Multivariate Profiling of Metabolites in Human Disease
Method evaluation and application to prostate cancer

Elin Thysell
To my family
Abstract

There is an ever increasing need of new technologies for identification of molecular markers for early diagnosis of fatal diseases to allow efficient treatment. In addition, there is great value in finding patterns of metabolites, proteins or genes altered in relation to specific disease conditions to gain a deeper understanding of the underlying mechanisms of disease development. If successful, scientific achievements in this field could apart from early diagnosis lead to development of new drugs, treatments or preventions for many serious diseases.

Metabolites are low molecular weight compounds involved in the chemical reactions taking place in the cells of living organisms to uphold life, i.e. metabolism. The research field of metabolomics investigates the relationship between metabolite alterations and biochemical mechanisms, e.g. disease processes. To understand these associations hundreds of metabolites present in a sample are quantified using sensitive bioanalytical techniques. In this way a unique chemical fingerprint is obtained for each sample, providing an instant picture of the current state of the studied system. This fingerprint or picture can then be utilized for the discovery of biomarkers or biomarker patterns of biological and clinical relevance.

In this thesis the focus is set on evaluation and application of strategies for studying metabolic alterations in human tissues associated with disease. A chemometric methodology for processing and modeling of gas chromatography-mass spectrometry (GC-MS) based metabolomics data, is designed for developing predictive systems for generation of representative data, validation and result verification, diagnosis and screening of large sample sets.

The developed strategies were specifically applied for identification of metabolite markers and metabolic pathways associated with prostate cancer disease progression. The long-term goal was to detect new sensitive diagnostic/prognostic markers, which ultimately could be used to differentiate between indolent and aggressive tumors at diagnosis and thus aid in the development of personalized treatments. Our main finding so far is the detection of high levels of cholesterol in prostate cancer bone metastases. This in combination with previously presented results suggests cholesterol as a potentially interesting therapeutic target for advanced prostate cancer. Furthermore we detected metabolic alterations in plasma associated with metastasis development. These results were further explored in prospective samples attempting to verify some of the identified metabolites as potential prognostic markers.

Keywords

Metabolite profiling, metabolomics, predictive metabolomics, mass spectrometry, GC-MS, biomarkers, chemometrics, design of experiments, multivariate data analysis, prostate cancer, bone metastases, plasma
Sammanfattnings på svenska


Metabolomik eller global metabolitprofilerings avser att detektera metabola fingeravtryck i biologiska prover (t.ex. blodplasma, urin eller vävnad) genom analys med känsliga bioanalytiska instrument. Bakgrunden till metodiken är att halter av metaboliter förändras enligt specifika mönster vid exempelvis sjukdom eller miljöpåverkan. Dessa mönster kan liknas vid ett unikt fingeravtryck karaktäristiskt för det studerade systemets fysiologiska tillstånd. Trenden inom metabolomikområdet går mot användandet av allt mer känsliga analytiska tekniker, vilka skapar stora mängder informationsrika data. Till de mest använda analytiska teknikerna hör kärrkromatografi kopplat till masspektrometri (GC-MS) och vätskekromatografi kopplat till masspektrometri (LC-MS). De komplexa fingeravtryck som dessa tekniker genererar, bestående av halterna av hundratals metaboliter, analyseras sedan med kemometriska eller multivariata analysmetoder (t.ex. PCA, PLS och OPLS) för att skapa tolkningsbara kartor över förändringar i metabolismen.

Syftet med studierna i denna avhandling var att utveckla metoder för att kunna använda metabolomik inom medicinska frågeställningar med målet att öka möjligheterna attagnostisera, förstå och i förlängningen behandla allvarliga sjukdomar.

För att skapa förutsättningar för detektion och identifiering av metaboliter har vi inom vår forskningsgrupp utvecklat en strategi för kurvupplösning kallad hierarkisk multivariat kurvupplösning (H-MCR). Förenklad är kurvupplösning en matematiskt metod som gör det möjligt att kvantifiera och identifiera metaboliter som inte kan separeras analytiskt genom exempelvis kromatografi. Vi har i detta arbete visat att i kombination med multivariata analysmetoder gör H-MCR det möjligt att skapa system för tolkning av metabola processer, detektion och identifiering av biomarkörer eller biomarkörmönster samt prediktion av oberoende prover. Dessutom visar resultaten att detta kan göras tidseffektivt i stora provserier.

De utvecklade metoderna har använts i studier som syftar till att förstå de mekanismer som ligger bakom att vissa fall av prostatacancer utvecklas och blir aggressiva och livshotande för patienten, medan andra förblir långsamväxande och ofarliga. Ofta är det närvaron av dottertumörer, metastaser, som bestämmer allvaret vid en cancersjukdom. Inom ramen för denna avhandling försöker vi förstå vad som reglerar tillväxt av bennmetastaser vid aggressiv prostatacancer för att i förlängningen förstå hur dessa bör förebyggas och behandlas. Det övergripande syftet är att utveckla känsliga metoder för att hitta tumörer och för att förutsäga
hur farliga de är, så att patienter kan få rätt behandling i ett botbart skede av sjukdomen. Förhoppningsvis kan dessa markörer även bidra till utvecklingen av nya behandlingsmetoder för prostatacancer.


List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals. In the list of papers Thysell E and Johansson ES refers to the same person.


*Paper I and II* are reprinted with permission from the *American Chemical Society* and *Mary Ann Liebert, Inc.* respectively.
Other papers by the author not appended in the thesis


Notations

The following notations will be used in this thesis. Matrices are represented by bold, capital letters e.g. C and vectors are denoted by bold, lower case letters, e.g. p. Vectors are column vectors unless stated otherwise.

C Matrix of weight vectors for Y, [MxA]
E Residual matrix of predictor variables, [NxK]
F Residual matrix of response variables, [NxM]
P Matrix of loading vectors for Y, [KxA]
T Matrix of score vectors for X, [NxA]
X Matrix of predictor variables, [NxK]
Y Matrix of response variables, [NxM]
p Loading vector for X, [Kx1]
t Score vector for X, [Nx1]
p predictive
o orthogonal
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>CoA</td>
<td>coenzyme A</td>
</tr>
<tr>
<td>CRPC</td>
<td>castration-resistant prostate cancer</td>
</tr>
<tr>
<td>CV</td>
<td>cross validation</td>
</tr>
<tr>
<td>CYP11,</td>
<td>cytochrome P450, family 11</td>
</tr>
<tr>
<td>CYP17</td>
<td>cytochrome P450, family 17</td>
</tr>
<tr>
<td>DA</td>
<td>discriminant analysis</td>
</tr>
<tr>
<td>DIMS</td>
<td>direct infusion mass spectrometry</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DoE</td>
<td>design of experiments</td>
</tr>
<tr>
<td>FTIR</td>
<td>fourier transform infrared spectroscopy</td>
</tr>
<tr>
<td>GC</td>
<td>gas chromatography</td>
</tr>
<tr>
<td>HSD3B2</td>
<td>hydroxy-delta-5-steroiddehydrogenase, 3 beta- and steroid delta-isomerase 2</td>
</tr>
<tr>
<td>H-MCR</td>
<td>hierarchical multivariate curve resolution</td>
</tr>
<tr>
<td>IGF1</td>
<td>insulin-like growth factor 1</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MVA</td>
<td>multivariate data analysis</td>
</tr>
<tr>
<td>NIPALS</td>
<td>non-linear iterative partial least squares</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance spectroscopy</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>OSC</td>
<td>orthogonal signal correction</td>
</tr>
<tr>
<td>OPLS</td>
<td>orthogonal projections to latent structures</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PLS</td>
<td>partial least squares</td>
</tr>
<tr>
<td>PSA</td>
<td>prostate specific antigen</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>reverse transcription polymerase chain reaction</td>
</tr>
<tr>
<td>SVD</td>
<td>single value decomposition</td>
</tr>
<tr>
<td>TIC</td>
<td>total ion current</td>
</tr>
<tr>
<td>TOF</td>
<td>time of flight</td>
</tr>
</tbody>
</table>
# Table of Contents

