Metabolic Variation in Autoimmune Diseases

Rasmus K. Madsen
Table of Contents

Table of Contents i
Abstract iii
List of papers iv
Additional work v
Abbreviations and notations vi
Enkel sammanfattning på svenska viii
Metabolisk Variation i Autoimmuna Sjukdomar viii
Introduction 1
Aims of this work 3
Background 4
Immunity 4
Autoimmunity and autoimmune diseases 5
Rheumatoid Arthritis 6
Diabetes Mellitus Type 1 7
Metabolism 8
Anabolism and catabolism 8
Properties of metabolism and metabolites 8
Body fluids 9
Metabolic profiling 10
Nomenclature and definitions 10
Analytical techniques 11
Metabolic profiling data from hyphenated instruments 14
Gas Chromatography (GC) coupled MS-method 15
Liquid Chromatography (LC) coupled MS-method 17
Future developments 18
Statistical techniques 18
Principal Component Analysis (PCA) 19
Orthogonal Projection to Latent Structures (OPLS) 20
Discriminant analysis (DA) 21
Data centering and scaling 21
Model validation and evaluation of model quality 22
Validation strategies 24
Univariate techniques 25
Considerations regarding data analysis 26
Discussion on classification, biomarkers and disease diagnosis 26
Results – Discussion of papers 28
Paper 1 28
Paper 2 30
Paper 3 31
Paper 4 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions and future perspectives</td>
<td>33</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>35</td>
</tr>
<tr>
<td>References</td>
<td>37</td>
</tr>
</tbody>
</table>
Abstract

The human being and other animals contain immensely complex biochemical processes that govern their function on a cellular level. It is estimated that several thousand small molecules (metabolites) are produced by various biochemical pathways in humans. Pathological processes can introduce perturbations in these biochemical pathways which can lead to changes in the amounts of some metabolites.

Developments in analytical chemistry have made it possible measure a large number metabolites in a single blood sample, which gives a metabolic profile. In this thesis I have worked on establishing and understanding metabolic profiles from patients with rheumatoid arthritis (RA) and from animal models of the autoimmune diseases diabetes mellitus type 1 (T1D) and RA.

Using multivariate statistical methods it is possible to identify differences between metabolic profiles of different groups. As an example we identified differences between patients with RA and healthy volunteers. This can be used to elucidate the biochemical processes that are active in a given pathological condition.

Metabolite concentrations are affected by a many other things than the presence or absence of a disease. Both genomic and environmental factors are known to influence metabolic profiles. A main focus of my work has therefore been on finding strategies for ensuring that the results obtained when comparing metabolic profiles were valid and relevant. This strategy has included repetition of experiments and repeated measurement of individuals’ metabolic profiles in order to understand the sources of variation.

Finding the most stable and reproducible metabolic effects has allowed us to better understand the biochemical processes seen in the metabolic profiles. This makes it possible to relate the metabolic profile differences to pathological processes and to genes and proteins involved in these.

The hope is that metabolic profiling in the future can be an important tool for finding biomarkers useful for disease diagnosis, for identifying new targets for drug design and for mapping functional changes of genomic mutations. This has the potential to revolutionize our understanding of disease pathology and thus improving health care.
List of papers

These four papers are part of this thesis. Madsen RK and Madsen R refer to the same person.


3) Madsen R, Rantapää-Dahlqvist S, Lundstedt T, Moritz T, Trygg J: Metabolic Responses to Change in Disease Activity during Tumor Necrosis Factor Inhibition in Patients with Rheumatoid Arthritis. *Journal of Proteome Research* 2012, **11:**3796-3804.


Paper 3 reprinted with permission from the American Chemical Society. Copyright 2012
**Additional work**

The author contributed to the following papers during the thesis work, but they are not appended to the thesis.


Abbreviations and notations

Vectors are denoted in bold.
Matrices are denoted by capital letters and in bold.

RA Rheumatoid Arthritis
T1D Diabetes Mellitus type 1
Ncf1 Neutrophil cytosolic factor 1 (gene)
TNF Tumor Necrosis Factor
IL-1β Interleukine 1β
IL-6 Interleukine 6
DNA Deoxyribonucleic acid
HLA Human Leukocyte Antigen
PTPN22 Protein Tyrosine Phosphatase, Non-receptor type 22 (gene / protein)
ACPA Antibodies against Citrullinated peptides/proteins
NOX NADPH oxidase (protein complex)
NADPH Nicotinamide Adenine Dinucleotide Phosphate (reduced form)
ROS Reactive Oxygen Species
HDL High Density Lipoproteins
GAD Glutamic Acid Decarboxylase (enzyme)
IA-2 Tyrosine phosphatase-related islet antigen 2
NOD Non-Obese Diabetic (mouse strain)
BB Biobreeding (rat strain)
ATP Adenosine Tri-Phosphate
RNA Ribonucleic acid
KEGG Kyoto Encyclopedia of Genes and Genomes
AA Arachidonic Acid
NMR Nuclear Magnetic Resonance
MS Mass Spectrometry
GC Gas Chromatography
LC Liquid Chromatography
EI Electron Impact
ESI Electrospray ionization
m/z mass / charge-ratio
ToF Time of Flight
QqQ Triple quadropole
N Number of samples
K Number of variables (metabolites)
NxK A dataset with N samples characterized by K variables
MSTFA N-methyl-N-(trimethylsilyl) trifluoroacetamide
TCMS Trimethylchlorosilane
RT Room Temperature
HMCR Hierarchical Multivariate Curve Resolution
AR                 Alternating Regression
NIST                National Institute of Standards and Technology
UPLC                Ultra Performance Liquid Chromatography
LV                  Latent Variable
PCA                 Principal Component Analysis
OPLS                Orthogonal Projections to Latent Structures
p                  Loading value
t                  Score value
X                  Matrix of observation values
T                  Matrix of score values
P                  Matrix of loading values
T                  Transpose
E                  Matrix of X residuals
SVD                Single Value Decomposition
NIPALS             Non-linear Iterative Projections by Alternating Least Squares
PLS                Partial Least Squares
Y                  Matrix of response values
T_P                Matrix of predictive scores (OPLS)
P_P                Matrix of predictive loadings (OPLS)
T_O                Matrix of orthogonal scores (OPLS)
P_O                Matrix of orthogonal loadings (OPLS)
C                  Matrix of y-related loadings
F                  Matrix of Y residuals
P_P                Vector of predictive loadings (OPLS)
y                Vector of responses
w                Vector of weights
t                Vector of score values
c                Vector of y-related loadings
p                Vector of loadings
w_O               Vector of orthogonal weights (OPLS)
t_O               Vector of orthogonal scores (OPLS)
p_O               Vector of orthogonal loadings (OPLS)
XE                Residual X matrix
DA                Discriminant Analysis
p(corr)            Correlation scaled loading vector
CV                Cross-Validation
Q2                Goodness of prediction
RMSECV             Root Mean Square Error of Cross-Validated prediction
CV-ANOVA           Cross-validation analysis of variance
R2                Goodness of Fit
CRP                C-Reactive Protein
Idd                Insulin dependent diabetes
r                Pearsons correlation value
Enkel sammanfattning på svenska

Metabolisk Variation i Autoimmuna Sjukdomar
Människan och andra djur innehåller oerhört komplexa biokemiska processer som styr deras funktion på cellulär nivå. Det uppskattas att flera tusen små molekyler (metaboliter)produceras av olika biokemiska vägar hos människor. Patologiska processer kan leda till störningar i dessa biokemiska vägar som kan leda till förändringar i mängden av vissa metaboliter.

Utvecklingen inom analytisk kemi har gjort det möjligt att mäta ett stort antal metaboliter i ett enda blodprov, vilket ger en metabolisk profil. I denna avhandling har jag arbetat med att etablera och förstå metaboliska profiler från patienter med reumatoid artrit (RA, också kallat ledgångsreumatism) och från djurmodeller av de autoimmuna sjukdomarna diabetes mellitus typ 1 (T1D) och RA.

