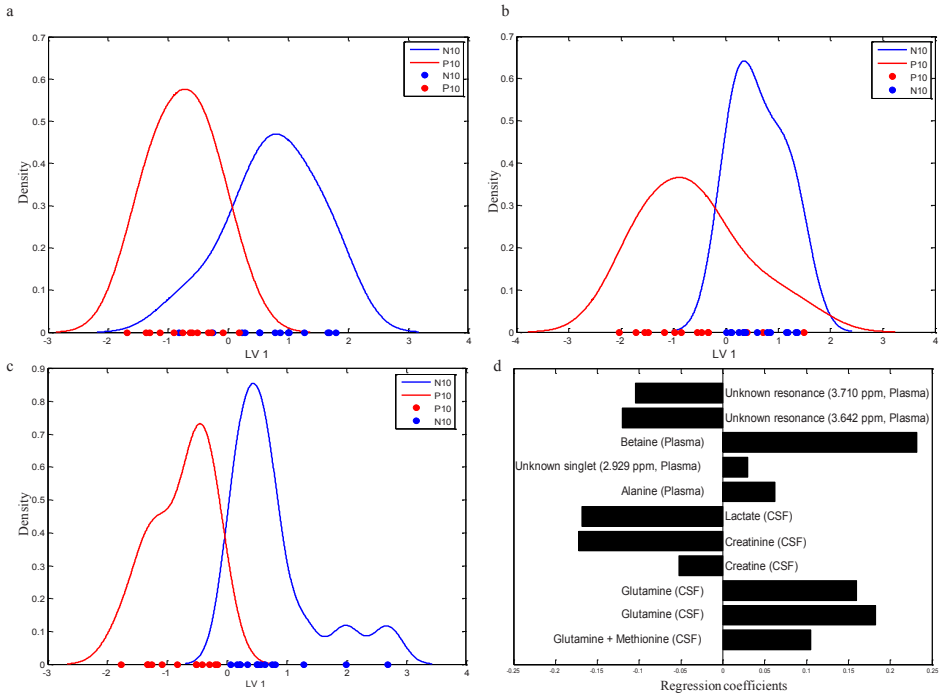


Figure 4. Density distribution of PLS-DA scores of: (a) “P10” vs. “N10” for CSF data, the amount of y variance for 1 LV is equal 77.5%; (b) “P10” vs. “N10” for plasma data, the amount of y variance for 1 LV is equal 63.5%; (c) “P10” vs. “N10” for fused data, the amount of y variance for 1 LV is equal 61.3%; (d) Regression coefficients of fused PLS-DA model.

Table 2 summarizes the results of the PLS-DA models for P10 versus N10, and for two other pairs of groups, namely “C10” vs. “P10”, and “N10” vs. “N14”. These models were used in the HMF. Table 2 displays, apart from the degree of correct classification for independent test sets, also the number of selected variables by RFE-SVM used in PLS-DA models for the fused data. We find that the binary PLS-DA model for “P10 vs. N10” (early onset of neuroinflammation) results in a 100% correct classification. The same holds for “N10” vs. “N14” (progression of neuroinflammation), while for “C10” vs. “P10” a 93 % correct classification is achieved. The score plots and regression coefficients of models “C10” vs. “P10” and “N10” vs. “N14” are shown in Figures 5Sa-5Sd in the supplementary material.

In Table 2 only a few of many possible pairs of groups for PLS-DA have been presented. Nevertheless, in Table 2S in the supplementary material the correct classification rate for the independent test set obtained for individual analysis of plasma data, CSF data and fused datasets by PLS-DA for different pairs of groups can be found. To achieve a full or nearly full description of the fused data set, without having to use all pairs of groups, we apply the hierarchical data fusion model, HFM in the next section.

Table 2. Correct classification rates for an independent test set obtained for fused datasets, number of variables coming from plasma and CSF used in a PLS-DA model and number of samples in training set and test set.

PLS-DA model	Correct classification	# variables in PLS-DA model		# samples	
		Plasma	CSF	training	test
C10 vs. P10	93%	4	13	20	8
P10 vs. N10	100%	5	6	20	8
N10 vs. N14	100%	8	3	18	7

5.3.3 Hierarchical Models Fusion

The predictive power of the individual PLS-DA models is by itself already satisfactory (see Table 2 and in the supplementary material Table 2S). At this point one could stop the analysis and start the biological interpretation of the results. However each PLS-DA model only looks at two groups at a time and therefore is not able to predict a completely unknown sample. Thus, it is necessary to combine the different models. The idea of HMF is to join them in a meaningful order. To perform the HMF on the datasets of CSF and plasma, we used the PLS-DA models of “C10” vs. “P10”, “P10” vs. “N10” and “N10” vs. “N14” (shown in Table 2). They characterize, respectively, the effect of peripheral inflammation, neuroinflammation and progress of neuroinflammation. They are therefore consistent with the experimental design shown in Table 1. The HMF approach used in this paper is represented in Figure 5. Note that we present here HMF on the fusion of two datasets but the same principle could be applied on a single dataset. As explained in Materials and Methods, the HMF is validated in two ways. First, all individual PLS-DA models were statistically validated with independent test sets. Secondly, the complete

scheme of HMF was validated with a set including all test sets used in the binary PLS-DA models from Table 2 and additionally, some samples belonging to classes “C14” (4 samples in test set and 10 in training set) and “P14” (5 samples in test set and 10 in training set). The graphical representation of HMF for training and test set samples is shown in the supplementary material in Figure 5S. As can be seen all test samples are correctly predicted. Additionally the permutation test was performed for all 6 classes as extra check. The p-value for 40000 permutations was equal to 0.0006.

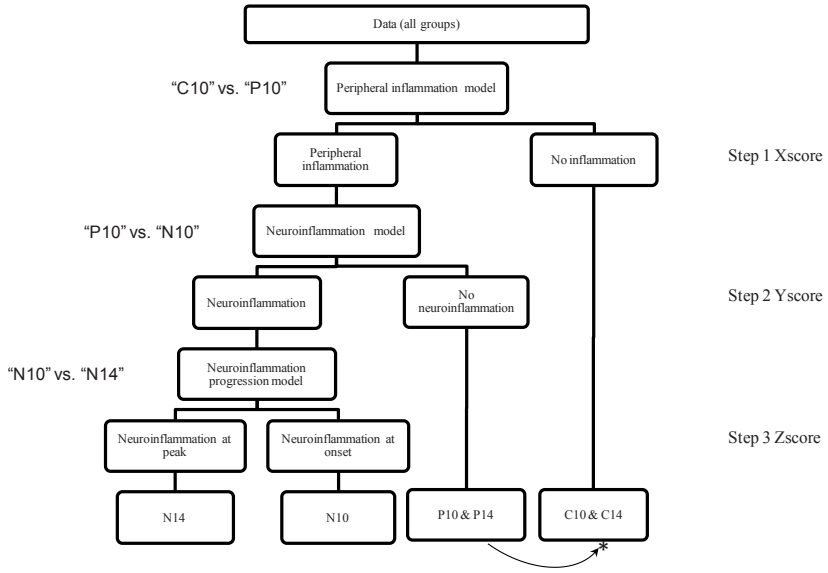


Figure 5. Representation of hierarchical models fusion for fused plasma and CSF NMR datasets. * Note that the “P14” is classified as control (the inflammation is gone, see section below).

We started with the PLS-DA model of peripheral inflammation, i.e. “C10” vs. “P10” (shown in Figure 5 as step 1). This allows one to separate healthy objects, i.e. the one without any type of inflammation (see Table 1) from all those with peripheral inflammation. The latter also includes groups which have undergone neuroinflammation, because in accordance with the experimental design shown in Table 1 these groups were injected with CFA. This step enables creating a first new score (i.e. Xscore) for all

samples in the data. In other words, this model separates the healthy groups from the ones presenting any form of inflammation (neuro- or peripheral).

In a second step, we used a PLS-DA model of “P10” vs. “N10”, shown in Figure 4d. This model distinguishes peripheral inflammation from neuroinflammation at the onset of EAE. Therefore, by using this model we are able to separate neuroinflamed animals from animals that were only peripherally inflamed (i.e. “P10” and “P14” see Figure 5 step 2). Similar to step 1, a second score is generated, the Yscore. At this level, we have separated samples belonging to groups with peripheral inflammation (i.e. “P10” and “P14”) from the neuroinflamed groups (“N10” and “N14”).

The last step (number 3) considers the separation of the onset of the disease from the peak of EAE. In order to achieve this separation, we applied the PLS-DA model of “N10” vs. “N14”, i.e. the model describing the severity of neuroinflammation. At this level, a third new score is created, the Zscore. After iterative application of these simple 1LV models to fused plasma and CSF datasets, we can integrate the three new scores, i.e. Xscore, Yscore and Zscore. They are then used to visualize the outcome. They represent the relation between the groups and their separation. The corresponding graph is shown in Figure 6. As can be observed, full separation of the different groups is achieved. It is worthwhile to mention that samples belonging to healthy groups “C10” and “C14” mostly overlap. However, there is a small shift along the x-axis observable, probably due to sampling time (Day 10 vs. Day 14). As can be noticed samples belonging to group “P14” overlap with healthy groups, which is in agreement with our previous finding, that peripheral inflammation has vanished by day 14[11].

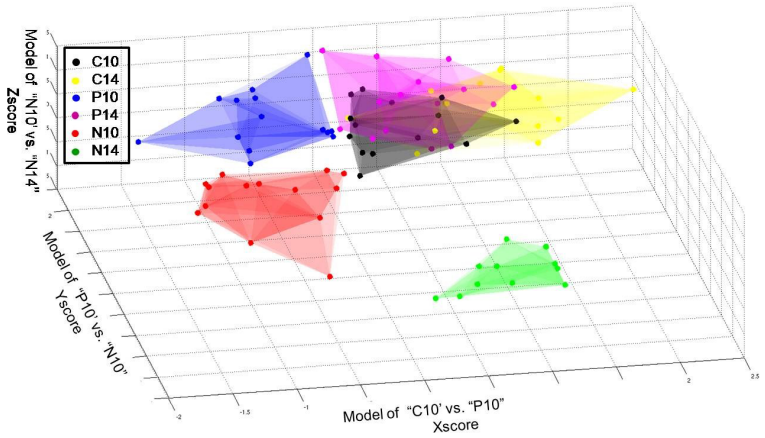


Figure 6. Graphical representation of HMF applied to fused data of plasma and CSF.

5.4 DISCUSSION AND CONCLUSIONS

Using a mid-level fusion architecture we were able to identify a set of metabolites that revealed significant changes in plasma and CSF of neuroinflamed animals. Based on the regression coefficients of the PLS-DA model of fused datasets (see e.g. Figure 4d and supplementary materials) the importance of the individual metabolites in each PLS-DA model as well as a direction of elevation/reduction of concentration can be evaluated. Based on this information the biological interpretation of these metabolites and their connection to EAE, neural inflammation and/or MScl can be performed. Therefore, the first aspect to be discussed is the nature of the selected metabolites. It should be mentioned that the main aim of the paper is not to provide a biological explanation, but to present the methodology for fusion and analysis of ^{1}H NMR metabolomics datasets. Therefore biological conclusions are not stressed.

One can notice that many amino acids (mostly neutral ones) were found to be discriminatory for EAE groups and therefore we focus on them. The transport of neutral amino acids through BBB is significant for the overall regulation of cerebral metabolism and neurotransmitters production [27]. BBB amino acids transport plays an important role in the regulation of several pathways of brain amino acids metabolism. It is known that EAE affects the BBB. It causes disruption in the BBB and affects the saturable transport system of substances involved in the disease process [28]. Injection of CFA can itself lead to increased BBB permeability to small molecules and even certain serum proteins [29].

We found that tyrosine concentration is reduced in plasma of groups “P10”, “N10” and “N14”. It has been reported previously that tyrosine has a role in BBB permeability [30]. In accordance with our results Monaco et al. detected a reduced level of plasma tyrosine in Mscl [31]. Another neutral amino acid related to EAE groups is alanine. This metabolite was found as a relevant metabolite in both plasma and CSF. Its concentration is reduced in CSF and plasma of EAE groups in comparison to healthy controls and peripheral inflamed group “P10”. Alanine is associated with energy metabolism and is known to be used as a source for pyruvate for energy metabolism and macromolecules within neural and immune cells [11]. Similarly, lysine concentration was elevated in CSF and plasma related to neuroinflamed groups “N10” and “N14”. Qureshi and co-workers

in a study on the role of neurotransmitters amino acids in CSF of MScl patients have reported increased levels of lysine in CSF and plasma of MScl patients [32].

We found the combination of glutamate and proline signals in plasma decreased in the EAE groups compared to the other groups. In a previous study a change of glutamate concentration in CSF was reported in a clinical study of MScl [33]. Glutamate is a very important neurotransmitter and the most abundant free amino acid in the brain. A metabolite that is closely interconnected with glutamate is glutamine. This metabolite was found in plasma as discriminatory for groups injected with immune booster (i.e., "P10" and "N10") when compared to the healthy groups and its concentration was elevated in these groups. Additionally, it was found as discriminatory when comparing "P10" and "N10" groups. In CSF, its level was found to be down regulated in group "P10" in comparison to healthy controls. This metabolite is involved in energy metabolism. It was shown that glutamine is a necessary nutrient for cell proliferation, serving as a specific fuel for inflammatory cells and enterocytes and, when present in appropriate concentrations, enhancing cell function [34]. The last amino acid that is discussed here is phenylalanine. This metabolite was diminished in EAE groups. This metabolite is the precursor to Tyrosine, it is necessary in the function of catecholamine neurotransmitters epinephrine, norepinephrine, dopamine and tyramine. In the previous study by Monoco et al. a reduced level of phenylalanine in MScl was found.[31]

One aspect which was not emphasized is the importance of a proper preprocessing. Here the use of AI binning ensures that one bin corresponds to one peak, therefore preventing signals of different metabolites to be mixed within one bin. Normalization is the second important preprocessing aspect. We compared the influence of the classical total area normalization to the probabilistic quantum normalization. No strong differences were observed (data not shown) therefore we decided to use the simplest approach. However one should be aware that the total area normalization could be suboptimal due to the large influence of highly abundant multiplets (e.g. glucose).

The third aspect to be discussed is connected to the data analysis strategy used in this manuscript, i.e. mid-level data fusion and HMF. Firstly, the approach of mid-level data fusion performed here enabled individual variable selection and thus to discard irrelevant

information. Secondly, the HMF method, shown in this paper, represents a novel, simple strategy for multi-class analysis. Each PLS-DA model only looks at two groups at a time and therefore a single model is not able to predict a completely unknown sample. This is a minor advantage of HMF over multiple PLS-DA models. One should keep in mind that the outputs of this method are statistically accurate, since they are based on validated binary predictive models. Moreover, the complete scheme of HMF was validated as well. The output of HMF (i.e. new scores) can be used for visualization or prediction of new samples. However, it is good practice to check if these new scores are orthogonal.

When comparing HMF and PLS2-DA, it is important to mention that it is possible that if some groups do not behave in accordance to experimental design, the optimal solution for class separation can be flipped. In other words, if one or more groups cannot be distinguished, PLS2-DA still tries to separate them, which may affect the solution for the whole PLS2-DA model. In the case of EAE datasets, there are two groups (i.e. “N10” and “P14”) that are characterized with behaviour different than was assumed by experimental design. In case of the “N10” group we have shown previously, that animals are heterogeneous regarding to disease response [11]. Further, the second group “P14” was not (or no longer) peripherally inflamed at day 14. This causes the results obtained by PLS2-DA to be sub-optimal for groups “N10”, “C14” and “P14”. In the method proposed here, HMF, the situation described for PLS2-DA cannot happen. HMF leads to the optimal solution, because it includes relevant sources of variance between groups individually rather than all at the same time. This suggests that if two groups are not separable this can be easily detected during the HMF and does not influence the separation between other groups.

In our study the HMF was shown for fused plasma and CSF NMR datasets. However this approach can be also used for one platform (as shown in the supplementary material). Obviously, the individual PLS-DA models can be developed for any platform and then HMF can be applied.

Conclusions

In this study, we have demonstrated the feasibility of fusion of metabolomics ^{1}H NMR datasets from different biofluids. From the data analysis point of view multiple challenges had to be addressed. One of them had to do with the biological variation usually encountered in omics experiments. Another issue was linked to the number of variables recorded by NMR, which first is greatly superior to the number of samples and secondly most of them are probably unrelated to the studied problem or redundant. We successfully solved these problems using a new architecture for data fusion, where SVM-RFE is used as variable selection method and PLS-DA to focus on the information of interest through a training procedure.

We analyzed CSF and plasma metabolomics data of the EAE model for MScl using mid-level data fusion. The procedure was represented by constructing predictive model for neuroinflamed group "N10", i.e. before physical symptoms have appeared, versus peripherally inflamed group "P10". Prediction models based on either CSF or plasma metabolomics data alone could not separate the immune booster and EAE groups at day 10, whereas the predictive model using a fused set of variables from CSF and plasma managed to separate the two groups with a 100% correct classification rate for the independent test set. One should be aware that these results do not imply that all new samples will be always correctly classified. However validation with the independent test set and the permutation test set indicate that the results are meaningful. This shows that by using bio-molecular information (metabolomic data), a diagnosis can be made before physical symptoms arise. Our results also demonstrate that plasma can play a significant role in diagnosis of neuroinflammation. Therefore, we believe that plasma should be considered when investigating neuroinflammation.

Finally, we have introduced a new multi-class method HMF, which aims to describe relevant sources of variance connected to groups' description by fusing individual binary models. We have shown that by using HMF we are able to separate groups in our data by using simple, easily interpretable, one-component predictive models.

From a biological point of view, the selected metabolites appears to be relevant, because the metabolites described in this study were previously found in relationship to

the EAE and/or Mscl. Therefore, they provide a biological validation for the fusion of data from two different biofluids.

Further research will focus on the deeper interpretation and absolute quantification of newly detected metabolites in plasma and CSF and their relation to BBB. These two steps are time consuming but would bring more insights on disease mechanism. The pattern and concentrations defined by these variables could also be studied by themselves and put into a systems biology context. Absolute quantification would be crucial for obtaining advanced biological conclusions and conformation using completely different analytical method (e.g. Mass Spectrometry)

ACKNOWLEDGEMENTS

This work was performed within the framework of Dutch Top Institute Pharma, project "The CSF proteome / metabolome as primary biomarker compartment for CNS disorders" (project nr. D4-102: AS, JP, LB, KA, LB, AA, TT, TL, LMCB, SSW). This work was further supported by the Dutch ministry of Economic Affairs and the Provinces Gelderland and Overijssel via their financing support of the project HYPHEN-ID (Ref. PID06014) to Spinnovation (MD, PJM, FCG) and the Radboud University (SSW).

REFERENCES

1. Pilz, G.; Wipfler, P.; Ladurner, G.; Kraus, J., Modern multiple sclerosis treatment - what is approved, what is on the horizon. *Drug Discov. Today* **2008** 13, (23-24), 1013-1025.
2. Compston, A.; Coles, A., Multiple sclerosis. *Lancet* **2008,** 372, (9648), 1502-17.
3. Hughes, R. A.; Cornblath, D. R., Guillain-Barre syndrome. *Lancet* **2005,** 366, (9497), 1653-66.
4. Miller, D. H.; Ormerod, I. E.; Gibson, A.; du Boulay, E. P.; Rudge, P.; McDonald, W. I., MR brain scanning in patients with vasculitis: differentiation from multiple sclerosis. *Neuroradiology* **1987,** 29, (3), 226-31.
5. Miller, D. H.; Kendall, B. E.; Barter, S.; Johnson, G.; MacManus, D. G.; Logsdail, S. J.; Ormerod, I. E.; McDonald, W. I., Magnetic resonance imaging in central nervous system sarcoidosis. *Neurology* **1988,** 38, (3), 378-83.
6. Boone, K. B.; Miller, B. L.; Lesser, I. M.; Mehringer, C. M.; Hill-Gutierrez, E.; Goldberg, M. A.; Berman, N. G., Neuropsychological correlates of white-matter lesions in healthy elderly subjects. A threshold effect. *Arch. Neurol.* **1992,** 49, (5), 549-54.
7. Minagar, A.; Alexander, J. S., Blood-brain barrier disruption in multiple sclerosis. *Mult. Scler.* **2003,** 9, (6), 540-9.
8. Kabat, E. A.; Wolf, A.; Bezer, A. E., Rapid Production of Acute Disseminated Encephalomyelitis in Rhesus Monkeys by Injection of Brain Tissue With Adjuvants. *Science* **1946,** 104, (2703), 362-3.
9. Schwentker, F. F.; Rivers, T. M., The Antibody Response of Rabbits to Injections of Emulsions and Extracts of Homologous Brain. *J. Exp. Med.* **1934,** 60, (5), 559-74.
10. Hendricks, J. J. A.; Alblas, J.; van der Pol, S. M. A.; van Tol, E. A. F.; Dijkstra, C. D.; de Vries, H. E., Flavonoids influence monocytic GTPase activity and are protective in experimental allergic encephalitis. *J. Exp. Med.* **2004,** 200, (12), 1667-1672.
11. Smolinska, A.; Attali, A.; Blanchet, L.; Ampt, K.; Tuinstra, T.; van Aken, H.; Suidgeest, E.; van Gool, A. J.; Luider, T.; Wijmenga, S. S.; Buydens, L. M., NMR and pattern recognition can distinguish neuroinflammation and peripheral inflammation. *J Proteome Res* **2011,** 10, (10), 4428-38.
12. Wevers, R. A.; Engelke, U.; Wendel, U.; de Jong, J. G.; Gabreels, F. J.; Heerschap, A., Standardized method for high-resolution 1H-NMR of cerebrospinal fluid. *Clin. Chem.* **1995,** 41, (5), 744-51.
13. ACD/1D HNMR Manager, v., Advanced Chemistry Development, Inc, Toronto On, Canada. *www.acdlabs.com* **2003**.
14. Eilers, P. H. C., A perfect smoother. *Anal. Chem.* **2003,** 75, (14), 3631-3636.
15. Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R., Improved parametric time warping for proteomics. *Chemom. Intell. Lab. Syst.* **2010**.
16. de Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Tsiporkova, E.; Rietzschel, E. R.; De Buyzere, M. L.; Gillebert, T. C.; Bekaert, S.; Martins, J. C.; Van Criekinge, W., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem.* **2008,** 80, (10), 3783-3790.
17. Walczak, B.; Daszykowski, M.; Serneels, S.; Kaczmarek, K.; Van Espen, P.; Croux, C., TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemom. Intell. Lab. Syst.* **2007,** 85, (2), 269-277.
18. Snee, R. D., Validation of regression models: Methods and examples. *Technometrics* **1977,** 19, (4), 415-428.
19. Guyon, I.; Weston, J.; Barnhill, S., Gene selection for cancer classification using Support Vector Machine. *Machine Learning* **2002,** 46, 389-422.
20. Trygg, J.; Holmes, E.; Lundstedt, T., Chemometrics in metabonomics. *J. Proteome Res.* **2007,** 6, (2), 469-479.
21. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van-der Vat, B. J. C.; Jellema, R. H., Fusion of mass spectrometry-based metabolomics data. *Anal. Chem.* **2005,** 77, (20), 6729-6736.
22. Cristianini, N.; Shawe-Taylor, J., *An introduction to support vector machines and other kernel-based learning methods*. The University of Cambridge: 2000.

23. Giskeodegard, G. F.; Grinde, M. T.; Sitter, B.; Axelson, D. E.; Lundgren, S.; Fjosne, H. E.; Dahl, S.; Gribbestad, I. S.; Bathen, T. F., Multivariate Modeling and Prediction of Breast Cancer Prognostic Factors Using MR Metabolomics. *J. Proteome Res.* **2010,** 9, (2), 972-979.
24. Wold, S.; Martens, H.; Wold, H., The Multivariate Calibration-Problem in Chemistry Solved by the Pls Method. *Lecture Notes in Mathematics* **1983,** 973, 286-293.
25. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., *Classification and Regression Trees*. Monterey, California, 1984.
26. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M., Targeted profiling: Quantitative analysis of H-1 NMR metabolomics data. *Anal. Chem.* **2006,** 78, (13), 4430-4442.
27. Pardridge, W. M., Blood-brain barrier carrier-mediated transport and brain metabolism of amino acids. *Neurochem. Res.* **1998,** 23, (5), 635-644.
28. Pan, W.; Banks, W. A.; Kennedy, M. K.; Gutierrez, E. G.; Kastin, A. J., Differential permeability of the BBB in acute EAE: enhanced transport of TNT-alpha. *Am. J. Physiol.* **1996,** 271, (4 Pt 1), E636-42.
29. Pan, W.; Banks, W. A.; Kennedy, M. K.; Gutierrez, E. G.; Kastin, A. J., Peripheral injections of Freund's adjuvant in mice provoke leakage of serum proteins through the blood-brain barrier without inducing reactive gliosis
Brain Res. Bull. **1999,** 832, (1-2), 84-96.
30. Staddon, J. M.; Herrenknecht, K.; Smales, C.; Rubin, L. L., Evidence That Tyrosine Phosphorylation May Increase Tight Junction Permeability. *J. Cell Sci.* **1995,** 108, 609-619.
31. Monaco, F.; Fumero, S.; Mondino, A.; Mutani, R., Plasma and cerebrospinal fluid tryptophan in multiple sclerosis and degenerative diseases. *J. Neurol. Neurosurg. Psychiatry* **1979,** 42, (7), 640-1.
32. Qureshi, G. A.; Baig, S. M., Role of Neurotransmitter Amino-Acids in Multiple-Sclerosis in Exacerbation, Remission and Chronic Progressive Course. *Biog. Amines* **1993,** 10, (1), 39-48.
33. Sarchielli, P.; Greco, L.; Floridi, A.; Gallai, V., Excitatory amino acids and multiple sclerosis: evidence from cerebrospinal fluid. *Arch. Neurol.* **2003,** 60, (8), 1082-8.
34. Noga, M. J.; Dane, A.; Shi, S.; Attali, A.; van Aken, H.; Suidgeest, E.; Tuinstra, T.; Muilwijk, B.; Coulier, L.; Luider, T. M.; Reijmers, T. H.; Vreeken, R. J.; Hankemeier, T., Metabolomics of cerebrospinal fluid reveals changes in central nervous system metabolism in a rat model of multiple sclerosis. *Metabolomics* **2011**.

CHAPTER 6

Interpretation and Visualization of non-linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis

A. Smolinska, L. Blanchet, L. Coulier, K. Ampt, T. Luider, R. Q. Hintzen, S. S. Wijmenga, and L.M.C. Buydens

PLoS One (2012), 7(6), pp. e38163

ABSTRACT

Background: In the last decade data fusion has become widespread in the field of metabolomics. Linear data fusion is performed most commonly. However, many data display non-linear parameter dependences. The linear methods are bound to fail in such situations. We used proton Nuclear Magnetic Resonance and Gas Chromatography-Mass Spectrometry, two well established techniques, to generate metabolic profiles of Cerebrospinal fluid of Multiple Sclerosis (MScl) individuals. These datasets represent non-linearly separable groups. Thus, to extract relevant information and to combine them a special framework for data fusion is required.

Methodology: The main aim is to demonstrate a novel approach for data fusion for classification; the approach is applied to metabolomics datasets coming from patients suffering from MScl at a different stage of the disease. The approach involves data fusion in kernel space and consists of four main steps. The first one is to extract the significant information per data source using Support Vector Machine Recursive Feature Elimination. This method allows one to select a set of relevant variables. In the next step the optimized kernel matrices are merged by linear combination. In step 3 the merged datasets are analyzed with a classification technique, namely Kernel Partial Least Square Discriminant Analysis. In the final step, the variables in kernel space are visualized and their significance established.

Conclusions: Conclusions: We find that fusion in kernel space allows for efficient and reliable discrimination of classes (MScl and early stage). This data fusion approach achieves better class prediction accuracy than analysis of individual datasets and the commonly used mid-level fusion. The prediction accuracy on an independent test set (8 samples) reaches 100 %. Additionally, the classification model obtained on fused kernels is simpler in terms of complexity, i.e. just one latent variable was sufficient. Finally, visualization of variables importance in kernel space was achieved.

6.1 INTRODUCTION

Currently, due to the increasing amount of data generated from different analytical platforms for a single studied system, for instance in fingerprinting a disease in the metabolomics and proteomics fields, optimal data concatenation, or data fusion, has become an issue that needs to be addressed. Each analytical technology demonstrates different strengths and limitations regarding its capability to distinguish between different biological conditions, depending upon factors such as sensitivity, sample preparation, analytical stability, and analytical reproducibility. The jointed use of two or more analytical technologies gives then a more robust strategy for data analysis than the use of a single platform [1].

Data fusion is widely applied in the pattern recognition field [2]. For example, in chemistry, biology, medicine and many others fields linear techniques are used to construct a mathematical model that relates spectral responses from different techniques to analyte concentrations [3-6]. In the omics related fields, data fusion is performed in different ways and on different data levels [7]. To date, data fusion methods are organized in three levels: low-level, mid-level and high-level fusion [8, 9]. In low-level fusion, different data sources are concatenated at the data level. In the mid-level fusion, data from different sources are combined at the data level by selection of variables or at the latent variables level. In high-level data fusion, different model responses (for instance prediction for each available data set) are joined to produce a final response. Currently, several linear techniques, such as Principal Component Analysis (PCA) or Partial Least Squares Discriminant Analysis (PLS-DA), are used for the above mentioned types of data fusion. These different linear data fusion approaches have been applied with good success in recent times in the different omics fields, including metabolomics [8, 10-12]. To our knowledge non-linear methods have not been applied to data fusion in for instance metabolomics. However, some chemical systems and problems are inevitably non-linear and reveal characteristics in a non-linear fashion. The assumption of a linear response is then incorrect and non-linear description is appropriate [13]. Of course, to follow Occam's razor principle, it is common practice to first apply linear methods and only if they fail to move to non-linear techniques like kernel-based methods. Kernel-based methods transform the data to a high dimensional feature space by means of a kernel function.

This generates a new data matrix, which can be viewed as a similarity matrix. The kernel function takes relationships that are implicit in the data and makes them explicit, so that patterns are easier to detect. Moreover, they have been designed to deal with datasets where many variables are present. Kernel-based methods have already been demonstrated to form powerful tools and therefore are widely applied to various statistical problems due to their flexibility and good performance [14, 15]. A major disadvantage of these Kernel-based methods has been that information on the importance of variables is lost. However, recently an approach has been proposed for representing the importance of variables in kernel space, a method based on the principles of so-called pseudo samples [16, 17].

Nowadays, proton Nuclear Magnetic Resonance (^{1}H-NMR) and Gas Chromatography-Mass Spectrometry (GC-MS) are well-established powerful analytical methods for generating metabolomics profiles. For analysis of complex, biological samples like those from Cerebrospinal fluid (CSF) both techniques have their advantages and disadvantages. For instance, ^{1}H-NMR requires limited sample preparation, is quantitative, non-destructive and unbiased. ^{1}H-NMR may detect compounds that are too volatile for GC, while metabolites without proton (phosphoric acid) are not detected by ^{1}H-NMR. GC-MS requires derivatization and thus more time consuming sample preparation. On the other hand, GC-MS yields a higher sensitivity than NMR and therefore may detect metabolites that are present in a concentration below the detection limit of ^{1}H-NMR. Therefore, these analytical platforms give wide and complementary views of the studied system. To obtain the maximum/optimal amount of relevant information about the complex biological system, the data from these powerful analytical techniques need to be combined and analyzed with advanced multivariate statistical tools.