Abstract iii  
Sammanfattning på svenska v  
List of papers vii  
Notations ix  
Abbreviations x  
Table of Contents xi  
Background 2  
Metabolomics 3  
*Metabolomics in prostate cancer research* 6  
Multivariate data processing and analysis 8  
*Hierarchical multivariate curve resolution* 9  
Chemometrics 11  
*Predictive metabolomics* 14  
Aims of the study 17  
Results 18  
Paper I 18  
Paper II 20  
Paper III 22  
Paper IV 24  
Paper V 26  
Conclusion and future perspectives 28  
Acknowledgements 31  
References 33
Background

What can you learn from a simple blood test during a routine medical check-up? For instance you can be told if you have an infection, determine your mineral content, and monitor or assess the effectiveness of a drug. The level of glucose in serum can be used to diagnose diabetes, elevated levels of human chorionic gonadotrophin confirm early pregnancy, and the level of creatinine in serum provides information about renal function. Imagine, if you could also diagnose cancer, at an early stage of development, with a blood test within minutes.

The physiological and biochemical composition of a blood sample can be regarded as a signature of biomarkers, reflecting the biological state of an individual. When the balance in a living system is disturbed it will be reflected in the biological components of that system. Biomarkers are naturally occurring components of the cell that can be used as indicators of a specific physiological state, such as disease or other aspects of health.¹ The clinical applications of biomarkers are, for example, disease detection, identification of people at risk for developing disease, detection of recurrence after therapy, or to guide personalized treatment.²⁻⁴
Cancer is a term used for diseases where transformed cells divide without control and are able to invade other tissues (The National Cancer Institute, NCI\(^5\)). Over recent years, knowledge of the complex processes that act to overrule normal biological regulation during cancer progression has increased dramatically. Today, one of the great challenges in cancer research is to gain mechanistic understanding of cell transformation that can further lead to detection of suitable biomarkers for cancer. Such biomarkers could represent e.g. altered patterns of gene expression, inflammation, hyperplasia, or hyperproliferation.\(^6\) Detecting cancer at as early stage as possible is directly related to the effectiveness and outcome of the treatment of the patient.\(^7\) However, many of the existing markers are not sensitive enough to detect early stages, but instead work well for the detection of late stage tumors, and prognostic information is often definitive only for patients with an already fatal disease.\(^8\)-\(^10\) In order to work also for the detection of cancer at early stages, biomarkers need to be released in to the circulation in measurable amounts early in disease progression, even by a small asymptomatic tumor. It is also important that the level of the marker molecule is not affected by non-cancer disease, since the ability to detect cancer is then compromised greatly. It is also crucial that the marker have a high specificity for the tissue of origin. If not, it is likely that the levels in healthy individuals produced by other tissues will be high and overlapping with the levels measured in patients.\(^11\) The discovery of biomarkers for early disease detection in cancer is extremely difficult and researchers are putting a lot of effort and resources into the discovery of cellular components that can be used to diagnose disease.\(^12\)-\(^15\)

This thesis focuses on development and application of strategies for studying metabolic alterations in human tissues and plasma associated with disease. The developed strategies are applied to the study of prostate cancer, with the aims to identify possible prognostic metabolite markers associated with aggressive disease and development of bone metastases. The resulting metabolites or metabolite patterns could be of value as novel biomarkers but also as providers of new mechanistic insights for prostate cancer progression.

**Metabolomics**

At the same time as cells grow and divide, they consume and produce small chemical molecules known as metabolites. Metabolites are often referred to as low molecular weight compounds (\(<1\) kDa) involved in chemical reactions that occur inside the cells of living organisms to uphold life, i.e. the metabolism. The chemical diversity of the metabolome, defined as the
The complement of all detectable metabolites, is large and includes a wide range of compound classes, e.g. carbohydrates, amino acids, organic acids, sterols, nucleosides (Figure 1). The quantity and number of metabolites vary with changing conditions such as environment, diet and in response to disease. The research field of metabolomics seeks to understand the relationship between expressed metabolites in human tissues and biological mechanisms by studying differences between samples in relation to their metabolic composition. In order to detect these differences, the hundreds to thousands of metabolites that are present in a sample need to be quantified. This is done using sensitive bioanalytical techniques. In this way a unique chemical fingerprint is generated for each sample that represents the metabolic composition of the sample.

![Figure 1](image.png)

**Figure 1** Metabolites representative for the chemical diversity of the metabolome.

Metabolites are end products of the hierarchy starting with genes (genome) and ranging to the collection of gene transcripts (transcriptome) and proteins (proteome) (Figure 2). In the field of systems biology, information from 'omics' analyses is combined to elucidate the interactions between genes, proteins and metabolites. Although genomics, transcriptomics, proteomics, and metabolomics should be considered as complementary techniques, metabolomic profiles are regarded as containing integrated information about the events that take place at different levels in the organism as a result of their particular location downstream of the genome, transcriptome and proteome. The collection of metabolites in a sample, i.e. the sample metabolome, is highly dynamic, changing over time. Global metabolic profiling by metabolomics, therefore, provides a direct picture of the current state and phenotype of an organism, allowing discovery of clinical biomarkers or biomarker patterns. The size of the metabolome varies from a few hundred to a few thousand metabolites depending on the...
organism studied (not including lipids). In addition, over 100,000 molecules are believed to be present in humans due to consumption of food, drugs etc. In recent times, metabolomics has evolved from a conventional profiling technique into one used to study biological systems as an interacting system of genes, proteins, metabolites and cellular events. By combining information from genes, proteins and metabolites, global models of biosystem function can be generated. This global analysis can help to further identify possible diagnostic and prognostic biomarkers as well as uncover altered pathways and thereby understand disease processes.