Genom att använda multivariata statistiska metoder är det möjligt att identifiera skillnader mellan metaboliska profiler set inom olika grupper. Som ett exempel har vi identifierat skillnader mellan patienter med RA och friska försökspersoner. Detta kan användas för att belysa de biokemiska processer som är verksamma vid ett givet patologiskt tillstånd.

Metabolitkoncentrationer påverkas av en många andra saker än närvaro eller frånvaro av en sjukdom. Både genomiska och miljömässiga faktorer är kända för att påverka metaboliska profiler. En huvudinriktning under mitt arbete har därför varit att hitta strategier för att säkerställa att de resultat som erhålls vid en jämförelse av metaboliska profiler var giltiga och relevanta. Denna strategi har omfattat upprepning av experiment och repeterad mättning av individers metaboliska profil för att förstå källorna till variation.

Att hitta de mest stabila och reproducerbara metabola effekter har tillåtit oss att bättre förstå de biokemiska processerna som syns i de metaboliska profilerna. Detta gör det möjligt att relatera de metaboliska skillnader till patologiska processer samt till gener och proteiner som är inblandade i dessa.

Förhoppningen är att metabolisk profilering i framtiden kan bli ett viktigt verktyg för att hitta biomarkörer användbara för sjukdomsdiagnostik, för att identifiera nya mål för läkemedelsutveckling och för kartläggning av funktionella förändringar beroende på genomiska mutationer. Detta har
potential att revolutionera vår förståelse av sjukdoms patologi och därmed förbättra framtidens sjukvård.
**Introduction**

Metabolic profiling is a relatively new field of research that focuses on measuring low molecular weight compounds (metabolites) in biological material such as blood plasma or sera. Multivariate statistical analyses of these measurements can be used to identify metabolic differences between groups – for instance patients with a disease and healthy controls. These metabolic differences can then be used to determine which group a newly-tested individual belongs to and can also provide information on the biological / pathological differences between the groups.

![Idealized workflow in metabolic profiling studies.](image)

While the method may sound easy to implement when described in this way, it is an unfortunate fact that dealing with organisms as complex as human beings is never as easy as it sounds. Competence in analytical chemistry, univariate and multivariate statistics, design of experiments, biochemistry, pathology and medicine is required to perform metabolic profiling studies. Most of all, it requires good collaborations between specialists in these diverse subjects.

This thesis describes the methods used, the problems encountered, and the results obtained in my work on mapping and understanding metabolic variation in autoimmune diseases. It starts with an introduction to the aims of the four studies of the thesis, followed by an introduction to the relevant biology: the immune system, autoimmunity, type 1 diabetes, rheumatoid
arthritis and general metabolism. This is followed by descriptions of the methods used in the studies, first analytical and then statistical. The thesis concludes with a discussion of some questions I have found to be important as well as the results of the four composing studies and the future of metabolic profiling as I would like to see it.
Aims of this work

The main aim when commencing this work was to explore the possibilities of applying metabolic profiling techniques in medical research, specifically aimed at the autoimmune diseases rheumatoid arthritis and diabetes mellitus type 1.

One focus was to be on strategies for validating the relevance of variation exhibited by metabolic profiles and for gaining biological knowledge and understanding from this variation. Another focus was assessing the potential for developing clinical tools based on metabolic profiling that can aid in diagnosis and prognosis of patients.

In retrospect, the aims of each study can be defined more clearly:

1) To explore the potential for diagnosing Rheumatoid Arthritis using metabolic profiling.
2) To identify functionally important gene products in the progression to Type 1 Diabetes by analyzing perturbed metabolic pathways.
3) To develop methods for predicting disease activity in patients with Rheumatoid Arthritis at the group and individual levels.
4) To identify the physiological effects of a mutation in Ncf1, a gene that affects susceptibility to autoimmunity.
Background

Immunity

A well functioning immune system is of paramount importance to humans and other animals. Without the ability to defend themselves against pathogens and parasites, complex multi cellular animals could never have evolved. As a result, several components of the immune system have been highly conserved throughout evolutionary history [1]. The immune system can be loosely divided into two parts – the innate and the adaptive immune system – although they work together and many cell types have roles in both. The innate immune system is based on receptors capable of recognizing specific patterns of structures (primarily on the cell surface) that are common on pathogens. The main effector cells of innate immunity are neutrophils, monocytes / macrophages and natural killer cells. Neutrophils and monocytes / macrophages are phagocytes, meaning that they can engulf and kill pathogens upon recognition. Natural killer cells are cytotoxic and kill off damaged cells and cells infected with pathogens. The innate immune response is capable of eradicating many commonly encountered pathogens. Sometimes a pathogen cannot be cleared from the body directly, in which case the innate immune system functions to limit the spread of the pathogen until an even more specific adaptive immune response have time to develop.

Macrophages are the primary cells involved in detecting pathogenic infections. When a pathogen is detected, the macrophages initiate inflammation. Inflammation is primarily controlled by signaling through chemokines and cytokines such as Tumor Necrosis Factor (TNF), Interleukine 1β (IL-1β), and IL-6. The effects of inflammation are seen throughout the body and include increases in body temperature, increased proliferation and maturation of immune cells, and synthesis (in the liver) of acute phase proteins that participate in the defense against the pathogen and induce metabolic changes that mobilize energy reserves to fuel the body’s defense mechanisms. At the infected site, the main effect of inflammation is to increase the permeability of blood vessels and promote the migration of fluid, proteins and immune cells towards the site to facilitate the immune response and containment of the pathogen. Together these effects lead to the four classical clinical features of inflammation: heat, pain, redness and swelling. If the antigen that provokes inflammation is not cleared inflammation can become chronic. Chronic, inappropriate inflammation is one of the main features of rheumatoid arthritis as discussed later.
Persistent pathogens are ultimately killed off by the adaptive immune system. The adaptive immune response has two main branches: the humoral immune response and the cell-mediated immune response. Both are highly target-specific due to special processes for genomic DNA rearrangement and somatic hypermutation that create an enormous repertoire of cells capable of recognizing diverse pathogens. The most important cells involved in the humoral immune response are the B-cells, which secrete antibodies that tag pathogens and assist phagocytes and the complement system in killing them, and T-helper-cells, which are important for the differentiation of the B-cell. The humoral immune system defends the extracellular spaces of the body. However, many pathogens infect host cells in addition to being present in the extracellular spaces. Infected host cells are killed by the cell-mediated immune response, which is primarily mediated by the cytotoxic T-cells. An exhaustive introduction to the immune system and the maturation of immune responses can be found in references [2] and [3].

Autoimmunity and autoimmune diseases

The immune system’s ability to target specific pathogen-derived structures is almost unlimited. This is important for minimizing the spread of pathogens in the body, but an equally important function is to ensure that the immune system does not target the body’s own structures. There are several mechanisms to ensure that the immune system does not react to self antigens, but these sometimes fail, leading to autoimmunity [4-7]. Two well known autoimmune diseases are Rheumatoid Arthritis (RA) and Type 1 Diabetes Mellitus (T1D). The primary structures targeted in RA are within small synovial joints, where inflammation causes pain, stiffness and swelling. In T1D, the β-cells of the pancreas are targeted, reducing insulin secretion [8]. Despite the very different clinical presentations of RA and T1D, they have many mechanisms in common. For instance, the presence of specific alleles of the Human Leukocyte Antigen (HLA) locus are the strongest genetic proponent for development of both RA and T1D [9]. Moreover, genes involved in modulating immune responses (such as PTPN22, coding a phosphatase involved in controlling signaling in T- and B-cells) are associated with several autoimmune diseases [10].