This paper presents a novel framework for integrating data from different analytical sources by applying non-linear kernel-based statistical learning methods. We demonstrate this non-linear kernel fusion approach on ^{1}H-NMR and GC-MS metabolomics datasets obtained from CSF of patients with Multiple Sclerosis (MScl) [18]. These data display non-linear response characteristics. The proposed approach for non-linear Kernel-based data fusion consists of four steps. The first step aims to extract

relevant variables from both datasets separately. Variable selection is performed by means of Support Vector Machine Recursive Feature Elimination (SVM-RFE) for non-linear kernels [19]. The second step is designed to fuse the relevant information of both datasets by using linear combinations of kernel matrices [20]. This kernel fusion falls outside the range of the classical low-, mid- and high-level fusion. The next step (step 3) consists of applying PLS-DA on the fused kernels as classification method. In step 4, the visualization of the relative contribution of each variable to K-PLS-DA model (variable importance) was achieved by applying and extending the recently developed pseudo samples principle [16, 17]. Consequently, in our approach the importance of variables is visualized. The variables can then be interpreted in terms of the underlying biology of system. Application of our non-linear Kernel-based data-fusion methodology to the ^{1}H-NMR and GC-MS metabolomic datasets from samples of CSF of MScl individuals and individuals in the early stage of the disease enabled better classification than using the data from the two sources separately. More importantly, the biological interpretation can now be done based on the joined data from the two platforms. The approach proposed here can be extended to other types of datasets such as to MS or NMR data from proteomics or data from microarrays and Liquid Chromatography. The number of samples used to study the progression of MScl is relatively small. Therefore, some limitations with respect to biological interpretation as well as prediction of future samples may exist, e.g. due to biological variation. In order to use the findings in the clinic they should be validated in a new cohort with a larger number of samples. This issue will be further addressed in the discussion section.

5.2 MATERIALS AND METHODS

5.2.1 CSF sampling and patients

The CSF patients involved in this study were all followed by the Rotterdam Multiple Sclerosis Center and the department of Neurology at Erasmus University Medical Center (Rotterdam, The Netherlands). The Medical Ethical Committee of Erasmus University Medical Centre in Rotterdam, The Netherlands, approved the study protocol and all study patients gave written consent. All CSF samples were specifically collected from patients that were not under any drug treatment.

All CSF samples were taken from patients via lumbar puncture. Immediately after sampling, the CSF samples were centrifuged to remove cells and cellular elements (10 minutes at 3000 rpm). Subsequently, a fraction of the CSF samples were used for diagnosis purpose and the remaining amounts were aliquoted and stored at -80^0 C.

The CSF samples were classified into two groups. The first group consisted of CSF samples collected from patients diagnosed with MScl. The second group of CSF samples was taken from patients who were diagnosed with clinically isolated syndrome of demyelination (CIS), which represents an early stage of MScl. It is worthwhile to mention that all patients diagnosed with CIS have later developed MScl. The overview of the available CSF samples for NMR and GC-MS is presented in Table 1, while clinical information is described in the supplementary material. It is important to mention that the set of samples analysed by NMR and GC-MS only partly overlap (Table 1).

Table 1. The number of samples included in a training and independent test set.

Group	No. samples NMR			No. samples GC-MS			Overlap NMR and GC-MS		
	Training	Test	Total	Training	Test	Total	Training	Test	Total
MScl	19	7	26	18	6	24	7	5	12
CIS	15	5	20	10	4	14	7	3	10

5.2.2 NMR samples preparation and data acquisition

The CSF samples of the CIS and MScl classes were prepared as follows. An aliquot of 20µL of the stored frozen human CSF sample (-80 °C) was thawed at room temperature. Subsequently, 200 µL D_2O was added to biofluid in order to obtain sufficient sample volume for NMR measurements. We used 3-(Trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP-d_4 99 at.%D) as internal standard for chemical shift reference (δ 0.00 ppm) and metabolite quantification. For this and buffering, 70µL of buffer solution was added to the 220 µL of human CSF sample. The buffer solution solvated in a mixture of water and D_2O consists of 2,85mM TSP, 6.92 mM sodium azide (NaN_3) and 42.08 mM sodium phosphate dibasic dehydrate ($Na_2HPO_4 \bullet 2H_2O$). The addition of mixture solution to 220 µL of CSF sample leads to a final concentration of 0.66 mM TSP-d_4 and corresponding concentrations of buffer solution components. The pH of the CSF NMR sample was adjusted to around 7 (7.0 – 7.1) by the buffering capacity of the phosphate in the buffer solution. The final CSF NMR sample (290 µL) was transferred to a SHIGEMI microcell tube for NMR measurements.

All spectra were recorded by using a standard pulse sequence (1D-NOESY; recycle delay-90°-t1-90°-t_m-90°) at a temperature of 25 °C. The water suppression was achieved by presaturation during the relaxation delay (8 s) and mixing time (100 ms). All ^{1}H NMR spectra were acquired at 600 MHz Bruker NMR Spectrometer equipped with cryo-cooled probe. For each 1D ^{1}H NMR spectrum 256 scans were accumulated with a spectral width of 7200 Hz resulting in a total of 16K data points. The acquisition time for each scan was 2.2s. Prior to spectral analysis, all Free Induction Decays (FIDs) were multiplied with a 0.3 Hz line broadening function, Fourier transformed and manually phased. In addition, the TSP internal reference peak was set to 0 ppm. This initial processing was done using ACD/SpecManager software version 12.02 [21].

All 46 human CSF spectra were acquired and pre-treated as described above and subsequently, transferred to Matlab, version 7.6 (R2008b) (Mathworks, Natick, MA) for further analysis.

5.2.4 Preprocessing of NMR spectra

The NMR spectral data of human CSF was pre-processed, which typically involves baseline correction, alignment, binning, normalization and scaling. Asymmetric Least Square method was used for baseline correction of NMR spectra [22]. Next, in order to remove variations in peak position, NMR spectra were aligned by using correlation optimized warping [23]. A further problem is the high dimensionality of the data (*circa* 15000 variables). To reduce the number of variables associated with the NMR spectra, we performed binning via adaptive intelligent binning [24]. Before binning data were normalized to total area. The chemical shift ranges of δ 0.75 – 4.15 and δ 8.65 – 8.85 were used for the binning procedure. The binning procedure led to 233 bins in total. In the final step of preprocessing data were scaled to unit variance.

5.2.5 GC-MS samples preparation and data acquisition

The GC-MS method applied here is a non-targeted GC-MS method which uses a derivatization step that has frequently been applied for metabolomics studies [25]. With this method it is possible to analyse simultaneously various classes of (polar) metabolites, e.g. amino acids, organic acids, fatty acids, sugars.

Human CSF samples (100 µL) were deproteinized by adding 400 µL methanol and subsequently centrifuged for 10 min at 10000 rpm. The supernatant was dried under N_2 followed by derivatization with methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) in pyridine similar to Koek et al.[25]. During the different steps in the sample work-up, i.e. prior to deproteinization, derivatization and injection, different (deuterated) internal standards were added at a level of circa 20 ng/µL. The end volume was 135 µl and 1 µl aliquots of the derivatized samples were injected in splitless mode on a HP5-MS 30 m x 0.25 mm x 0.25 mm capillary column (Agilent Technologies, Palo Alto, CA) using a temperature gradient from 70 °C to 320 °C at a rate of 5 °C/min. GC-MS analysis was performed using an Agilent 6890 gas chromatograph coupled to an Agilent 5973 quadrupole mass spectrometer. Detection was carried out using MS detection in electron impact mode and full scan monitoring mode (m/z 15-800). The electron impact for the generation of ions was 70 eV.

A total of 38 human CSF samples were analysed by GC-MS. The samples were randomly distributed over batches and each sample was injected once. A pooled CSF sample was prepared from the study samples for quality control (QC). Aliquots of this QC sample were analysed in sextuplicate in each batch according to the procedure described by van der Greef et al. [26].

Data-pre-processing was performed by composing target lists of peaks detected in the samples based on retention time and mass spectra. Peaks were characterized by retention time and *m*/*z* ratio and identified by comparison with a spectral database. These peaks were integrated for all samples. The peak areas were subsequently normalized using internal standards and corrected for intra- and inter-batch effects using the QC samples according to the procedure described by Verheij et al. [27]. The final step of preprocessing was unit variance scaling.

5.2.6 Data analysis

5.2.6.1 Explorative analysis

The first step of our data analysis strategy consists of a data exploration by means of Robust – Principal Component Analysis (R-PCA) [28] and PCA. R-PCA was employed on the autoscaled data to detect the outliers in both datasets. To extract and display the systematic variation in the two datasets PCA was also carried out on the autoscaled data.

5.2.6.2 Selection of training set and independent test set

In order to validate the performance of the classifier an independent test set was used. Dividing the data into training and test sets is a widely accepted approach for this purpose [29, 30]. The commonly used leave-one-out cross-validation (LOOCV) is biased to assess the predictive ability of the classification model. External validation using test sets provides a means to establish a more reliable predictive performance of the classification model [31-33].

The training set and an independent test set were selected separately for NMR and GC-MS datasets using the Kennard-and-Stone algorithm [34] in such a way that the number of samples in the test set in every group (i.e. MScl or CIS, see Table 1) was equal to 25% of the total number of samples in a group). The training sets were used for all

optimization steps and for developing a classifier, while the independent test was utilized to assess the predictive ability of the classification model. The Kennard-and-Stone algorithm is one of many possible approaches for data division [32, 35, 36]. The use of Kennard-and-Stone algorithm for data division is justified by the advantage of obtaining representative training set and the reproducibility of the selection. Nevertheless, since it is Euclidean based algorithm it might be influenced by noisy variables. Therefore, in addition the training and independent test sets were selected randomly. The results of presented fusion approach for random division is shown in the supplementary material.

The number of samples included in the training set and independent test set is shown in Table 1. Since the number of overlapping samples between NMR and GC-MS is relatively low (22) this puts limitations on the accuracy of the predictions when using a relatively small independent test. Therefore, as an additional check of the meaningfulness of the classification model, a permutation test with 10000 permutations was performed. Using a permutation test, we checked if the assessment of the classification of objects into the original classes is significantly better than any random classification in two arbitrary classes.

5.2.6.3 Supervised analysis: linear and non-linear approaches

The supervised analysis is carried out in order to extract class related information. Below, we briefly describe the overall strategy. First, the supervised data analysis involving linear methods is described followed by the proposed kernel based non-linear methods. In the next sections detailed information on specific technical aspects of the supervised data analysis strategy is provided.

The most straightforward approach in data analysis is to first use a linear method. Therefore, the linear method by means of the cross model validation (CMV) PLS-DA is applied [37]. In this technique, two cross validation procedures are included in the variable selection procedure based on jack-knifing. This approach enables removal of irrelevant variables and optimizes the model for accurate prediction of group memberships. This technique was first applied to individual datasets and then the selected variables were fused and analysed by the linear classifier.

Next, if the considered classification problem is suspected to be non-linear (e.g. when prediction accuracy of linear model is low), more sophisticated algorithms can be applied. Here, a non-linear technique based on kernel methods was utilised. The strategy is shown in Figure 1. The steps 1 and 2 were carried out on the training set. The first step consists of a variable selection method, which aims to obtain meaningful information from each individual data set. We used SVM-RFE for the non-linear kernel as variable selection method [19]. For both datasets the radial basis function (rbf), i.e. a Gaussian function, is used to map the original input data into a feature space [38]. The choice of kernel function is performed both by means of visual inspection of PCA score plot and using the root mean square error of cross validation (RMSECV).

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}}$$

Here, Y is a real class label, while $\hat{Y}$ is the predicted class label; n indicates the number of observations.

The kernel parameter sigma (σ) is optimized by LOOCV. More specifically, in each iteration, one object from the training set is removed and a model is constructed on the remaining objects for different values of σ. This is repeated until each object has been removed once. The RMSECV is calculated for each iteration. The optimal σ value is selected based on the first minimum in the RMSECV. SVM-RFE is performed for both datasets separately. Only the selected variables are then used in the subsequent steps. In the final part of step 1 (see Figure 1), the data with only significant variables are analysed by K-PLS-DA, which is an alternative to the SVM technique. This part is employed to tune the kernel parameters and to estimate the classification accuracy of the separate datasets. The optimal model complexity (i.e. number of latent variables (LV's)) was selected based on RMSECV. Note that the selected variables per dataset can be concatenated in classical mid-level fusion and analysed with K-PLS-DA.[8] In our procedure, SVM-RFE was selected as variable selection and K-PLS-DA as classification method. The use of SVM-RFE it justified by the fact that it is a well-established method, able to find significant variables in non-linear space. The binary classifier (PLS-DA) is a popular alternative to SVM. Our choice was guided by the fact that SVM offers sparse

solutions based on a limited number of observations, i.e. the support vectors. Since the obtained hyperplane can be based on outlying objects, this brings a question about the robustness of SVM [38]. The main benefits of K-PLS-DA are its efficiency and simplicity. In addition, it has convenient visualisation options in the latent variable space. Nevertheless, as it will be shown latter, in terms of prediction K-PLS-DA and SVM perform similarly (see Data fusion by MKL).In the second step (see Figure 1), the kernels of the individual datasets are concatenated by linear combination of their kernel matrixes. In step 3 the combined kernels are analyzed with K-PLS-DA. The accuracy of the K-PLS-DA model is validated by the independent test set and by the permutation test. In order to obtain more robust classification model in the final step (number 4) the K-PLS-DA model is reconstructed using all available samples (i.e. both training and test sets) and all previously optimized parameters, namely number of variables, sigma for rbf kernel, coefficients μ and nr. of LV's (see later). Moreover in this step variable importance in kernel space is evaluated and visualized.

5.2.7 Variable selection by SVM- RFE

The first step of our approach (Figure 1) consists of extracting the most relevant information from the datasets by using SVM- RFE variable selection. SVM is a powerful, supervised method and since this technique is extensively discussed in the literature we do not focus on its description [39]. SVM- RFE is an application of RFE in the SVM algorithm and was introduced by Guyon [19]. RFE is a backward elimination algorithm that ranks variables on the basis of the smallest change in a cost function minimized in the SVM algorithm. In the specific case of a non-linear kernel, used in this manuscript, the cost function to be minimized takes the form:

$$J = (1/2)\alpha^T \mathbf{H}\alpha - \alpha^T 1 \qquad (1)$$

Here, **H** is a matrix with elements $y_i y_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, **K** is a kernel function, y_j and y_j denotes the class labels, α's are the Lagrangian multipliers and 1 is a vector of ones. The algorithm begins by using all training data to train SVM. The matrix **H** is than recomputed for every variable being removed, while the α's values remain unchanged. The elimination of the input variable "*i*" causes the change in cost function, J. The change in the cost function is calculated according to equation 2:

$$\Delta J(i) = (1/2)\alpha^T \mathbf{H}\alpha - (1/2)\alpha^T \mathbf{H}(i)\alpha \qquad (2)$$

Here, **H**(-*i*) indicates the matrix **H** calculated when the input component "*i*" is removed. All the ΔJ(i) is calculated and the values are sorted accordingly. A subset of variables corresponding to the end of the sorted list of ΔJ (i.e. those with small ΔJ) is then removed. In our case, the subset is formed by only one variable in each iteration.

In order to select an optimal set of variables LOOCV approach is used. We used RFE with cross-validation since it increases the likelihood that relevant variables are selected. Averaging over cross-validation iterations ensures that the variables that were significant in each run are selected. This gives a better estimation of the important variables than performing variable selection only once using all training samples. Moreover, using a variable selection procedure with cross-validation, overly optimistic results (solely valid for the training models) can be avoided. In LOOCV in each iteration, one object of the training set is left out and a ranking is obtained. Next, the total ranking is obtained by sorting the variables based on the amount of times it is selected in the LOOCV. All variables that appear twice or more in the "top ten" of the rankings are selected. Although the number ten is somewhat arbitrary, exploration of other options (e.g. "top fifteen" or median +1 of the amount it is selected in the LOOCV) did not affect the outcome.

5.2.8 Data fusion by Multiple kernel learning

The second step of our approach (step 2, Figure 1) aims to combine the kernels. This is done by means of Multiple Kernel Learning (MKL), which was pioneered by Lanckriet *et al.* [40]and Bach *et al.* [41] as extension of single kernel to integrate multiple kernels in SVM. They integrated multiple kernels in classification problems. The essence of MKL is to combine kernel matrices into a single kernel using basic algebraic operations such as addition or multiplication. For example, given two (positive semi-definitive) kernels $\mathbf{K_1}$ and $\mathbf{K_2}$ it is possible to define the new kernel **K**, which is a parameterized linear combination of $\mathbf{K}_1$ and $\mathbf{K_2}$. In particular, given a set of kernels **K** it is possible to combine them by linear combination according to equation 3:

$$\mathbf{K} = \sum_{i=1}^{m} \mu_i \mathbf{K}_i \qquad (3)$$

Here m is a number of kernels and coefficients μ are non-negative to assure positive semi-definiteness of **K**: $\mu_i \geq 0$. Note that the dimensions of the kernels have to be equal (i.e. the number of samples in the datasets has to match). The coefficients μ_i in equation 3 can be tuned to weight the importance of the different kernels. The weights can be obtained in multiple ways, i.e. by applying different regularization method such as the L_1 or L_2-norm. L_1 regularization on the kernel coefficients corresponds to the requirement that the sum of μ_i equals one ($||\mu_i||_1=1$). L_1 regularization can lead to sparse optimal solution and diminishing one of the platforms.. In the current problem of deriving metabolite profiles from NMR and GC-MS datasets both datasets are relevant and complementary (the sets of measured metabolites are partially different). In order to avoid the possibility of shrinking the importance of any platform the L_2-norm was used as regularization parameter. In the L_2-norm approach, different constraint on the coefficients are used, i.e. the sum of squares of μ_i equals one ($||\mu_i||_2=1$). The L_2-norm yields a non-sparse solution and it distributes the coefficients over multiple kernels [20]. Using the L_2-norm MKL we try to find the best separation between classes by solving the objective as follows:

$$\min \sqrt{\left(Y-\hat{Y}\right)^2}$$

$$\hat{\mathbf{Y}} = \mathbf{XB}; \mathbf{X} = \sum_{i=1}^{m} \mu_i \mathbf{K}_i \qquad 4)$$

$$\mu_i \geq 0, i = 1,...,m$$

$$\| \mu_i \| = 1$$

The weights μ_i were optimized by LOOCV performed on the training set. The optimal weights were selected based on the minimal error of the root mean square error of cross-validation (RMSECV).

5.2.9 K-PLS-DA and variables visualisation

In the non-linear architecture presented in Figure 1, the fused kernels are analyzed with a classification method, K-PLS-DA (step 3). This means that PLS-DA is applied on to the combined kernel matrix.

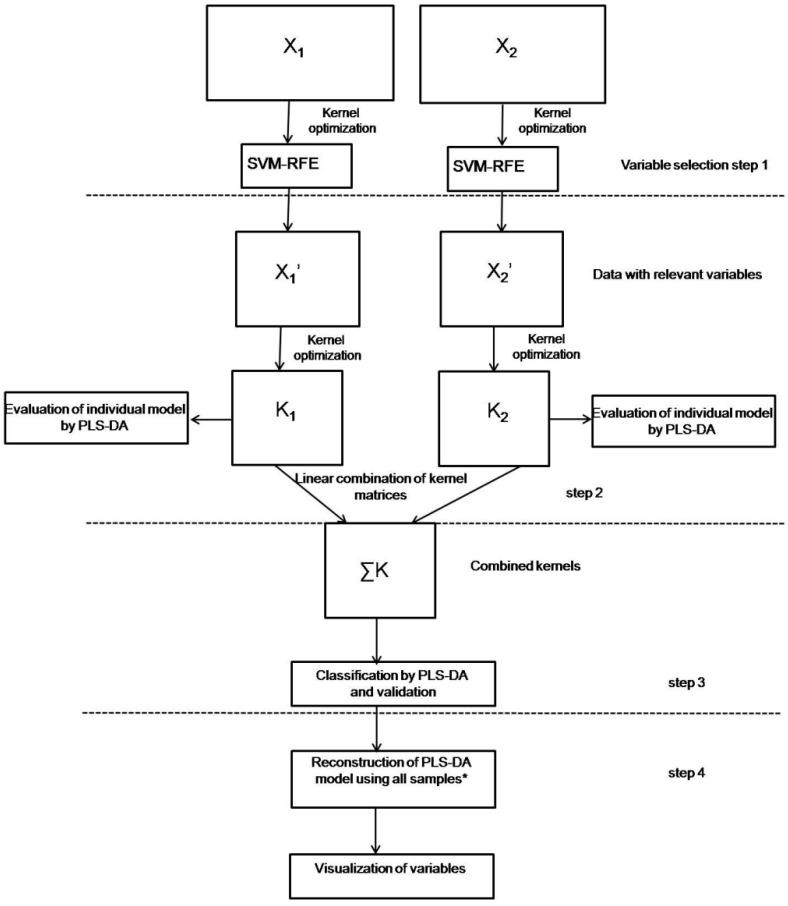


Figure 1. Conceptual flowchart of kernel-based data fusion. X_1 and X_2 are two blocks of data. * Note that all optimized parameters, i.e. number of variables, sigma for the rbf kernel, coefficients μ and nr. of LV's are kept during the model reconstruction using all available samples. The particular steps are described in sections data analysis.

The classification model is statistically validated, i.e. it is based on the prediction accuracy of the K-PLS-DA model, on the independent test and on the permutation test. Therefore, in the final fourth step of the approach shown in Figure 1 K-PLS-DA is first reconstructed using all available samples and next variables importance is established.

To represent the importance of the original variables, the pseudo samples principle, recently proposed by Krooshof *et al.* [16] and based on the non-linear plot principle described by Gower [42], was applied (step 4; Figure 1). As shown in Figure 2a, the matrix **X** (with n number of objects and p number of variables) is mapped by the kernel function $K(x_i,x_j)$ (where x_i and x_j are samples from matrix **X**). The obtained kernel matrix **K** is a square matrix of size "n x n" (where n is a number of samples). The application of PLS-DA on the kernel matrix leads to a linear model, i.e. **y=Kb+r**, where **y** a vector of group memberships, **b** regression coefficients and **r** a model residual. It is possible to obtain predicted ŷ-values for all training samples of matrix **X,** but the information about the variables (i.e. metabolites) involved in the discrimination is lost. In our approach every original variable is represented as a set of pseudo samples. The pseudo samples are artificial samples constructed as follows: every pseudo sample contains a certain value (e.g. 1) for only one variable and zeros for all the others. It is possible to check the influence of these pseudo samples in a K-PLS-DA model by predicting their corresponding ŷ-values or projecting them into latent variable space.
The graphical representation of the pseudo samples principle is shown in Figure 2b. It is possible to construct for each original variable a series of pseudo samples containing different values. These different values permit to describe a complete trajectory for each variable. In that way, for every variable a matrix of size "k x p" (where k is the number of pseudo samples used to span the complete range of the original variable and p the number of original variables) is created. From now on, we call this set of pseudo samples describing a single original variable a pseudo samples matrix. For data matrix **X** (shown in Figure 2a) "p" pseudo samples matrices are created, each of size "k x p". Once all pseudo samples matrices are constructed, one can apply the K-PLS-DA model to estimate the influence of the original variables. The pseudo samples are first mapped into the kernel space in relation to the original data matrix **X** using the same kernel

function as derived for data matrix **X** (Figure 2a), i.e. $K(x_i,ps_j)$ where x_i is an object of matrix **X** and ps_j is one pseudo sample. This leads to "p" kernel pseudo samples matrices (Figure 2b). Next the ŷ-values of pseudo samples can be estimated using regression vector "**b**" of K-PLS-DA model or they can be projected into LV space using loading vector of K-PLS-DA model. It has been shown that for linear kernel predicted ŷ-values of pseudo samples can be directly related to the regression coefficient of the original variables [17]. The projections of pseudo samples into the regression vector "**b**" of K-PLS-DA model from now on will be called "regression coefficient"; while the projection of the pseudo samples in the LV space will be referred to as a loading plot.

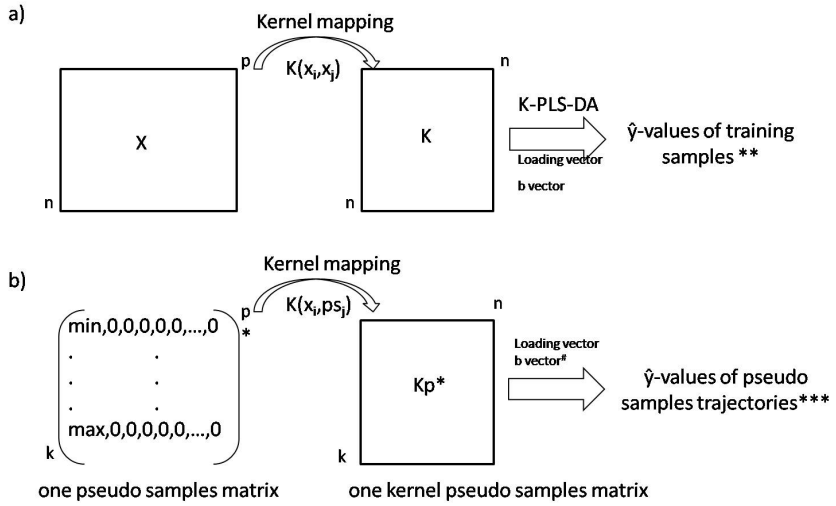


Figure 2. Representations of the a) kernel mapping of data matrix **X** into kernel space; b) pseudo samples principle in K-PLS-DA. k indicates the range of pseudo sample values (uniformly distributed); * Note that there are "p" pseudo sample matrixes and "p" kernel pseudo samples matrixes. **The ŷ-values can be projected into latent variable space. #Note that for kernel pseudo samples the loading and **b** vector of K-PLS-DA model are used. *** These ŷ-values can be represented as "regression coefficients" shown later in Figure 4 or loading plot shown in Figure 5.

The first graphical representation (Figure 3a) permits one to investigate how the original variables evolve as a function of the studied response as well as their global and local importance in the model. As described above, the kernels pseudo samples (see Figure 2b) are projected into the K-PLS-DA model to visualize the importance and behaviour of the original variables. A schematic example is provided in Figure 3a. The "regression coefficients" of four variables trajectories are displayed, each one illustrating a different case. If the influence of a given variable to the model is linear the corresponding pseudo samples trajectory should form a straight line, as variable 1 in Figure 3a. Variable 2 behaves linearly in the low variable range but becomes non-linear in the high range as can be observed from the corresponding curvature. Variable 3 represents a more complex sigmoid shape. This variable has big importance in the model in the low range and in high range but less in intermediate range. Note that in the high range variable 3 shows a plateau, which indicates that after passing a certain concentration value its importance stays constant. Finally, the last variable shown in Figure 3a, variable 4, has very little influence on the model. Note that, if the optimal K-PLS-DA model complexity is one LV, information contained in the regression coefficient and the loading vector (obtained from K-PLS-DA model) is equivalent. Therefore it is possible to use the loading vector instead of the regression vector "**b**" for obtaining the predicted ŷ-values of pseudo samples (the y-axis of Figure 3a represent 1LV). This kind of plot in linear PLS-DA is called loading plot. Therefore, in the rest of paper it will also be called loading plot.
Another piece of information delivered from Figure 3a, is the change of variable value between studied groups. Positive predicted ŷ-values of pseudo samples indicates group A and a negative indicates group B. For instance the value of variable 1 increases from group A to group B, while the value of variable 2 decreases from group A to group B.
Figure 3b is an enhanced version of the figure presented in reference [17]. It allows direct visualisation of the importance of each variable on the K-PLS-DA model. Figure 3b is constructed as follows: the absolute value of the maximal "regression coefficients" (i.e. predicted values of pseudo samples) is used as the relative importance of each variable. Note that this approach can be used when the original variables are scaled to unit variance. Note further, an alternative to estimate/visualize the relative importance would be by taking the absolute value of the difference between maximum value and minimum

value. The result can be graphically represented using the traditional regression plot obtained in any regression method [8]. Note that the importance of the variables can be also directly read off from Figure 3a, i.e. from the absolute max values along the horizontal axis. The 4 variables in Figure 3b correspond thus to the ones shown in Figure 3a.

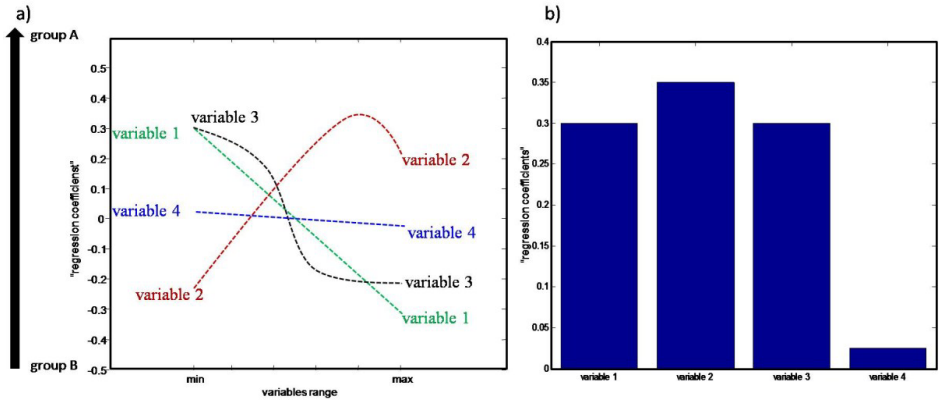


Figure 3. (a) A schematic example of "regression coefficients" of original variables trajectories plotted versus their range; (b) The maximum absolute value of "regression coefficients" of original variables trajectories shown in a.

5.2.10 Data

Every NMR spectrum of CIS and MScl groups was divided into 233 bins, corresponding to relative metabolites concentrations. These bins are equivalent to approximately 50 identified metabolites and some unidentified resonances. The GC-MS data consists of 66 metabolites and their corresponding relative concentrations. It is important to mention that 20 metabolites were measured by both NMR and GC-MS. Some metabolites are identified only by NMR (e.g. methanol) or only by GC-MS (e.g. urea) [43].

These two datasets are used as case study to represent the proposed architecture for non-linear data analysis and fusion. After variables selection by SVM-RFE the NMR data and GC-MS data are reduced to 47 bins and 29 metabolites, respectively. In case of

NMR these 47 informative bins correspond to 20 identified metabolites and some unidentified resonances.

It is important to keep in mind that σ, i.e. the parameter controlling the smoothness of the function, has to be tuned correctly, since it impacts the model performance. The σ parameter used for rbf kernel function is optimized separately for NMR dataset and GC-MS dataset and again before kernel fusion. An overview of σ parameters optimized in particular steps in Figure 1 is summarized in Table 2.

Table 2. Summary of σ parameter for rbf kernel function.

σ parameter at:	NMR	GC-MS
Step 1 (variable selection)	0.5	0.55
Step 3 (kernel fusion)	0.3	0.3

5.2.11 Metabolites identification

After selection and visualization of the most important variables, the corresponding metabolites were identified (NMR). Metabolite identification for NMR data was carried out by using the 600 MHz library of metabolite NMR spectra from the Chenomx NMR Suite 7. The library of metabolite spectra is obtained based on a database of pure compound spectra acquired using particular pulse sequence and acquisition parameters, the tn-noesy-presaturation pulse sequence with 4s acquisition time and 1s of recycle delay [44].