Metabolomics analyses involve several important steps before a reliable biological interpretation can be carried out. These steps include planning the experiment, sampling, sample handling and preparation, chemical analytical analysis, data processing, statistical analysis or modeling, and validation. The output of this chain of events is imperative for drawing correct conclusions about the biological system under investigation. Hence, many of these steps need to be carefully standardized, optimized, and/or validated. This includes standardized protocols for sampling, sample handling and instrumental analysis, optimized strategies for sample preparation, and validation of statistical models. Guidelines on reporting of studies and methods have been recently suggested by the Metabolomics Standardization Initiative (MSI).

![Diagram](image_url)

**Figure 2.** Summary of hierarchical levels in the organization of the cell involved in the progression from genotype (hereditary information) to phenotype (structure and function).
**Metabolomics in prostate cancer research**

Cancer is one of the leading causes of death in economically developed countries and the incidence is increasing as a result of population aging and an adoption of lifestyle choices associated with cancer, such as smoking, diet and physical inactivity. Prostate cancer is one of the most common cancers in the Western world and roughly 10,000 men are diagnosed with the disease every year in Sweden according to Swedish official statistical data (Socialstyrelsen). For those diagnosed with prostate cancer, prediction of clinical outcome is primarily performed by digital rectal examination, transrectal ultrasound guided biopsy and measurement of serum levels of prostate specific antigen (PSA). If increased PSA levels are detected, prostate biopsies are taken and cancer diagnosis is confirmed or refuted microscopically. At present, most of the men diagnosed have PSA levels between 4-10 ng/mL. In this range, the specificity of PSA testing is only 20-30 %. The low specificity is due to the fact that PSA is prostate, not prostate cancer, specific and leaks into the circulation not only due to cancer but also during inflammation, prostatitis and benign prostate hyperplasia (BPH). As a result many men that undergo a prostate biopsy examination will not have prostate cancer. To reduce unnecessary biopsies, tests that can rule out cancer alone or that can increase specificity in combination with PSA are needed.

The prognosis of prostate cancer in early stages is highly variable. A large group of men may have tumors that are highly unlikely to cause symptoms for many years. If such indolent tumors could be identified at the time of diagnosis, many men could be spared unnecessary treatments. On the other hand, about one fourth of the cases suffer from aggressive, fatal cancer (Socialstyrelsen) eventually spreading to distant sites in the body. Current diagnostic methods cannot separate indolent tumors from the aggressive forms, nor predict metastatic prostate cancer at a curable stage of the disease. The standard therapy for patients with advanced prostate cancer is to reduce the amount of circulating androgens, either by surgical or medical castration. Castration therapy decreases androgen receptor signaling and causes decreased tumor cell proliferation and increased tumor cell apoptosis, which is probably preceded by decreased blood supply to the tumor. After a period of initial remission tumors relapse and by that time they are no longer responsive to castration treatment, and are so termed castration-resistant prostate cancer (CRPC). Therapeutic options for patients with CRPC are limited and in general the best outcome for these patients is to maintain, or to improve, their quality of life.
Limitations of current diagnosis methods have led to an intense focus on the potential of molecular biomarkers to improve both diagnostics and individual prognostics.

When cancer cells grow their metabolism is altered and optimally the metabolites that are produced could be detected and used as biomarkers for the disease. Tumors cells are known to have different metabolic profiles compared to normal cells. These general properties include, for instance, increased aerobic glycolysis, production of lactate and higher bioenergetic demands. Metabolic alterations occur early in the neoplastic transformation of prostate cells. Over the past decade, changes in choline phospholipid metabolism have been associated to aggressive forms of cancer including brain, prostate and breast cancers (reviewed by Glunde). Choline serves as precursor for cell membrane phospholipids and as a methyl donor in DNA methylation and DNA repair mechanisms, however the complete role of choline metabolism in prostate tumorigenesis is still unknown.

Prostate cells have a unique metabolite profile among human organs by producing high concentrations of the major components in prostate fluid; citrate, PSA, polyamines e.g. spermine and myo-inositol. In the process of abnormal and uncontrolled growth, prostate cells shift from accumulating and secreting citrate into an active oxidation state where chemical energy is generated from citrate through the citric acid cycle by the production of ATP. The increased level of ATP is used to nourish the large bioenergetic demand in malignant prostate cells. Another metabolic change related to high proliferation rate found in malignant cells is increased lipid biosynthesis, used to produce building blocks for membrane formation and intercellular signaling. This is achieved, for example, by increased synthesis of acetyl CoA from citrate. Acetyl CoA can then be used as a precursor in lipogenesis and cholesterolgenesis. Key enzymes involved in the synthesis of cholesterol and fatty acids have been reported to increase activity in prostate cancer cells.

The major interest in the potential application of metabolomics for identification of biomarkers for prostate cancer followed a recent paper by Sreekumar and colleagues. In this study the metabolite composition in tissue, urine, and plasma from prostate cancer patients was profiled using liquid (LC) and gas chromatography (GC) – time of flight mass spectrometry (TOFMS). They identified the amino acid sarcosine as a potential marker for prostate cancer cell invasion, migration, and aggressiveness. A number of recent studies have collectively indicated that mass spectrometry-based methods could be used to characterize metabolomic changes during cancer progression and to further identify possible diagnostic and prognostic
biomarkers, or biomarker patterns, as well as increase our knowledge about disease progression.\textsuperscript{56-62} Metabolic profiling of blood plasma, or serum, has been used to unveil metabolic alterations associated with prostate cancer.\textsuperscript{55,62-65}

**Multivariate data processing and analysis**

In metabolomics, techniques that separate molecules by way of their varying size and charge are often used to characterize the metabolites present in a sample. The resulting chemical fingerprint should ultimately represent the complete sample metabolome. However, metabolites come from many different compound classes and are present in a large range of concentrations (Figure 1). Thus, there is not one single analytical technique that can cover the whole metabolome. Instead many complementary analytical tools must be used in order to get the most comprehensive representation possible.

Today the most frequently used methods for analysis of the metabolome are nuclear magnetic resonance (NMR) spectroscopy and hyphenated techniques such as gas chromatography (GC) and liquid chromatography (LC) coupled to mass spectrometry (MS). In addition, Fourier transform InfraRed spectroscopy (FTIR) have been used together with direct infusion mass spectrometry (DIMS) as well as other complementary techniques (reviewed by Dunn\textsuperscript{66}). NMR and FTIR require minimal sample preparation, however, detection limits are higher compared to the MS-based techniques and elucidation of spectra composed of many metabolites can be problematic. For that reason, hyphenated techniques are generally preferred to allow both quantification and identification of as many metabolites as possible. In the selection of a specific analytical method it has to be assessed whether the method of choice is suitable for answering the particular biological question. However, in metabolomics studies it is often not known which metabolites are the most interesting beforehand, or at what concentrations they are present. Therefore, general and more global analytical methods are often used initially, followed by validation by means of targeted analysis using more high resolution techniques.