Since autoimmune diseases involve the destruction of tissues by the immune system, it is theoretically possible to prevent their progression by using immune modulating therapy [11-12]. There is therefore great interest in identifying markers for susceptibility to autoimmunity, as was done in papers 2 and 4 of this thesis.
**Rheumatoid Arthritis**

RA is a systemic autoimmune disease in which the synovial tissues, especially in the smaller joints of the hands and feet, are infiltrated by immune cells [13-14]. RA is a severe burden to both individual patients and society, due to both morbidity from joint destruction and the increased prevalence of co-morbidities such as anemia, cardiovascular events and certain cancers [15-16]. RA has a chronic, relapsing progression; patients are therefore followed on a lifelong basis using the Disease Activity Score in 28 joints (DAS28) to assess disease severity and adjust the applied therapeutic strategy accordingly [17-18]. In paper 3 we investigated if there is a link between metabolic profiles and DAS28 score.

The pathogenesis of RA is only partly understood, but recent studies have yielded important new results. RA has a substantial genetic component, with an estimated heritability of 50-60 %, and around 30 genes are currently known to be associated with RA. Two of the most strongly RA-associated genes are HLA and PTPN22, both of which are discussed above [19-20]. In addition, the roles of cytokines in driving disease progression is now better understood, and this has led to the development of successful treatments such as biological TNF inhibitors [21]. RA was considered to be an autoimmune disease without an auto-antigen for a long time. However, the discovery of antibodies against citrullinated peptides/proteins (ACPA) have suggested plausible antigens for the disease [22]. ACPA are now important in the diagnosis of RA, and around 75 % of RA patients are ACPA-positive (at 98 % specificity). ACPA can often be detected years before the onset of other symptoms [23-25].

The exact mechanisms that lead to the failure of self-tolerance in RA are still poorly understood. Functional characterization of the involved genes should go some way to addressing this issue. Another gene that was shown to be important in the development of arthritis in rodents and to correlate with RA in humans is the *Neutrophil cytosolic factor 1 (Ncf1)* gene [26-27]. Ncf1 regulates the activity of the NADPH oxidase (NOX) complex and thus helps to control the production of reactive oxygen species (ROS). It has been shown that ROS production by macrophages is important in regulating the activities of T-cells and thus for modulating immune responses and preventing autoimmunity [28-29]. In paper 4 we studied metabolic effects of a mutation in the *Ncf1* gene under physiological conditions.

In an early metabolic profiling study, it was found that synovial fluid of RA patients had slightly higher amounts of triglycerides than that from osteoarthritis patients and patients with traumatic effusions but that the
variability within each group was large [30]. RA is also known to change the blood lipid composition, with RA patients having a more atherogenic lipid profile than controls [31-32]. This finding is consistent with those from the first systematic metabolic profiling study of plasma from RA patients, which showed that there were significant metabolic differences between RA patients with active disease, RA patients in remission, and healthy controls. RA was characterized by elevated levels of lactate, acetylated glycoproteins and cholesterol together with reduced levels of high density lipoproteins (HDL) [33]. Reduced serum histidine levels have also been observed in several studies [34-36]. In paper 1 we did a systematic investigation of the diagnostic properties of metabolic perturbations in patients with RA.

**Diabetes Mellitus Type 1**

Diabetes is a group of diseases characterized by hyperglycemia due to a failure to secrete or respond to insulin. In T1D, the underlying abnormality is autoimmune destruction of the insulin-secreting β-cells of the islets of Langerhans in the pancreas. Patients typically present with symptoms related to hyperglycemia such as increased thirst, polyuria and increased appetite in combination with weight loss. At least three different antigens exist: insulin itself, the enzyme glutamic acid decarboxylase (GAD), and the protein tyrosine phosphatase-related islet antigen 2 (IA-2) [37]. Unlike most other autoimmune diseases, T1D symptoms typically become apparent before adolescence. Immunological abnormalities are often present years before the onset of symptoms, showing that the development of autoimmunity and islet destruction is a gradual process [38].

Using metabolic profiling, it was found that metabolic dysregulation in children who subsequently developed T1D preceded both the development of diabetes and traditional serological signs of autoimmunity. Changes in amino acid and lipid metabolism were detected throughout the follow-up period and distinct changes were seen before seroconversion [39]. In a subsequent investigation, it was found that these autoimmunity-related metabolic changes were not age-dependent, but distinct metabolic profiles were seen depending on when the child first developed serological markers of autoimmunity. This suggests that different pathways are active in development of T1D in early childhood and in puberty [40]. As such, analyses of these changes may provide important information that can be applied for the early detection of autoimmune diseases and to better understand their diverse pathology.

The non-obese diabetic (NOD) mouse is one of the most common models of autoimmune diabetes [41]. It shares several immunological features with
human T1D, including pancreas-specific auto-antibodies, and its disorder involves the same cell types and genetic linkage patterns [42]. Pre-diabetic metabolic changes have been demonstrated in another animal model of T1D, the congenic BB rat [43]. In paper 2 we investigated whether metabolic changes resembling those seen in children were present in NOD mice.

**Metabolism**

**Anabolism and catabolism**

Metabolism is a collective term for the biochemical processes relating to energy expenditure in cells and organisms and can thus be seen as one of the foundations of all life. The importance of metabolism is demonstrated by the high degree of conservation of the metabolic pathways, enzymes and structural motifs across a wide range of species [44-45]. Breakdown of chemical bonds in energy harvesting is termed catabolism, whereas synthesis of more complex molecules is termed anabolism.

In animals the primary substrates for catabolic metabolism are small energy-rich compounds such as carbohydrates, amino acids and lipids produced by digestion of food. The energy of the chemical bonds in these compounds is converted into high-energy bonds in general purpose energy storing compounds, notably ATP. During metabolism other small compounds are produced – all the small compounds in an organism are termed metabolites and collectively they have been termed the metabolome, a term analogue to the genome and the proteome.

The energy harvested during catabolism is then used to drive various biochemical processes such as contraction of the sacromere, enabling locomotion, and synthesis of building blocks for anabolic cellular processes, such as specific lipids that are important for lipid bilayers, nucleic acids (DNA and RNA), and the many different proteins necessary in animals.

**Properties of metabolism and metabolites**

Although metabolism occurs in the cell, it is regulated both at the tissue and organism levels as well as at the cellular level [46]. It is governed both by the local availability of substrates and products and the overall energy balance, but also via local and systemic signaling by species such as hormones [47-48]. Metabolism is thus even more complex than it might seem based on data from sources such as the KEGG encyclopedia [49-50].
A further complication of relevance to the research in this thesis is the strong integration between metabolism and immunity [51]. Lipid compounds in particular are important signaling molecules that regulate important aspects of the immune response. The enzyme phospholipase A2, which releases fatty acids from triglycerides in cell membranes and blood lipoproteins, is particularly noteworthy in this context [52]. Its most well-characterized and important effect is the release of arachidonic acid (AA), but there are several related isoenzymes with differing specificities [53-54]. Once released, AA is oxidized into prostaglandins and leukotrienes, which regulate many cellular processes related to inflammation and wound healing [55].

One of the distinctive properties of metabolism that is not encountered to the same extent when studying things such as genes or proteins is that metabolite behavior is extremely dynamic [56]. Metabolite levels are affected by both genetic and environmental factors, meaning that there is both inter- and intra-individual variation in metabolite profiles and levels. The former is based on both environmental and genetic factors, while the latter reflects changes in environmental factors such as health status, nutrition, daily rhythms and pharmacological treatment [57]. Depending on one’s perspective, these complex patterns of variation may represent problems or opportunities. Either way, it is important to be aware of them when designing experiments, analyzing results and drawing conclusions from metabolic profiling studies.

Another defining trait of metabolites is that most of them are present in all subjects and all tissues. Congenital metabolic diseases may lead to the complete absence of a metabolite, but this is an exception to the rule [58]. This means that it is unlikely to find a single metabolite that can be used a specific marker of any given condition. Instead, biological and pathological processes are characterized by studying patterns consisting of several metabolites, thus the study of metabolites is inherently multivariate.