6.3 RESULTS

6.3.1Linear methods

The analysis of the data can be first performed per analytical method. This is particularly significant not only during exploratory phase but also during supervised analysis, where relevance of individual sets is investigated. Both datasets were first analyzed with R-PCA and PCA for presence of outliers and to detect potential trends. In total 4 NMR spectra and 3 GC-MS samples were detected as outliers and removed from further analysis. Since PCA score plots did not reflect any groupings and the variations did not separate according to groups CIS and MScl, the results of this analysis are not shown. Next, the linear method, CMV-PLS-DA, was employed on separate platforms and on fused datasets in mid-level fashion. In our case, the application of linear methods provided disappointing results for the separate datasets as well as for the datasets fused in the mid-level fashion. The degree of correct classification for a validation set obtained for the individual data-set analysis and for the concatenated sets can be seen in Table 3 and the corresponding figures are shown in the supplementary material (Figure 2Sa-2Sc).

Table 3. An overview of prediction accuracy for the validation set using linear methods, non-linear methods and MKL.

	Correct classification rate		
	NMR	GC-MS	fusion (NMR +GC-MS)
Linear method	61%	63%	65%
Non-linear method	93%	85%	89%
MKL			100%

6.3.2 Non-linear analysis

Since linear methods did not lead to satisfactory results (see Table 3), we used more sophisticated methods (i.e. non-linear) to find differences in metabolic profiles of CSF of CIS and MScl groups. As pointed out previously (see Materials and Methods), our approach is based on four steps. The first one consists of a variable selection performed on each dataset. We used SVM-RFE in order to get good predictive group membership ability and a meaningful interpretation of the model. After the first step, we analysed the separate datasets by K-PLS-DA. After variable selection every dataset can be assessed in terms of complexity and prediction accuracy. The overall correct classification for independent test sets, left out during model optimization and construction, is 93% for NMR data and 85% for GC-MS data, respectively (Table 3). Both K-PLS-DA models were constructed for 2 latent variables (LV's). These results suggest that both data sources hold relevant information concerning discriminating CIS and MScl groups. The overview of the prediction of each K-PLS-DA model is presented in Table 3.

The most straightforward approach for data fusion is to analyse the two datasets together by simply concatenating the selected variables from two data sources together (mid-level fusion). It is expected that the two types of information from the NMR and GC-MS datasets should complement each other and improve the class separation. However, this mid-level fusion provides very similar results in terms of complexity of the K-PLS-DA model and correct classification (i.e. 2LV's and 89% correct classification, see Table 3).

6.3.3 Data fusion by Multiple Kernel Learning

Since the analysis of both sets as unique matrices does not provide a better separation of the groups, we decided to apply kernel-based fusion by MKL (Materials and Methods). Note that this kernel fusion architecture, as applied by us here, falls outside the range of the classical low-, mid-, or high-level fusion. It uses the specific property of the kernel matrix in the data fusion, i.e. its dimensions and its nature comparable to a similarity matrix.

The MKL approach used here is composed of optimizing weights for each kernel matrix. The optimized weights were equal 0.75 for NMR and 0.661 for GC-MS. This indicates

that both datasets are almost equally important. After weighted kernel-based fusion, the newly formed kernel matrix can by analysed by PLS-DA. The kernel fusion leads to correct prediction of 100% on the independent test set (versus 89% for mid-level fusion, see Table 3). The K-PLS-DA model was constructed by using 1 LV. As an additional check, we performed a permutation test. The p-value for 10000 permutations was equal to 0.0013. The accuracy of K-PLS-DA was further compared to SVM. The correct prediction was as well 100% on the independent test set.

Since the model shows good predictive ability on the independent test set, we consider it as statistically validated and as shown in Figure 1 (step 4) the K-PLS-DA model is then reconstructed using all available samples. The resulting model can be graphically assessed using a score plot (here not shown). Obviously, this kind of plot is the normal visual representation of kernel method.

Variables importance visualization

As shown in Materials and Methods, it is possible to visualize the original variables in discriminating the groups. For that purpose the maximum absolute value of the predicted values of pseudo samples was calculated. The obtained values are shown in Figure 4. This figure demonstrates that there are several variables having very high importance. For instance, variables number 67 (sucrose), 76 (urea) and 50 (3-methyl-2-hydroxybutanoic acid) have the highest values of the predicted values of pseudo samples, demonstrating the relevance of these variables. There are just few variables seen as less significant, for example variable 57 (glycerol) or 63 (phenylalanine). The complete list of names of metabolites corresponding to variable number is given in the supplementary material (Table 1S).

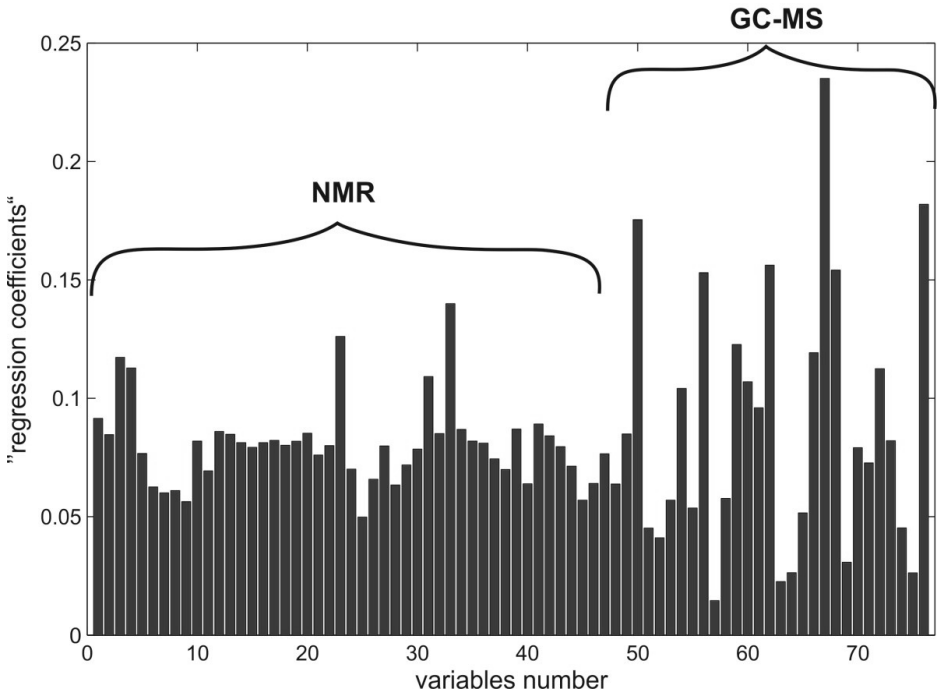


Figure 4. The maximum absolute value of "regression coefficients" of original variables.

As was explained in Materials and Methods, to investigate the relation between individual variables and changes of metabolite concentration (i.e. elevation or reduction) the trajectory of predicted values of pseudo samples (representing individual variables) can be studied. Since the optimal model complexity is 1LV we used loading vector delivered from K-PLS-DA model to project the pseudo samples into LV1. The obtained trajectories are shown in Figure 5. Because presenting trajectories for all variables makes the plot unreadable, in Figure 5, only a few of them are given. Trajectories for all variables are given in the supplementary material (Figure 1S).

Besides showing the importance of variables in discriminating, Figure 5 also reveals the linear or non-linear trend and/or monotonicity of the variables in certain concentration ranges. A variable which shows a non-linear trend is glutamine and is derived from

NMR. Valine is characterized by linearity and monotonicity in its low range, and non-linearity in its high range. Urea and sucrose demonstrate linearity over the whole concentration range.

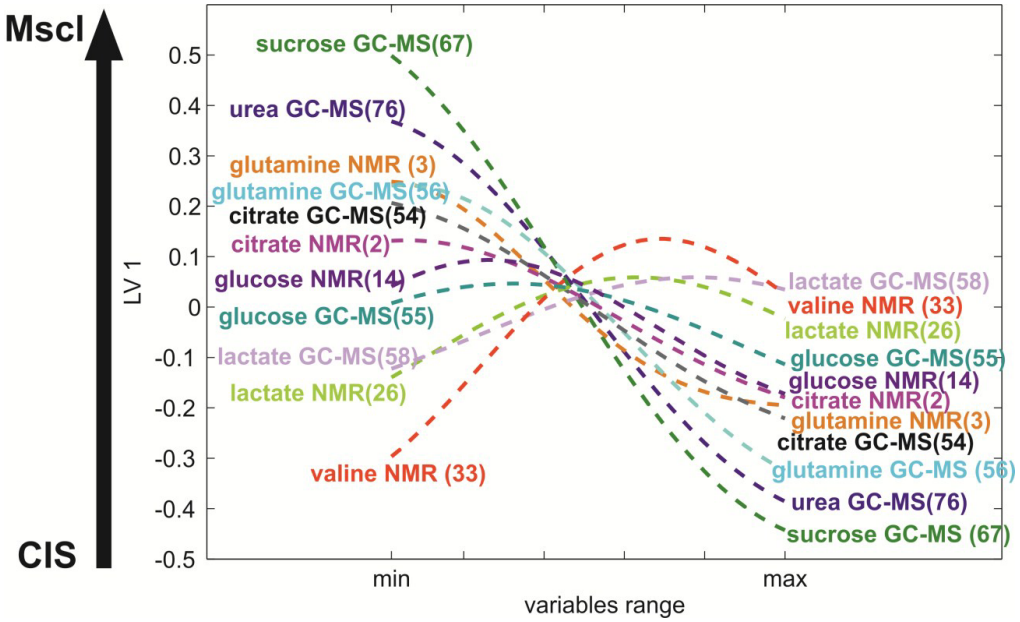


Figure 5. Loading plot of pseudo samples trajectories for selected variables. Numbers in the brackets correspond to variable numbers in Figure 4.

As mentioned before the change in metabolite concentrations across groups can be assessed. The horizontal axis in the Figure 5 represents the range of every original variable (scaled to its min to max value). The levels of lactate and of valine both increase, while the concentration of glutamine and citrate is reduced with disease progression. To make the change of metabolite concentrations more evident we included the direction of groups along vertical y-axis. More specifically, the negative values of the predicted values of pseudo samples correspond to CIS and the positive values to MScl.

At this point one should remember that some metabolites were measured both by NMR and GC-MS. It is therefore interesting to check how the corresponding variables compare with each other. For instance, pseudo sample trajectories for glutamine derived from NMR and GC-MS reveal very similar evolution upon disease progression. Correspondingly, pseudo sample for lactate, glucose and citrate measured by NMR and GC-MS display comparable trajectories along concentration range. This suggests that even after non-linear transformation the same variables measured by two different analytical methods are correlated and demonstrate their akin behaviour.

6.4 DISCUSSION

In this paper we have described a procedure for kernel-based data fusion. We have demonstrated an application of the proposed procedure to the classification problem of metabolomics datasets of CIS and MScl individuals. We have proposed a framework based on four steps, where the first one is focused on optimization of individual datasets. This is relevant, since we want to make sure that accurate information extracted from both data sources is included in fusion. We applied the L_2 MKL framework, demonstrated by Yu *et al.* for SVM, to K-PLS-DA, since it is characterized by the non-sparse integration of multiple kernels. Indeed, the optimization of the L_2 norm showed that both datasets, i.e. NMR and GC-MS, are valuable for discriminating CIS and MScl individuals.

The application of SVM-RFE allowed one to reduce both datasets significantly and select a set of informative variables. The classification performance of K-PLS-DA performed on fused kernels was better than single analysis and common mid-level fusion. Additionally, the K-PLS-DA model was simpler in terms of complexity, i.e. just one LV was sufficient to obtain optimal classification model.

The visualization of the variables relative contribution to K-PLS-DA model was achieved by applying and extending the pseudo samples approach demonstrated by [16, 17]. This allowed us to show that even after non-linear kernel transformation two different analytical methods are consistent with the results. The pseudo samples trajectories of the same variable measured by NMR or GC-MS demonstrate very similar trends.

Several potential challenges remain in the proposed framework. One possibility is to apply it to larger datasets with more different sources, for instance lipids and metabolites.

Since the number of available samples was limited, the potential impact of over-fitting of statistical model must be considered. We used here an independent test set and permutation test, which yielded results that clearly show that over-fitting, is highly unlikely. Note that in individual analysis the total number of samples was larger than in the fused set.

Classification with different types of non-linear functions in the original space can be achieved using diverse types of kernels. Of course, the choice of kernel function has to be done beforehand. A correct selection of kernel function has a significant influence on classification accuracy. However, no rules can be defined. In our experiment different kernel functions were tried. The Gaussian kernel was chosen as the one to fit the data properly. If the sigma value is too small over-fitting can easily occur. It has also influence on pseudo samples trajectories. Small sigma value (e.g. 0.1) might lead to very non-linear and hardly interpretable trajectories.
The case study described here represents indeed a non-linearly separable problem. As was shown, linear methods gave poor classification performance. The class separation was possible after application of a non-linear kernel function and PLS-DA. Non-linearity is also visible in the pseudo samples trajectories. There are several variables that are characterized by curved trajectories. The curvatures of the trajectories illustrate the effect of the original data. Importantly, even if these trajectories are non-linear, they are still simple enough to be interpretable.

The example given on the metabolites analysis of CSF gave very interesting results. However the number of used samples was relatively low. It should then be pointed out that due to this small number of samples, this study may have several limitations. Obviously, the size of the training and testing sets has an influence on the accuracy assessment of the classification method. Small sample size may result in detecting only the largest differences. Indeed the data size and classification rate are correlated.
The bigger the groups size the more representative and robust the results become. It has been shown [45] that as the sample size is increased prediction accuracy overcomes local minima and next stabilizes and therefore become more reliable and accurate. Obviously classification performance of a classifier is influenced by the natural difficulty of the studied problem, however there are possibilities where the performance of a classifier is degraded because of small training cases. Therefore one should be aware that results (with 100% correct prediction on test set) shown here do not imply that all new samples will be always correctly classified (due to e.g. biological variance). Hence, more studies involving a larger cohort will be necessary to fully establish and assess the

findings. However, despite the drawbacks of the study, it seems that validation with the independent test set and the permutation test set attest that the results are meaningful. In general the accuracy of the classification of objects into the original classes is significantly better assessed than any random classification in two arbitrary classes (p-value of 0.0013). Additionally the random division of data to training and test set was performed, showing that the average correct prediction over 10000 different runs supports our results.

From a biological point of view the metabolites having a relatively high contribution in the K-PLS-DA model, e.g. urea, glutamine, lactate, citrate, valine, are consistent with biological knowledge. These metabolites, described in this study, were previously found in relationship to the MScl [6]. They, therefore, provide a biological validation for the fusion of data. However, the full interpretation of the presented models in terms of biology still remains to be made. Therefore, future work will focus on the interpretation of newly detected metabolites and on highlighting pathways involved in the MScl disease process. The pattern defined by these variables must also be studied by itself and put into context in a system biology approach.

The kernel fusion approach presented in this paper assumes the dimensionalities of the kernels to be equal, i.e. the samples present in each dataset have to come from the same subjects at the same time points. Although, extending our method for missing values appears valuable and would be an interesting subject for further research.

ACKNOWLEDGEMENTS

The authors would like to thank Bas Muilwijk and Raymond Ramaker for the GC-MS data.

REFERENCES

1. Barbas, C.; Garcia-Perez, I.; Alves, A. C.; Angulo, S.; Li, J. V.; Utzinger, J.; Ebbels, T. M. D.; Legido-Quigley, C.; Nicholson, J. K.; Holmes, E., Bidirectional Correlation of NMR and Capillary Electrophoresis Fingerprints: A New Approach to Investigating Schistosoma mansoni Infection in a Mouse Model. *Analytical Chemistry* **2010,** 82, (1), 203-210.
2. Kuncheva, L. I., *Combining Petterns Classifiers*. Hoboken, New Jersey, 2004.
3. de Haan, J. R.; Wehrens, R.; Bauerschmidt, S.; Piek, E.; van Schaik, R. C.; Buydens, L. M. C., Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* **2007,** 23, (2), 184-190.
4. Ghauri, F. Y. K.; Nicholson, J. K.; Sweatman, B. C.; Wood, J.; Beddell, C. R.; Lindon, J. C.; Cairns, N. J., Nmr-Spectroscopy of Human Postmortem Cerebrospinal-Fluid - Distinction of Alzheimers-Disease from Control Using Pattern-Recognition and Statistics. *NMR in Biomedicine* **1993,** 6, (2), 163-167.
5. Constantinou, M. A.; Papakonstantinou, E.; Spraul, M.; Sevastiadou, S.; Costalos, C.; Koupparis, M. A.; Shulpis, K.; Tsantili-Kakoulidou, A.; Mikros, E., H-1 NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine. *Analytica Chimica Acta* **2005,** 542, (2), 169-177.
6. Sinclair, A. J.; Viant, M. R.; Ball, A. K.; Burdon, M. A.; Walker, E. A.; Stewart, P. M.; Rauz, S.; Young, S. P., NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases--a diagnostic tool? *NMR in Biomedicine* **2010,** 23, (2), 123-32.
7. Roussel, S.; Bellon-Maurel, V.; Roger, J. M.; Grenier, P., Fusion of aroma, FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. *Chemometrics and Intelligent Laboratory Systems* **2003,** 65, (2), 209-219.
8. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van-der Vat, B. J. C.; Jellema, R. H., Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry* **2005,** 77, (20), 6729-6736.
9. Steinmetz, V.; Sevila, F.; Bellon-Maurel, V., A methodology for sensor fusion design: Application to fruit quality assessment. *Journal of Agricultural Engineering Research* **1999,** 74, (1), 21-31.
10. Richards, S. E.; Dumas, M. E.; Fonville, J. M.; Ebbels, T. M. D.; Holmes, E.; Nicholson, J. K., Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework. *Chemometrics and Intelligent Laboratory Systems* **2010,** 104, (1), 121-131.
11. Jacobsson, S. P.; Forshed, J.; Idborg, H., Evaluation of different techniques for data fusion of LC/MS and H-1-NMR. *Chemometrics and Intelligent Laboratory Systems* **2007,** 85, (1), 102-109.
12. Blanchet, L.; Smolinska, A.; Attali, A.; Stoop, M. P.; Ampt, K. A.; van Aken, H.; Suidgeest, E.; Tuinstra, T.; Wijmenga, S. S.; Luider, T.; Buydens, L. M., Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics* **2011,** 12, (1), 254.
13. Liang, Y. Z.; Cao, D. S.; Xu, Q. S.; Hu, Q. N.; Zhang, L. X.; Fu, G. H., Exploring nonlinear relationships in chemical data using kernel-based methods. *Chemometrics and Intelligent Laboratory Systems* **2011,** 107, (1), 106-115.
14. Pekalska, E.; Haasdonk, B., Kernel Discriminant Analysis for Positive Definite and Indefinite Kernels. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **2009,** 31, (6), 1017-1031.
15. Ben-Hur, A.; Noble, W. S., Kernel methods for predicting protein-protein interactions. *Bioinformatics* **2005,** 21, I38-I46.
16. Krooshof, P. W.; Ustun, B.; Postma, G. J.; Buydens, L. M., Visualization and recovery of the (bio)chemical interesting variables in data analysis with support vector machine classification. *Analytical Chemistry* **2010,** 82, (16), 7000-7.
17. Postma, G. J.; Krooshof, P. W.; Buydens, L. M., Opening the kernel of kernel partial least squares and support vector machines. *Analytical Chimica Acta* **2011,** 705, (1-2), 123-34.
18. Compston, A.; Coles, A., Multiple sclerosis. *Lancet* **2008,** 372, (9648), 1502-17.
19. Guyon, I.; Weston, J.; Barnhill, S., Gene selection for cancer classification using Support Vector Machine. *Machine Learning* **2002,** 46, 389-422.
20. Yu, S.; Tranchevent, L. C.; De Moor, B.; Moreau, Y., *Kernel-based Data Fusion for Machine Learning. Methods and applications in Bioinformatics and Text mining*. Springer: Berlin 2011.
21. ACD/1D HNMR Manager, v., Advanced Chemistry Development, Inc, Toronto On, Canada. *www.acdlabs.com* **2003**.
22. Eilers, P. H. C., A perfect smoother. *Analytical Chemistry* **2003,** 75, (14), 3631-3636.

23. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **2004,** 18, (5), 231-241.
24. de Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Tsiporkova, E.; Rietzschel, E. R.; De Buyzere, M. L.; Gillebert, T. C.; Bekaert, S.; Martins, J. C.; Van Criekinge, W., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry* **2008,** 80, (10), 3783-3790.
25. Koek, M. M.; Muilwijk, B.; van der Werf, M. J.; Hankemeier, T., Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical Chemistry* **2006,** 78, (4), 1272-81.
26. van der Greef, J.; Hankemeier, T.; McBurney, R. N., Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics* **2006,** 7, (7), 1087-94.
27. van der Kloet, F. M.; Verheij, E. R.; Bobeldijk, I.; Jellema, R. H., Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. *Journal of Proteome Research* **2009,** 8, (11), 5132-5141.
28. Croux, C.; Ruiz-Gazen, A., High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis* **2005,** 95, (1), 206-226.
29. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., Assessment of PLSDA cross validation. *Metabolomics* **2008,** 4, (1), 81-89.
30. Rubingh, C. M.; Bijlsma, S.; Derks, E. P. P. A.; Bobeldijk, I.; Verheij, E. R.; Kochhar, S.; Smilde, A. K., Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2006,** 2, (2), 53-61.
31. Golbraikh, A.; Tropsha, A., Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design* **2002,** 16, (5-6), 357-369.
32. Stanimirova, I.; Kubik, A.; Walczak, B.; Einax, J. W., Discrimination of biofilm samples using pattern recognition techniques. *Analytical and Bioanalytical Chemistry* **2008,** 390, (5), 1273-1282.
33. Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A., Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design* **2003,** 17, (2), 241-253.
34. Kennard, R. W., Computer Aided Design of Experiments. *Technometrics* **1968,** 10, (2), 423-&.
35. Daszykowski, M.; Walczak, B.; Massart, D. L., Representative subset selection. *Analytica Chimica Acta* **2002,** 468, (1), 91-103.
36. Galvao, R. K. H.; Araujo, M. C. U.; Jose, G. E.; Pontes, M. J. C.; Silva, E. C.; Saldanha, T. C. B., A method for calibration and validation subset partitioning. *Talanta* **2005,** 67, (4), 736-740.
37. Gidskehaug, L.; Anderssen, E.; Alsberg, B. K., Cross model validation and optimisation of bilinear regression models. *Chemometrics and Intelligent Laboratory Systems* **2008,** 93, (1), 1-10.
38. Czekaj, T.; Wu, W.; Walczak, B., About kernel latent variable approach and SVM. *Journal of Chemometrics* **2005,** 19, 341-354.
39. Cristianini, N.; Shawe-Taylor, J., *An introduction to support vector machines and other kernel-based learning methods*. The University of Cambridge: 2000.
40. Lanckriet, G. R. G.; Cristianini, N.; Bartlett, P.; El Ghaoui, L.; Jordan, M. I., Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* **2004,** 5, 27-72.
41. Bach, F. G.; Lanckriet, G. R. G.; Jordan, M. I., Multiple kernel learning, conic duality, and the SMO algorithm. *In Proceedings of the 21 st International Conference on Machine Learning : Association for Computing Machinery* **2004**.
42. Gower, J. C.; Harding, S. A., Nonlinear biplots. *Biometrika* **1988,** 78, (3), 445-455.
43. Stoop, M. P.; Coulier, L.; Rosenling, T.; Shi, S.; Smolinska, A. M.; Buydens, L.; Ampt, K.; Stingl, C.; Dane, A.; Muilwijk, B.; Luitwieler, R. L.; Smitt, P. A.; Hintzen, R. Q.; Bischoff, R.; Wijmenga, S. S.; Hankemeier, T.; van Gool, A. J.; Luider, T. M., Quantitative proteomics and metabolomics analysis of normal human cerebrospinal fluid samples. *Molecular and Cellular Proteomics* **2010,** 9, (9), 2063-75.
44. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M., Targeted profiling: Quantitative analysis of H-1 NMR metabolomics data. *Analytical Chemistry* **2006,** 78, (13), 4430-4442.
45. Sord, M.; Zeng, Q., Although classification performance of a classifier is influenced by the natural difficulty of the studied problem, however there are possibilities that the performance of a classifier is degraded because of small training cases. *ISBMDA* **2005,** LNBI 3745, 193-201.

CHAPTER 7

SUMMARY AND FUTURE PERSPECTIVES

SUMMARY

In this thesis we investigated the metabolomics of biofluids by Nuclear Magnetic Resonance (NMR) in combination with chemometrics analysis. Apart from the development of new methods, the main aim was to identify potential metabolomic biomarkers of Multiple Sclerosis (MScl).

MScl is a disease of the Central Nervous System (CNS) and is characterized by a combination of many factors, such as inflammation, demyelination, remyelination and axonal damage, most probably caused by a disturbance of the autoimmune system. It is the most common chronic disease in young adults. Since the early 20th century, constant research has been performed to understand the origin and the pathology of the disease, to set up diagnostic criteria and to come up with a cure for this disease. However, the cause(s) of MScl still remains unknown. Moreover, a cure has not yet been found. MScl is a very heterogeneous disease and thus it is difficult to diagnose especially at an early stage. Therefore, there is still a need for new molecular biomarkers, which would allow an early diagnosis and consequently improve preventative action. Since MScl is a CNS disorder, analyzing Cerebrospinal fluid CSF is the most suitable and interesting compartment, because of its proximity to the brain, besides the brain itself. Therefore, our main focus was to investigate the metabolic profile of CSF. Nevertheless, we also examined blood plasma in the animal model of MScl and we showed that not only CSF, but blood plasma as well contains significant information about neuroinflammation.

In this thesis four main aspects are covered and discussed, namely: (i) sample treatments and measuring by NMR, (ii) data preprocessing of NMR data, (iii) chemometric analysis of NMR metabolomics data, and (iiii) data interpretation in biomarker discovery of MScl disease. The joint analysis of NMR of biofluids and pattern recognition methods has driven forward the relevance of metabolomics in biomarker discovery field. In **chapter 1** a review is given of the data acquisition and multivariate analysis of NMR-based metabolomics data, with particular emphasis on CSF and MScl. We demonstrate recent developments in biofluid sample handling and measuring, crucial steps in preprocessing of NMR spectra, i.e. baseline correction, alignment, binning and scaling, and different approaches in multivariate data analysis. Moreover,

the application of current developments of NMR and chemometrics methods in metabolic biomarker discovery for MScl is addressed.

A proper biomarker search first requires standardization in sample handling. Several aspects, such as sample collection, sample storage and time between collection and storage can influence the results. Therefore, in order to obtain reliable results and avoid false biomarker candidates the stability of metabolites identified by NMR was investigated. The metabolites that were identified in human CSF by NMR showed negligible changes in concentration when left at room temperature for 30 and 120 minutes before freezing and storing the samples at -80^{o}C. The results are presented in **chapter 2** of the thesis.

Investigating the biological and analytical variations in metabolite concentrations in CSF of "healthy" (i.e. with no neurological disease) individuals is brought up in the **third chapter** of the thesis. The knowledge of inter-individual variation of metabolite concentrations in CSF of "healthy" humans is essential for evaluating a potential up-regulated/down-regulated metabolite when confronted with CSF samples of subjects with a disease. The results clearly showed that variations in metabolite levels range from circa 8% to 53% for majority, while the analytical variation was found to be less than 9%. Moreover, using Principal Component Analysis no clear relation between variation in metabolite concentrations and gender and age range was found.

In biomarker discovery studies, it is a common approach to start the investigation of a certain disease with a control experiment, i.e. an animal model. Animal models allow a better understanding of the underlying (molecular) processes in a disease. As presented in **chapter 4**, we first investigated MScl by means of Experimental Autoimmune Encephalomyelitis (EAE), an animal model which mimics a certain aspect of MScl, namely neuroinflammation. In this study, we found that by using NMR spectra of rat CSF in combination with state of the art pattern recognition methods (PLS-DA and ANOVA-PCA), a set of metabolites relevant for neuroinflammation can be established. We demonstrated that the CSF metabolic profile of neuroinflamed animals is distinct from that of healthy and peripherally inflamed individuals. Moreover, we have shown that

neuroinflamed animals at the onset of the EAE are heterogeneous regarding to the disease response. More importantly, our findings were validated with a second independent set of animals, showing the relevance of the metabolites specific for neuroinflammation.

Our first study involving CSF of EAE affected rats indicated that it is difficult to distinguish neuroinflamed rats from peripherally inflamed animals at the onset of the disease. This was mostly due to heterogeneous response of neuroinflamed animals to disease. For this reason metabolic profiles of blood plasma and CSF were combined by the mid-level-fusion strategy. **In chapter 5,** the combined analysis of blood plasma and CSF of EAE models is detailed. It describes that the combined metabolomics information from blood plasma and CSF enables a more efficient and reliable discrimination of the onset of the EAE. Several metabolites in blood plasma are relevant in discriminating neuroinflammation at its onset. At this point new chemometric developments were required. Therefore, we present and introduce a new chemometric method: Hierarchical Models Fusion (HMF). This approach hierarchically combines previously developed classification two-class models. We show that HMF allows one to distinguish neuroinflamed rats also on the day of onset from either healthy or peripherally inflamed rats as well as to investigate the progression of EAE.

Following the studies of the two previous chapters we investigated the metabolic profile of MScl in humans. Recently, a number of studies aiming to identify metabolic components of MScl pathology have been published. However, in most of these studies a single metabolomic analytical platform (e.g. NMR or Gas Chromatography-Mass Spectrometry (GC-MS)) is used to find metabolic patterns characteristic for MScl. In **chapter 6,** we discuss a novel approach for data fusion, which is subsequently applied to CSF metabolic profiles of patients suffering from MScl and obtained by NMR and GC-MS. Due to complexity of these datasets, new challenges needed to be tackled. The data fusion architecture involves fusion in kernel space, which permits one to establish non-linear discrimination models. Using this approach we were able to study the progression of MScl in humans. We unveiled that fusion of datasets in kernel space provides more efficient discrimination between MScl individuals and patients

representing early stage of MScl in comparison to mid-level fusion approach and analysis of individual datasets. Moreover the visualization tools implemented for this method open the gate for a thorough biological interpretation. It is important to point out that the main focus was the statistical approach not the biological interpretation and translation. Nevertheless, the majority of metabolites specific for neuroinflammation found in CSF of EAE-affected rats, were also significant for describing progression of MScl in human CSF.

Conclusively, the metabolic biomarker discovery of MScl disease was investigated in different species (humans and rats) as well as in different biofluids (CSF and plasma). The thesis also presents new analytical/statistical methods for the metabolics research, methods which are able to improve the metabolic biomarker discovery. We found that the metabolic profile of MScl in an early stage can be recognized for humans and in a rat model. We also found that the majority of MScl related metabolites in rats are similar to those in humans. This is very interesting since it suggests that the results found on rats could be transmitted to humans. However for a good “translation” study further research is needed. Another interesting observation was the strong similarity of the metabolic profile in blood plasma and CSF of EAE rats, which suggest that the plasma metabolites could be used as a source for biomarker discovery of MScl instead of CSF. However, the similarity between blood plasma and CSF of EAE rats had only been cursory glanced, so in the future a detailed comparison has to be made to validate these results.