GC-MS has become an important analytical platform for generating metabolic fingerprints and comparing samples in global metabolite analysis.\textsuperscript{58,67-70} The reason for its frequent application is due to its relatively high sensitivity and reproducibility, but particularly because identification of detected metabolites by mass spectral databases is rather straightforward. However, the requirement to derivatize non-volatile metabolites increases the sample preparation time and often complicates the identification of
unknown compounds. The GC method used in this thesis relies on derivatization by oximation followed by silylation prior to analysis, in order to provide coverage of the largest range of metabolites. Detected compound classes include alcohols, aldehydes, amino acids, amines, organic acids, sugars, sugar acids, sugar amines, sugar phosphates, purines and pyrimidines. The efficiency of derivatization is an important factor as it has an effect on the reproducibility of the analyzed metabolites. Isotopically labeled standards, eluting with an even spread over the chromatogram, are added to each sample to monitor extraction and derivatization effects.

Before statistical analysis the analytical data needs to be processed so that the same identity is assigned to the same metabolite in each sample. The generation of high quality data for further statistical analysis, interpretation, and identification is central in the search for new markers as well as for a deeper understanding of underlying biological processes. In this method, it is the quality of the information that will be decisive for successful results and a key issue is to create data that are representative for the metabolic composition of the studied samples. For hyphenated data this can be achieved by data deconvolution or curve resolution. Curve resolution can be seen as a mathematical method for generating chromatographic profiles and mass spectra for the detected metabolites, as well as enhancing analytical resolution. The chromatographic profiles are then used for quantification, while the corresponding mass spectra are used to identify the metabolites, either by means of spectral database comparisons or by de novo interpretation.

**Hierarchical multivariate curve resolution**

Hierarchical multivariate curve resolution (H-MCR) extracts chromatographic profiles and mass spectra from GC-MS data of many biological samples simultaneously, as compared to other curve resolution methods that are performed on an individual sample basis. The improvements using this approach are that no matching of resolved compounds is needed and that the same compound will be quantified identically over all samples, using the shared spectral profile. Another advantage with the H-MCR method is that completely overlapping compounds, eluting at identical retention times, can be resolved as long as the concentration ratio between them and their mass spectra differs.

The H-MCR method includes the following processing steps: (1) smoothing of each sample in all mass channels by a moving average; (2) alignment by finding maximum covariance between the samples total ion current (TIC) chromatograms; (3) division of data in to time windows; and (4) multivariate
curve resolution of each time window separately. This generates a chromatographic profile for each compound in each sample with a corresponding common spectral profile (Figure 3). Next, a data matrix $X$ is created that can be used for any kind of multivariate data analysis (MVA) or other statistical analysis with the aim of comparing samples. In the data matrix each row represents one sample and each column represents one compound. The value in each cell of the matrix represents the integrated area under a resolved chromatographic profile in one sample.

![Figure 3](image)

**FIGURE 3.** An overview of the main steps included in H-MCR. (a) Smoothing of mass channels and alignment of sample chromatograms. (b) Division into time windows, edges set at low intensity points. (c) Each time window is then resolved separately by H-MCR, resulting in a chromatographic and corresponding mass spectral profile for each resolved compound.

An extension to the H-MCR method also showed that this curve resolution could be carried out predicatively for independent sample sets (Paper I), which can prove to be of great value in biomarker pattern verification and validation as well as various important applications, such as diagnosis and high-throughput analysis (Papers II-V).
Chemometrics

Chemometrics is the computational field for planning experiments and extracting information from high-dimensional data, i.e. the information aspect in the study of complex systems. Chemometrics can be seen as a toolbox containing statistical methods that can handle many dependent variables, so that multivariate thinking can be present in every step from setting up studies to verifying the results. Per definition, there are two main branches of chemometrics i) Design of experiments (DOE) and ii) Multivariate data analysis (MVA). By combining DOE and MVA, an efficient tool for studying systems that include multiple variables and correlated responses is created. One important issue when dealing with human subjects is the need to minimize confounding variation or, at best, being able to detect and handle this variation. Here, the chemometric methodology can be a valuable contribution together with previous experience and knowledge related to studies.

Design of experiments (DOE)

The main aim of DOE is to maximize the information output from a minimum number of experiments. In DOE, variation is introduced into the data in a systematic fashion so that the effects of investigated variables (for example sex, age and disease) and their interactions on one or many responses (e.g. response to drug treatment) can be deciphered by means of statistical analysis. DOE has been utilized in metabolic studies for optimization of analytical procedures and for detecting statistically significant changes and related biomarker patterns. Study design is indispensable in biomarker discovery and also pointed out to be a main reason why biomarkers fail to be accepted for clinical use. In this work, DOE was used (i) for diversity-based selections of representative sample subsets (Papers I, II, III and V), (ii) to investigate and optimize the outcome of the H-MCR method (Paper II) and (iii) to create a dataset consisting of standard compounds that was varied in concentration (Paper II).

Multivariate data analysis (MVA)

Multivariate data analysis is the branch of chemometrics concerned with analysis and interpretation of complex data structures. Multivariate projection methods for data exploration (e.g., principal component analysis, PCA) and regression (e.g., partial least squares, PLS) have been shown to be very useful for interpreting the systematic changes that exist
between many samples characterized by the relative concentrations of a multitude of metabolites.\textsuperscript{89-94} Several features are associated with multivariate projection-based model systems, such as separation of systematic variation from noise, managing many and correlated variables, outlier detection, handling of missing values, possibilities for model validation, and prediction of independent samples. These characteristics have been specifically attractive for pushing the global metabolite analysis forward in terms of understanding complex interactions in biological systems.\textsuperscript{95-98}

**PCA**

Principal component analysis (PCA) is the most widely used, unsupervised projection method for exploratory analysis and data compression. This central multivariate data analysis tool extracts the systematic variation in the original data matrix $X$ by a reduced set of latent variables. A latent variable is a concealed variable that cannot be directly measured or observed, but is described by other measured variables.\textsuperscript{99} Latent variables can for instance, be hidden metabolite structures in a blood sample related to gender, age or disease. The two most commonly used methods for calculating a PCA model are non-linear iterative partial least squares (NIPALS)\textsuperscript{91} or single value decomposition (SVD)\textsuperscript{92} algorithms. The NIPLS algorithm is used throughout this work due to its ability to handle missing values that are often present in large metabolomics datasets.

$$X = TP^T + E$$

The PCA algorithm compresses the systematic information in the original data $X$ into a latent variable representation, $TP^T$, so that it explains the largest amount of variation in the original data. The scores ($T$) gives an overview of the samples and how they relate to each other. Interpretation of the patterns in the scores is found in the corresponding loadings ($P$). The loadings express how each variable contributes to the separation among samples and reveal relative importance of each variable. Unsystematic information, e.g. noise, is stored in the residual matrix, $E$. PCA gives a summary of the data and essential systematic effects can be easily visualized in a separate latent variable and interpreted individually.