**Body fluids**

The human body consists of around 60% water that is continuously circulated between different body compartments. The circulation of body fluids is important because it provides the smallest functional units of the body – its cells - with nutrients. Fluids such as the plasma thus contain several thousand chemical entities [8]. Some of the body fluids are relatively easy to sample in living humans, such as plasma in the central circulation, synovial fluid in the joints, and the central spinal fluid. This stands in contrast to individual cells or intact organs.
In the works in this thesis both plasma and serum have been sampled. The plasma is the liquid part of the blood – that is the blood’s cellular content has been removed. In serum also clotting factors have been removed as compared to plasma.

**Metabolic profiling**

The idea that diseases can affect body fluid composition is old, but it was not until the 1970s and 1980s that it became possible to use untargeted analytical techniques for studying changes in metabolite composition in body fluids and thereby identify biochemical differences between healthy patients and those with specific diseases. It was rapidly recognized that these differences could in turn be useful for diagnosing patients and understanding pathological changes [30, 59-60].

These early studies were conducted at a time when the analytical techniques for identifying and quantifying metabolites were relatively rough. Furthermore there was a lack of methods for dealing with the many variables involved, and so metabolic profiling approaches did not become generally applicable until around the turn of the millennium. The development of what could be called modern metabolic profiling techniques was largely driven by the toxicology research group lead by Jeremy K. Nicholson [61] and the plant sciences group of Oliver Fiehn [62]. Medical applications followed shortly after (see reference [63] for an overview of metabolic profiling results related to disease diagnosis).

Numerous different methods and approaches have been employed in studies of metabolism. It seems that since metabolic profiling gained prominence relatively quickly, it drew in researchers with a diverse range of scientific backgrounds who sought to offer their expertise and obtain new insights into their research topics. As a result, the following sections does not aim to provide a comprehensive discussion of the various methodologies that have been adopted in metabolic profiling research but instead focuses on the methods that were used in the studies appended to this thesis.

**Nomenclature and definitions**

There are three different traditional approaches that have been adopted in studies of the metabolome [64]. Most current studies can be regarded as being based on one of these:

**Metabolic fingerprinting** is the rapid classification of samples based on a detected chemical profile such as a chromatogram or a spectrum. This can be
achieved using pattern recognition methods. No regard is paid to the identities of the chemical entities behind the profile, and although the approach could in theory be useful for diagnosing diseases, the lack of transparency as to the biochemistry behind the classification means that this approach is not widely used any more.

**Metabolic profiling** is described as the measurement of a pre-defined set of metabolites. These are often members of a specific class of compounds such as amino acids or certain lipid species, but could also be metabolites produced by a particular biochemical pathway.

**Metabolomics** is (for all practical purposes) synonymous with **metabonomics** and refers to the identification and quantification of all metabolites in a biological system. This is obviously not possible with today’s analytical techniques, and in practice these terms are often used to describe so-called unbiased analytical approaches. Obviously all analytical techniques are biased towards something and so “unbiased” in fact means that the methods used are designed to provide an overview of several compound classes and so might better be described as being “untargeted”. Metabolomics datasets often include information on the levels of “metabolites” that have not been identified.

Throughout this thesis and the related manuscripts, I have referred to what I did as “metabolic profiling.” This was done because, with the exception of some species discussed in study 1, all of the metabolites that were studied were identified. Moreover, both targeted and untargeted approaches were employed in all of the experimental work I performed. The main aim of the project was always to employ the analytical techniques that would give the best possible coverage of small metabolites possible in order to maximize the biological information in the obtained data sets.

Metabolites are extremely diverse with regard to physical and chemical properties. Analyzing metabolites in a complex mixture containing perhaps thousands of compounds is therefore a challenging undertaking and requires the use of state of the art analytical techniques.

**Analytical techniques**

The analytical techniques that have been most widely used in metabolite profiling are Nuclear Magnetic Resonance (NMR) spectroscopy and chromatography coupled to mass spectrometry (MS). While other methods have been used in some cases, these two techniques remain dominant in metabolic profiling research [65]. While NMR data are generally more
reproducible than MS data, the number of compounds that can be identified in complex samples is relatively limited. MS is more sensitive and offers greater scope for identifying and quantifying a wide variety of compounds in the complex samples associated with metabolic profiling studies. All of the studies described in this thesis were based on MS, and so mass spectrometric and chromatographic techniques suitable for metabolite profiling are discussed in more detail below.

Coupling chromatography to mass spectrometry, so-called hyphenated analytical equipment, provides a powerful analytical combination. Compounds are chromatographically separated based on chemical properties which provide resolution even with complex samples. By varying instrument setup a very diverse range of compounds can be reliably analyzed. Eluting compounds can be identified based on their mass spectra and quantified based on intensity. The structure of a hyphenated gas chromatography (GC)-MS instrument is seen in figure 2.

Figure 2: The structure and function of a hyphenated GC-MS instrument.

Chromatography is a family of chemical separation techniques and is often used in combination with MS during the metabolic profiling of complex samples. Human plasma is assumed to contain several thousand metabolites; if these were all to enter the detector of the mass spectrometer simultaneously, it would be impossible to know what was in the sample. By separating the compounds in the sample using chromatography, it becomes possible to identify individual constituents.

In GC separation is achieved by letting the samples flow with a carrier gas (Helium in GC-MS) through a silica column that is gradually heated to allow even slow-migrating compounds to elute. Analyzed compounds needs to be
volatile, so non-volatile compounds are derivatized to make them more volatile. A common derivatization method is based on reaction with silica containing compounds (See more in figure 4). Analytes are separated based on the strengths of their interactions with the column, which primarily depend on how similar their polarity is to that of the stationary phase, but also depending on their volatility [66].

Separation by Liquid Chromatography (LC) is achieved by pumping the analytes in a liquid solution through a column with a (usually) hydrophobic stationary phase while gradually changing the polarity of the eluent. The analytes will migrate through the column at different rates depending on their ability to form stabilizing hydrophobic interactions with the stationary phase [66-67].

The principal structure of a mass spectrometer is seen in figure 2. The main components and functions of a mass spectrometer are described below [68].

The sample is introduced into the mass spectrometer in a sample inlet. In metabolic profiling studies this is often coupled to a chromatographic step as described above.

Analytes are ionized in the ionization source. When combined with a GC ionization is usually done using electron impact (EI) ionization. EI is a hard ionization technique where the analyte is fragmented producing a characteristic fragmentation pattern that is useful for identifying the compound. When combined with LC the ionization techniques used in the works of this thesis was electrospray ionization (ESI). ESI is a soft ionization technique and the analyte is thus not split to the same extent was with EI. Instead a high resolution mass analyzer can be used which allows identification of the analyte or ions can be further fragmented as explained later.

When placed in magnetic or electric fields ions exhibit characteristics that are dependent on their mass/charge (m/z) ratio – this is utilized for characterizing them using a mass analyzer. The GC coupled MS was equipped with a time-of-flight (ToF) analyzer; in ToF the generated ions are accelerated in an electric field and allowed to fly a distance in vacuum. The m/z-ratio can be calculated from the flight time. The LC coupled MS had a tripple quadrupole (QqQ) mass analyzer. In a quadrupole ions are deflected from their flight path by exposing them to an electric field. Only ions with certain m/z-ratios will get through the quadrupole. In the QqQ instrument certain selected ions are additionally fragmented in a second quadrupole and the produced ions analyzed by a third quadrupole, a technique known as
tandem MS. This can increase specificity and the amount of structural information.

The detector counts the number of ions hitting it at a certain time. This allows quantification of ions and thus of the analytes that produced them. Since ionization is not equally efficient for all compounds the ratio between different compounds are not necessarily completely quantitative, but using standard compounds for an external calibration curve allows quantification.