FUTURE PERSPECTIVES

The metabolomics analysis of biofluids by means of Nuclear Magnetic Resonance (NMR) and chemometrics in the investigation of neurological diseases is presented. The main focus was metabolomic biomarker search in Multiple Sclerosis (MScl) disease. Several aspects connected to suitable biofluids sampling and prior measurement samples handling, data pretreatment and data preprocessing, and proper analyzing are brought up, described and presented. Consequently, as it was pointed out in the introduction and summary, this thesis highlights four main points: coefficient

(i) analytical, i.e. samples handling and NMR measuring

(ii) preprocessing of NMR metabolomics data

(iii) data analysis of NMR metabolomics data

(iiii) interpretation of findings and their biological relevance

Obviously, each of these aspects opens doors for future research. However, the synergy between them is potentially more interesting. The most straightforward step is the validation of the findings on larger cohort of individuals, because the amount of samples is very crucial for precise statistical multivariate analysis. The larger the groups, the more representative and accurate the analysis becomes. Unfortunately, in many animal and human studies large study groups are difficult to reach and thus uncommon, e.g. due to ethical reasons. In biomarker discovery studies the number of individuals is generally low, because of the explorative nature of this type of research. To compensate for this pitfall, our statistical approaches were always validated with an independent test set. When it was possible, an extra validation was performed with additional subjects coming from an independent experiment. Nevertheless, the validation on the target population using hundreds or thousands of samples ensures the validity and reliability of discovered markers. Besides that, this kind of validation is mandatory if a biomarker is going to be used as diagnostic tool.

In parallel to validation on large population one could use these putative biomarkers to define drugs targets or effects. This is the basic concept behind

pharmacometabonomics. This is currently the new field where drug efficiency and/or toxicity are predicted from the metabolic profile of an individual prior to treatment. Concretely, the biomarkers could be utilized to study the pharmaceutical effects of different drugs in e.g. rats or mice using EAE model. The metabolite profiles would be used to predict a response to potential drug treatment. In that way the quantitative pharmaceutical properties of drugs could be measured. The concept of investigating the pharmaceutical effects can be extended into personalization of drug treatments, so called personalized medicine, which combines pre-dose metabolite profiling and chemometrics to model and predict the responses of individual subjects. From a chemometrics point of view it would be interesting to combine the information present in the covariance matrix to the information in the network analysis. Information contained in the network might be used directly to weight the covariance matrix, for instance if two metabolites share a reaction it can be assumed that they should be correlated. This could bring together two approaches, the usual bottom up approach of chemometrics (starting from data to build hypotheses) and the top down approach generally used in biochemical modeling. This would allow reaping the benefits of both concepts.

1D NMR spectra are very rich in information and completely untargeted. Therefore, one can expect to obtain a complex biomarker pattern based on multiple metabolites. Some of these metabolites can be straightforwardly assigned, while others remain challenging in terms of identification. Moving to higher dimensional NMR would be beneficial by reducing signals overlap and thus entail that more metabolites can be assigned. The application of homonuclear techniques like COSY and TOCSY could help in identifying unknown resonances in complex biofluid mixtures. Moreover, improving the sensitivity of NMR by applying recent developments in the area of hyperpolarization are potentially very useful in metabolomics. Finally, improved applications of so-called pure-shift NMR, in which NMR spectra are simplified by removing the resonance splittings due J-couplings, may considerably improve the spectral resolution. These advances in combination with higher magnetic fields open the doors for considerably better coverage of the metabolome of biofluids by means of NMR.

Once all signals have been assigned, the next step of the metabolomics analysis is to investigate the relationships between the metabolites. This allows finding others metabolites involved in a disease and consequently, to improve our understanding of a disease. In this respect, the integration of metabolomics with others 'omics' data, i.e. in a systems biology approach, is essential to come to a better understanding of the multi-factorial interactions of causes and effects of a disease. Such an approach will provide system-wide properties of a studied disorder. For this point, developing further data fusion methods is essential. One of the major problems in fusion methods is the aspect of missing values. This aspect would be of particular interest to explore in non-linear space.

In this thesis MScI was studied metabolomics in two different systems, namely in rats and in humans. Fortunately, in comparison to other 'omics' fields, metabolomics is far less species dependent. Therefore, it has potential to relatively easily perform translation from animals to humans. However, there is at present little known about direct comparison between metabolic profiles of animals, e.g. rats (used so often in animal models), and humans. It has been shown that there is still considerable difference in metabolite composition between human and rat urine. This is most probably triggered not only by dissimilarity in metabolic pathways in humans and rats, but as well by nutritional and food source differences. Therefore, an interesting issue would be to first investigate the qualitative and next quantitative difference in metabolic profile of humans and e.g. rats CSF. This would simplify the translation studies. The next step forward would consist of translating the results found in pre-clinical study into humans. In the EAE studies we have discovered candidate markers in CSF and blood plasma that are related to neuroinflammation. In our research metabolic profile of CSF of humans affected with MScI was also examined. It was necessary to develop novel chemometric approach, since the standard available ones did not fully enable to extract class-related variance. It ought to be mention that therefore the focus was the statistical approach not the biological interpretation and translation. Nevertheless, metabolites found previously in EAE model were identified in human CSF as disease related. However, still more exhaustive translation study is necessary.

The philosophy of translation does not apply only between species but also between different biofluids. CSF is in closest interaction with the brain therefore it is the obvious choice for analyzing CNS disorders. The biochemical composition of CSF may indicate the malfunction occurring in the brain and more generally the abnormal status of the brain. However, the collection of CSF requires a lumbar puncture, which is invasive and might result in clinical complications. Moreover the total volume of CSF is far lower than the total volume of blood, so this circumscribes the amount that can be sampled. Therefore the detection of markers in blood would be more optimal. The CSF metabolites are absorbed into the blood via the blood-brain barrier (BBB) and thus effects of CNS disorders can presumably also be detected in the biochemical composition of blood. Since BBB is often damaged in MScl causing “leakage”, this presumes that that blood may comprise significant information about the disease. In our research blood plasma was investigated in an animal model of MScl, namely EAE. We demonstrated the usefulness of blood plasma for discriminating the early onset of EAE. However an in depth investigation of relation between markers found in CSF, plasma and vice versa has not been done. Therefore the translation of markers found in CSF into blood is a possibility of future research.

For any translation study the chemometrics approach is essential. One can imagine adapting the calibration transfer methods to adjust a model constructed for rodents into a model suitable for humans, or a model for CSF to plasma.

SAMENVATTING

SAMENVATTING

In dit proefschrift worden metabolomics studies beschreven van lichaamsvloeistoffen waarin de data gemeten zijn met behulp van Nucleaire Magnetische KernspinResonantie (NMR) en geanalyseerd met behulp van chemometrie. Naast het ontwikkelen van nieuwe methoden, was het identificeren van potentiële metabole biomarkers voor Multiple Sclerose (MScl) een van de belangrijkste doelstellingen.

Multiple Sclerose is een aandoening van het centrale zenuwstelsel en wordt gekenmerkt door een combinatie van verschijnselen, zoals ontstekingen, beschadiging van het myeline en aantasting van de axonen, waarschijnlijk voorzaakt door een verstoring van het immuun systeem. Bij jong volwassenen is het de meest voorkomende chronische ziekte. Sinds het begin van de 20^{e} eeuw wordt onderzoek gedaan naar het het ontstaan en de pathologie van de ziekte. Ook wordt onderzoek gedaan om te komen tot betere criteria voor het diagnosticeren van de ziekte. Dit alles gericht op het ontwikkelen van een behandeling van de ziekte. Hoewel het duidelijk is geworden dat MScl een auto-immuun ziekte is, blijft het nog altijd niet geheel duidelijk waardoor MScl wordt veroorzaakt en, vooral, is er nog geen effectieve behandeling gevonden. MScl kan zich bovendien heel verschillend uiten, vooral in de beginstadia, en is dan ook moeilijk te diagnosticeren in een vroeg stadium. Daarom is er nog altijd een grote behoefte aan nieuwe moleculaire biomarkers die een vroege diagnose mogelijk maken en dientengevolge preventief handelen kunnen verbeteren. Omdat MScl een aandoening is van het centrale zenuwstelsel, is, naast onderzoek naar de hersenen zelf, het analyseren van hersen vloeistof- oftewel cerebrospinale vloeistof (CSF) - het meest voor de hand liggende en interessantste gebied. Daarom hebben wij ons onderzoek vooral gericht op het metabole profiel van CSF. Daarnaast hebben we ook bloed plasma van het dierlijke MScl model onderzocht en konden we aantonen dat niet alleen hersenvloeistof (CSF) maar ook bloed plasma belangrijke informatie over inflammatie van het centrale zenuw stelsel bevat.

Vier belangrijke aspecten worden in dit proefschrift beschreven en bediscussieerd, te weten: (i) behandeling van monsters en meten met behulp van NMR, (ii) bewerken van NMR data zodat deze bruikbaar worden voor de volgende stap, te weten: (iii) chemometrische analyse van NMR metabolomics data, en (iiii) interpretatie van data bij

het zoeken naar biomarkers voor MScl. Dit proefschrift laat via de combinatie, NMR metaboliet spectra van lichaamsvloeistoffen en analyse daarvan met patroon herkennings methoden,het grote belang en potentie van metabolomics zien en heeft daarmee metabolomics op het gebied van de ontdekking van biomarkers een impuls gegeven. In **hoofdstuk 1** wordt een overzicht gegeven van de data acquisitie en de multivariate analyse van de op NMR gebaseerde metabolomics gegevens, waarbij de nadruk wordt gelegd op CSF en MScl. We laten recente ontwikkelingen op het gebied van het behandelen en meten van monsters van lichaamsvloeistoffen de revue passeren, behandelen de cruciale stappen in het voorbehandelen van NMR spectra, zoals basislijncorrectie, alignment, binning en schaling van data. Daarnaast worden verschillende moderne methodes voor de multivariate data analyse beschreven. Verder wordt ook de toepassing van actuele ontwikkelingen van de NMR en chemometrische methoden bij het zoeken naar metabolische biomarkers voor MScl belicht.

Een goede zoektocht naar biomarkers vraagt allereerst om standarisatie in de behandeling van monsters. Verschillende aspecten, zoals het verzamelen en opslaan van de monsters, en de tijdsduur tussen verzameling en opslag, kunnen de resultaten beïnvloeden. Om betrouwbare resultaten te krijgen en valse biomarker kandidaten te vermijden, is daarom de stabiliteit van door NMR geïdentificeerde metabolieten onderzocht. De metabolieten die door NMR in menselijk CSF werden geïdentificeerd, lieten verwaarloosbare veranderingen in concentratie zien als ze tussen 30 en 120 minuten aan kamertemperatuur werden blootgesteld alvorens de samples in te vriezen en op te slaan bij -80˚C. Deze resultaten worden in **hoofdstuk 2** van dit proefschrift gepresenteerd.

Onderzoek naar de biologische en analytische variaties in metaboliet concentraties in het CSF van “gezonde” (d.w.z. zonder neurologische aandoeningen) individuen wordt in het **derde hoofdstuk** van dit proefschrift gepresenteerd. Kennis van de variatie van metaboliet concentraties in het CSF van en tussen “gezonde” mensen is van groot belang om de verhoging of verlaging van de concentratie van een metaboliet op zijn waarde te kunnen schatten wanneer men wordt geconfronteerd met CSF samples van zieke subjecten. De resultaten geven duidelijk aan dat de variatie in metabole niveaus

verschillen van ongeveer 8% tot 53% voor de meerderheid, terwijl voor de analytische variatie werd aangetoond dat deze kleiner dan 9% is. Door gebruik van hoofdcomponentenanalyse (PCA) werd bovendien zichtbaar dat er geen relatie is tussen variatie in metabole concentraties enerzijds en geslacht en leeftijd anderzijds.

Bij onderzoek naar biomarkers is het gebruikelijk het onderzoek naar een bepaalde ziekte te starten met een controle experiment, b.v. via een diermodel. Diermodellen maken het mogelijk tot een beter begrip van onderliggende (moleculaire) processen bij een ziekte te komen. Zoals gepresenteerd in **hoofdstuk 4**, hebben we MScl eerst onderzocht met behulp van Experimentele Auto-immune Encephalomyelitis (EAE), een diermodel dat een bepaald aspect van MScl, te weten neuro-inflammatie, nabootst. In dit onderzoek ontdekten we dat door het gebruik van NMR spectra van het CSF van ratten in combinatie met 'state-of-the art' patroon herkennings methoden (PLS-DA en ANOVA-PCA), een set van metabolieten relevant voor neuro-inflammatie kon worden vastgesteld. We toonden aan dat het metabole CSF profiel van dieren met neuro-inflammatie verschilt van dat van gezonde dieren en van dieren met perifere inflammatie. Ook toonden we aan dat de groep dieren met neuro-inflammatie bij de start van de EAE een heterogene reactie op de ziekte heeft. Van belang is dat onze resultaten zijn gevalideerd door een tweede onafhankelijke groep dieren, die de relevantie van metabolieten specifiek voor neuro-inflammatie liet zien.

Ons eerste onderzoek op het gebied van CSF van met EAE aangetaste ratten liet zien dat het moeilijk is om in de beginfase van de ziekte een onderscheid te maken tussen enerzijds ratten met neuro-inflammatie en anderzijds ratten met perifere inflammatie. Dit was vooral te wijten aan de heterogene respons op de ziekte van de neuro-geïnflammeerde dieren. Daarom hebben we de metabole profielen van bloed plasma en CSF gecombineerd met behulp van zogenaamde 'mid-level' data fusie. In **hoofdstuk 5** wordt de gecombineerde analyse van metabolite data van bloed plasma en CSF van EAE modellen gedetailleerd beschreven. Hieruit komt naar voren dat de gecombineerde metabolomics informatie uit bloed plasma en CSF een meer efficiënte en betrouwbare bepaling van de start van de EAE mogelijk maakt. Verschillende metabolieten in bloed plasma blijken van betekenis voor het bepalen van neuro-inflammatie in de beginfase.

Om tot dit resultaat te kunnen komen moest een nieuwe chemometrische methode worden ontwikkeld, namelijk Hierarchical Model Fusion (HMF). In HMF worden eerder ontwikkelde twee-groeps classificatie modellen hiërarchisch gecombineerd. We laten zien dat HMF het mogelijk maakt neuro-geïnflammeerde ratten vanaf het begin te onderscheiden van zowel gezonde als perifeer geïnflammeerde ratten. Op deze manier kan men de progressie van EAE volgen en onderzoeken vanaf een vroeg stadium.

Volgend op het onderzoek in de twee voorafgaande hoofdstukken, onderzochten we het metabole profiel van MScl bij mensen. Recentelijk is een aantal onderzoeken gepubliceerd waarin wordt geprobeerd de metabole componenten van de MScl pathologie te identificeren. In de meeste van deze publicaties wordt echter een enkelvoudig metabolomics platform (zoals NMR of Gas Chromatografie-Massa Spectrometrie (GC-MS)) gebruikt om de voor MScl karakteristieke metabole patronen te vinden. In **hoofdstuk 6** bespreken we een nieuwe aanpak van data fusie, welke wordt toegepast op door NMR en GC-MS verkregen metabole profielen van het CSF van MScl patiënten. Vanwege de complexiteit van deze verzamelingen van gegevens (datasets), moesten nieuwe analyse methoden worden onwikkeld, i.h.b. nieuwe data fusie methodes, namelijk in data fusie en analyse in kernel ruimte. Deze datafusie architectuur in de kernel ruimte, maakt het mogelijk om niet lineaire verbanden zichtbaar te maken. Dankzij deze kernelruimte (niet-lineaire) benadering konden we de progressie van MScl bij mensen onderzoeken. We laten zien dat fusie van datasets in de kernelruimte een effectieve methode is om onderscheid te maken tussen MScl patienten in een vroeg en later stadium, dwz tussen MScl patienten (laat MScl stadium) en CIS patienten (Clinically Isolated Syndrome; vroeg MScl stadium). Deze kernel fusie methode werkt aanzienlijk beter dan de mid-level fusie benadering en de analyse van individuele datasets. De visualisatie handvaten die voor deze methode zijn geïmplementeerd effenen het pad voor een diepgaande biologische interpretatie. Wij benadrukken dat in dit onderzoek de nadruk lag op de ontwikkeling van de statistische analyse methoden en niet op de biologische interpretatie en translatie. Desalniettemin was de meerderheid van de metabolieten, die karakteristiek waren voor neuro-inflammatie in het CSF van EAE ratten, ook significant voor het beschrijven van de progressie van MScl in menselijk CSF.

Dit proefschrift beschrijft een zoektocht naar metabole biomarkers voor MScl in verschillende lichaamsvloeistoffen (CSF en plasma) bij mens en dier (rat, EAE model). Het proefschrift presenteert ook nieuwe analytische/statistische methoden op het onderzoeksgebied van de metabolics, methoden die kunnen helpen bij deze zoektocht. Deze nieuwe methoden kunnen een impuls geven aan de verdere ontwikkeling van het veld van de metabolomics. Het is gevonden dat metabolite profiel MScl in een vroeg stadium is te herkennen in mensen en in een rat model (Hoofdstuk 5). Het is verder gevonden dat de meerderheid van aan de MScl-ziekte gerelateerde metabolieten in ratten overeenkomt met die in mensen. Dit is zeer interessant omdat dit suggereert dat resultaten gevonden voor ratten overdraagbaar zijn naar mensen. Voor een volledige 'translatie' studie moet evenwel een uitgebreidere en diepere analyse worden uitgevoerd. Een andere interessante observatie is de sterke overeenkomst in het metabole profiel van bloed plasma en CSF van EAE ratten (ratten die door EAE zijn beïnvloed). Dit suggereert dat het plasma metaboloom gebruikt zou kunnen worden als bron voor biomarkers voor MScl diagnostiek in plaats van de veel minder makkelijk toegankelijke CSF. Echter, de overeenkomst tussen bloed plasma en het CSF van met EAE aangetaste ratten is maar kort besproken in Hoofdstuk 4 en een gedetailleerde vergelijking moet in de toekomst worden uitgevoerd om dit punt te valideren.

SUPPLEMENTARY MATERIAL CHAPTER 2

SUPPLEMENTARY MATERIAL CHAPTER 2

Table S1. Peptide isotope standards used for SRM quantification

Protein (accession number)	Peptide	Modification	Mass difference (Da)	Concentration (mg/L)
Albumin (P02768)	^{427}FQNALLVR^{434}	L-Arg-$^{13}C_6$, $^{15}N_4$	10	27.7
Cystatin C (P01034)	^{52}ALDFAVGEYNK^{62}	L-Lys-$^{13}C_6$, $^{15}N_2$	8	6.6

Table S2. Settings used for SRM quantification.[1]

Peptide	Q1 mass	Q3 mass	Fragment ion	Collision energy (V)
FQNALLVR	480.8	685.4	y6	23.1 (2^+ ion)
ALDFAVGEYNK	613.8	709.4	y6	35.0 (2^+ ion)
ALDFAVGEYNK	613.8	780.3	y7	35.0 (2^+ ion)

Table S3. Cystatin C concentrations measured by SRM MS[2]

	Cystatin C concentration (mg/l)		
Sample	t = 0 minutes	t = 30 minutes	t = 120 minutes
H 1	2.71	2.96	2.98
H 2	2.91	2.77	3.15
H 3	2.41	2.55	2.19
H 4	5.24	6.08	5.83
H 5	2.32	2.17	2.60
H 6	2.28	2.30	2.40
Average	2.98	3.14	3.19

Table S4. Albumin concentrations measured by SRM MS[3]

	Albumin concentration (mg/l)		
Sample	t = 0 minutes	t = 30 minutes	t = 120 minutes
H 1	195.40	187.47	193.93
H 2	240.04	232.62	239.81
H 3	227.30	222.02	213.57
H 4	225.62	232.52	218.75
H 5	214.44	208.08	219.13
H 6	192.38	197.57	192.63
Average	215.87	213.38	212.97

[1] For all peptides the dwell time was set to 100 ms, entrance potential was set to 10 V and declustering potential was set to 66.2 V.

[2] No significant differences were observed between the time points (paired t-test, t = 0 vs. t = 30: p = 0.34, t = 0 vs. t = 120; p = 0.11, and t = 30 vs. t = 120: p = 0.71).

[3] No significant differences were observed between the time points (paired t-test, t = 0 vs. t = 30: p = 0.40, t = 0 vs. t = 120; p = 0.32, and t = 30 vs t = 120: p = 0.92).

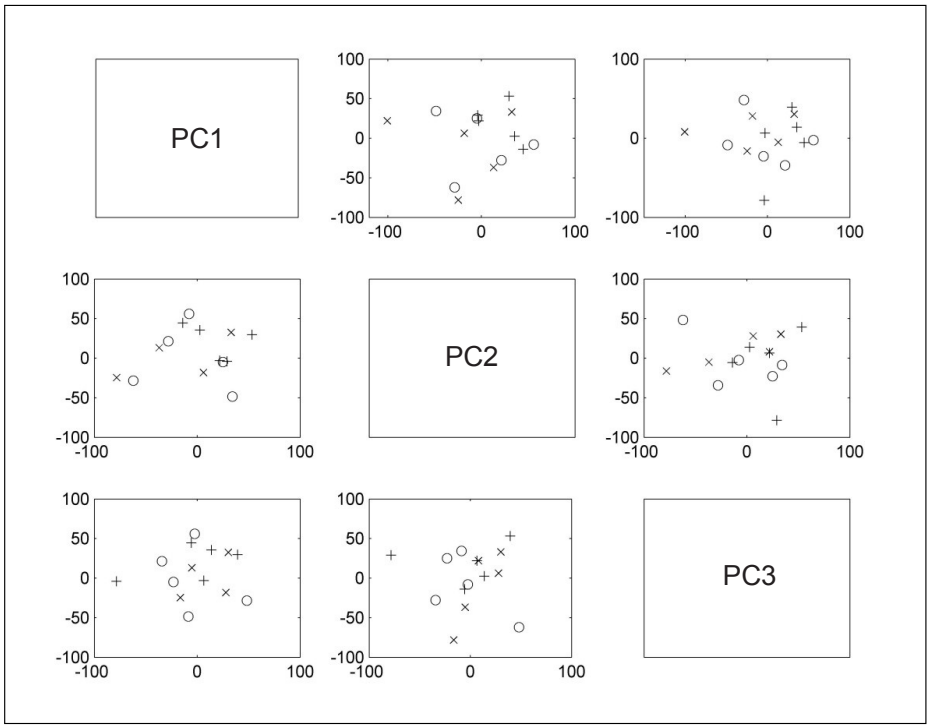


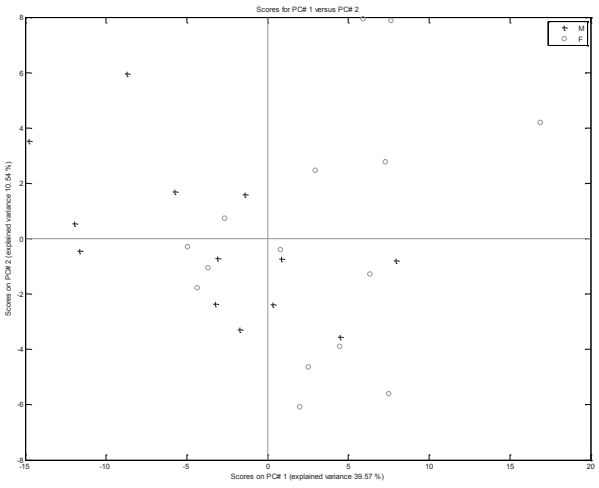
Figure S1. Score plots for PCs 1-3 of the NMR spectral data using vast scaling and meancentering per patient (0 min: o; 30 min: +; 120 min: ×). Total variance explained per first three PCs was 21%, 18% and 13%, respectively.

SUPPLEMENTARY MATERIAL CHAPTER 3

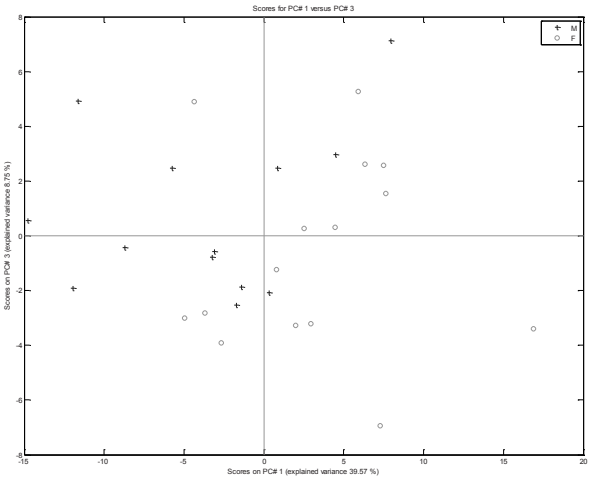
SUPPLEMENTARY MATERIAL CHAPTER 3

These figures all illustrate the lack of clustering of the different subgroups in a PCA analysis, indicating that the total variation is not unevenly influenced by one of these subgroups.

(a)



(b)



(c)

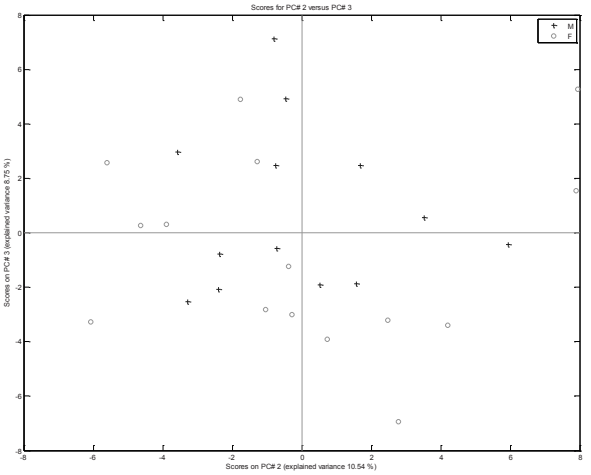
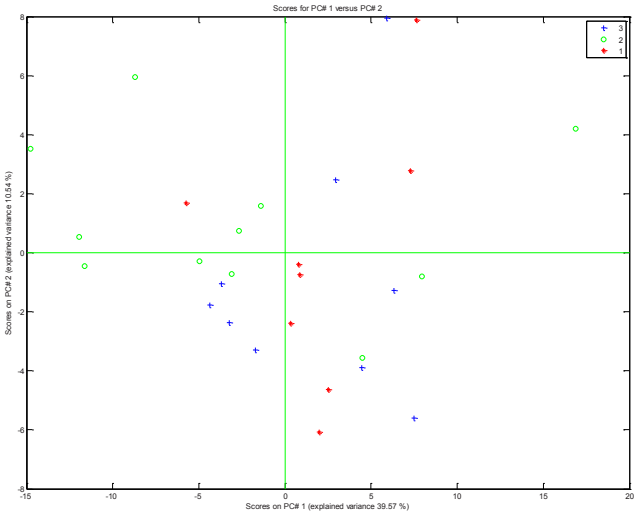
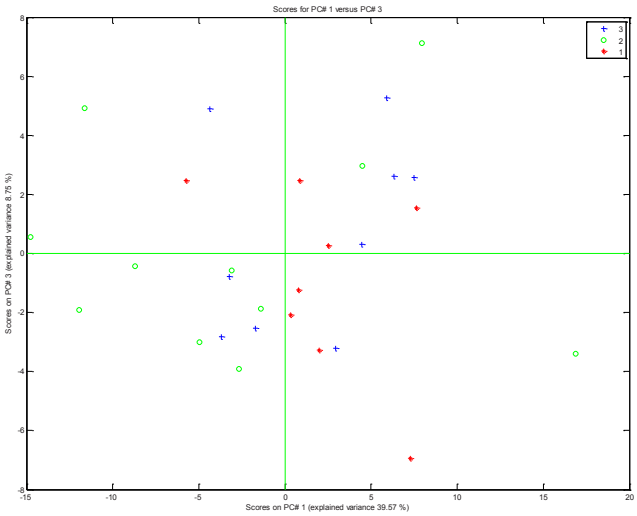


Figure S1. PCA plots proteomics ESI-Orbitrap, gender effect (Male vs. Female); (a), PC1 vs. PC2; (b) PC1 vs. PC3 ; (c) PC2 vs. PC3.