In this work PCA has been used to: (i) provide an overview of the data and for detection of deviating samples (Papers I-V), (ii) create a basis for selections of representative sample subsets in the compressed multivariate space (Papers I and III) and (iii) predict analytical replicates, revealing good analytical conditions (Paper III).
PLS and OPLS

In metabolomics the requirement to transform complex data into knowledge is reached by the integration of chemistry, biology and medicine with statistics. The supervised regression method partial least squares (PLS) is a particularly important data analysis tool for these kinds of applications. PLS extracts latent variables by maximizing the explained variation in $X$ as well as the covariation between two sets of data i.e. $X$ and $Y$.

The extension of PLS i.e., orthogonal PLS methodology (OPLS)\textsuperscript{100}, has greatly improved the interpretation of large complex datasets, such as those generated within global metabolite analyses.\textsuperscript{77,101,102} OPLS is a specialized PLS algorithm, where the systematic variation in data can be separated into correlated (predictive) variation, $T_p P_p^T$ and uncorrelated (orthogonal) variation $T_o P_o^T$. For single-$y$ OPLS models, the multidimensional variation in the metabolite profiles is compressed into a single dimension that maximizes the differences between groups, for instance patients and controls. When $Y$ is qualitative, e.g. denoting class membership, the OPLS method is called OPLS discriminant analysis (OPLS-DA).\textsuperscript{103} For two classes, the model representation produces a predictive score vector, $t_p$ that capture between class variation, and the uncorrelated (orthogonal) score vectors, $T_o$ capture any within-class variations. The sources of orthogonal variation is often related to specific study conditions such as sample handling, sample storage, experimental problems, instrument drift, age, gender and environmental factors.

\[ X = T_p P_p^T + T_o P_o^T + E \]
\[ Y = T_p C_p^T + F \]

In metabolomics studies the number of variables is generally high compared to the number of samples. To reduce the probability to find correlations by chance, methods such as cross validation, permutation test and external test sets for validation are applied.\textsuperscript{104-106}

OPLS is the selected regression method for multiple sample comparisons and sample predictions in the work included in this thesis. OPLS has been used to correlate the metabolic information against the day of sample collection in a study of rat urine toxicity and to classify aspen leaf extracts and human blood plasma samples (Paper I), (ii) model and the predict systematic metabolic variation related to the acute effect of strenuous exercise (Paper III) (iii) extract and interpret the systematic variation in tissue and plasma profiles related to prostate cancer disease progression (Papers IV-V).
**Predictive metabolomics**

Metabolomics has attracted great interest as a sensitive technique for elucidation of biomarkers, or biomarker patterns, in biofluids or tissues associated with disease-related processes. However, even though many recent results have provided proof of the ability of the technique, there are still challenges ahead in order to move from isolated study results to predictive or diagnostic systems based on verified marker patterns.

Among the main challenges identified for the future of metabolomics are; i) diagnosis/prognosis, ii) result verification, iii) screening of large sample sets, and iv) linking metabolic pattern changes to mechanistic information. To meet these challenges a number of issues need to be addressed including e.g., generation of high quality data in a high throughput fashion, development of predictive systems for sample characterization and classification and elucidation of robust biomarker patterns of biological relevance. Key factors in this are of course the availability of representative samples, as well as sensitive and robust analytical characterization. However, issues such as study design, data processing and data analysis are equally important in order to address the aforementioned challenges.

Predictive metabolomics is a concept developed in our lab, by applying chemometrics to metabolomics studies as a means to help in this development towards a more comprehensive utilization of the metabolomics technique in solving biological questions. The concept involves strategies for sample selection, planning of studies and development of analytical protocols all based on DOE, followed by strategies for data processing, data analysis and validation based on multivariate statistical analysis (Figure 4).

Today we have a predictive system generating high quality data in a high throughput fashion based on GC-TOFMS data (Paper I)\(^{107}\). The crucial step in this development was the extension to the H-MCR algorithm making it possible to resolve data from independent samples predictively. Thus, by connecting predictive H-MCR processing to multivariate projection analysis it became possible to obtain a predictive chain of events including both data processing and sample characterization, or classification. As a consequence, validation of all steps in the process could be accomplished but more importantly, diagnostic modeling based on high quality metabolite profiles also became possible.
1) DESIGN OF EXPERIMENTS
   - to maximise information output

   Sample collection, preparation and chemical characterisation according to:
   STANDARDIZED PROTOCOLS

2) HIERARCHICAL MULTIVARIATE CURVE RESOLUTION
   - to resolve overlapping peaks

3) MULTIVARIATE ANALYSIS
   - Sample comparison and prediction

4) VALIDATION:
   - Pattern verification and Biological interpretation

**Figure 4.** Schematic diagram detailing the predictive metabolomics strategy. (1) Design of experiments and standardized protocols should be used throughout the study. (2) Model samples (all or a representative selection of samples) are resolved by H-MCR. Obtained H-MCR parameters can then be used to resolve new samples or remaining samples not selected in the H-MCR processing. (3) The resulting data obtained from the H-MCR processing and treatment according to the H-MCR parameters is saved in two separate data tables. (4) Sample comparisons by means of multivariate data analysis can then be carried out by modelling the processed data and the model can be used to make predictions of the treated samples. Alternatively, the processed and treated data can be merged and used for calculating a model based on all samples.
The additional benefit of predictive H-MCR, being much faster compared to the original H-MCR processing, enabled the processing phase to become high throughput. This allowed screening of large sample sets without compromising the data quality, something that potentially will be highly valuable for future applications, e.g. in screening large populations for specific diseases.

Since the introduction of the predictive metabolomics concept is has been applied to a number of studies with medical or clinical focus. These include applications to cancer, exercise and nutrition and neurodegenerative disease.
Aims of the study

The overall aims of the studies included in this thesis were to:

Evaluate a chemometric methodology for processing and modeling of GC-MS based metabolomics data with the ambitions of generating a predictive system for; i) generation of representative data, ii) validation and result verification, iii) diagnosis and iv) screening of large sample sets.

Apply the methodology to studies of prostate cancer in humans to;

i) understand disease development and ii) reveal metabolic markers for disease aggressiveness.
Results

Paper I

Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data - A potential tool for multi-parametric diagnosis

This study presents an extension of the hierarchical multivariate curve resolution (H-MCR) method that makes it possible to import new independent samples into a predictive framework for processing of metabolic GC-MS data and subsequent multivariate modeling. Application of the strategy is proven to facilitate result validation and verification in terms of biological interpretation, as well as metabolite changes in new independent samples, e.g. analytically characterized at a different point in time. As a consequence of this, diagnostic modelling and metabolic screening of large sample sets based on high quality data is made feasible.

Within an organism or cell, it is often combinations of metabolites that are responsible for the variation of interest. In these situations it is essential to be able to use the whole metabolite profile to extract biomarker patterns, as opposed to a single biomarker. Predictive models of metabolite pattern changes in biological systems will potentially be of great value with regard to
early disease diagnosis, and disease prognosis, as well as in clinical monitoring for personalized treatment and healthcare.

When comparing metabolite compositions in different samples, the resolved compounds from one sample need to be identified in the other samples, if they exist. The matching of resolved compounds in new samples that have not been part of the curve resolution procedure by H-MCR includes the following steps: Smoothing of the data in the same way as the model samples; alignment of samples using the target for the model samples; division of the data into time windows using the same edges; resolving the new data for each time window, using the spectral profiles found in the model samples.