MS is a good technique for measuring the abundance of complex molecules such as metabolites because it can be used to both quantify molecular concentrations (using external standards) and provide structural information that permits the identification of the molecule. Depending on the chosen ionization method, it can provide either a high-resolution mass for the analyte or convert it into fragments whose distribution and mass can be useful in identifying the molecule from which they derived.

**Metabolic profiling data from hyphenated instruments**

The aim of the analytical procedures used in metabolic profiling studies is to measure a number of metabolites in all samples, so every sample is characterized by a semi-quantitative or quantitative concentration measure for each metabolite. This thus produces data tables containing one row for each of N samples and one column for each of K metabolites, generating an NxK structure.

The natural output from analysis of one sample on hyphenated analytical equipment such as GC-MS or LC-MS is a time resolved chromatogram with a mass spectrum recorded at each time point as illustrated in figure 3. The desired NxK structure can be achieved by building chromatograms with selective ions that provide pure chromatographic peaks and quantifying these peaks or if sufficiently selective ions are hard to find mathematical resolution of compounds can be done.
Figure 3: Structure of data from hyphenated analytical equipment. Each sample is characterized by a list of ion masses hitting the detector at a given time point.

**Gas Chromatography (GC) coupled MS-method**

The studies reported in this thesis were performed using a GC-MS based workflow developed primarily by Jonsson, Gullberg and A [69-72], who tested and optimized a range of commonly-used extraction protocols and derivatization procedures using experimental design during the method’s development. An overview of the resulting method is presented in figure 4.
Figure 4: Overview the protocol used for extracting metabolites and derivatizing them for GC-MS analysis for a commonly used sample amount.

GC-MS analyses performed in this way generate huge amounts of data because 30 mass spectra are recorded every second, and so extracting metabolite identities and quantities from the resulting data becomes particularly challenging. Jonsson solved this problem by developing a method based on Hierarchical Multivariate Curve Resolution (HMCR) [71], a mathematical technique that can be used to distinguish between signals arising from two or more compounds that elute simultaneously provided that they have different mass spectra and concentration profiles [73]. The computational demand of the curve resolution process is drastically reduced by adopting a hierarchical approach; an overview of this method for data processing method is presented in figure 5. The method is implemented, together with baseline correction and alignment, in an in-house software written in Matlab [74].
Figure 5: Overview of the HMCR procedure: The chromatogram is divided in time windows and each window is resolved using alternating regression. The advantage of the method is that information from all samples is used in identifying compounds while still being computationally efficient. Alternating regression resolves chromatographic and mass spectral profiles of simultaneously eluting compounds.

In my work, I have used Jonsson’s HMCR method for initial extraction of metabolite data from GC-MS spectra. I found the method useful for identification of metabolites, but the quantification often split peaks from more abundant compounds. Therefore, whenever possible, metabolites were quantified using ions that gave pure chromatographic peaks.

Metabolites identities’ were established based on their retention indices and by using the NIST Mass Spectral Search Program to compare their mass spectra to those of known compounds stored in commercial and in-house spectral libraries [75-76].

**Liquid Chromatography (LC) coupled MS-method**

Amino acids were profiled using a method based on the reagent 6-Aminoquinolyl-N-Hydroxysuccinimidyl Carbamate [77], which is now commercially available and widely used. Compounds were separated using UPLC (ultra performance-LC) and quantified using MS or tandem MS, a technique where ions are additionally selected and fragmented to provide
structural information. The chromatographic resolution in this method combined with quantification of selected characteristic ions provides excellent separation of compounds so that no additional mathematical resolution is necessary. The main advantage of this method is that it can be combined with an external standard curve to provide quantitative concentrations. The ability to obtain concentrations for each metabolite rather than the traditional semi-quantitative data provides some obvious advantages – it is easier to check whether the obtained results seem reasonable by comparison to literature values, and most importantly, it makes it possible to compare results from different experiments.

**Future developments**

One of the things I found most frustrating while working on metabolic profiling is that it was very difficult to compare results from different experiments, even when the experimental designs used were similar. This makes proper validation and contextualization difficult and also makes it difficult to identify and understand experimental biases, which can lead to erroneous conclusions. Luckily, because of the steady improvement of analytical instrumentation and strategies, a growing number of studies are using quantitative analytical techniques [78-80]. I believe this development is absolutely critical for the future credibility of the field. Molecular concentrations were measured in one of the studies in paper 3.

Another point of focus for analytical development should be improving metabolic pathway coverage. This will increase the amount of biological information obtained from each experiment, thus providing more knowledge and leaving less room for speculation. This goal can for instance be achieved by using targeted methods for analyzing a relevant pathway or a pre-defined compound class.

**Statistical techniques**

Between recording metabolic profiling data tables on analytical equipment and harvesting new knowledge from the experiments lies the crucial step of data analysis. There are many statistical techniques that can be used for this purpose. Chemometric methods such as latent variable (LV) multivariate approaches have been widely used in metabolic profiling since the field was first established and continue to be used extensively today [63, 81].

Latent variable approaches are attractive in metabolic profiling for a number of reasons, but especially because they are designed for use with datasets describing correlated variables. Such correlations are common in
metabolism, since metabolites interact with each other via numerous biochemical pathways [44]. The use of latent variables also makes it easy to visualize multivariate data sets by reducing their complexity [82]. Two LV-techniques, principal component analysis (PCA) and orthogonal projections to latent structures (OPLS) were employed in the studies comprising this thesis and are described below.

An important distinction is between supervised and unsupervised modeling methods. With unsupervised methods no a priori information about the observations is used during model building and this type of methods can thus be seen as “objective”. When modeling with supervised techniques the model is steered towards a specific solution by including sample information such as group or a particular response value of interest. This allows the study of effects that may not be the largest source of variability in the data.

**Principal Component Analysis (PCA)**

PCA is an unsupervised technique used for compressing high-dimensional multivariate data sets into a smaller number of dimensions (LVs) and thereby making it easier to obtain an overview of trends and patterns. The K variables are summarized using LVs that are uncorrelated with each other and are ordered so that the first variable captures as much of the variation in the data set as possible, with each subsequent LV describing as successively less of the remaining unexplained variation. Each of the original variables has a characteristic influence on each LV that is described in terms of its loading (p) and each sample is described in terms of a score value (t) that is based on the loadings. Several LVs are usually required to capture the relevant variation and so the complete PCA solution (disregarding any pretreatment of the data) has the form [82-83]:

\[ X = TP^T + E \]

With E being the residuals not described by the LV-model.

There are several algorithms for performing PCA. The most common are single value decomposition (SVD) and Non-linear Iterative Projections by Alternating Least Squares (NIPALS). All PCA calculations reported in this work were performed using the NIPALS algorithm. The main feature of the NIPALS algorithm, which I consider an advantage, is that the LVs are calculated sequentially and can thus be added or removed as required [84]. This makes it relatively easy to determine the number of LVs required to describe a given data set since this can be done based on visual inspection of
data such as those presented in score plots rather than by relying on mathematical analyses alone.

PCA was used extensively in papers 2 and 4 and for exploratory data analysis in all of the studies comprising this thesis.

**Orthogonal Projection to Latent Structures (OPLS)**

OPLS is a supervised calibration technique based on the NIPALS algorithm for Partial Least Squares (PLS) modeling [85]. OPLS was originally intended as a pre-processing filter for PLS but was subsequently developed into a full modeling method [86-87]. As with PLS, the purpose of OPLS is to identify correlations between latent variable representations of a data matrix $X$ that contains observations and a second matrix $Y$ that contains responses.

The distinctive feature of OPLS method is that it separates the systematic variation into two components: predictive variation, which overlaps between $X$ and $Y$ matrices and is thus useful for predicting the modeled properties; and orthogonal variation, which is systematic variation that is unrelated (orthogonal) to the modeled properties. In a metabolic profiling study, this orthogonal variation could be information relating to patient age, sex, and dietary habits, or to analytical differences.