(a)


Scores for PC# 1 versus PC# 2
3
2
1
Scores on PC# 2 (explained variance 10.54 %)
Scores on PC# 1 (explained variance 39.57 %)

(b)


Scores for PC# 1 versus PC# 3
3
2
1
Scores on PC# 3 (explained variance 8.75 %)
Scores on PC# 1 (explained variance 39.57 %)

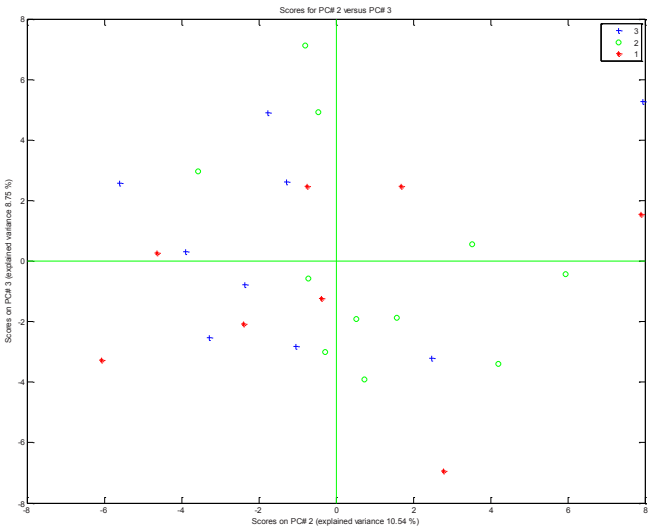
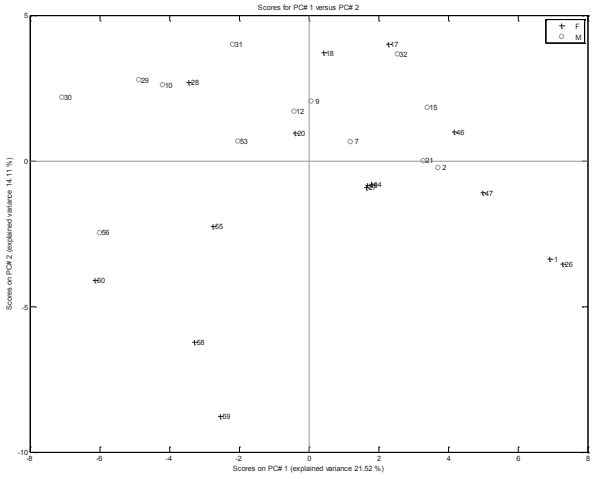
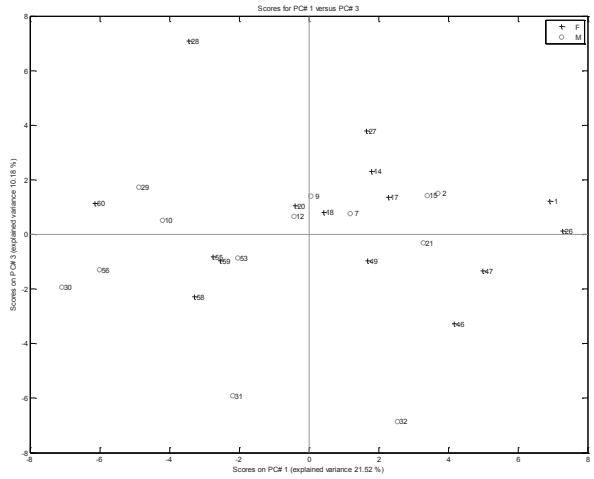


Figure S2. PCA plots proteomics ESI-Orbitrap, age effect (1 = <35, 2 = >35 and < 50, 3 = >50); (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

(a)


Scores for PC# 1 versus PC# 2
F
M
Scores on PC# 2 (explained variance 14.11 %)
Scores on PC# 1 (explained variance 21.52 %)

(b)


Scores for PC# 1 versus PC# 3
F
M
Scores on PC# 3 (explained variance 10.18 %)
Scores on PC# 1 (explained variance 21.52 %)

(c)

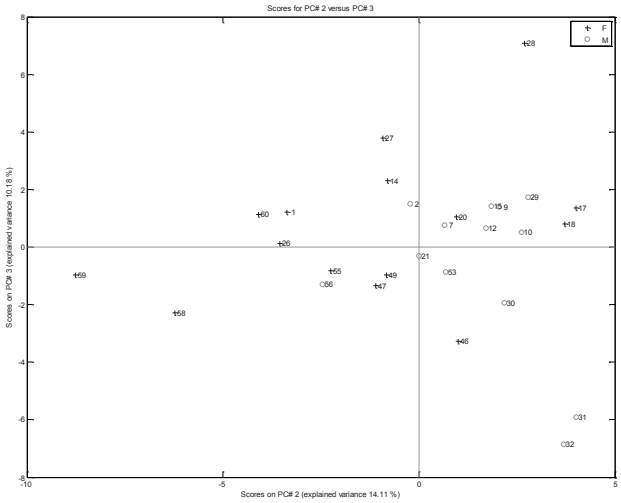
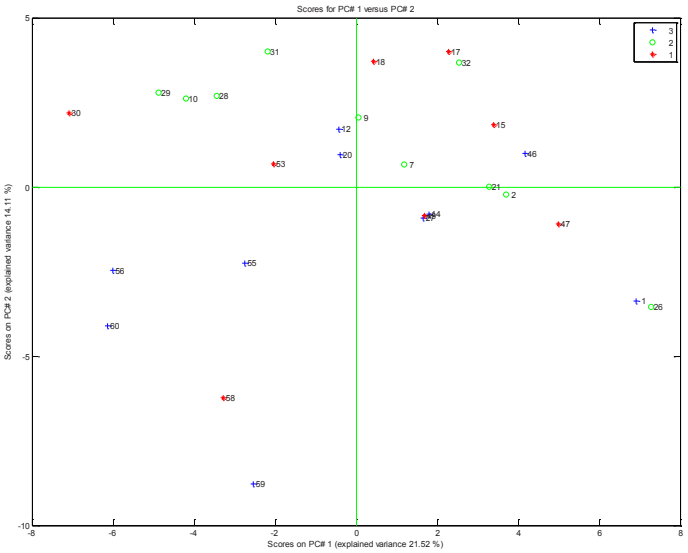
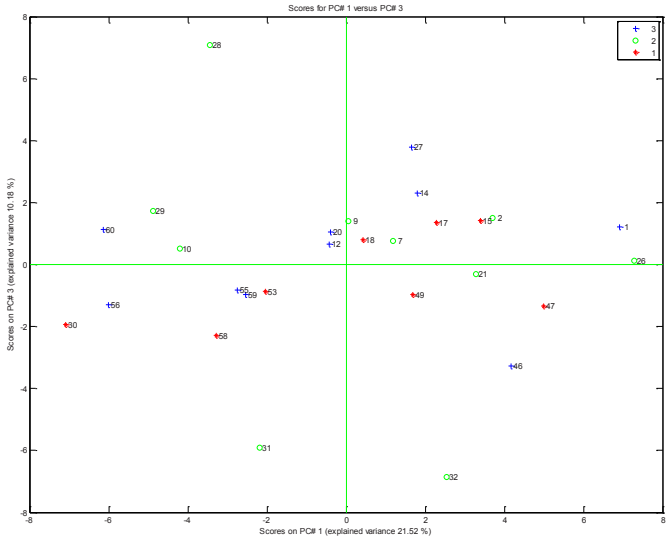


Figure S3. PCA plots metabolomics GC-MS, gender effect (Male vs. Female);(a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

(a)


Scores for PC# 1 versus PC# 2
Scores on PC# 2 (explained variance 14.11 %)
Scores on PC# 1 (explained variance 21.52 %)
3
2
1

(b)


Scores for PC# 1 versus PC# 3
Scores on PC# 3 (explained variance 10.18 %)
Scores on PC# 1 (explained variance 21.52 %)
3
2
1

(c)

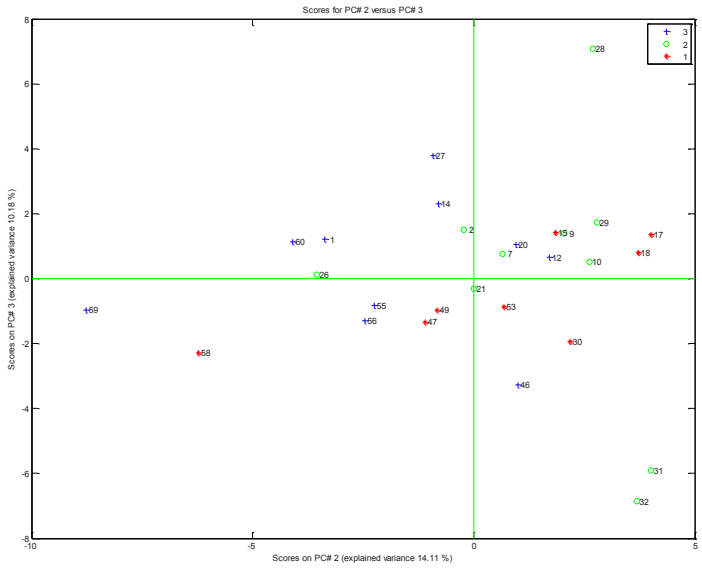
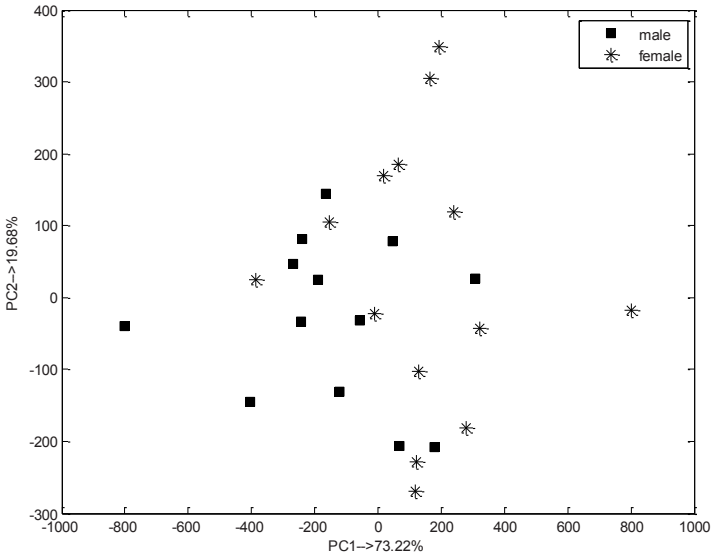
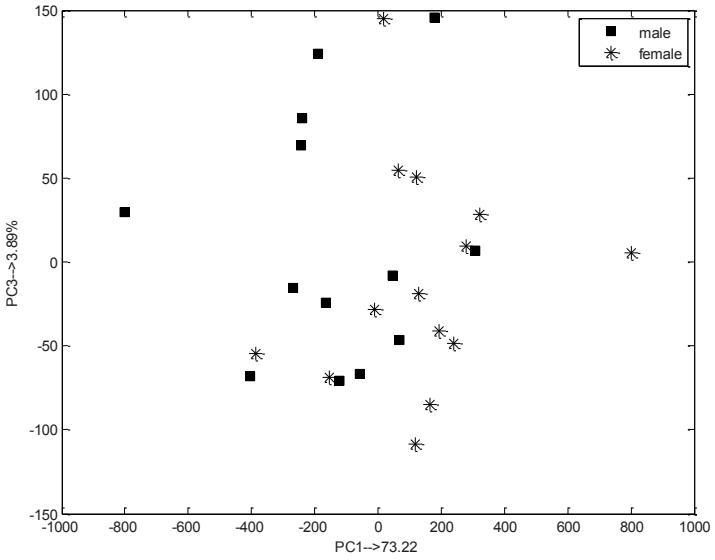


Figure S4: PCA plots metabolomics GC-MS, age effect (1 = <35, 2 = >35 and < 50, 3 = >50); (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

(a)



(b)



(c)

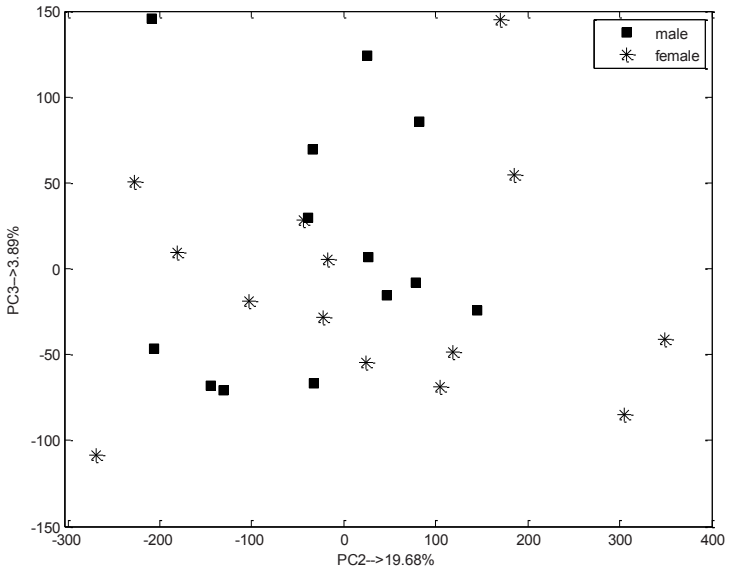
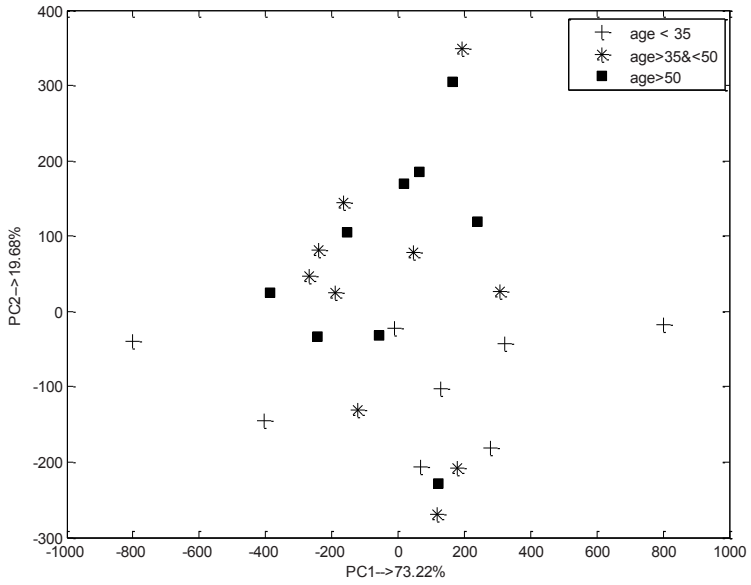
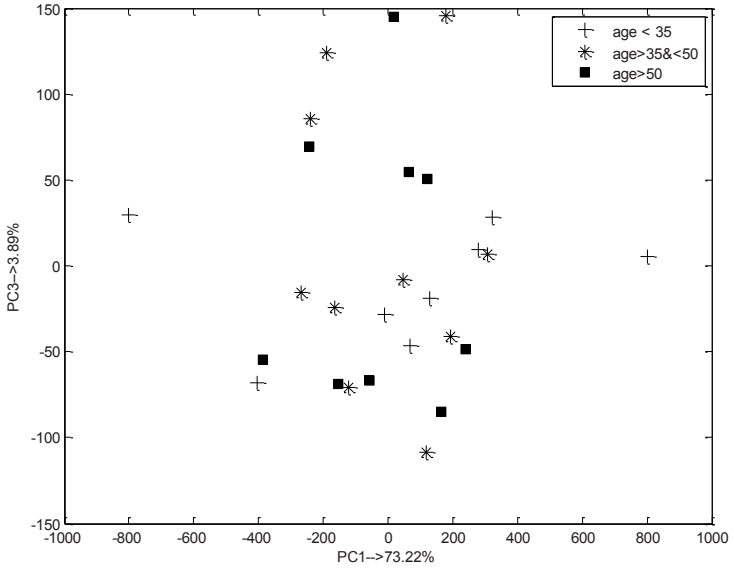


Figure S5. PCA plots metabolomics NMR, gender effect (Male vs. Female); (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

(a)


age < 35
age>35&<50
age>50
PC2-->19.68%
PC1-->73.22%

(b)


age < 35
age>35&<50
age>50
PC3-->3.89%
PC1-->73.22%

(c)

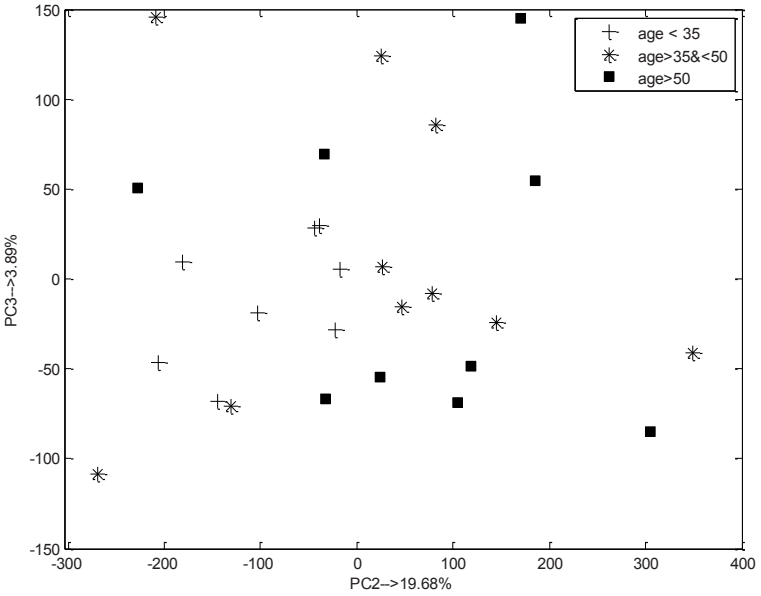
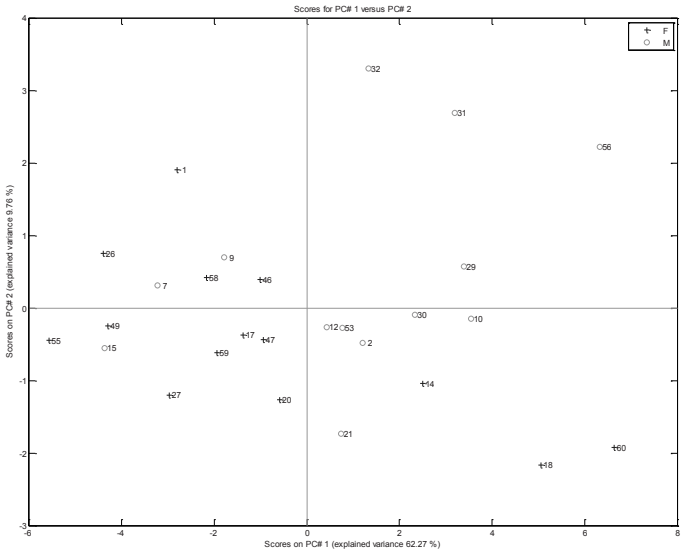
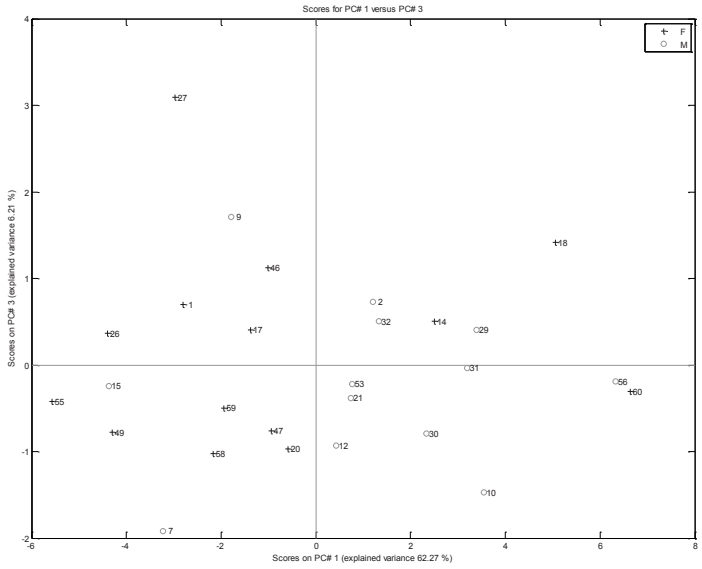


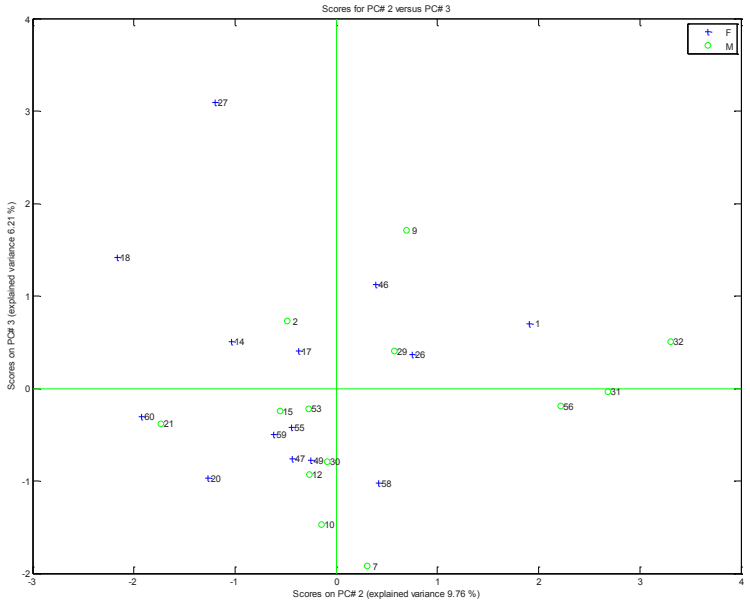
Figure S6. PCA plots metabolomics NMR, age effect (1 = <35, 2 = >35 and < 50, 3 = >50); (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

(a)


Scores for PC# 1 versus PC# 2
F
M
Scores on PC# 2 (explained variance 9.76 %)
Scores on PC# 1 (explained variance 62.27 %)

(b)

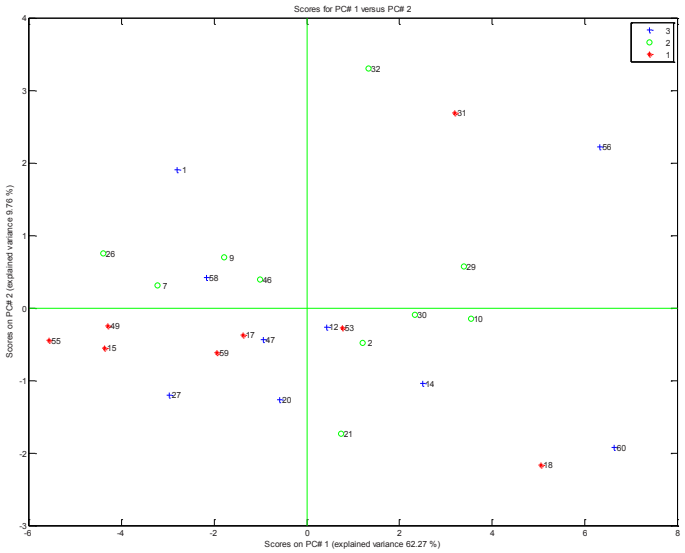

Scores for PC# 1 versus PC# 3
F
M
Scores on PC# 3 (explained variance 6.21 %)
Scores on PC# 1 (explained variance 62.27 %)



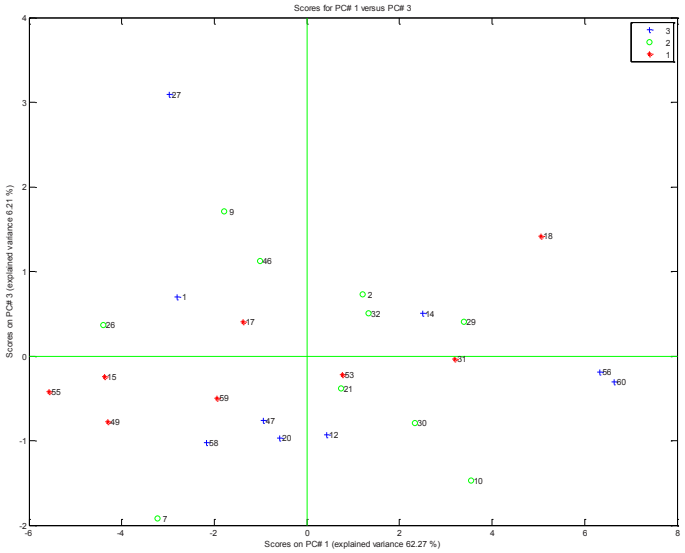
(c)

Figure S7. PCA plots metabolomics LC-MS/MS, gender effect (Male vs. Female); (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

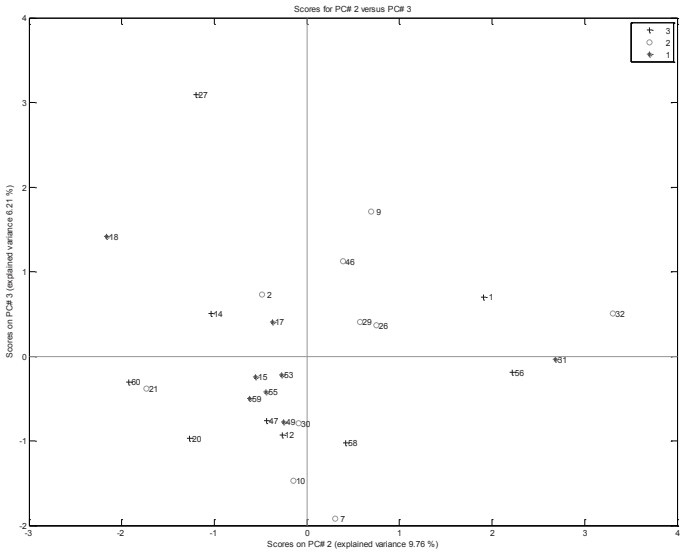
(a)


Scores for PC# 1 versus PC# 2
Scores on PC# 2 (explained variance 9.76 %)
Scores on PC# 1 (explained variance 62.27 %)
3
2
1

(b)


Scores for PC# 1 versus PC# 3
Scores on PC# 3 (explained variance 6.21 %)
Scores on PC# 1 (explained variance 62.27 %)
3
2
1

(c)



PCA plots metabolomics LC-MS/MS, age effect, age effect (1 = <35, 2 = >35 and < 50, 3 = >50), (a) PC1 vs. PC2; (b) PC1 vs. PC3; (c) PC2 vs. PC3.

Original samples information				
Sample	Gender	age	Protein (g/L)	Albumin (g/L)
61454	M	59	0.45	0.244
61525	M	30	0.25	0.118
61281	M	44	0.4	0.221
61430	M	51	0.62	0.403
61435	M	57	0.4	0.173
61349	M	62	0.33	0.179
61373	F	72	0.46	0.241
61378	M	57	0.31	0.097
61607	F	27	0.26	0.171
AVG	7M/2F	51	0.3867	0.2052
STDV		14.849	0.1162	0.0897
AVG Male		51.4	0.39	
STDV Male		11.1	0.12	
AVG Female		49.5	0.36	
STDV Female				

Validation samples information			
Sample number		**Gender**	**protein concentration (g/L)**
SFG01	F	64	0.32
SFG02	M	44	0.38
SFG07	M	43	0.34
SFG09	M	49	0.44
SFG10	M	40	0.65
SFG12	M	52	0.34
SFG14	F	76	0.4
SFG15	M	27	0.33
SFG17	F	17	0.21
SFG18	F	34	0.38
SFG20	F	74	0.48
SFG21	M	42	0.49
SFG26	F	49	0.19
SFG27	F	54	0.32
SFG28	F	42	0.35
SFG29	M	42	0.54
SFG30	M	39	0.35
SFG31	M	34	0.56
SFG32	M	36	0.45
SFG46	F	39	0.31
SFG47	F	50	0.29
SFG49	F	35	0.19
SFG53	M	32	0.36
SFG55	F	22	0.33
SFG56	M	62	0.23
SFG58	F	64	0.3
SFG59	F	29	0.28
SFG60	F	54	0.46
AVG	13M/15F	44	0.3668
STDV		14.513	0.1112
AVG Male		41.7	0.42
STDV Male		9.1	0.12
AVG Female		46.9	0.32
STDV Female		17.9	0.09

Experimental samples information metabolomics					
GC-MS	**MALDI-FT-ICR MS**	**Gender**	**Diagnosis**	**Age**	**protein concentration (g/L)**
2	171	M	PP MS	51	0.41
4	199	M	CIS	41	0.36
5	163	M	RR MS	62	0.47
6	198	M	RR MS	36	0.31
7	186	M	OIND	24	0.44
8	192	F	OIND	39	0.24
11	197	F	PP MS	60	0.29
12	120	M	OND	44	0.29
14	182	M	PP MS	36	0.44
15	200	F	CIS	29	0.24
17	159	F	CIS	27	0.27
18	161	M	OND	56	0.63
20	162	F	OND	46	0.36
22	150	M	CIS	22	0.29
23	148	M	OIND	62	0.77
24	147	F	RR MS	41	0.4
26	140	M	PP MS	50	0.62
27	103	F	PP MS	49	0.29
28	138	F	OIND	46	0.31
30	98	M	RR MS	16	0.53
31	125	F	RR MS	44	0.28
33	89	M	OIND	70	0.44
35	106	M	OND	40	0.27
36	87	F	OND	76	0.29
37	92	F	PP MS	48	0.35
40	57	M	RR MS	39	0.63
41	105	F	RR MS	38	0.33
42	102	F	OND	39	0.23
43	107	F	CIS	50	0.4
44	101	F	RR MS	46	0.33
45	50	M	PP MS	36	0.45
46	97	M	CIS	57	0.39
49	62	F	CIS	26	0.24
50	54	M	CIS	30	0.28
51	A261	M	RR MS	41	0.43
52	60	F	OND	57	0.4
53	68	M	OIND	36	0.27

56	64	F	OIND	54	0.42
59	69	M	OND	54	0.53
61	A66	M	RR MS	37	0.44
62	40	F	RR MS	36	0.19
63	129	M	PP MS	37	0.35
AVG	23 M/19 F			43.5	0.38
STDV				12.8	0.13
AVG Male				42.5	0.44
STDV Male				13.6	0.13
AVG Female				44.8	0.31
STDV Female				12.1	0.07

Experimental samples information proteomics

Sample number	Gender	Age	Protein concentration(g/L)	Albumin concentration(g/L)	Diagnosis
42	F	25	0.27	0.154	Headache
106	M	40	0.27	?	Headache
126	F	55	?	?	Headache
128	F	34	0.36	?	Headache
195	M	42	0.35	?	Headache
A69	M	40	0.22	?	Headache
A82	M	41	0.24	0.149	Headache
C1063	M	36	0.51	?	Headache
C1111	M	46	0.6	?	Headache
C1117	M	27	0.44	?	Headache
C1210	F	21	0.35	0.229	Headache
92	F	48	0.35	0.211	PP MS
103	F	49	0.29	0.159	PP MS
197	F	60	0.29	0.186	PP MS
A4	M	52	0.44	0.169	PP MS
A5	F	60	0.29	0.186	PP MS
A24	M	36	?	0.314	PP MS
A91	M	37	0.35	0.163	PP MS

A134	F	51	0.49	0.337	PP MS
A146	F	55	?	0.271	PP MS
A169	M	50	0.62	0.477	PP MS
A179	M	51	0.41	0.238	PP MS
C1251	F	51	0.49	0.337	PP MS
47	F	31	0.33	0.2	RR MS
91	F	36	0.61	0.405	RR MS
100	F	38	0.41	0.259	RR MS
105	F	38	0.33	0.2	RR MS
115	F	29	0.2	0.12	RR MS
122	F	33	0.45	0.196	RR MS
125	F	44	0.28	0.154	RR MS
147	F	41	0.4	0.245	RR MS
155	F	30	0.34	0.2	RR MS
163	M	62	0.47	?	RR MS
A102	F	39	0.35	0.189	RR MS
A167	F	22	0.3	0.116	RR MS
C1222	F	52	?	0.26	RR MS
AVG	13M/23F	41.722	0.378125	0.2268	
STDV		10.841	0.110173983	0.0858	
AVG Male		43.1	0.41		
STDV Male		9.0	0.13		
AVG Female		41.0	0.36		
STDV Female		11.9	0.09		

9 Original samples				
Name	**RSD (%)**	**Analytical error (%)**	**Av**	**SD**
Complement C4-A	18	25	27138450	4866115
Complement factor H	18	19	1700013	312547
Complement C3	19	24	87009580	16659566
Antithrombin-III	20	21	4526377	906119
Alpha-2-HS-glycoprotein	20	27	14545015	2928096
Insulin-like growth factor-binding protein 6	20	21	866435	174989
Tetranectin	21	24	323073	67465
Complement C1q subcomponent subunit B	24	26	399373	95186
Gelsolin	24	26	33507052	8044402
Alpha-2-macroglobulin	24	19	27648642	6694200
Complement C1s subcomponent	24	20	216100	52462
Ceruloplasmin	25	23	17944231	4425677
Serotransferrin	25	28	369677777	91474621
Clusterin	25	18	55161172	13968349
Apolipoprotein D	25	24	14147657	3602011
Ig gamma-1 chain C region	25	25	96127790	24489601
Alpha-1-antichymotrypsin	26	17	24319660	6224496
Cathepsin D	26	17	759239	194412
Ig kappa chain C region	26	18	82086700	21228205
Prothrombin	26	24	654402	172085
Plasma protease C1 inhibitor	26	24	1996211	524950
Beta-2-microglobulin	27	26	1332327	353236
Alpha-1B-glycoprotein	28	16	5949602	1581320
Epididymal secretory protein E1	29	26	12728294	3426895
Vitamin D-binding protein	29	24	12170699	3418187
Ganglioside GM2 activator	29	23	193682	56813
Hemopexin	29	22	32879169	9684949
Beta-2-glycoprotein 1	30	18	2313020	697871
Fibronectin	30	23	5098155	1539160
Mimecan	30	26	865639	261395
Apolipoprotein A-I	31	15	7056838	2185665
Transthyretin	32	24	26458600	8343161
Vitronectin	32	18	3310278	1065162
Plasminogen	33	15	870518	285846
Ig lambda chain C regions	34	26	26621885	9039823
Extracellular superoxide dismutase	35	17	302183	104416