We show that the proposed strategy is efficient for describing metabolic changes in many different biological sample matrices. Here, presented for data sets of aspen leaf extracts from a plant development study, rat urine samples from a toxicity study, and human blood plasma samples from male and female subjects. By this we have achieved greatly reduced times for data processing, since only a smaller representative subset is processed by curve resolution. In addition it addresses problems with biased classification modeling in cases where one class contains more samples than another. Using the proposed strategy, the same number of samples from defined classes can be processed and modeled simultaneously, whereas remaining samples can be predicted and thus not affect the classification in any way.

For the strategy to work optimally it is important to use experimental design introducing all known sources of variation into the model and to optimize analytical procedures to create reproducible data over time and between labs. A requirement for achieving reliable prediction result is clearly that the diagnostic information in the deciding model is founded on biologically relevant variation for the study of interest. Additionally, it is vital that the selection of diverse samples for building models includes variation that is representative for future prediction samples. Methodologies for selecting a representative set of samples intended for building models is presented and investigated in Paper III.
Reliable profile detection in comparative metabolomics

In global metabolite profiling the quality of the metabolic information is key in finding biomarker patterns that are indicative of a specific physiological status. Low quality data are more prone to produce spurious correlations resulting in over-fitting and misinterpretation. This is problematic in the search for diagnostic marker patterns that ought to be verified over multiple studies and ultimately used for predictions in new samples. For these purposes, representative data, reflecting the biology of the system in a reliable fashion is a requirement in order for global metabolite analysis to become a tool for clinical diagnosis.

As previously described, all compounds in a complex sample cannot be separated on the GC column. Thus, signals for different compounds will be recorded at the same time, resulting in a mix of them in the mass spectrometer. Methods for curve resolution are often used to separate overlapping metabolites. However, for complex samples this is not always straightforward. Instead, profiles will be estimates of the pure compounds, whereas others will be left unresolved. These issues can occur because of e.g., difficulties with minor compounds vs. noise, high concentration differences for a single metabolite in different samples and errors in the estimation of the number of metabolites to be resolved. To deal with these problems the curve resolution method H-MCR is controlled by several factors, which can be adjusted by the user.

In this paper, a strategy for optimizing and understanding the data processing is suggested with the overall aim to produce a framework for generating representative data with regards to both quantitative and qualitative issues. The considered characteristics of the resolved GC-MS data were the number of resolved compounds, the quality and reproducibility of the resolved chromatographic profiles, and the quality of the resolved mass spectra. These characteristics were investigated in relation to the user adjustable factors by DOE. As a first step, all combinations of the factor settings were studied in a $2^4$ full factorial design and correlated to responses describing the above characteristics. This systematic analysis provided knowledge about the effects of adjustable factors, as well as on how to select settings beneficial for different data characteristics.
By altering the H-MCR factor settings in a systematic way we show that the possibilities for extracting data with different characteristics are extensive. The results show that by only using the number of resolved metabolite profiles as the response to optimize against, the quality of the output data is partly neglected. This will induce negative effects on issues such as sample comparisons, predictions, and metabolite identifications being the cornerstones of metabolomics. Thus the importance of producing high quality data cannot be underestimated.

By considering both quantitative and qualitative features of the data it was possible to understand and optimize the performance of the H-MCR method. In this way a framework for resolving a high number of metabolite profiles exhibiting both high chromatographic and spectral quality was obtained. This is a general strategy that can be applied to any type of data processing provided that the important processing parameters and key output data characteristics can be defined. The results, however, are not general, meaning that the optimal settings will vary depending on the data, suggesting that the presented approach should be used as an integrated step in global metabolite profiling.
Paper III

Processing of mass spectrometry based metabolomics data for large scale screening studies and diagnostics

Efficient screening of large sample sets, or sample banks, for informative patterns of metabolites is an issue of major concern for the progression of metabolomics. From our point of view, efficient metabolomic screening should take advantage of the higher sensitivity of the analytical techniques to create data of higher quality by using sophisticated data processing. To do this in a high throughput fashion the strategy needs be based on the selection and use of representative samples relevant for the question at hand. By selecting representative sample subsets by chemometric approaches and using these subsets to perform curve resolution (H-MCR) and multivariate classification analysis (OPLS-DA), it was possible to form a predictive metabolomics screening strategy for GC-TOFMS data on human blood serum samples. The samples were collected in a study examining the effect of strenuous physical exercise. Healthy and regularly training male subjects performed four identical tests of strenuous ergometer cycling exercise. Blood samples were collected before and immediately after each exercise session, to improve the understanding of human metabolism in connection to acute physical exercise. The selection of sample subsets was performed according to two different principles, the first being a selection based on property data and the other being based on already acquired analytical data processed using a fast and crude processing method.

![Figure 5](image-url)

**Figure 5.** Visualization of the representative sample subset selection based on analytical data (marked as black spheres and remaining test samples as gray spheres). The subset was selected by space-filling design which maximizes the minimum Euclidean distance between the nearest neighbors of the selected observations.

The results showed that the presented strategy provide an organized approach, which could be applied to efficient screening of biobanks, or other
large sample sets, while retaining data quality and interpretation. Also, the methodology could be applied to verify biomarker patterns in independent sample sets, in this case, analytically characterized eight months later than the model samples, which is the basis for working and validated diagnostic systems. The time and utility for producing representative data is very important for an efficient screening of large sample sets. Generally, curve resolution methods are limited in the number of samples that can be processed. This is mainly due to time-consuming processing and limitations in computer capacity. However, the predictive feature of the H-MCR method solved this issue as processing of 16 model samples took 6h and 29 min., while predictive processing of remaining 77 samples took merely 10 min. This shows that as long as the selected subset is representative in terms of retained variation and that the predictions can be performed in a robust and reliable way, this is an efficient strategy for producing high quality data, with no limitations concerning the number of samples that can be treated.

The proposed method can be used for:
- Sample bank mining where sample availability is constrained and it is very important to extract as much information as possible from a limited number of samples.
- High throughput screening of large sample sets producing high quality data for interpretation and biomarker identification.
- Fast and high quality metabolic screening to help select samples for more expensive and time consuming analyses in other types of -omics studies.
- Developing systems for biomarker pattern verification, being an extremely important issue from a clinical perspective and for validation of findings in independent studies.

Even though the results are promising there are still challenges that need to be addressed to reach a complete and robust method for screening and predicting samples over time. For example, this study includes a homogenous human cohort, which is not representative for the whole human population. This is clearly something that needs to be considered in, e.g. disease diagnosis modeling. Also, it would be valuable to further evaluate the strategy by performing a completely independent study, applying the same test to another population and make predictions for these subjects into the existing model. Another issue that needs to be investigated is strategies for continuous updating of models to assure robust and reliable end results. Also, of utmost importance is the continuous investigation and optimization of protocols and quality control of sample handling and analytical characterization.
Paper IV

Metabolomic characterization of human prostate cancer bone metastases reveals increased levels of cholesterol

In this study we applied the predictive metabolomics strategy to characterize metabolomic alterations associated with prostate cancer progression. The overall aim was to identify possible prognostic metabolite markers associated with aggressive disease and bone metastases, which could, furthermore, increase our understanding about disease progression.