In OPLS the $X$ and $Y$ matrices are modeled separately:

$$X = T_P P_T^T + T_O P_O^T + E$$

$$Y = T C_T^T + F$$

The $T$ matrix contains score vectors that (as in PCA) describe the observations, whereas the $X$ and $Y$ loadings in the $P$ and $C$ matrices describe the original variables in relation to the projection. By removing orthogonal variation, it is possible to simultaneously maximize the correlation and covariation between the $X$ and $Y$ matrices. When only a single $y$ exists, all predictive variation is modeled by a single predictive vector ($p_p$) that is directly related to the response. The advantage of this is that it facilitates interpretation without sacrificing predictive ability, which can be difficult to achieve using PLS or other similar methods [86]. As above, OPLS is based on the NIPALS PLS algorithm. The first five steps of both processes are similar, but the OPLS procedure contains additional subsequent steps to remove orthogonal variation. The entire procedure is shown below for a single $y$ case, as used in this work:
1) \( w = X^T y (y^T y)^{-1} \)
2) \( w = w(\|w\|)^{-1} \)
3) \( t = Xw \)
4) \( c = y^T (t^T t)^{-1} \)
5) \( p = X^T t (t^T t)^{-1} \)
6) \( w_0 = p - w(\|p - w\|)^{-1} \)
7) \( t_0 = Xw_0 \)
8) \( p_0 = X^T t_0 (t_0^T t_0)^{-1} \)
9) \( XE = X - t_0 p_0^T \)
10) Return to step 1, using \( XE \) in place of \( X \)

The basis for the OPLS method is the difference between the \( p \) and \( w \) loadings, as discussed extensively by Stenlund [88-89].

OPLS was used as a calibration technique (i.e. with continuous response variables) in paper 3.

**Discriminant analysis (DA)**

OPLS can be used as a classification technique by training the model with a binary dummy matrix describing the class to which each observation belongs [90]. This technique is called OPLS-Discriminant Analysis (DA) [91] and was used extensively throughout this work. Its main advantage for metabolic profiling studies is that in addition to reliable classification results, it generates correlation-scaled loading (\( p(corr) \)) that provide a direct measure of metabolite concentrations in each class.

OPLS-DA was used in all of the papers of this thesis.

**Data centering and scaling**

Multivariate data is typically centered and scaled before modeling [92]. Centering means that the average value for a variable is subtracted from each observation so that the modeled values become distributed around zero. Centering was used in all of the studies that comprise this thesis unless otherwise stated. Scaling is a bit more controversial, and different methods can lead to vastly different modeling results [93]. Commonly used alternatives include not scaling; Pareto scaling, which is achieved by dividing the observed value by the square root of the standard deviation of the variable; and unit variance scaling, which is achieved by dividing each observed value by the standard deviation. The effects of data scaling is illustrated in figure 6.
Unit variance scaling of the data ensures that all variables have the same *a priori* influence on the model since each has the same standard variation. Since metabolite concentrations are not necessarily directly related to the intensity measured in metabolic profiling studies, I think that unit variance scaling is the only approach that can be justified, and so this method was used in all of the studies presented herein. A disadvantage of this approach is that it can cause the importance of metabolites with very little variance to be blown out of proportion. As more metabolite profiling data is recorded, quantitative scaling may become more important, and it is likely that other scaling techniques will have important advantages for identifying relevant biomarkers.

**Model validation and evaluation of model quality**

An important topic in multivariate modeling and metabolic profiling is model validation, which is performed to ensure the reliability of the results obtained. Many different strategies and opinions exist on this topic; some of the flavor of the ongoing debate can be obtained by reading references [94-96]. In this section I discuss the statistical methods and measures used in my own work and their strengths and drawbacks, followed by a discussion of practical considerations and strategies.
Cross-validation (CV) is routinely used to determine the correct number of components for predictive multivariate models and to test their predictive properties [97]. While a number of implementations of this method have been developed, seven-fold CV was used throughout this work as implemented in the software package Simca-P+ [98]. This involves dividing the observations in the data set into seven separate groups, removing one of these groups from the data set, and constructing a model based on the remaining six. This process is repeated seven times, excluding a different group in each case. The resulting models are then used to predict the values of the corresponding excluded observations, thereby providing a measure of the model’s predictive ability. Measures obtained from CV include the Goodness of Prediction (Q₂) value and the Root Mean Square Error of Cross-Validated prediction (RMSECV):

\[ Q^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \]

\[ RMSECV = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{N}} \]

Here, \( \hat{y}_i \) is the response value calculated by the model with the observation excluded, \( y_i \) is the observed response value, \( \bar{y} \) is the average response value, and \( N \) is the total number of samples.

The Q₂ value ranges from negative infinity to 1, with a value of 1 indicating perfect prediction for all sample values and 0 or negative values indicating a predictive precision that is comparable to or worse than guessing the average value. It is difficult to identify a specific value between 0 and 1 that is indicative of a good model because any such value would be highly application-dependent. RMSECV gives a measure of precision in the original units and can thus be interpreted as the expected average error of each prediction.

Another recent diagnostic based on CV is the CV-ANalysis Of VAriance (ANOVA) test [99]. Here the mean square of the regression is tested against the mean square of the residuals using the F-distribution. This gives a formal and easily understandable measure of the quality of the model.

A widely used measure that is not based on CV is the Goodness of Fit (R₂) value, which is calculated using the formula:
Where $x_i$ is the measured value, $\hat{x}_i$ is the modeled value and $\bar{x}$ is the average value.

$R^2 = \frac{\sum_i(x_i - \bar{x})^2}{\sum_i(x_i - \bar{x})^2}$

$R^2$ can be interpreted as the share of total variation in the dataset that is described by a given model or model component. $R^2$ values are particularly useful in the context of OPLS because they provide a measure of how much variation is attributable to a certain trait. This property was utilized and discussed extensively in paper 3. $R^2$ values are also useful for PCA models because they give a measure of how much of the variation is captured in each component of the model.

**Validation strategies**

CV is an example of what is called internal validation, meaning that the samples used to build the model are also, at some point, used to validate it. A more strict way of testing metabolic profiling models entails testing their ability to predict the properties of completely new samples; this is known as external validation since the observations used for model validation are not used for training in any way [92, 95]. The relative reliability of internal and external validation has been discussed extensively, albeit often from a purely philosophical perspective.

In purely statistical terms, it has been shown that well-performed internal validation is superior to splitting a given dataset into separate test and training sets [96, 100]. However, this argument is only valid in cases where only a single or limited dataset is available. Often, when validating using a single dataset, the primary concern is to validate the statistical properties of the model (e.g. to demonstrate that it is not over-fitted), regardless of whether internal or external validation is used. In metabolic profiling, it may be more important to validate the properties of the dataset, i.e. to test things such as how reproducible the measurements were, how representative the sampling was, and how stable metabolic profiles are over time. Answering these questions was one of the greatest challenges encountered in the course of this work, and the only realistic way of doing it is to perform repeated experiments with multiple samplings from the same individual. This achieves validation that is external in biological and analytical terms as well as from a statistical viewpoint.
Univariate techniques

The univariate statistical methods employed throughout this work are all commonly used methods such as Student’s t-test for assessing differences in metabolite levels between groups [101] and Pearson correlation (r), which was primarily used to compare the similarities of metabolic profiles to one-another [102].

Measures of specificity and sensitivity were used in paper 1 to describe the models’ diagnostic properties:

\[
Specificity = \frac{\text{#Correctly classified as having disease}}{\text{#Patients having disease}}
\]

\[
Sensitivity = \frac{\text{#Correctly classified as healthy}}{\text{#Healthy persons}}
\]

Specificity is thus a measure of the model’s ability to identify patients as having the disease, while sensitivity indicates the model’s ability to identify healthy persons as being healthy [103].