[Cu-Zn]				
Fibrinogen beta chain	35	22	829886	289591
Apolipoprotein A-IV	35	15	462888	162289
Chromogranin-A	35	29	13666560	10277369
Fibrinogen alpha chain	35	23	1069593	377280
Apolipoprotein E	36	28	21931476	7818894
Angiotensinogen	36	26	39827610	14414312
Insulin-like growth factor-binding protein 7	36	28	819068	297200
Complement factor B	37	28	1206238	441960
Zinc-alpha-2-glycoprotein	37	18	2710354	995077
Secretogranin-1	37	29	13441778	4976916
Complement component C7	37	24	336342	125909
Galectin-3-binding protein	38	26	459980	173737
Apolipoprotein A-II	39	20	197709	76502
Alpha-1-antitrypsin	39	24	67673227	26229891
Serum albumin	39	23	4381612279	969042168
Inter-alpha-trypsin inhibitor heavy chain H4	39	24	286511	112307
Alpha-2-antiplasmin	40	21	51774	20609
Kallikrein-6	40	29	7760011	3115900
Complement C1r subcomponent	40	27	314186	126650
Prostaglandin-H2 D-isomerase	40	21	422446514	170728098
Cystatin-C	41	20	95242329	39101877
Histidine-rich glycoprotein	41	29	2399445	986981
Calsyntenin-1	41	24	1093188	451402
Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	41	28	2441554	1008732
Ig gamma-3 chain C region	41	22	7832786	3245788
Pigment epithelium-derived factor	42	29	20021202	8428956
Ribonuclease pancreatic	42	24	1660649	701515
Biotinidase	43	24	83830	35754
Ig gamma-4 chain C region	44	19	9375918	4169811
Neural cell adhesion molecule 1	45	28	3438240	1560063
Ig kappa chain V-III region SIE	46	25	1449706	662829
Afamin	46	23	59458	27186
Osteopontin	46	23	6064788	2800927
Kininogen-1	47	26	927469	435902
Carboxypeptidase E	47	25	698574	328487
Peptidyl-glycine alpha-amidating monooxygenase	47	27	155663	73622
Alpha-1-acid glycoprotein 1	48	18	13926929	6654663

Alpha-1-acid glycoprotein 2	48	29	3470429	1664010
Inter-alpha-trypsin inhibitor heavy chain H1	49	20	210609	103093
Chitinase-3-like protein 1	49	17	250984	123304
Dickkopf-related protein 3	50	24	17936224	8987787
Pyruvate kinase isozymes M1/M2	50	15	187792	94322
Amyloid beta A4 protein	50	18	2207802	1109417
Lumican	51	27	243809	123767
Beta-Ala-His dipeptidase	51	24	6140453	3134697
Ig gamma-2 chain C region	51	26	30692062	15691973
Neuroendocrine protein 7B2	52	16	723035	377600
ProSAAS	52	28	1453298	760500
Inter-alpha-trypsin inhibitor heavy chain H2	53	23	598461	315987
Fibrinogen gamma chain	54	20	698561	377501
Amyloid-like protein 1	54	19	5115411	2775572
Brevican core protein	55	23	749172	412662
Cartilage acidic protein 1	56	29	1589651	888129
Procollagen C-endopeptidase enhancer 1	56	21	163431	91350
Protein FAM3C	57	17	1526470	875357
Collagen alpha-1(VI) chain	59	28	566701	337022
N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase	60	30	4212554	2513146
Insulin-like growth factor-binding protein 2	60	23	30120	18025
Cadherin-2	60	27	488433	295467
Major prion protein	61	27	1525232	924607
Reelin	61	28	26557	16274
Secretogranin-3	62	28	4392667	2729659
Protein kinase C-binding protein NELL2	63	22	141105	89249
Neural cell adhesion molecule 2	65	22	252609	163981
SPARC-like protein 1	65	26	7946151	5169792
Tyrosine-protein phosphatase non-receptor type substrate 1	66	19	84342	55393
Ubiquitin	68	25	224325	151670
Neuronal cell adhesion molecule	69	22	5730668	3957629
Superoxide dismutase [Cu-Zn]	71	18	535106	378451
Cell adhesion molecule 3	71	30	220836	157002
Secretogranin-2	72	27	1613520	1155587
Neuroserpin	72	26	123199	88440
Thy-1 membrane glycoprotein	74	29	211863	157129
Neurotrimin	75	26	172232	129000

Neural cell adhesion molecule L1-like protein	76	16	2988910	2269783
Neurosecretory protein VGF	76	23	2497365	1906055
IgGFc-binding protein	78	20	56777	44510
Receptor-type tyrosine-protein phosphatase zeta	81	26	27691	22364
Cadherin-13	82	28	124803	102958
Limbic system-associated membrane protein	83	29	273267	228088
L-lactate dehydrogenase B chain	85	27	28031	23767
Disintegrin and metalloproteinase domain-containing protein 22	86	22	47948	41253
Neuronal pentraxin receptor	97	19	1531542	1491664
Seizure 6-like protein	102	21	17922	18231
Voltage-dependent calcium channel subunit alpha-2/delta-1	103	19	38474	39755
Ephrin type-A receptor 4	107	19	71103	76301
Contactin-2	124	19	1740303	2166015
Basement membrane-specific heparan sulfate proteoglycan core protein	134	24	421578	563833
Haptoglobin	135	28	21359232	28936658
Hornerin	148	28	9722	14363

Variation analysis proteomics

28 validation samples						
Name	RSD %	RSD % Age >50	RSD % Age >35 & <50	RSD % Age<35	RSD % Male	RSD % Female
Complement C4-A	14	15	13	13	14	14
Complement factor H	18	18	20	18	18	19
Complement C3	15	14	12	17	13	17
Antithrombin-III	14	16	15	14	14	16
Alpha-2-HS-glycoprotein	15	13	15	16	15	16
Insulin-like growth factor-binding protein 6	17	18	19	17	17	18
Tetranectin	18	17	23	18	42	26
Complement C1q subcomponent subunit B	16	16	17	16	17	16
Gelsolin	19	20	20	16	15	20
Alpha-2-macroglobulin	20	22	18	20	21	19
Complement C1s subcomponent	20	20	23	21	21	20

Ceruloplasmin	19	16	21	20	16	23
Serotransferrin	18	19	19	21	20	16
Clusterin	21	18	22	25	22	21
Apolipoprotein D	20	23	15	16	17	21
Ig gamma-1 chain C region	22	21	26	20	26	19
Alpha-1-antichymotrypsin	17	16	17	21	16	18
Cathepsin D	20	22	19	30	18	22
Ig kappa chain C region	23	25	25	21	26	16
Prothrombin	22	21	21	25	21	22
Plasma protease C1 inhibitor	23	20	23	25	28	18
Beta-2-microglobulin	23	24	22	19	28	19
Alpha-1B-glycoprotein	23	20	26	23	21	26
Epididymal secretory protein E1	25	26	27	21	23	25
Vitamin D-binding protein	21	18	23	22	24	18
Ganglioside GM2 activator	26	31	28	19	24	25
Hemopexin	21	21	23	16	21	21
Beta-2-glycoprotein 1	24	16	27	26	25	24
Fibronectin	17	17	20	14	17	18
Mimecan	21	22	26	19	23	20
Apolipoprotein A-I	29	32	27	25	29	28
Transthyretin	24	27	28	20	25	23
Vitronectin	25	27	23	31	27	24
Plasminogen	24	20	22	16	17	29
Ig lambda chain C regions	26	33	31	20	28	26
Extracellular superoxide dismutase [Cu-Zn]	30	27	26	35	27	35
Fibrinogen beta chain	29	30	32	28	30	28
Apolipoprotein A-IV	29	32	29	28	27	29
Chromogranin-A	32	31	33	26	30	33
Fibrinogen alpha chain	30	34	32	29	29	31
Apolipoprotein E	27	29	27	23	25	29
Angiotensinogen	25	29	24	25	26	23
Insulin-like growth factor-binding protein 7	26	32	28	19	24	27
Complement factor B	26	22	28	29	23	28
Zinc-alpha-2-glycoprotein	26	27	29	24	21	30

Secretogranin-1	27	27	32	21	34	23
Complement component C7	28	32	23	33	30	27
Galectin-3-binding protein	27	23	23	35	35	21
Apolipoprotein A-II	42	46	44	36	39	47
Alpha-1-antitrypsin	31	33	27	36	28	32
Serum albumin	30	27	29	32	29	30
Inter-alpha-trypsin inhibitor heavy chain H4	27	30	29	24	27	27
Alpha-2-antiplasmin	36	34	44	35	44	28
Kallikrein-6	31	28	23	27	31	31
Complement C1r subcomponent	27	21	34	21	30	25
Prostaglandin-H2 D-isomerase	28	28	27	29	30	25
Cystatin-C	30	27	38	32	33	28
Histidine-rich glycoprotein	29	32	38	24	30	28
Calsyntenin-1	33	23	38	27	35	30
Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	31	28	32	31	21	38
Ig gamma-3 chain C region	32	30	38	27	30	33
Pigment epithelium-derived factor	30	26	27	35	27	32
Ribonuclease pancreatic	30	32	32	28	25	36
Biotinidase	32	32	26	36	27	36
Ig gamma-4 chain C region	34	34	33	35	30	39
Neural cell adhesion molecule 1	32	31	32	33	23	42
Ig kappa chain V-III region SIE	35	34	33	38	32	37
Afamin	34	27	35	24	33	34
Osteopontin	35	30	38	32	33	36
Kininogen-1	31	38	31	24	26	35
Carboxypeptidase E	26	24	27	25	24	27
Peptidyl-glycine alpha-amidating monooxygenase	45	50	42	49	43	47
Alpha-1-acid glycoprotein 1	33	37	34	33	33	34

Alpha-1-acid glycoprotein 2	32	28	31	37	21	43
Inter-alpha-trypsin inhibitor heavy chain H1	44	37	45	52	56	32
Chitinase-3-like protein 1	33	37	53	28	40	45
Dickkopf-related protein 3	33	33	39	30	31	35
Pyruvate kinase isozymes M1/M2	36	32	46	26	38	33
Amyloid beta A4 protein	37	41	39	30	39	36
Lumican	39	45	34	36	29	48
Beta-Ala-His dipeptidase	40	39	46	33	45	35
Ig gamma-2 chain C region	40	37	40	42	42	37
Neuroendocrine protein 7B2	39	41	43	34	39	37
ProSAAS	39	47	32	39	32	47
Inter-alpha-trypsin inhibitor heavy chain H2	43	42	45	36	38	47
Fibrinogen gamma chain	40	36	42	42	35	44
Amyloid-like protein 1	41	39	47	40	37	45
Brevican core protein	45	42	50	42	45	43
Cartilage acidic protein 1	40	40	46	34	35	45
Procollagen C-endopeptidase enhancer 1	41	47	35	44	42	41
Protein FAM3C	42	44	44	41	42	42
Collagen alpha-1(VI) chain	44	47	46	42	37	49
N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase	44	49	44	37	45	43
Insulin-like growth factor-binding protein 2	41	54	38	33	39	43
Cadherin-2	42	46	47	35	34	50
Major prion protein	45	45	45	45	32	62
Reelin	46	46	47	46	46	48
Secretogranin-3	45	53	45	41	45	46
Protein kinase C-	47	45	48	46	40	52

binding protein NELL2						
Neural cell adhesion molecule 2	48	58	43	43	44	51
SPARC-like protein 1	48	55	48	42	42	55
Tyrosine-protein phosphatase non-receptor type substrate 1	35	29	50	22	40	30
Ubiquitin	51	60	57	37	61	42
Neuronal cell adhesion molecule	51	56	45	48	41	60
Superoxide dismutase [Cu-Zn]	49	56	47	54	49	49
Cell adhesion molecule 3	42	38	46	41	31	52
Secretogranin-2	47	40	54	49	53	42
Neuroserpin	51	64	47	41	42	62
Thy-1 membrane glycoprotein	52	66	46	46	49	57
Neurotrimin	56	60	52	56	57	54
Neural cell adhesion molecule L1-like protein	52	62	49	47	54	50
Neurosecretory protein VGF	52	63	47	48	49	54
IgGFc-binding protein	52	55	52	50	46	58
Receptor-type tyrosine-protein phosphatase zeta	55	65	50	50	47	61
Cadherin-13	60	58	59	63	61	58
Limbic system-associated membrane protein	59	60	60	59	53	66
L-lactate dehydrogenase B chain	69	68	69	69	72	67
Disintegrin and metalloproteinase domain-containing protein 22	68	67	68	68	59	75
Neuronal pentraxin receptor	63	65	61	63	60	65
Seizure 6-like protein	86	98	84	72	81	92
Voltage-dependent calcium channel subunit alpha-2/delta-1	61	64	59	59	56	66
Ephrin type-A receptor 4	72	82	67	64	75	68

Contactin-2	80	86	82	70	83	75
Basement membrane-specific heparan sulfate proteoglycan core protein	86	102	70	88	80	93
Haptoglobin	84	91	84	78	74	94
Hornerin	184	209	156	185	141	228

Variation analysis proteomics

28 Validation samples	
Name	RSD %
Complement C4-A	79
Complement factor H	30
Complement C3	84
Antithrombin-III	53
Alpha-2-HS-glycoprotein	47
Insulin-like growth factor-binding protein 6	49
Tetranectin	52
Complement C1q subcomponent subunit B	58
Gelsolin	39
Alpha-2-macroglobulin	39
Complement C1s subcomponent	50
Ceruloplasmin	50
Serotransferrin	49
Clusterin	81
Apolipoprotein D	54
Ig gamma-1 chain C region	98
Alpha-1-antichymotrypsin	62
Cathepsin D	53
Ig kappa chain C region	124
Prothrombin	46
Plasma protease C1 inhibitor	57
Beta-2-microglobulin	64
Alpha-1B-glycoprotein	88
Epididymal secretory protein E1	81
Vitamin D-binding protein	55
Ganglioside GM2 activator	54
Hemopexin	62
Beta-2-glycoprotein 1	67
Fibronectin	65
Mimecan	70
Apolipoprotein A-I	70

Transthyretin	73
Vitronectin	84
Plasminogen	106
Ig lambda chain C regions	73
Extracellular superoxide dismutase [Cu-Zn]	66
Fibrinogen beta chain	89
Apolipoprotein A-IV	70
Chromogranin-A	72
Fibrinogen alpha chain	108
Apolipoprotein E	64
Angiotensinogen	69
Insulin-like growth factor-binding protein 7	69
Complement factor B	86
Zinc-alpha-2-glycoprotein	81
Secretogranin-1	106
Complement component C7	56
Galectin-3-binding protein	69
Apolipoprotein A-II	85
Alpha-1-antitrypsin	72
Serum albumin	60
Inter-alpha-trypsin inhibitor heavy chain H4	73
Alpha-2-antiplasmin	59
Kallikrein-6	71
Complement C1r subcomponent	100
Prostaglandin-H2 D-isomerase	59
Cystatin-C	80
Histidine-rich glycoprotein	98
Calsyntenin-1	88
Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	73
Ig gamma-3 chain C region	151
Pigment epithelium-derived factor	80
Ribonuclease pancreatic	89
Biotinidase	94
Ig gamma-4 chain C region	148
Neural cell adhesion molecule 1	110
Ig kappa chain V-III region SIE	126
Afamin	72
Osteopontin	71
Kininogen-1	99
Carboxypeptidase E	82
Peptidyl-glycine alpha-amidating monooxygenase	88
Alpha-1-acid glycoprotein 1	83

Alpha-1-acid glycoprotein 2	92
Inter-alpha-trypsin inhibitor heavy chain H1	104
Chitinase-3-like protein 1	87
Dickkopf-related protein 3	94
Pyruvate kinase isozymes M1/M2	84
Amyloid beta A4 protein	89
Lumican	111
Beta-Ala-His dipeptidase	106
Ig gamma-2 chain C region	172
Neuroendocrine protein 7B2	130
ProSAAS	120
Inter-alpha-trypsin inhibitor heavy chain H2	107
Fibrinogen gamma chain	120
Amyloid-like protein 1	119
Brevican core protein	92
Cartilage acidic protein 1	101
Procollagen C-endopeptidase enhancer 1	68
Protein FAM3C	115
Collagen alpha-1(VI) chain	91
N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase	118
Insulin-like growth factor-binding protein 2	100
Cadherin-2	96
Major prion protein	103
Reelin	119
Secretogranin-3	99
Protein kinase C-binding protein NELL2	108
Neural cell adhesion molecule 2	91
SPARC-like protein 1	111
Tyrosine-protein phosphatase non-receptor type substrate 1	85
Ubiquitin	106
Neuronal cell adhesion molecule	126
Superoxide dismutase [Cu-Zn]	112
Cell adhesion molecule 3	126
Secretogranin-2	127
Neuroserpin	134
Thy-1 membrane glycoprotein	127
Neurotrimin	145
Neural cell adhesion molecule L1-like protein	156
Neurosecretory protein VGF	114
IgGFc-binding protein	120
Receptor-type tyrosine-protein phosphatase zeta	134
Cadherin-13	143

Limbic system-associated membrane protein	127
L-lactate dehydrogenase B chain	126
Disintegrin and metalloproteinase domain-containing protein 22	159
Neuronal pentraxin receptor	165
Seizure 6-like protein	160
Voltage-dependent calcium channel subunit alpha-2/delta-1	139
Ephrin type-A receptor 4	161
Contactin-2	156
Basement membrane-specific heparan sulfate proteoglycan core protein	159
Haptoglobin	182
Hornerin	159

Variation analysis metabolomics LC-MS

	Original samples				
concentration (uM)	Av (n=8)	SD (n=8)	RSD (%, n=8)	Analytical error (%)	pooled CSF sample
alanine	36.2	14.7	40	4	44.9
arginine	25.1	6.9	28	8	31.6
asparagine	8.7	3.0	34	4	10.6
citrulline	-	-	-	-	-
glutamine	-	-	-	-	-
glycine	10.3	3.2	31	5	12.0
histidine	21.7	6.9	32	9	27.1
iso-leucine	5.0	1.7	34	6	6.1
leucine	17.4	5.0	29	11	22.1
lysine	35.9	10.9	30	8	43.8
methionine	4.9	2.3	46	9	6.0
phenylalanine	12.3	5.0	41	8	17.0
proline	0.66	0.34	52	7	0.78
serine	30.3	9.1	30	9	36.2
threonine	36.3	11.6	32	4	44.5
tryptophan	2.6	1.2	44	9	3.4
tyrosine	12.0	5.1	42	7	16.0
valine	22.5	8.6	38	8	29.0

Variation analysis metabolomics LC-MS (Validation samples)

concentration (uM)	Av (n=27)	SD (n=27)	RSD (%, n=27)	Analytical error (%)
alanine	30.4	9.7	32	1
arginine	24.8	6.2	25	9
asparagine	6.7	1.5	22	2
citrulline	2.1	0.7	34	7
glutamine	537	77	14	7
glycine	6.3	1.9	30	5
histidine	17.2	2.8	16	9
iso-leucine	6.0	1.8	30	3
leucine	14.7	4.0	27	2
lysine	32.6	7.2	22	7
methionine	3.8	1.3	34	8
phenylalanine	9.0	2.1	23	8
proline	1.0	0.5	49	5
serine	27.0	4.5	17	3
threonine	30.4	8.1	27	2
tryptophan	2.3	0.6	24	9
tyrosine	11.2	3.3	30	6
valine	18.4	5.2	28	2

Variation analysis metabolomics LC-MS (Validation samples)

concentration (uM)	**MALE**			**FEMALE**		
	Av (n=13)	SD (n=13)	RSD (%,n=13)	AV (n=14)	SD (n=14)	RSD (%, n=14)
alanine	32.0	9.3	29	28.9	10.2	35
arginine	27.9	6.6	24	21.9	4.2	19
asparagine	7.1	0.9	13	6.4	1.9	29
citrulline	2.5	0.8	31	1.8	0.5	26
glutamine	565	74	13	511	73	14
glycine	6.7	2.2	32	5.8	1.5	25
histidine	18.1	2.8	16	16.3	2.5	15
iso-leucine	6.6	1.2	18	5.4	2.2	40
leucine	16.0	2.6	16	13.4	4.7	35
lysine	34.3	7.2	21	30.9	7.1	23
methionine	4.1	1.1	26	3.6	1.5	40
phenylalanine	9.3	1.8	20	8.7	2.3	26
proline	1.1	0.6	50	0.8	0.3	40
serine	25.7	3.8	15	28.1	4.9	17

threonine	31.4	6.5	21	29.5	9.4	32
tryptophan	2.4	0.6	26	2.2	0.4	21
tyrosine	11.7	2.7	23	10.7	3.9	36
valine	19.4	4.0	20	17.4	6.0	35

Variation analysis metabolomics LC-MS (Validation samples)

concentration (uM)	**AGE <35**			**AGE >35 and <50**			**AGE <50**		
	Av n=8	SD n=8	RSD %, n=8	Av n=10	SD n=10	RSD %, n=10	Av n=9	SD n=9	RSD %, n=9
alanine	24.9	5.0	20	31.1	9.6	31	34.5	11.5	33
arginine	23.3	7.5	32	27.2	5.4	20	23.5	5.5	23
asparagine	6.4	2.0	31	6.8	1.1	17	6.9	1.4	21
citrulline	1.9	0.8	41	2.2	0.5	23	2.2	0.9	41
glutamine	504	83	16	557	82	15	545	63	12
glycine	5.6	2.5	45	6.3	1.4	22	6.8	1.7	25
histidine	17.2	2.2	13	17.1	2.1	12	17.2	3.9	23
iso-leucine	5.9	1.8	30	5.9	1.6	28	6.2	2.2	36
leucine	13.5	4.0	30	15.2	3.3	22	15.1	4.8	32
lysine	32.9	7.0	21	32.8	9.0	27	31.9	5.9	18
methionine	3.4	1.5	43	3.8	1.0	25	4.2	1.4	34
phenylalanine	8.3	2.4	29	8.9	1.6	18	9.6	2.3	24
proline	0.9	0.3	33	0.9	0.6	62	1.0	0.5	51
serine	26.7	4.1	16	25.8	4.1	16	28.5	5.2	18
threonine	30.9	10.4	34	30.0	6.5	22	30.5	8.3	27
tryptophan	2.0	0.6	28	2.4	0.5	22	2.5	0.5	22
tyrosine	9.5	3.2	34	11.1	3.0	27	12.8	3.4	27
valine	16.8	5.0	30	18.2	3.9	21	19.9	6.5	33

Variation analysis metabolomics NMR (Original samples)

concentration (uM)	Av n=5	SD n=5	RSD %, n=5	Analytical error (%)	pooled CSF sample
2-aminobutyric acid	3.4	0.6	17	8	3.2
2-hydroxybutyric acid	21.1	3.4	16	4	19.4
2-hydroxyisovaleric acid	4.1	1.8	43	8	3.0
3-hydroxyisovaleric acid	32.6	13.8	42	4	40.2
3-hydroxybutyric acid	6.2	3.1	50	9	4.7
Acetic acid	37.2	12.4	33	6	79.0
Acetoacetic acid	5.1	1.4	27	8	7.0
Acetone	12.6	2.5	20	4	97.7
Alanine	21.8	4.1	19	6	25.6
Aconitic acid	19.6	9.0	46	9	28.4
Arginine	22.0	2.0	9	5	23.1
Choline	3.3	0.7	22	9	4.4
Citric acid	178	33	19	5	186.5
Creatine	32.8	4.2	13	6	33.7
Creatinine	42.7	5.2	12	5	49.5
Dimethylamine	2.3	0.3	15	8	3.1
Formic acid	18.7	1.5	8	5	21.7
Fructose	230	68	29	9	262.5
Galactitol	5.6	0.8	13	7	5.6
Glucose	1895	308	16	6	2404
Glutamine	428	86	20	4	424
Glycine	9.3	3.2	34	9	11.2
Histidine	8.6	2.0	23	5	9.8
Isoleucine	3.6	0.7	21	6	4.2
Lactic acid	848	100	12	4	915
Leucine	9.1	2.3	26	3	9.5
Lysine	13.0	4.8	37	9	12.8
Methanol	36.4	9.9	27	6	39.9
Methionine	4.4	1.6	37	7	4.4
1-methylhistidine	5.6	2.8	50	8	6.0
3-methylhistidine	5.2	1.3	25	7	4.5
Myo-Inositol	59.1	20.2	34	4	54.2
Phenylalanine	10.6	2.1	20	5	12.7
Pyruvic acid	46.3	7.4	16	7	50.9
Succinic acid	2.9	0.7	23	7	2.7
Trimethylamine-N-oxide	4.8	1.2	25	9	4.8
Threonine	21.5	3.6	17	6	25.1
Tyrosine	7.1	1.6	22	6	9.9
Urea	2758	1472	53	8	1975
Valine	11.9	2.1	18	9	14.4
Xanthine	7.8	2.7	35	8	7.6

Variation analysis metabolomics NMR (Validation samples)

concentration (uM)	Av n=27	SD n=27	RSD %, n=27	Analytical error %
2-aminobutyric acid	3.6	1.0	28	5
2-hydroxybutyric acid	21.0	6.1	29	4
2-hydroxyisovaleric acid	4.0	1.2	30	8
3-hydroxyisovaleric acid	11.5	1.7	15	2
3-hydroxybutyric acid	7.1	1.0	15	3
Acetic acid	77.0	39.7	52	2
Acetoacetic acid	4.9	1.3	26	4
Acetone	8.7	1.7	20	5
Alanine	29.7	8.0	27	3
Aconitic acid	22.6	6.2	28	6
Arginine	19.1	3.7	19	4
Choline	1.9	0.5	24	6
Citric acid	204	31	15	2
Creatine	38.5	5.9	15	3
Creatinine	56.8	9.5	17	2
Dimethylamine	1.9	0.5	25	6
Formic acid	37.9	7.3	19	5
Fructose	135	32	24	7
Galactitol	5.5	1.3	23	4
Glucose	2621	303	12	8
Glutamine	399	70	18	2
Glycine	7.5	1.8	24	2
Histidine	11.5	1.8	15	3
Isoleucine	5.1	1.3	25	4
Lactic acid	1127	155	14	2
Leucine	12.1	3.1	26	4
Lysine	18.2	3.0	16	2
Methanol	47.1	10.0	21	2
Methionine	3.0	0.9	31	6
1-methylhistidine	3.9	1.9	49	4
3-methylhistidine	1.0	1.5	143	8
Myo-Inositol	122	30	25	9
Phenylalanine	8.8	1.99	23	5
Pyruvic acid	62.6	14.4	23	5
Succinic acid	2.5	0.6	23	9
Trimethylamine-N-oxide	5.1	0.9	18	3
Threonine	26.5	5.3	20	4
Tyrosine	9.5	2.6	27	6
Urea	-	-	-	-
Valine	16.5	4.6	28	2
Xanthine	3.4	2.9	86	8

Variation analysis metabolomics NMR (Validation samples)

	MALE			FEMALE		
concentration (uM)	Av n=13	SD n=13	RSD %, n=13	AV n=14	SD n=14	RSD %, n=14
2-aminobutyric acid	3.8	1.3	34	3.3	0.5	16
2-hydroxybutyric acid	22.4	6.6	29	19.6	5.8	30
2-hydroxyisovaleric acid	3.7	0.8	22	4.3	1.5	34
3-hydroxyisovaleric acid	7.3	1.2	16	6.8	0.7	11
3-hydroxybutyric acid	11.7	1.1	9	11.3	2.2	20
Acetic acid	82.7	32.8	40	71.0	47.0	66
Acetoacetic acid	4.7	1.2	25	5.1	1.4	27
Acetone	9.0	1.7	19	8.3	1.7	20
Alanine	29.0	8.1	28	30.6	8.2	27
Aconitic acid	19.0	2.7	14	19.1	2.7	14
Arginine	2.0	0.5	24	1.8	0.4	24
Choline	24.7	6.7	27	20.3	4.9	24
Citric acid	206	37	18	201	25	13
Creatine	38.7	5.8	15	38.3	6.5	17
Creatinine	58.6	10.5	18	54.8	8.5	15
Dimethylamine	1.8	0.5	27	2.0	0.4	23
Formic acid	39.3	6.9	18	36.3	7.2	20
Fructose	128	28	22	142	35	25
Galactitol	5.7	1.3	22	5.2	1.2	23
Glucose	2487	265	11	2764	279	10
Glutamine	384	68	18	415	75	18
Glycine	7.9	2.2	27	7.0	1.1	16
Histidine	11.5	2.1	18	11.5	1.4	13
Isoleucine	5.1	1.6	30	5.0	1.0	19
Lactic acid	1161	186	16	1090	107	10
Leucine	11.8	2.5	22	12.3	1.6	13
Lysine	18.1	2.5	14	18.3	3.6	20
Methanol	47.0	12.7	27	47.1	5.5	12
Methionine	3.1	1.1	35	2.9	0.7	26
1-methylhistidine	3.3	2.2	66	4.5	1.4	31
3-methylhistidine	1.0	1.5	150	1.0	1.5	142
Myo-Inositol	125	31	25	117	28	24
Phenylalanine	9.4	1.4	15	8.3	2.3	28
Pyruvic acid	67.9	13.8	20	56.9	13.4	23
Succinic acid	2.4	0.6	25	2.6	0.6	23
Trimethylamine-N-oxide	27.0	6.0	22	25.9	4.4	17
Threonine	4.9	1.0	19	5.3	0.9	17
Tyrosine	8.3	1.5	17	10.7	3.0	28
Urea	-	-	-	-	-	-
Valine	16.1	5.7	36	17.1	3.3	19
Xanthine	2.4	3.1	130	4.4	2.3	53

Variation analysis metabolomics NMR (Validation samples)

concentration (uM)	**AGE <35**			**AGE >35 and <50**			**AGE <50**		
	Av n=8	SD (n=8)	RSD %, n=8	Av n=10	SD n=10	RSD %, n=10	Av n=9	SD n=9	RSD %, n=9
2-aminobutyric acid	3.9	1.0	26	3.1	0.6	19	3.7	1.3	34
2-hydroxybutyric acid	18.0	3.5	19	20.6	5.0	24	24.1	8.2	34
2-hydroxyisovaleric acid	4.5	1.8	39	4.0	0.8	21	3.5	0.7	22
3-hydroxyisovaleric acid	7.5	1.1	14	6.8	1.0	15	7.0	0.9	13
3-hydroxybutyric acid	12.4	1.3	10	11.0	2.3	21	11.2	1.0	9
Acetic acid	90.4	31.9	35	72.3	52.3	72	70.6	30.8	44
Acetoacetic acid	5.4	1.4	26	4.9	1.1	22	4.5	1.3	29
Acetone	8.8	1.9	21	8.4	1.7	21	8.9	1.7	19
Alanine	25.5	4.4	17	30.1	9.8	33	33.1	7.2	22
Aconitic acid	21.8	4.7	22	19.9	5.5	27	26.2	6.9	26
Arginine	19.3	2.9	15	19.0	2.9	15	18.9	2.5	13
Choline	1.8	0.4	21	1.7	0.4	22	2.2	0.5	21
Citric acid	194	21	11	199	18	9	218	45	21
Creatine	36.4	4.4	12	39.3	6.8	17	39.4	6.6	17
Creatinine	53.1	8.2	15	55.9	9.0	16	61.0	10.7	17
Dimethylamine	1.9	0.4	19	1.9	0.5	28	1.9	0.5	28
Formic acid	38.6	5.8	15	37.4	7.7	21	37.7	8.2	22
Fructose	136	37	27	127	26	20	142	35	24
Galactitol	5.2	1.0	19	5.0	1.0	20	6.3	1.4	22
Glucose	2571	482	19	2636	206	8	2648	201	8
Glutamine	375	42	11	412	84	20	406	80	20
Glycine	7.4	2.1	28	7.3	1.9	25	7.7	1.5	20
Histidine	12.3	1.0	8	11.0	1.2	11	11.3	2.6	23
Isoleucine	4.6	0.6	13	4.8	0.8	16	5.9	1.8	31
Lactic acid	1038	76	7	1139	176	15	1193	157	13
Leucine	12.3	2.0	16	11.9	2.1	17	12.0	2.5	21
Lysine	19.1	2.3	12	17.5	3.5	20	18.1	3.1	17
Methanol	48.6	12.5	26	46.0	8.0	17	46.8	9.8	21
Methionine	3.5	0.8	22	2.5	0.6	24	3.0	1.2	39
1-methylhistidine	3.9	1.1	29	3.7	1.8	49	4.1	2.6	64
3-methylhistidine	0.6	1.1	194	1.3	1.5	115	1.1	1.7	158