The study was performed with the hypothesis that potential markers for aggressive prostate cancer could be found by discovering metabolites at high levels in bone metastases and then investigate if these factors also were increased in primary tumours and in blood samples from patients with metastatic disease. We performed a GC-TOFMS-based metabolomics study of prostate cancer bone metastases in comparison to corresponding normal bone, primary prostate cancer tumours and normal prostate tissue. The tissue samples were obtained from patients operated for metastatic spinal cord compression or pathologic fractures at Umeå University Hospital 2003–2009. In addition, characterized blood samples from patients with and without diagnosed bone metastases were included, to identify metabolites that could be used to improve prognostication and aid in therapy of advanced prostate cancer. The findings in the bone metastasis tissue were verified in a separate test set, also including metastatic bone tissue from cancers of different origin.

The result showed that prostate cancer bone metastases have a different metabolic signature, compared to normal bone and to bone metastases from other cancers. Among the detected metabolites we specifically noted the high levels of cholesterol when differentiating prostate cancer bone metastases from normal bone tissue and from other bone metastases. Cholesterol has previously been linked to prostate cancer disease progression (reviewed by Solomon) and was, therefore, selected for further analysis.

In order to examine possible causes for the high levels of cholesterol in prostate cancer bone metastases, paraffin embedded tissue sections were immunostained for enzymes involved in the influx of exogenous cholesterol to the cell, as well as de novo synthesis. The results showed that prostate cancer bone metastases have the possibility for both uptake and the de novo synthesis of cholesterol. Other studies have also indicated that increased availability of cholesterol may be of relevance for the growth and
development of bone metastases, including apoptosis, cell proliferation, migration and invasion. \textsuperscript{114-117}

In a resulting study, cholesterol synthesis was targeted in an in vitro model for bone metastases \textsuperscript{118}. The aim was to further understand the role of cholesterol in the interaction between prostate cancer cells and bone and it was shown that cholesterol is important for prostate cell growth in vitro and in co-culture with bone. In the process of tumor growth, prostate cancer cells induced a lytic response in bone and release of the insulin-like growth factor 1 (IGF1), which acts as a powerful survival factor. \textsuperscript{119} Interestingly, it was further shown that by targeting cholesterol synthesis and the IGF1R simultaneously, prostate cancer cell growth was more efficiently inhibited than by either therapy alone. As a result, inhibition of IGFR1 in combination with existing apoptosis-inducing treatments, such as statins, castration and chemotherapy are now being studied in animal models for their effects on prostate cancer bone metastases. Taken together, the presented results and previous findings in the literature indicate the prospect of using cholesterol inhibitors as treatment or chemopreventive agents for prostate cancer metastases. \textsuperscript{113} However, novel drugs are then probably needed that efficiently target peripheral organs.

We also hypothesized that cholesterol, by its conversion in to androgens via metabolic enzymes, possibly increases androgen receptor signaling as well as castration resistant tumor growth in patients treated with androgen-deprivation therapy. \textsuperscript{120,121} This did not however, seem to be the case in a subsequent study examining the pathways involved in synthesis of androgens from cholesterol using gene expression arrays, RT-PCR, and immunohistochemistry. In this study, neither of the steroid-converting enzymes in the early steps of cholesterol conversion into testosterone; CYP11, CYP17, or HSD3B2, showed significantly higher levels in CRPC bone metastases compared to non-castrated (hormone-naïve) bone metastases (Jernberg E \textit{et al.}, unpublished data).

In addition we discovered metabolic differences between primary prostate tumor tissues from high-risk patients, with and without established bone metastases. Finally, differentiating metabolite profiles in blood plasma from patients diagnosed with high risk tumors, with and without established bone metastases could be identified. The pattern in blood plasma associated to aggressive disease is further examined in \textbf{Paper V}. 

\textsuperscript{113}
Paper V

Evaluation of metabolic alterations in patient plasma associated with disease aggressiveness in prostate cancer

At the time when the prostate cancer has spread outside the prostate organ and metastasized, only palliative therapies are available. In this study, predictive metabolomics was used to identify metabolites in clinical series of plasma samples associated with high risk prostate cancer and biochemical relapse after radical prostatectomy. The findings could possibly aid in the search for new biomarkers for detection of clinically significant tumors at a curable time-point.

Multivariate data analysis were carried out with the strategy of first discovering plasma metabolites associated with prostate cancer disease progression in a series of patients stratified according to prostate cancer risk groups (The National Comprehensive Cancer Network. Practice Guidelines in Oncology-Version.1.2010. Prostate cancer), including patients at five different stages ranging in disease severity from benign to metastatic disease, and secondly, by investigating these metabolites for association with biochemical relapse after radical prostatectomy in a second series of patients. The fact that detected metabolic markers were evaluated in two different sample cohorts facilitated cross study verification for relevance and robustness.

13 metabolites were highlighted as important in OPLS-DA models including all detected plasma metabolites both for i) difference between patients with benign and metastatic disease and ii) difference between sample groups related to prostate cancer risk. Within the extracted metabolite patterns for metastatic disease two metabolites were detected which were possibly consumed by aggressive prostate cancer (decreased plasma levels with increased PCa risk and increased after surgery in the relapse but not in the non relapse group). We further found increased levels of four metabolites in plasma from patients with metastatic disease, while no supposed tumor-derived metabolite was detected at an early stage of the disease (increased plasma levels in low to intermediate prostate cancer risk and decreased levels after surgery). In addition, verification of metabolite markers for metastatic disease detected previously by us and others was performed. By this we were able to confirm decreased plasma levels of stearic acid and increased levels of pseudouridine with metastatic disease, while others were not confirmed or not detected (e.g. phenylalanine, taurine, glutamate) or detected (e.g. sarcosine).
An interesting finding was also that we could identify differences related to biochemical relapse by evaluation of multivariate metabolite patterns before and after radical prostatectomy. The results highlighted an interesting phenomenon that by introducing an intervention to the biological system, in this case the surgery, metabolic differences between the relapse and non-relapse groups seemed to be magnified. This is a methodologically interesting event that might be of assistance in diagnostic or prognostic modeling and in gaining mechanistic understanding of metabolic events such as variations in disease sub groups.

In this study the majority of the metabolites pointed out as interesting markers in the different comparisons were unknowns. This highlights the fact that identification of metabolites is still an issue in metabolomics studies. Without comprehensive spectral libraries the usefulness of the data is limited. However, strategies for identification of unknown metabolites will be applied in terms of fractionation, parallel spectroscopic analyses and structural determination by NMR.
Conclusion and future perspectives

This thesis describes research directed towards solving current challenges in metabolomics including result verification, screening of large sample sets and working applications for disease diagnosis and prognosis.