One of the challenges of metabolic profiling studies is to find a way of showing results for several measured metabolites for a large number of patients in a relatively confined space. In paper 1, this was achieved by calculating the Standard Score (Z-score) for each metabolite and plotting these in heat maps:

\[
Z = \frac{x - \mu}{\sigma}
\]

Where x is the measurement value and \( \mu \) and \( \sigma \) are the average and standard deviation of the group being standardized against.

The primary objective of the study described in paper 2 was to detect differences between different groups and over time. This was achieved by calculating the fold change:

\[
Fold \ change = \frac{\mu_{group \ 1}}{\mu_{group \ 2}}
\]

The main advantages of these methods are that they are well known and understood by most researchers and there is a relative consensus on how
they should be used. They are thus useful supplements to multivariate methods that facilitate the interpretation of results.

**Considerations regarding data analysis**

One of the main disadvantages surrounding multivariate statistical methods in my opinion is that there are too many methods for doing very similar things. The consequence can be that relatively few people are familiar with all of the available methods, which can make it difficult to communicate results to others. A related problem is lack of standardization concerning how a given method is best used. For instance with supervised LV-methods it can have a very different visual impact whether normal score plots or cross-validated score plots are presented. This problem can be alleviated by showing predicted results rather than scores, and supplementing analysis with well-known measures such as correlation coefficients, t-tests or loadings with jack-knifed confidence intervals [104] instead of just unscaled loadings.

**Discussion on classification, biomarkers and disease diagnosis**

One of the central terms used in all profiling studies is “biomarker”. The “Biomarkers Definitions Working Group” agreed on the following definition:

- **Biological marker (biomarker):** A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [105].

What is good about this definition is that it stresses that the biomarker is a measure of a process, meaning that it has a biological or pathological meaning. This stands in contrast to the term’s usage in many profiling studies whose primary objective is to classify subjects [106]. Statistical validation is definitely important, but understanding biological and pathological processes is even more important. For example, C-reactive protein (CRP) is a very commonly-measured biomarker that is used when studying conditions such as pneumonia, RA and atherosclerosis [107]. In all of these conditions, CRP exhibits low specificity but it is of great value because the biological / pathological interpretation of the measured values is straightforward: increases in CRP abundance indicate the activity of inflammatory processes. In combination with clinical examination, this constitutes valuable information that can form the basis of a medical intervention. A hypothetical marker profile discovered by fingerprinting techniques would not have any understandable relation to disease processes and so it would need almost perfect sensitivity and specificity to provide a
useful basis for medical intervention. This demonstrates that it can be more important to understand the biological factors that make a given candidate a potentially useful biomarker than for it to have particularly attractive statistical properties, even though such an understanding may be difficult to achieve.

An understanding of biological and pathological mechanisms makes it possible to formulate testable hypotheses, which is necessary to link profiling studies and clinical applicability. It is also probably the most effective way of avoiding problems with bias and false discoveries due to multiple hypothesis testing [108-109].

A major aim throughout all of the studies presented in this thesis was to understand the metabolic processes behind the discovered perturbations, but this proved to be a much more difficult task than merely measuring metabolite levels and identifying statistically significant differences. Determining the biological relevance of the results obtained proved to be one of the most challenging and important aspects of the studies. This was done by extensive reproduction of the experimental results. In all of the studies, separate validation experiments were performed or repeated sampling was undertaken to ensure that the identified metabolic patterns were genuinely relevant and to provide focus for subsequent biological analyses. This proved to be an excellent way of avoiding spurious results.
Results – Discussion of papers

Paper 1

Objective: The aim of this study was to systematically investigate metabolic perturbations in patients with RA and assess the scope for developing a diagnostic method based on the profiled metabolites. An important aspect of the study was that the results were validated by classifying samples from a separate patient population.

Results: In a separate study, it was found that samples from patients with RA could be distinguished from those for healthy controls with 93% sensitivity and 70% specificity. Patients with RA could be distinguished from patients with psoriatic arthritis with 90% sensitivity and 94% specificity based on cross-validation. Compared to samples from patients in other groups, samples from patients with RA contained higher concentrations of glyceric acid, D-ribofuranose, hypoxanthine and several free fatty acids, along with lower concentrations of histidine, threonic acid, methionine, cholesterol, asparagine and threonine. An important component in this study was the use of quantile normalization, which made it possible to predict the results for samples from a separate analytical run [110].

Discussion and epilogue: Some of the identified metabolic changes in RA patients had previously been reported but there had been no systematic investigation into their diagnostic properties. The aim of this work was therefore to take a more systematic approach to this question. The sensitivity of the metabolite based method used was better than that for antibody-based tests, but its specificity was worse [111]. The most recent diagnostic criteria for RA place heavy emphasis on the detection of auto-antibodies [23]. Consequently, the most likely clinical use for a metabolite based approach would be to provide an “objective” measure in patients without auto-antibodies.

Should this method be introduced in a clinical setting, it would be necessary to test it in prospective studies. In practice, it would probably be best to focus on a small panel of metabolites. Another concern regarding this study was that some of the patients were receiving treatment for RA. The results for both patients with and without treatment were compared during the evaluation of the study and it was found that treatment did not define their metabolic profiles. However, treatment could certainly affect metabolic profiles in principle. Systematic tests should therefore be performed in patients with untreated RA. Doing so would obviously be very demanding,
and so the results obtained in this work are important for determining whether it would be worthwhile to perform such a study.

We have not so far done any more work aimed specifically at diagnosing RA. The main reason for this is that the currently available diagnostics work quite well and other clinical problems such as finding markers for disease activity and treatment efficacy were considered to be more interesting and important.
Paper 2

Objective: Based on reports of metabolic changes in children who subsequently developed T1D, we decided to investigate whether similar changes occurred in NOD mice, which are widely used models for T1D. While this work was being conducted, it was also reported that there were metabolic differences between NOD mice that progress to diabetes and NOD mice that do not [112]. Our approach was somewhat different in that we used C57BL/6 mice as a control group for the NOD mice. The idea behind this was that it would be possible to correlate any metabolic differences to genetic strain differences, which would provide insights into the genetic origins of the metabolic differences.

Results: Female NOD and C57BL/6 mice were sampled at 0, 1, 2, 3, 4, 5, 6, 7, 9, 11, 13 and 15 weeks and their serum was analyzed for metabolite content. Metabolic differences between strains were seen throughout the study, with both groups showing a larger development of metabolism during the first five weeks with subsequent stabilization. Inosine, glutamic acid, citric acid cycle compounds and nucleic acid metabolites were amongst the most discriminatory metabolites. Genes responsible for the metabolism of these compounds have been identified in previously-defined insulin dependent diabetes (Idd) regions.

Discussion and Epilogue: The metabolic differences between NOD and B6 mice were found to partially resemble those between progressing and non-progressing children. It would be helpful to build on these findings by performing modification studies, for instance by introducing transgenes for important metabolic genes or by supplementation / restriction of certain metabolites to see how the incidence of diabetes is affected. This will be necessary to fully understand the pathological relevance of the identified metabolic differences.
Paper 3

Objective: RA is a lifelong disease with periods of remission intermittent between periods of more severe symptoms. Disease activity is followed using the DAS28 score, but the discovery of laboratory markers that correlate with disease activity would be highly desirable as a complement to clinical examinations. The main aim of this study was to determine whether the metabolic profiles of RA patients beginning anti-TNF treatment correlated with DAS28 scores. Patients beginning anti-TNF treatment were selected because they were expected to show a large change in disease activity in the first years of treatment. An additional aim was to identify markers for treatment response.

Results: In two studies following a total of 56 patients over their first 18–24 months of anti-TNF treatment, it was found that the obtained metabolic profiles did indeed correlate with DAS28 scores. However, the most discriminatory metabolites identified in the two studies were quite different. The most stable pattern observed was a reduction in the levels of several amino acids (tryptophan, methionine, lysine, serine, leucine, isoleucine and valine) with concomitant increases in the levels of arginine, 2,5-diaminovalerolactam, homoserine and guanosine. This was believed to be an effect of increased inflammation with increased acute phase protein synthesis and increased cell turnover. Early markers for treatment response could not be reproduced in both studies.