Myo-Inositol	106	26	24	132	27	20	124	31	25
Phenylalanine	8.7	1.3	15	9.5	1.4	15	8.2	2.8	34
Pyruvic acid	55.5	12.0	22	59.4	13.1	22	72.5	13.6	19
Succinic acid	2.4	0.5	20	2.4	0.6	26	2.7	0.6	23
Trimethylamine -N-oxide	5.1	1.0	19	5.1	1.0	19	5.2	1.0	19
Threonine	25.1	5.4	22	26.4	5.0	19	27.9	5.5	20
Tyrosine	9.1	2.5	27	10.1	3.2	32	9.2	2.1	23
Urea	-	-	-	-	-	-	-	-	-
Valine	15.6	3.9	25	15.6	4.1	26	18.4	5.7	31
Xanthine	3.0	2.3	77	3.8	2.8	73	3.1	3.6	116

Biological variation in the validation CSF sample set vs experimental CSF sample set with GC-MS

	Validation sample set	**Experimental sample set**
Name	RSD (%, n=28)	RSD (%, n=420
1,5-Anhydro-D-Glucitol	7	31
Arabitol	12	
alanine	12	32
leucine	13	32
iso-leucine	13	32
2-Hydroxybutanoic acid	14	88
3,4-Dihydroxybutanoic acid	14	
3-Hydroxypropanoic acid	16	43
Aspartic acid	16	58
Aminomalonic acid	17	30
Benzoic acid	17	29
Myo-inositol	18	38
Sucrose	19	53
2-Hydroxypropenoic acid	19	36
Asparagine	19	36
Glucose	19	41
Urea	21	
Glutamic acid	21	36
Tryptophan	22	
Arabinose	22	36
Hypoxanthine	22	116
2-Aminobutyric acid	23	28
Creatinine	24	56
Proline	25	94
Tyrosine	26	

3-hydroxyisovaleric acid	26	43
Glycerol	26	34
Acetoacetic acid	26	42
Glycerol-galactopyranoside	26	
C18:1 fatty acid	28	33
Cysteine	28	89
Inositol	28	36
Erythronic acid	28	
Quinic acid	28	36
sn-Glycerol-3-Phosphate	30	45
5,6-dihydrouracil	30	44
C18:0 Fatty acid	30	40
Cholesterol	30	52
phenylalanine	31	
C16:0 Fatty acid	31	32
Gluconic acid	33	38
2-Oxo-butanoic acid	33	56
Ribitol	34	64
1-Monostearoylglycerol	35	28
Glyceric acid	35	72
C14:0 Fatty acid	36	31
Mannitol	36	36
Serine	38	
2,3-dihydroxybutanoic acid	38	39
Fructose	38	104
2-hydroxyisovaleric acid	43	34
Methionine	45	47
Xylonic acid	45	
3-Hydroxyhexanoic acid	46	42
Uric acid	47	
Fucose	48	67
2,4-Dihydroxybutanoic acid	54	129
N-Acetylaminomalonic acid	57	47
Phosphoric acid	58	
Amino-butyric acid isomer	63	29
Valine	64	
Threonic acid	68	71
Myo-inositol-1,2-cyclicphosphate	69	63
Xylose	82	-
Uridine	90	-

Phosphorylethanolamine	106	106
Threonine	113	
Inositol related compound	214	120

Concentrations comparison

	Validation sample set		Wishart et al. (13)		Literature values referred to by Wishart et al. (13)	
concentration (uM)	Av (n=27)	SD (n=27)	Av (n=35)	SD (n=35)	Av	SD
1-methylhistidine	3.9	1.9	ND	ND	ND	ND
2-aminobutyric acid	3.6	1.0	ND	ND	ND	ND
2-hydroxybutyric acid	21.0	6.1	40	24	35	24
2-hydroxyisovaleric acid	4.0	1.2	8	6	7	7
3-hydroxybutyric acid	7.1	1.0	34	31	46	24
3-hydroxyisovaleric acid	11.5	1.7	4	2	ND	ND
3-methylhistidine	1.0	1.5	ND	ND	ND	ND
Acetic acid	77.0	39.7	58	27	100	30
Acetoacetic acid	4.9	1.3	12	14	6	6
Acetone	8.7	1.7	20	21	67	24
Aconitic acid	22.6	6.2	ND	ND	ND	ND
Alanine	29.7	8.0	46	27	37	7
Arginine	19.1	3.7	ND	ND	ND	ND
Choline	1.9	0.5	3	1	8	5
Citric acid	204	31	225	96	176	50
Creatine	38.5	5.9	44	13	ND	ND
Creatinine	56.8	9.5	43	12	65	25
Dimethylamine	1.9	0.5	2	1	ND	ND
Formic acid	37.9	7.3	32	16	ND	ND
Fructose	135	32	160	91	240	20
Galactitol	5.5	1.3	ND	ND	ND	ND
Glucose	2621	303	2960	1110	5390	1650
Glutamine	399	70	432	204	444	80
Glycine	7.5	1.8	ND	ND	ND	ND
Histidine	11.5	1.8	14	8	12	2
Isoleucine	5.1	1.3	7	5	8	3
Lactic acid	1127	155	1651	626	1590	330
Leucine	12.1	3.1	16	9	19	4
Lysine	18.2	3.0	29	13	28	8

Methanol	47.1	10.0	44	36	ND	ND
Methionine	3.0	0.9	5	4	6	3
Myo-Inositol	122	30	84	40	133	20
Phenylalanine	8.8	1.99	15	13	18	7
Pyruvic acid	62.6	14.4	53	42	71	30
Succinic acid	2.5	0.6	3	2	29	5
Threonine	26.5	5.3	30	12	28	5
Trimethylamine-N-oxide	5.1	0.9	ND	ND	ND	ND
Tyrosine	9.5	2.6	12	9	10	4
Valine	16.5	4.6	19	13	24	7
Xanthine	3.4	2.9	13	7	5	1

Av Average
SD Standard deviation
RSD Relative standard deviation

SUPPLEMENTARY MATERIAL CHAPTER 4

SUPPLEMENTARY MATERIAL CHAPTER 4

List of metabolites identified in rat CSF by ^{1}H-NMR

Metabolites		
3-Hydroxyisovalerate	Arginine	Creatine
Acetate	Butyrate	Creatinine
Acetoacetate	Choline	Dimethylamine
Acetone	Cis-aconitate	Formate
Alanine	Citrate	Fructose
Glucose	Methanol	Pyruvate
Glutamine	Methylmalonate	Succinate
Glycerol	N-acetylaspartate	Threonine
Lactate	Ornithine	Valine
Lysine	Pantothenate	myo-Inositol
Malonate	Propylene glycol	N-acetylcompounds

Sample preparation and data acquisition of CSF samples from the second EAE experiment

10µL of rat CSF were thawed at room temperature and 210 µL D_2O (99.96 at.%D) were added to the biofluid. TSP-d_4 (Sodium 3-(trimethylsilyl)propionate-2,2,3,3-d4) (99 at.%D) was used as internal standard for chemical shift reference (δ 0.00 ppm). For the latter, 70µL of buffer solution was added to 220 µL of rat CSF. The buffer solution solvated in a mixture of water and D_2O consists of 2,85mM TSP, 6.92 mM Sodium azide (NaN3) and 42.08 mM sodium phosphate dibasic dehydrate (Na2HPO4•2H2O). The addition of buffer solution to 220 µL of CSF sample leads to a final concentration of 0.66mM TSP. The pH of the CSF was adjusted to around 7 (7.0 – 7.1) by phosphate buffer in buffer solution. The final CSF NMR sample (290 µL) was then transferred to a SHIGEMI microcell tube for measurements.

The 1D ^{1}H NMR spectra of rat CSF samples were acquired on a 600 MHz Bruker with a 5 mm CPTCI z-gradient cold probe. Suppression of water was achieved by using

presaturation. For each 1D ^{1}H NMR spectrum 512 scans were accumulated with a spectral width of 7000 Hz resulting in a total of 16K points. The acquisition time for each scan was 2.2s. Between scans, a 8s relaxation delay was employed. Prior to spectral analysis, all acquired Free Induction Decays (FIDs) were zero-filled to 32K data points, multiplied with a 0.3 Hz line broadening function, Fourier transformed, manually phased, the TSP internal reference peak was set to 0 ppm and baseline corrected by using ACD/SpecManager software version 12.02. All 44 rat CSF spectra were acquired and preprocessed as described above and subsequently transferred to Chenomx NMR Suite 7.0 for metabolites quantification.

Table 1S. Groups description of the second EAE experiment; "n" indicates number of rats (samples) for each group.

Treatment / Day 0	Group description	Day 10
CFA	Peripheral inflammation	P10-2* n=15
CFA+MBP	Neuroinflammation + peripheral inflammation	N10-2** n=15

*2 sample were discarded due to blood contamination

**4 samples were discarded due to blood contamination

PLS-DA of ^{1}H-NMR CSF data

C10 vs. P10: Effect of peripheral inflammation

Figure S1 presents the score plot of the PLS-DA model for groups "C10" versus "P10". This model enables us to examine the effect of peripheral inflammation on the metabolic profile of CSF. This PLS-DA model is also performing well, since the overall correct classification for the test set is 100% (Table S2).

P10 vs. P14: Evolution of peripheral inflammation

By studying groups “P10” and “P14” the evolution of peripheral inflammation may be investigated. We concluded from the “C14” vs. “P14” model that group “P14” was not inflamed anymore. The PLS-DA model of “P10” vs. “P14” has perfect prediction ability. The key metabolites contributing to discrimination are mostly glutamine and citrate. This is indeed in agreement with the PLS-DA model of groups “C10” vs. “P10”, where the effect of peripheral inflammation was examined.

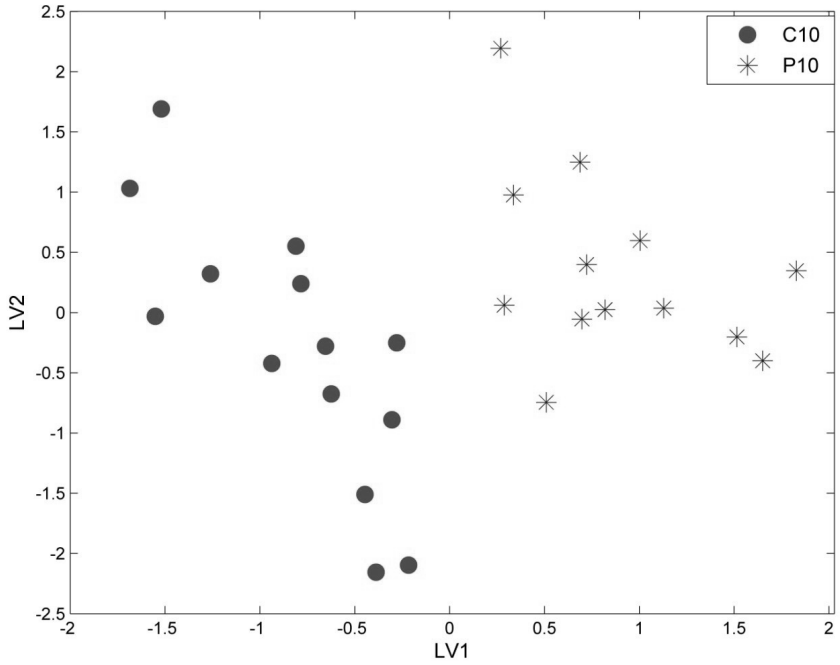


Figure 1S. PLS-DA score plots derived from ^{1}H-NMR spectra of rat CSF belonging to groups “C10” and “P10”. The amount of explained variance in Y for two latent variables was equal to 88.5%.

<u>P10 vs. P14: Evolution of peripheral inflammation</u>

By studying groups “P10” and “P14” the evolution of peripheral inflammation may be investigated. We concluded from the “C14” vs. “P14” model that group “P14” was not inflamed anymore. The PLS-DA model of “P10” vs. “P14” has perfect prediction ability.

The key metabolites contributing to discrimination are mostly glutamine and citrate. This is indeed in agreement with the PLS-DA model of groups “C10” vs. “P10”, where the effect of peripheral inflammation was examined.

Table 2S. Summary of multivariate modeling diagnostic accuracy (PLS-DA) for the test set.

Model	Group	Sensitivity [%]	Specificity [%]	Overall correct classification [%]	Number of variables in PLS-DA model
C10 vs. N10	C10	100	100	100	45
	N10	100	100		
C10 vs. P10	C10	100	100	100	25
	P10	100	100		
P10 vs. N10	P10	100	100	62.5	20
	N10	58	33		
N10 vs. N14	N10	100	100	100	45
	N14	100	100		
C14 vs. N14	C14	100	100	100	40
	N14	100	100		
P10 vs. N14	P10	100	100	100	25
	N14	100	100		

Table 3S. Key metabolites discriminating groups based on PLS-DA models.

PLS-DA model	Key metabolites
C10 vs. N10	fructose, glutamine, 3-hydroxyisovalerate, lysine, N-acetylaspartate
C10 vs. P10	citrate, glutamine, N-acetyl-compound and resonance at δ0.8983 which probably belongs to butyrate
C14 vs. N14	Lysine, arginine, alanine, pantothenate, malonate and unknown resonance at δ1. 18 (variable 126)

N10 vs. N14	arginine, dimethylamine, lysine, creatinie, pantothenate and unknown resonances at δ1.50 and δ1.18 (variable 126).
N10 vs. (C14&P14)	lysine, arginine, pantothenate, malonate
P10 vs. N14	lysine, pyruvate, choline, arginine, creatine, N-acetylaspartate

Table 4S. Particular effect observed based on PLS-DA results.

Groups	**Effect**	**Comments**		Number of variables in PLS-DA model
C10 vs. N10	Peripheral Inflammation& "neuroinflammation"	Neuroinflammation was only present in some animals	DAY 10	
C10 vs. P10	Peripheral inflammation			
P10 vs. N10	"Neuroinflammation" & peripheral inflammation			
C14 vs. N14	Neuroinflammation	Probably only the effect neuroinflammation is observed at Day 14 and no effect peripheral inflammation is anymore present	DAY 14	
C14 vs. P14	Lack of peripheral inflammation; effect observed unrelated to peripheral inflammation			
N14 vs. (C14&P14)	Neuroinflammation			
P14 vs. N14	Neuroinflammation			
P10 vs. P14	Peripheral inflammation		DAY 10 & DAY 14	
N10 vs. N14	Evolution of disease			
P10 vs. N14	Neuroinflammation			

ANOVA-PCA results obtained at the onset of disease

The first principal component (PC1), in Figure S2a, indicates variations, which distinguish between non-treated ("C10") and the EAE affected group ("N10"). The second PC accounts for the discrimination between group "P10" and groups C10&N10. By visual inspection of Figure 4a group specific metabolites are found. As an example, lactate (variable 116) and N-acetylaspartate (variable 90), have a high, positive interaction with group "N10, while citrate (variable 70) and glutamine (variable 72) interact negatively with this group.

ANOVA-PCA results obtained for the peak of disease

The first principal component of Figure S2b shows the variation, which differentiates the EAE-affected group at the second time point (group "N14") from the other two groups. The second principal component represents the variation distinguishing groups "C14" and "P14". However, the amount of variance explained by the second PC is low (12.3%), indicating that these groups are quite similar.

ANOVA-PCA results obtained for all diagnostic groups

Similarly we applied ANOVA-PCA to the metabolic profiles of all groups. The results are presented in Figures S2c and S2d. The first principal component of Figure S2c differentiates the group "N14" from other groups. It is also visible that the"N10" is different from the healthy group "C14" and group "P14". Furthermore, groups "C14" and "P14" are highly collinear, which confirms their similarity in metabolic profiles.

The score plot of the second and third principal component (Figure S2d) represents the distinction between the healthy group "C10" and group "N10". By inspection of the reciprocal location of groups "C14" and "P14" on the plane of PC2 and PC3 similar conclusions can be drawn. These groups show collinearity in the metabolic profiles.

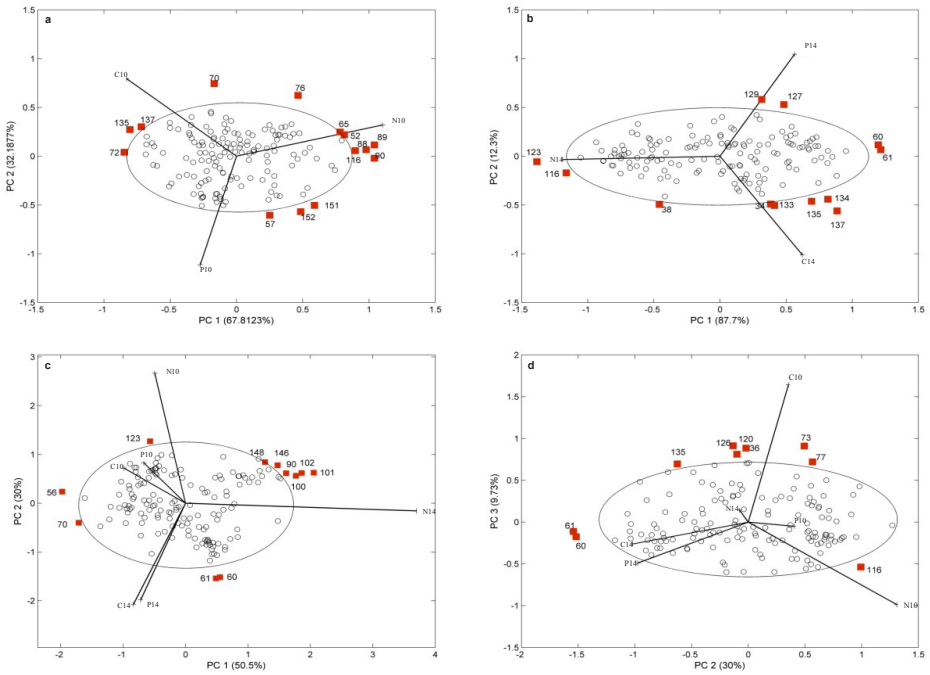


Figure 2S. Biplots of principal component analysis performed on the interaction between metabolites and treatments. Metabolites which are statistically significant are indicated with red squares. The arrows indicate the loadings of the treatments: **(a)** groups “C10”, “P10” and “N10”; **(b)** groups “C14”, “P14” and “N14”; **(c)** all groups plotted along two first principal components (PC1 and PC2); **(d)** all groups plotted along PC2 and PC3.

Table 5S. Group specific metabolites selected by mean of ANOVA-PCA.

Group	**Metabolites with significant interactions**	
C10 healthy Day 10	glutamine (variable 72)	citrate (variable 70)
	variables 135 and 137	threonine (variable 36)
	3-hydroxyisovalerate (variable 120)	succinate (variable 73)
	variable 77	variable 126
	variables 151 and 152 (probably butyrate)	
C14 healthy Day 14	myo-inisitol (variable 34)	propylene glycol (variables 133, 134)
	citrate (variable 70)	ornithine (variables 60 and 61)
	variables 135	variables 137
P10 peripheral inflammation Day 10	cis-aconitate (variable 57)	pyruvate (variable 76)
P14 peripheral inflammation Day 14	glycerol (variable 38)	ornithine (variables 60 and 61)
	variable 127	variable 129
N14 neuroinflammation Day 14	lactate (variable 116)	N-acetylaspartate (variable 90)
	N-acetyl-compounds (variables 88 and 89)	methylmalonate (variable 123)
	pyruvate (variable 76),	lysine (variable 65)
	arginine (variables 100, 101 and 102)	pantothenate (variables 146 and 148)
N10 neuroinflammation Day 10	citrate (variable 70)	malonate (variable 56)
	Methylmalonate (variable 123)	N-acetylaspartate (variable 90)
	lysine (variable 65)	arginine (variables 100, 101 and 102)
	pantothenate (variables 146 and 148)	

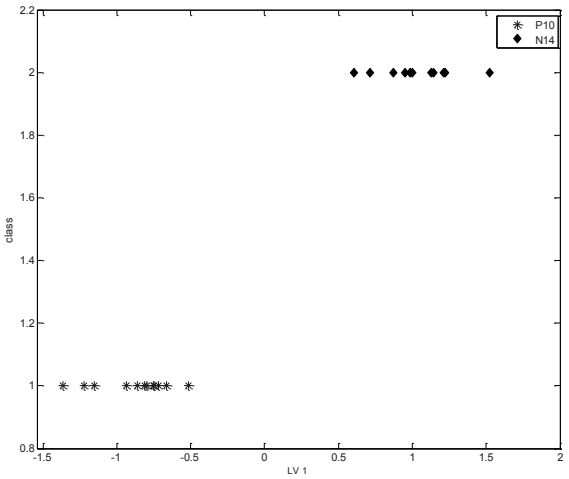


Figure 3S. PLS-DA score plots derived from absolute concentration of metabolites specific for neuroinflammation for groups “P10” and “N14”. The amount of Y explained variance for one latent variable was equal to 91.3%.

SUPPLEMENTARY MATERIAL CHAPTER 5

SUPPLEMENTARY MATERIAL CHAPTER 5

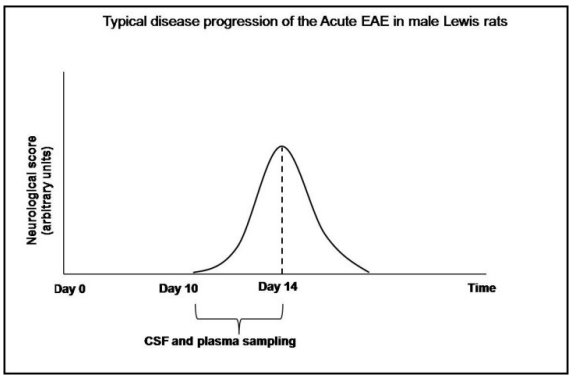


Figure 1S. Typical disease progression in EAE model.

Simulated data

In order to represent the use of Hierarchical Model Fusion (HMF) a simulated data set was constructed. This data set contains 280 samples (corresponding to four classes, i.e. 70 samples per class) and 105 variables. It was simulated in such way that the samples are represented by random vectors chosen from multivariate normal distribution. 5 variables are informative and 100 are irrelevant for the discriminating classes. The informative traits were incorporated with random noise (approximately 6% of informative signal). The data were simulated in such way that: classes 1 and 2 overlap with each other as well as classes 3 and 4, while classes 2, 3 and 4 are the most dissimilar one. In Figure 2S the scatters plot of the simulated data in the plane of variable 5 and 2 (both informative) and Principal Component Analysis (PCA) score plot are shown. As can be observed from Figure 2Sa these two variables allow for discriminating classes. On the contrary, the PCA score plot (of autoscaled data, see Figure 2Sb) does not reveal any groupings. Indeed, the four classes completely overlap in the plane defined by PC1 and PC2.

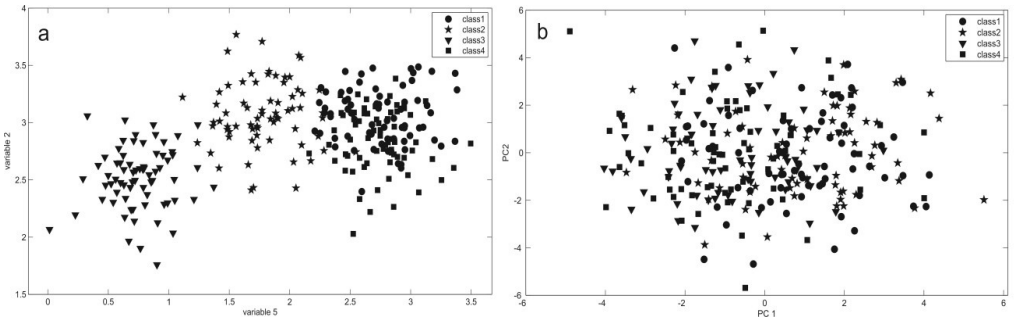


Figure 2S. Visualization of simulated data: (a) the scatter plot in the plane defined by the informative variables 5 and 2; (b) PCA score plot in the plane of PC1 and PC2.

The data were divided into model (52 samples per class) and independent test set (18 samples per class) using Duplex algorithm. In order to perform HMF three PLS-DA models were constructed, namely class 1 vs. class 2, class 2 vs. class 3 and class 3 vs. class 4 for autoscaled data. They were selected based on the information how the classes were simulated. The correct classification rate for the independent test set was equal 94.4%, 100% and 94.4%, respectively for the PLS-DA models.

These three PLS-DA models were next used to obtain three new scores following the HMF procedure. We started with PLS-DA model of class 1 vs. class 2. This step enables of creating Xscore. This allows one to separate class 1 from the rest. In the next step PLS-DA model of class 2 vs. class 3 was used. Similarly to the first step a second score is generated, i.e. Yscore. At this point class 2 is separated. In the final step PLS-DA model of class 3 vs. 4 was utilized leading to Zscore. By concatenating these three new scores a graphical representation is obtained, showing all classes at once. In Figure 3S the graphical representation of the HMF applied to simulated data is presented. Note that this figure shows where the test samples (18 samples per class) are projected into these new scores. As can be noticed the test samples are correctly projected (95.3% test samples are correctly predicted by HMF). This demonstrates the statistical relevance of the obtained results. To show if classification results are better than any other random classification a permutation test was made. Of 3000 runs none had correct classification bigger than 95.03%, leading to p-value 0.0003.

To evaluate the performance of HMF we constructed PLS2-DA model on the same training set used for HMF. The correct classification for the independent test set was equal to 83.3%. This result shows that PLS2-DA underperforms compared to HMF.

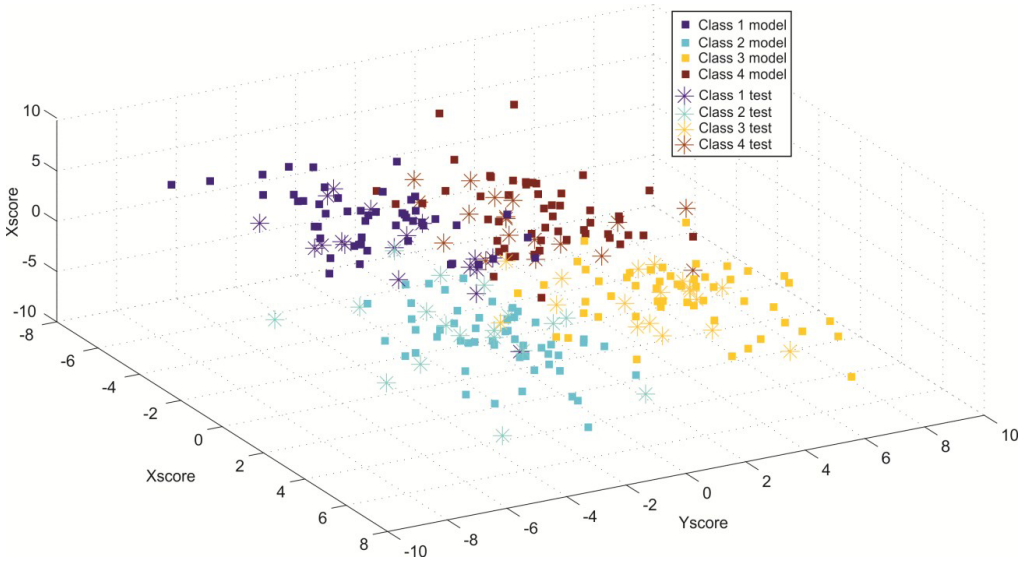


Figure 3S. Graphical representation of the results of the HMF obtained on the simulated data.

Addition of different noise levels

In order to represent the effect of noise we added different levels of homoscedastic and heteroscedastic noise to informative variables in the simulated data. We created 6 different levels and we repeated it 60 times. Every noise level comprised different percentage of informative traits (see Figure 4S on x-axis). 6 noise levels and 60 repetition leads to 360 runs. For each run PLS-DA models were recalculated and HMF was applied as described above. Moreover we performed permutation test (3000 runs) to reduce the possibility of random classification.

In Figure 4S the average correct classification and standard deviation obtained for HMF applied to simulated data with different levels of noise in informative traits is presented. As can be observed the higher the noise level the less accurate results have become. The algorithm performs well in terms of correct predictions for independent test set up to situation number 4 (i.e. data containing 35% of noise in comparison to informative traits). However, for data containing 35% of noise the results become unstable, since the outcome of permutation test shows that for 3000 runs some had higher correct classification than for the original classification. For the first three noise levels, of the 3000 permutations none had a number of correct classifications higher than for original classification, leading to a p-value of 0.0003.

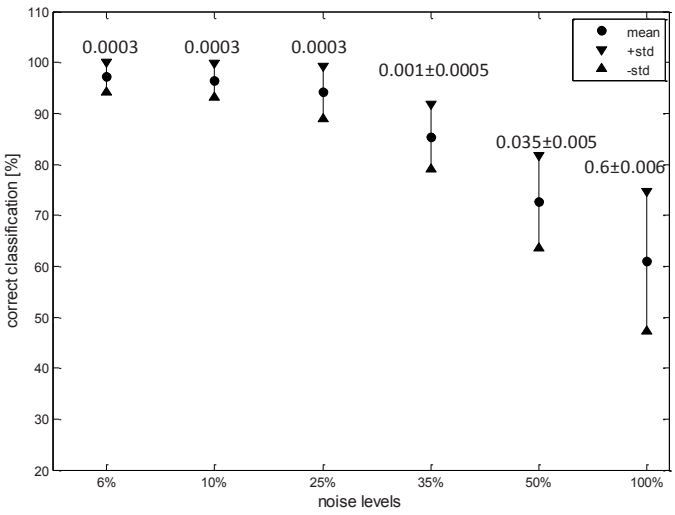


Figure 4S. Average correct classification rate and standard deviation obtained for simulated data with 6 different noise levels. In the x-axis amount of noise with respect to relevant variables is indicated. For each noise level the results from permutation test are included on top of the triangular.

Pseudo code for LOO CV used for variable selection and model optimization

Assuming a data matrix **X** for training set of size (n x p, where n is a number of samples and p a number of variables) a LOO CV is performed (separately for plasma and CSF) according to the following scheme, which was repeated for each sample (i) in **X**:

Beginning of LOO CV
*1. One object (i) is removed from training data matrix **X***
*2. Autoscaling of the data matrix **X** with remaining objects (of size m-1 x p)*
*3. RFE is performed on the autoscaled data matrix **X** with remaining objects (of size m-1 x p)*
4. A ranking of variables is obtained
End of LOO CV

In next step a final ranking for variables is obtained.
CSF and plasma training sets are concatenated and variables ranking is again performed with LOO CV; note that here **X** corresponds to concatenated plasma and CSF data:

Beginning of LOO CV
*1. One object (i) is removed from training data matrix **X***
*2. Autoscaling of the data matrix **X** with remaining objects (of size m-1 x p)*
*3. RFE is performed on the autoscaled data matrix **X** with remaining objects (of size m-1 x p)*
4. A ranking of variables is obtained
End of LOO CV

The final ranking for concatenated data is obtained.