The presented chemometric methodology for processing and modeling of GC-MS based metabolomics data creates a predictive system for sample characterization and classification as for linking changes in metabolic patterns to biological mechanisms. In order to meet the aforesaid challenges and to obtain essential conclusions about the biological system under investigation, methods for study design, data processing and data analysis are essential. Furthermore, standardized protocols to ensure sensitive and robust analytical characterization and availability of representative samples are key components.

The strategy can be used for efficient screening of biobanks, or other large sample sets, while retaining data quality and interpretation. Also, the methodology could be applied to verify biomarker patterns in independent sample sets, analytically characterized later in time than the model samples, which is the basis for working and validated diagnostic systems. Hence, the strategy can be used to create predictive models for biomarker pattern verification in biological systems that potentially will be of great value for early disease diagnosis and prognosis, as well as in clinical monitoring for personalized treatment and healthcare.
Metabolomics has several advantages when studying biological systems. Primarily, the metabolome describes the actual functional status related to the phenotype of the organism. Secondly, metabolomics in view of global metabolite profiling is a hypothesis free method, and therefore unexpected relationships and metabolite reactions can be discovered, which in itself is hypothesis generating.

One of the main issues for improvement of metabolomics is how to provide a complete coverage of the metabolism of a cell, tissue, organ or organism in order to get a mechanistic understanding of the biological systems. Today, no bioanalytical technique can generate a complete metabolomic description of even the simplest organism. However, only by combining technologies will we ever have the means to increase our coverage of the metabolome. Thus, there is a great need for improved methods for data integration in order to cross correlate the information produced by a variety of analytical approaches. Here chemometrics will have a great impact in developing strategies that efficiently can handle the large amounts of data of great complexity.

The challenge of metabolite identification has been known for several years, and has drawn attention to the need for comprehensive metabolite databases. In this way changes in identified metabolite levels can be related to global changes across classical metabolic pathways, so that significant understanding is reached.

Global metabolite profiling in its current form can efficiently be used as an exploratory tool to find interesting areas of the metabolism that can be further investigated by targeted analysis. This was presented in paper IV where we applied the predictive metabolomics strategy to characterize metabolomic alterations associated with prostate cancer progression. We discovered elevated cholesterol as one interesting metabolic marker for the development of bone metastases in prostate cancer. The finding has made us examine effects of cholesterol and cholesterol depletion on growth of prostate cancer and its metastases, and will hopefully lead to improved treatment of bone metastases in the future. Our study was the first to report on metabolomic characterization of prostate bone metastases in humans. Furthermore, we applied the predictive metabolomics strategy to detect new biomarkers for detection of clinically significant tumors by finding metabolites in plasma associated with high risk prostate cancer and biochemical relapse after radical prostatectomy. The metabolite markers suggested by us to be associated with aggressive prostate cancer in paper IV were to some extent verified in paper V (stearic acid and pseudouridine), while others were not. Furthermore, additional interesting metabolites were
detected in paper V. These metabolites need to be identified as well as carefully verified in separate patient cohorts to allow further interpretation of the results.

In conclusion, this work has confirmed the potential of metabolomics for finding new diagnostic and prognostic markers and more importantly to increase mechanistic knowledge about disease progression in general and for prostate cancer in particular.

Gaining mechanistic understanding about cancer and finding sufficient diagnostic tools is a problem that needs to be addressed in a multilevel fashion. Hopefully, the work included in this thesis has brought us one step closer to the utopia of diagnosing cancer with a simple blood test.

Now this is not the end.
It is not even the beginning of the end.
But it is, perhaps, the end of the beginning.

-Sir Winston Churchill
Acknowledgements

Jag vill tacka alla som på ett eller annat sätt bidragit till att min avhandling nu är klar.

Ett stort och innerligt tack till...


**Johan Trygg** och **Michael Sjöström**, mina biträdande handledare. Det har varit skönt att veta att jag har haft er i bakfickan. Ni har stor del i kemometrins framgång och jag är tacksam för den kunskap ni givit mig.

**Pernilla Wikström**, min sist i raden tillkomna biträdande handledare. Du är en sann förebild och jag vill speciellt tacka dig för att du berikat mitt arbete med spännande forskning och ovärdelig kunskap.

**Thomas Moritz**, för all din hjälp och för att du tagit dig tiden att svara på alla mina frågor. Tack till övriga hjälpsamma medarbetare på UPSC; Krister Lundgren, Inga-Britt Carlsson, Annika Johansson och Izabella Surowiec som alla bidragit till analyserna av proverna i det presenterade studierna.


**Nuvarande och forna kollegor**, som bidragit till att göra kemiska institutionen till en trevlig arbetsplats och inspirerande forskningsmiljö. Barbro Forsgren och Carina Sandberg och Lars-Göran Adolfsson för att ni alltid gett mig hjälpsamma svar på allt ifrån administrativa frågor till datorer. Speciellt tack till Patrik Andersson och Mattias Hedenström för mycket trevliga uppföljningsmöten, som skett i sista sekund var eviga år.

**Mia** och **Elin**, vart ska jag börja. Stöpta i samma form, kokade i samma gryta, grillade i samma ugn och tillsammans utgör vi en perfekt komponerad måltid. Ni är oskattbara!


**Min älskade släkt**, som gjort att jag varje söndag haft ett kalas att se fram emot. Det är underbart att ha er så nära och ni är ovärderliga för mig.

**Våra härliga grannar**, Carlsons och Wennerholms som återkommande visat omtanke och bidragit med ingredienser till våra barns pannkakor.

**Ralf**, för att du är en klippa och alltid ställer upp med allt ifrån kulinariska middagar till diverse layouter och trycksaker.

**Matilda**, min älskade systerister, för otaliga barnvakter med kort varsel och mysiga stunder.

Bror **Petrus**, för att du påminner mig om att allt går att ordna och att ingenting är omöjligt (i sann morfar-anda).

**Sofie**, för att du är ett härligt tillskott i familjen och för att din siluett pryder mitt omslag.

**Mamma** och **pappa**, för att ni är min trygga hamn och för att ni alltid ställer upp.

**Mina kärlaste**, Agaton, Vendela och Marcus, jag älskar er!
References


38. Sardana, G., Jung, K., Stephan, C. & Diamandis, E.P. Proteomic analysis of conditioned media from the PC3, LNCaP, and 22Rv1 prostate cancer cell lines: Discovery and validation of candidate


110. Wuolikainen, A., Andersen, M., Moritz, T., Marklund, S. & H, A. ALS patients with mutations in the SOD1 gene have an unique metabolomic profile in the cerebrospinalfluid compared with ALS patients without mutations. Molecular Genetics and metabolism accepted (2011).


120. Dillard, P.R., Lin, M.F. & Khan, S.A. Androgen-independent prostate cancer cells acquire the complete steroidogenic potential of synthesizing testosterone from cholesterol. *Molecular and Cellular Endocrinology* 295, 115-120 (2008).