Discussion and epilogue: The main take-home message from this study was that results should be carefully validated. Even though the patients in the two studies were being treated at the same clinic, we observed large differences that could not be linked to any clinical features.

Whether the method will be applicable in a clinical setting remains to be seen. The correlations for the identified markers were probably not strong enough to provide much additional value. The second of the studies reported in this paper yielded quantitative data on the concentrations of all of the studied amino acids, which would make it possible to directly compare the results obtained to those of potential future studies. This should make it easier to accumulate knowledge about metabolic variation over time in patients with RA.
**Paper 4**

**Objective:** To determine whether a mutation in *Ncf1* caused metabolic changes in B6 mice. Additional aims were to investigate the stability of the resulting metabolic changes and to understand their biological relevance.

**Results:** We showed that the *Ncf1* mutation has consistent metabolic effects outside a disease setting. It was found that *Ncf1* expression in monocytes restored the metabolic phenotype, showing that regulatory effects of this cell type are the major important factor controlling the metabolic effects. The metabolic effects were subtle and thus different throughout the studies. The most reproducible features were increases in the levels of the free fatty acids linoleic acid, oleic acid, docosahexanoic acid, palmitic acid, palmitoleic acid and stearic acid, along with uridine and inositol-1-phosphate in oxidizing genotypes. Levels of several amino acids (e.g. cysteine, cystine, phenylalanine, methionine, and tyrosine) as well as inositol and creatinine were reduced in oxidizing genotypes.

**Discussion and epilogue:** The observed differences between genotypes were consistent with previous findings from studies where disease models had been induced. Genotype differences have, however, been difficult to prove under physiological conditions. The most apparent metabolic changes occurred among lipid compounds, indicating that the differences may be related to the production of immunoregulatory compounds. As a follow-up study, it would be interesting to perform a more exhaustive profiling of lipids (including down-stream oxidation products of arachidonic acid) following *Ncf1* mutation.
Conclusions and future perspectives

The overall conclusions of these studies can be related to the title of this thesis. We certainly observed metabolic variation when studying patients with autoimmune diseases or models thereof. However, it is equally clear that not all of the observed variation was really relevant to the studied diseases. A major focus throughout this work therefore was to validate the results obtained by repeating experiments, and it was found that this is absolutely essential in order to be confident in the results and also to identify the most relevant biological features.

Looking at the stable patterns identified, it is difficult to see any common mechanism underpinning the metabolic differences seen in T1D and RA despite their genetic similarities and common pathological features. The differences seen between NOD and B6 mice were primarily related to energy metabolism. A recurring theme in RA was a reduction in the levels of amino acids, which we believe may reflect a redirection of protein synthesis towards the production of acute phase proteins. We have also seen examples of the importance of lipid metabolism in regulating the immune system. This demonstrates that although common genes govern predisposition important events inducing specific autoimmune reactions still remain to be understood.

In many ways, I feel that the work of understanding metabolic variation in autoimmune diseases has only just begun. On reflection, the results of the studies performed in this work often posed more questions than they answered. The type of metabolic screening that was performed is useful for generating hypotheses but it is often not possible to see the whole picture using only a relatively limited set of metabolites. As the field of metabolic profiling matures, I believe it will be obliged to move towards more targeted profiling methods. This will be necessary to get a full understanding of the mechanisms that underpin metabolic changes – one must know where the metabolite goes when its levels decrease. Another crucial development will be the development of quantitative profiling techniques, which will make it much easier to compare results from different studies and to determine whether results are reasonable.

The metabolic profiling community should also consider the merits of the myriad of techniques, methods, approaches and strategies that have been developed for every step in metabolic profiling studies. Having too many accepted methods makes it difficult for even experts to understand and evaluate others’ studies and results. An emphasis on quantitative rather than semi-quantitative results will increase the transparency of the actual data.
structures. As a tool for data analysis, OPLS has many desirable properties in that model statistics can be used to assess model quality and the amount of variation related to the studied features, but it is obviously necessary for authors to report these results when presenting OPLS analyses. The most important thing is to at least agree on standards for reporting results that facilitate transparency and comparability.

With the ongoing development of methods, metabolic profiling techniques could become an attractive tool for elucidating pathological mechanisms in a wide range of diseases. This could possibly lead to the discovery of new biological markers for diseases and treatment responses, although it remains challenging to identify metabolite patterns that are specific enough to serve as a foundation for making clinical decisions.

My personal view is that the philosophy of hypothesis-free investigation championed by some researchers should not be stretched too far. The studies conducted in this work clearly show that this approach can easily result in the identification of patterns with little relevance. Instead, I believe that the future lies in applying targeted methods to carefully defined research questions. This demands close collaboration between analytical chemists, data analysts and medical professionals. Such collaborations could be truly exciting for all involved parties and have the potential to revolutionize our understanding of many diseases.
Acknowledgements

Till min handledare Johan Trygg: tack för att du gav mig chansen på detta projekt och också tack för att du lärde mig att jobba självständigt och för att du låt mig följa mina egna idéer och drömmar under tiden!

Min biträdande handledare Thomas Mortiz – vi har kanske inte alltid haft så mycket kontakt, men du har alltid ställt upp om det behövdes och det känns tryggt. Tack!

Tack till Torbjörn Lundstedt för uppstart av jobbet med RA och tack för allt motspel du har givet mig under tiden.

En speciell tack till Solbritt Rantapää-Dahlqvist för att du ställde upp trots din skepsis. Det hade inte blivit mycket doktorand-projekt utan dig och du har lärt mig mycket om medicinsk tänkande som har varit båda viktigt och roligt.

Kristina Lejon: Tack för att du delade av dig med projektet och för ditt oerhört stora engagemang, samt för bra kommentarer och frågor – båda under och efter projektet. Also thanks to Viqar Showkat Banday for the coorporation on the NOD project.

Tack till Rikard Holmdahl med grupp (speciellt Tiina Kelkka, Bruno Raposo and Angela Pizzolla) – det har varit jättekul att få vara med på en hörna av eran forskning!

Stort tack till Inga-Britt Carlsson och Krister Lundgren för hjälp och expertis i labbet. Hade inte varit mycket forskning här utan er! I samma vända tack, Annika Johansson som också har hjälp med mycket praktisk under tiden.

Tack till Anna Linusson-Jonsson och Erik Björn för eran hjälp med att styra upp projektet när det behövdes.


Tak til mine forældre **Søren** og **Inger**, til mine brødre **Rune** og **Jakob** og til mormor **Kirsten** for at i altid har støttet mig igennem alt også når mine valg måske ikke var så lette at forstå.


Och till slut tack till alla som jag har träffat i samband med medicin studierna och i psykiatrin, känns som jag äntligen har hamnat rätt (med medicin alltså) och ni har en del av äran för det.
References


20. de Vries R: *Genetics of rheumatoid arthritis: time for a change!* *Curr Opin Rheumatol* 2011, **23**:227-232.


22. Raptopoulou A, Sidiropoulos P, Katsouraki M, Boumpas DT: *Anticitrulline antibodies in the diagnosis and prognosis of*


40. Pflueger M, Seppanen-Laakso T, Suortti T, Hyotylainen T, Achenbach P, Bonifacio E, Oresic M, Ziegler AG: Age- and Islet Autoimmunity-Associated Differences in Amino Acid and
**Lipid Metabolites in Children at Risk for Type 1 Diabetes.** *Diabetes* 2011, **60**:2740-2747.


74. www.mathworks.com


86. How to create an OSC filter for PLS and end up with a new generic modelling method, O-PLS [http://www.chemometrics.se/editorial/april2002.html]


98. [www.umetrics.com](http://www.umetrics.com)