Optimization of PLS-DA model complexity (number of latent variables) for data matrix **X** (repeated for each object (i) in **X**). Note that here **X** can correspond to CSF data, plasma data or fused data sets:

Beginning of LOO CV

*1. One object (i) is removed from training data matrix **X** (m x p)*

*2. Autoscaling of the data matrix **X** with remaining objects (of size m-1 x p)*

*3. Autoscaling of removed object (i) with mean and standard deviation delivered from the data matrix **X** with remaining objects*

*4. Fit PLS-DA model to autoscaled data matrix **X** with remaining objects (size m-1 x p)*

5. Classify removed object (i)

6. Calculation of the root mean square error of cross-validation (RMSECV).

End of LOO CV

Model complexity= min(RMSECV) -> first minimum

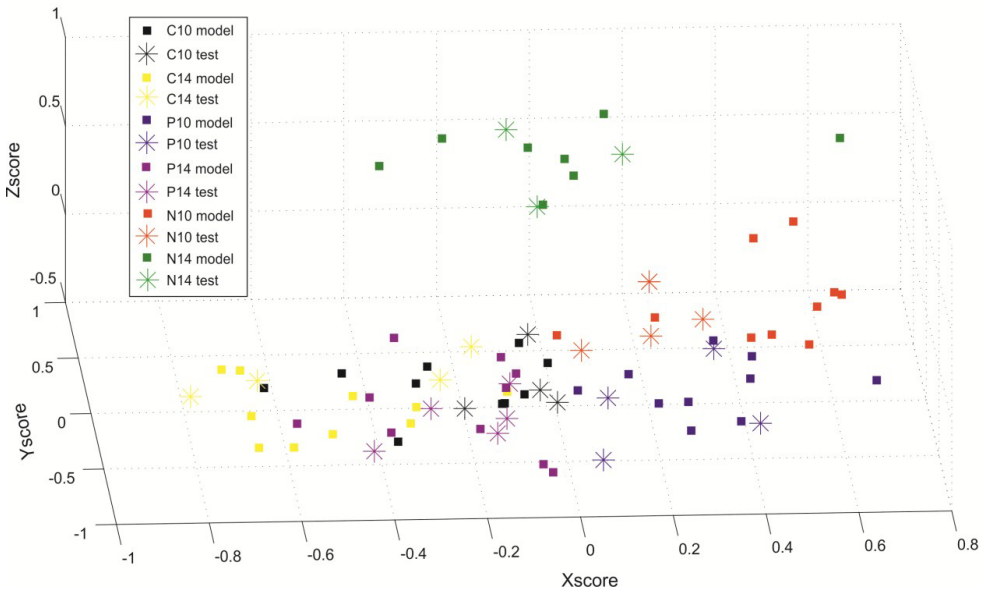


Figure 5S. Graphical representation of the results of the HMF obtained on the fused CSF and plasma datasets with indication of training and test samples.

Table 1S. Correct classification rate for independent test set obtained for individual analysis of plasma data, CSF data and fused datasets by PLS2-DA. Complexity of PLS2-DA models LV=3. Number of variables: 20 variables for plasma and 20 variables for CSF.

Groups	plasma	CSF	fused (plasma and CSF)
C10	50%	67%	80%
P10	100%	67%	80%
N10	0%	0%	67%
C14	0%	0%	0%
P14	80%	100%	67%
N14	100%	100%	100%

Table 2S. Correct classification rate for independent test set obtained for individual analysis of plasma data, CSF data and fused datasets by PLS-DA.

PLS-DA model	plasma	CSF	fused (plasma and CSF)
C10 vs. P10	62.50%	62.50%	93%
P10 vs. N10	75%	50%	100%
C10 vs. N10	62.5%	100%	100%
N10 vs. N14	62.5%	100%	100%
C14 vs. N14	100%	100%	100%
C14 vs. P14	100%	100%	100%

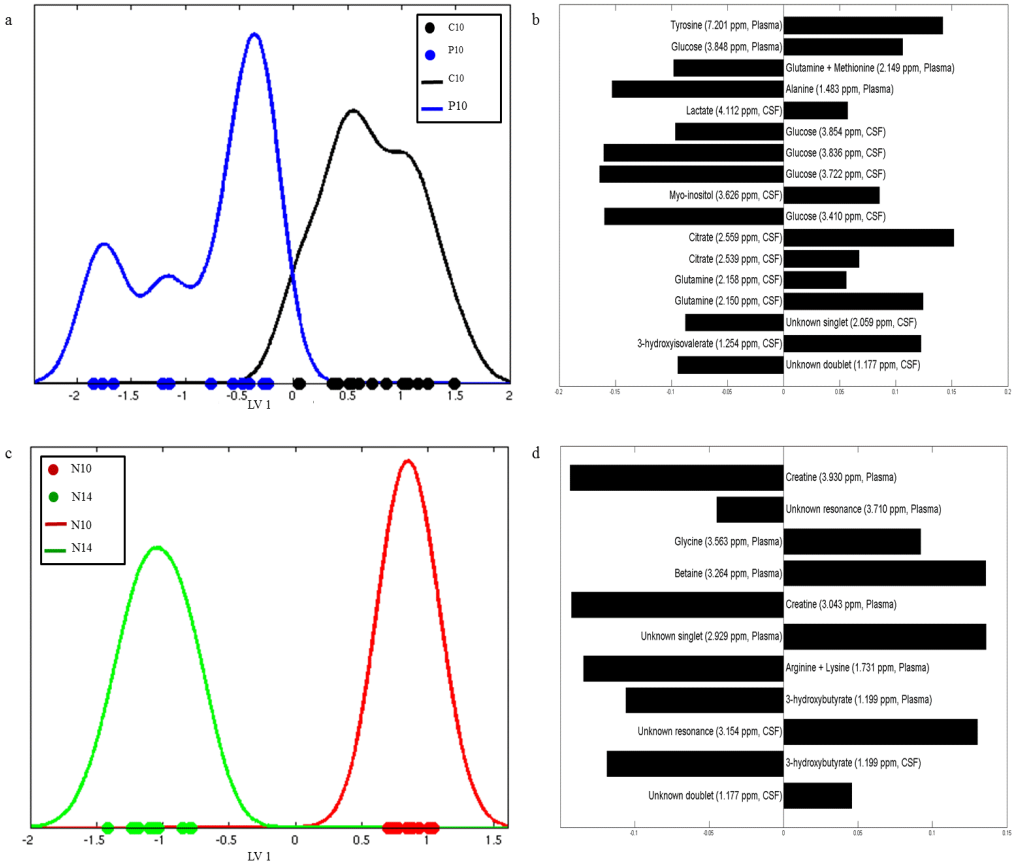


Figure 5S. Density distribution of PLS-DA scores of fused data: (a) “C10” vs. “P10”, the amount of y variance for 1 LV is equal 76.2%; (b) Regression coefficients of “C10” vs. “P10” PLS-DA model; (c) “N10” vs. “N14”, the amount of y variance for 1 LV is equal 94.04%; (d) Regression coefficients of “N10” vs. “N14” PLS-DA model.

SUPPLEMENTARY MATERIAL CHAPTER 6

SUPPLEMENTARY MATERIAL CHAPTER 6

Table 1S. Metabolites

Variable nr.	Metabolite	Variable nr.	Metabolite
1	citrate	39	leucine
2	citrate	40	leucine
3	glutamine	41	2-hydroxy-3-methylvalerate
4	glutamine	42	unknown
5	creatinine	43	unknown
6	creatine	44	2-hydroxybutyrate
7	lysine	45	2-hydroxybutyrate
8	lysine	46	2-methyl-2-oxovalerate
9	alanine	47	unknown
10	arginine	48	1,5-anhydroglucitol
11	choline	49	2,3-butanediol
12	glucose+glyceric acid+ascorbate	50	3-methyl-2-hydroxybutanoic acid
13	glucose+glyceric acid+ascorbate	51	Alanine
14	glucose+unknown	52	Arabinose
15	glucose+unknown	53	C16:0 fatty acid
16	glucose +unknown	54	Citric acid
17	glucose +unknown	55	Glucose
18	glucose +unknown	56	Glutamine
19	glucose+trimethylamine N-oxide	57	Glycerol
20	glucose+carnitine	58	Lactic acid
21	glucose+carnitine	59	Lysine
22	glucose+phosphocholine	60	Mannose
23	acetone	61	Myo-inositol
24	acetate	62	Ornithine
25	lactate	63	Phenylalanine
26	lactate	64	Phosphate
27	3-hydroxyisovaleriate	65	Pyruvic acid
28	2-methyl-2-oxovalerate	66	Ribitol or arabitol
29	unknown	67	Sucrose
30	2-oxobutyrate	68	Threonine
31	2-oxobutyrate	69	ascorbic acid derv 1
32	valine	70	butanediol isomer 2
33	2-oxobutyrate	71	erythronic acid
34	valine	72	fructose
35	unknown	73	inositol
36	valine	74	meso-erythritol
37	valine	75	sn-Glycerol-3-Phosphate
38	leucine	76	urea

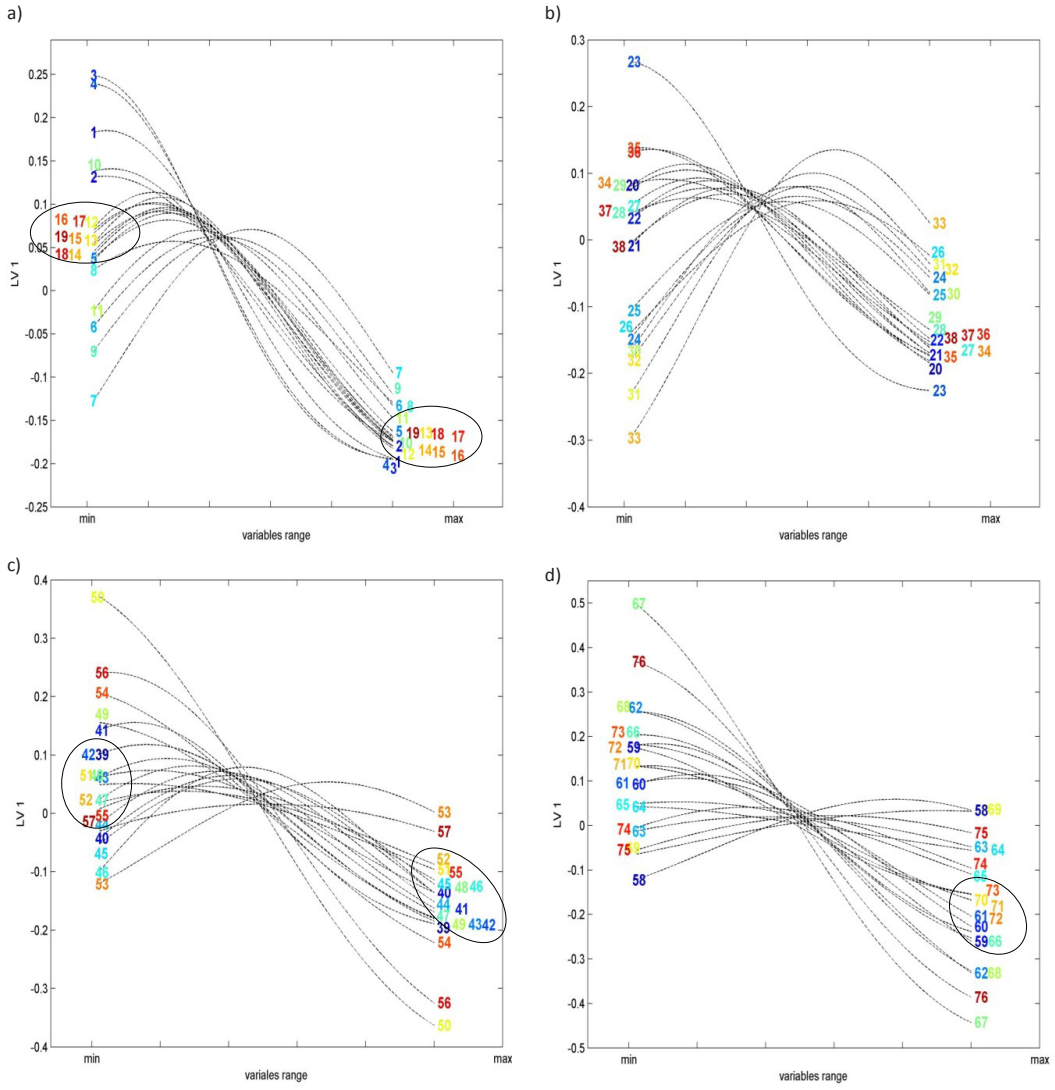


Figure 1S. Loading plot of pseudo samples trajectories for: (a) variables 1 till 19; (b) variables 20 till 36; (c) variables 37 till 57 and (d) variables 58 till 76. Numbers correspond to variable numbers in Table 1S.

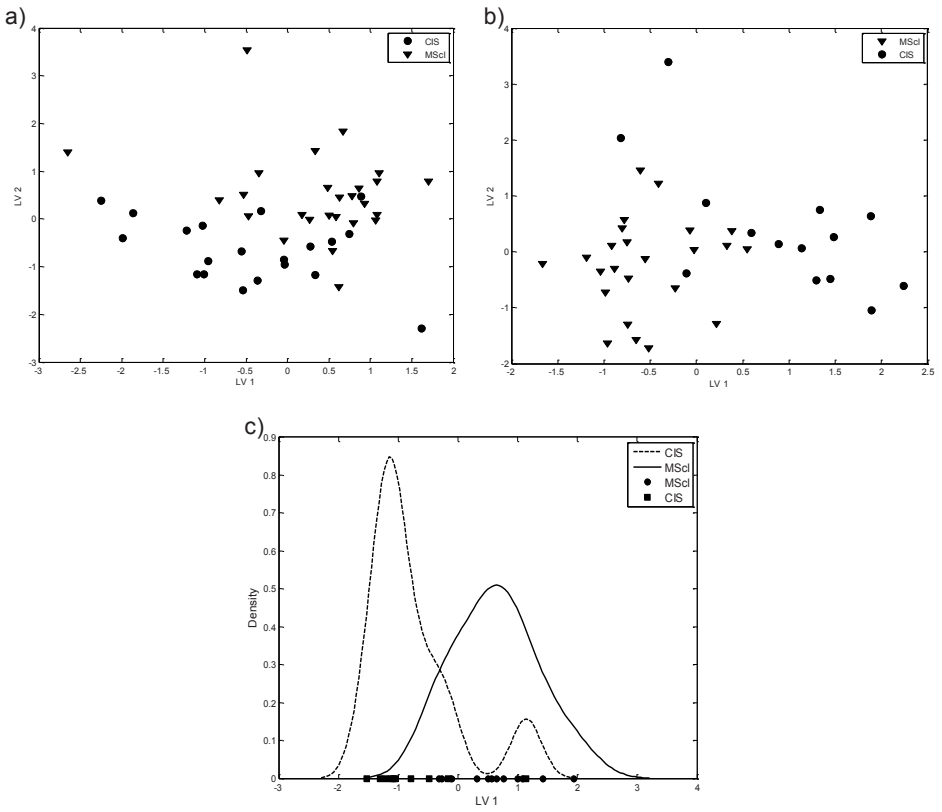


Figure 2S. The PLS-DA score plot of: (a)NMRdata; (b)GCMS data; (c) fused NMR and GC-MS in mid-level fashion.

Clinical information

Of the 26 patients in the MScl group in the NMR dataset, 19 patients had relapsing remitting (RR) MScl and 7 primary progressive (PP) MScl. In the GC-MS dataset of MScl patients, 7 patients were diagnosed with PP MScl and the others had PP MScl. The number of patients with PP MScl in the overlap NMR/GC-MS set is equal to 4. In the NMR dataset, the group of the MScl patients contains 6 males and 20 females, while in the GC-MS MScl dataset 5 males and 19 females are found. In the NMR/GC-MS

overlap set (MScl) 4 patients are male. In the MScl NMR/GC-MS overlap set the median of the time of disease duration for the MScl patients at the moment of CSF sampling was 4.5 years, while the average was 7.25 years with standard deviation of 6.6 years. For the complete NMR MScl set and GC-MS MScl set the disease duration was on average 6 years with standard deviation of 5 years and 7.3 years with standard deviation of 5.7 years, respectively. In the group of 20 patients with CIS, 5 are males and 15 are females in the NMR set. In the GC-MS dataset 4 patients are male. It is worthwhile to mention that all patients diagnosed with CIS have later developed MScl.

Kernel transformations

Different kernels transformations, namely linear and polynomial (2^{nd} and 3^{rd} degree), were studied. However the correct classification did not extent 60% for independent test set. Moreover, these kernels transformation presented the lowest RMSECV for training set.

Random division in training and independent test set

Random division (repeated 10^4 times) of the data was made and the MKL procedure was performed for each pair of training and independent test set. The average correct prediction of test set was equal to 90.5%. The average Receiver Operating Characteristics (ROC) curve of K-PLS-DA model is shown in Figure S3 with area under cure of 92.8%.

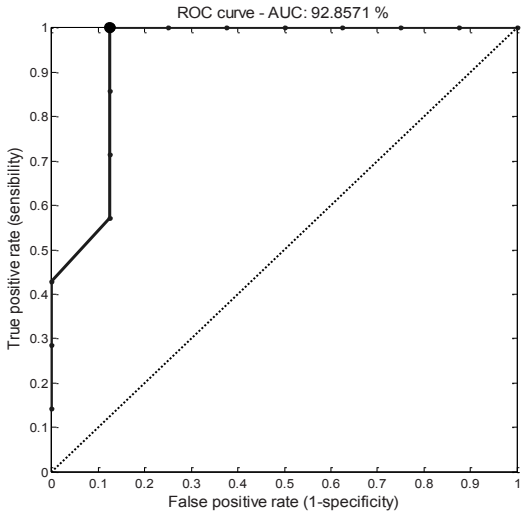


Figure 3S. Average Receiver Operating Characteristics derived from K-PLS-DA for random division of data.

LIST OF COMMUNICATIONS

LIST OF COMMUNICATIONS

ARTICLES

(1) Quantitative proteomics and metabolomics analysis of normal human cerebrospinal fluid samples.

Marcel P. Stoop, Leon Coullier, Therese Rosenling, Shanna Shi, ***Agnieszka M. Smolinska****, Lutgarde Buydens, Kirsten Ampt, Christoph Stingl, Adrie Dane, Bas Muilwijk, Peter A. E. Sillevis Smitt, Rogier Q. Hintzen, Rainer Bischoff, Sybren S. Wijmenga, Thomas Hankemeier, Alain J. van Gool, and Theo M. Luider*

Molecular Cellular Proteomics ***(2010)****, 9(9), pp. 2063-2075*

(2) The impact of delayed storage on the measured proteome and metabolome of human cerebrospinal fluid (CSF).

T. Rosenling, M. P. Stoop, ***A. Smolinska****, B. Muilwijk, L. Coulier, S. Shi, A. Dane, Ch. Christin, F. Suits, P. L. Horvatovich, S.S. Wijmenga, L. M.C. Buydens, R. Vreeken, T. Hankemeier, A. J. van Gool, T. M. Luider and R. Bischoff*

Clinical Chemistry **(2011)**, 57(12), pp. 1703-11

(3) Neuroinflammation and peripheral inflammation can be distinguished based on 1H-NMR and pattern recognition methods.

A. Smolinska*, A. Attali, L. Blanchet, K. Ampt, H. van Aken, E. Suidgeest, T. Tuinstra, A.L. van Gool, T. Luider, S.S. Wijmenga and L.M.C. Buydens*

Journal of Proteome Research **(2011)**, 10 (10), pp. 4428–4438

(4) Fusion of metabolomics and proteomics data for biomarkers discovery.

*L. Blanchet****, A. Smolinska****, A. Attali, M. Stoop, K. Ampt, H. van Aken, E. Suidgeest, T. Tuinstra, S.S. Wijmenga, T. Luider and L.M.C. Buydens*

*BMC Bioinformatics (****2011****), 12, 254*

(5) Simultaneous Analysis of Plasma and CSF by NMR and Hierarchical Model Fusion.

***A. Smolinska**, J. Posma, L. Blanchet, K.A.M. Ampt, A. Attali, T. Tuinstra, T. Luider, M. Doskocz, P.J. Michiels, F.C. Girard, L.M.C. Buydens, and S.S. Wijmenga*

*Analytical and Bioanalytical Chemistry **(2012)*** 403 (4), pp. 947-59

(6) Interpretation and Visualization of non-linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis

A.Smolinska, L.Blanchet, L.Coulier, K. Ampt, T. Luider, R. Q. Hintzen, S. S. Wijmenga, and L.M.C. Buydens

PLoS One **(2012)**, 7(6), pp. e38163

(7) NMR and Pattern Recognition Methods in Metabolomics. From Data Acquisition to Biomarker Discovery

A. Smolinska, L. Blanchet, L. M.C. Buydens and S. S. Wijmenga

Analytica Chimica Acta **(2012),** in press, (dx.doi.org/10.1016/j.aca.2012.05.049)

ORAL PRESENTATIONS

(1) Metabolome and proteome of human CSF studying by NMR and MS. Identification and quantification.

A. Smolinska, K. Ampt, A. van Gool, T. Luider, M. Stoop, L.Coulier, S. Shi, T. Rosenling, T. Hankemeier, R. Bischoff, R. Wehrens, L.M.C. Buydens and S.S. Wijmenga
*Spring Meeting 2009 TI Pharma, Utrecht, The Netherlands, 23 April **2009***

(2) Metabolomics investigation of CSF by Nuclear Magnetic Resonance.

A. Smolinska, A. Attali, K. Ampt , A. van Gool, T. Luider, L.M.C. Buydens and S.S. Wijmenga
*IMM Symposium, Nijmegen, Netherlands, 17-18 May **2010***

(3) ^{1}H-NMR spectroscopy and pattern recognition methods for metabolomics investigation of pre-clinical model of Multiple Sclerosis.

*A. Smolinska, L. Blanchet, K. Ampt, A. Attali, T.Luider, A. van Gool, S.S. Wijmenga and L.M.C Buydens Fachgruppentagung: Magnetische Resonanzspektroskopie - Joint Benelux/German MR Conference (FGRM10), Munster, Germany, 20-23 September **2010***

(4) ^{1}H-NMR spectroscopy and chemometrics for metabolomics investigation of pre-clinical model of Multiple Sclerosis.

*A. Smolinska, L. Blanchet, K. Ampt, A. Attali, T.Luider, A. van Gool, S.S. Wijmenga and L.M.C Buydens The Analytical Challenge 2010 (TAC 2010), Lunteren, Netherlands, 1-2 November **2010***

(5) Supervised analysis for biomarker discovery in metabolomics data of Multiple Sclerosis disease.

*A. Smolinska, L. Blanchet, K. Ampt, T. Tuinstra, A. Attali, S. Wijmenga and L. Buydens Scandinavian Symposium on Chemometrics, Billund, Denmark, 7-10 June **2011***

(6) Non-linear analysis in metabolomics data of Multiple Sclerosis disease

*A. Smolinska, L. Blanchet, S.S. Wijmenga and L.M.C. Buydens EUROanalysis 16, European Conference on Analytical Chemistry, Chalanges in Modern Analytical Chemistry, Belgrade, Serbia, 11-15 September **2011***

(7) Metabolomics fingerprinting in clinical study of Multiple Sclerosis disease

A.Smolinska, L. Blanchet, L. Coulier, K. Ampt, S.S. Wijmenga and L.M.C Buydens
*Metabolomics and System Biology, San Francisco, USA, 20-22 February **2012***

(8) Non-linear analysis of multiple metabolomics data

A.Smolinska, L.Blanchet, L. Coulier, S.S. Wijmenga and L.M.C. Buydens
XIII Chemometrics and Analytical Chemistry, Budapest, Hungary, 25-29 June **2012**

POSTERS PRESENTATIONS

(1) Stability studies on CSF using NMR

A. Smolinska, K. Ampt, A. Kolkman, A.J. van Gool, R.Wehrens, L.M.C. Buydens, S.S. Wijmenga
The Analytical Challenge, Lunteren, The Netherlands, 3-4 November ***2008***

(2) Stability study of human CSF by ^{1}H NMR

A. Smolinska, L. Blanchet, K. Ampt, A.J. van Gool, T. Luider, M. Stoop, R. Wehrens, S.S. Wijmenga and L.M.C. Buydens
Proteomics and Metabolomics Biomarker Discovery in CSF, Rotterdam, The Netherlands, May 14-15 ***2009***

(3) Identification and quantification of human CSF by Nuclear Magnetic Resonance.

A.Smolinska, K. Ampt, A. van Gool, T. Luider, M. Stoop, R. Wehrens, L.M.C. Buydens and S.S. Wijmenga
FIGON Dutch Medicine Days, Lunteren, The Netherlands, 12-14 October ***2009***

(4) Nuclear Magnetic Resonance of human CSF. Identification and quantification.

A. Smolinska, K. Ampt, A. van Gool, T. Luider, M. Stoop, R. Wehrens, L.M.C. Buydens, S.S. Wijmenga
2nd International Meeting on NMR *and Quantitative Analysis ,Stockholm Sweden, April 21-22* ***2009***

(5) Metabolomics investigation of Multiple Sclerosis by ^{1}H NMR.

A. Smolinska, A. Attali, K. Ampt, A. van Gool, T. Luider, L.M.C. Buydens and S.S. Wijmenga
1st International Metabolomics Symposium, München, Germany, March 10-12, ***2010***

(6) ^{1}H-NMR spectroscopy and pattern recognition methods for metabolomics investigation of pre-clinical model of Multiple Sclerosis.

A. Smolinska, L. Blanchet, K. Ampt, A. Attali, T. Luider, A. van Gool, S.S. Wijmenga and L.M.C Buydens
CAC 2010, Antwerpen, Belgium, 18-22 October ***2010***

(7) Multivariate analysis for biomarker selection in Multiple Sclerosis.

A. Smolinska, K. Ampt, T. Luider, A. van Gool, S.S Wijmenga and L.M.C Buydens
*CAC 2010 – Antwerpen, Belgium, 18-22 October **2010***

(8) Data fusion of ^{1}H-NMR metabolic spectra.

J.M. Posma, A. Smolinska, A. Attali, A. van Gool, T. Luider, S.S. Wijmenga and L.M.C. Buydens
*CAC 2010 – Antwerpen, Belgium, 18-22 October **2010***

(9) Multivariate analysis of 2D ^{1}H NMR spectra of pre-clinical model of Multiple Sclerosis.

R. ter Horst, A. Smolinska, A. Atali, P. Michiels, F. Girard, T. Luider, A. van Gool, S.S. Wijmenga and L.M.C. Buydens
*CAC 2010 – Antwerpen, Belgium, 18-22 October **2010***

(10) Supervised analysis for biomarker discovery in metabolomics data of Multiple Sclerosis disease.

A. Smolinska, L. Blanchet, K. Ampt, T. Tuinstra, A. Attali, S.S. Wijmenga and L.M.C. Buydens
Top Institute Pharma Spring Meeting, Utrecht, The Netherlands, April 14th **2011**

(11) Metabolomics investigation of Multiple Sclerosis disease.

A. Smolinska, L. Blanchet, K. Ampt, L.M.C Buydens and S.S. Wijmenga
*IMM Symposium, Nijmegen, The Netherlands, 17-18 May **2011***

(12) Biomarker discovery in clinical study of Multiple Sclerosis disease.

A. Smolinska, L.Blanchet, L. Coulier, K. Ampt, S.S. Wijmenga and L.M.C Buydens
*International Chemometrics Research Meeting, Berg and Daal, Netherlands, 25-29 September **2011***

(13) Non-linear analysis of clinical data of Mutiple Sclerosis disease.

A. Smolinska, L. Blanchet, L. Coulier, K. Ampt, L.M.C Buydens and S.S. Wijmenga
Chains 2011, Utrecht, The Netherlands, 28-30 November **2011**

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

This thesis rose out of four years of research that has been done since I came to Netherlands. Here I would like to thank all the people who helped me and were with me the past few years here in Netherlands, Poland and all over the world. Obviously, my thesis would not have been possible without the support of many people.

First of all I would like to thank my promoters, Lutgarde and Sybren whose expertise added considerably to my graduate experience. I appreciate their knowledge in many areas and their time.

I would like to express my gratitude and love to Lionel, my soul mate and my “quasi husband” for his support and all patience at all time, through the period of my studies, for waiting me in the middle of night with a cup of tea when I was coming back from the lab. Without his encouragement and constant motivation this thesis would have been never finished. You were always able to understand my feelings, problems and frustrations during my PhD. Thanks to always show me “the other side of the medal”. You have given me unequivocal support for which my mere expression of appreciations similarly does not suffice.

Chciałabymwyrazić mojąmiłośćiwdzięczność moim ukochanym rodzicom, Jolanciei Mirosławowi, i mojej siostrze Monice za niekończącą się miłość, zrozumienie i wiarę, która zawsze mi pomagała w przezwyciężeniu wszelkich trudności w czasie całego okresu moich studiów. Kochani rodzice jestem Wam bardzo wdzięczna za pomoc, oparcie, za nieustanną motywację, miłość, dobre rady, otuchę, wiarę we mnie i mówienie „ Nie martw się wszystko będzie w porządku”, albo „nie ma tego złego, co by na dobre nie wyszło”. To dzięki wam stałam się tym, kim jestem. Moni, są słowa, których nie należy mówić, słowa, które nic nie zrani i nie zgubi; dziękuję Ci po prostu za to, ze naprawdę byłaś i jesteś, za rozmowy, zrozumienie i wsparcie.

Je souhaite aussi remercier Evelyne, Michel et Jérôme pour leur gentillesse et leur soutien durant ces dernières années ; ainsi que de m’avoir forcé à me reposer. Evelyne, merci pour les conversations multilingues et pourtant compréhensible, pour les cours de français. J’espère que nous pourrons bientôt répéter ces cours. Merci à mon photographe officiel pour son aide dans le design de la couverture de cette thèse. Jérôme merci pour les bons petits plats préparés lors de chacune de nos visites.

Kirsten my office mate and friend, thanks for all discussion, fun and the support, especially before and during the meetings (you know which ones I mean, right!), to keep up the spirit; for teaching me NMR and being my paranymph.

Barbara, my second and favorite flat mate in Nijmegen, I would like to thank you for our great time we shared, for our dinners (I cooked and you cleaned) for our evening's discussion with good German beer and crazy shopping.

Special thanks to Jasper for being a good friend and taking time to practice Dutch with me after long working day. Jij hebt er aan bijgedragen dat ik geslaagd ben voor mijn NT2 examen. Bedankt voor jouw hulp.

I want to thank: Ard for being my very good colleague, for nice conversations not only about research; Ramon, thanks for good advises and sharing the doubts and blues about our PhD; Frank, thank you for all your help in the laboratory, especially with preparing Shigemi tube without breaking it; Thank you Marian and Brigitte for all administration help; Special thanks to Marian for writing my samenvatting and for releasing me and reducing the amount of stress in my last day at RU by taking care of binding my thesis.

I want to thank to my great master students, Joram, Jurn and Rob. It was great time having you guys around. Joram and Jurn, my dear friends, I hope you will like my country!

A very special thanks goes out to Prof. Dr. hab. Beata Walczak, which was in the first place the driving force behind getting me to Netherlands. I also would like to thank you for being part of the reading committee.

Next, thanks to Prof. Dr. J.P.M. van Duynhoven and Prof. Dr. A.P.M. Kentgens whose being as part of the reading committee they took their time to evaluate my thesis.

Thanks to the people in the TI Pharma consortium. I want to thank to all colleagues at the Chemometrics department (Tom, Patrick, Geert, Bulent, Ron) and Biophysical chemistry department (Margit, Jan van Oss, Gerrit, Sjaak, Chandrakala, Suresh, Vipin, Marco, Marc, Aafke, Jorge, Gijs, Dennis, Ernst, Jan van Bentum, Anne-Jo, Hans, Jacob). I have spent very nice time and I enjoyed your company a lot. Everybody was always so friendly and helpful.