Proteomics of Body Fluids

Lennard Dekker

ISBN-978-90-9022111-3

Printed by: Gildeprint Drukkerijen, Enschede, The Netherlands

Proteomics of Body Fluids

Proteomics van lichaamsvloeistoffen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof.dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties**.**

De openbare verdediging zal plaatsvinden op
woensdag 10 oktober 2007 om 13:45 uur

door

Leendert Johannes Marinus Dekker

geboren te Dirksland

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie:

Promotoren:	**Prof.dr. P.A.E. Sillevis Smitt** **Prof.dr. C.H. Bangma**
Overige leden:	**Prof.dr. J.M. Kros** **Prof.dr. P.J. van der Spek** **Prof.dr. R.P.H. Bischoff**
Copromotoren:	**Dr. T.M. Luider** **Dr.ir. G.W. Jenster**

Paranimfen:	Dr. Arzu Umar Eric Brouwer

Dit proefschrift is tot stand gekomen binnen de afdelingen Neurologie en Urologie van het Erasmus MC te Rotterdam.

Het in dit proefschrift beschreven onderzoek werd financieel ondersteund door The Netherlands Proteomics Centre, een subsidie van het Erasmus MC Revolving fund (top-down) en door een EU subsidie P-Mark, (Sixth EU Framework Program for Research and Technological Development contract number LSHC-CT-2004-503011).

In de drukkosten van dit proefschrift werd bijgedragen door het van Leersum Fonds, de Dr. Ir. Van de Laar Stichting en de J.E. Jurriaanse Stichting zowel in drukkosten als in experimentele expertise werd bijgedragen door Dionex Benelux bv.

Table of Contents

Part A: Introduction

1. General introduction 9

2. Peptide profiling in body fluids 15
Dekker, L. J., Burgers, P. C., Kros, J. M., Smitt, P. A., and Luider, T. M.
Expert Rev Proteomics 3 (2006): 297-309

Part B: Method development

3. A new method to analyze MALDI-TOF peptide profiling mass spectra 39
Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M.
Rapid Commun Mass Spectrom 19 (2005): 865-870.

4. A database application for pre-processing, storage and comparison of mass spectra 53
Titulaer, M.K., Siccama, I., Dekker, L.J., van Rijswijk, A.L., Heeren, R.M.,
Sillevis Smitt, P.A., and Luider, T.M.
BMC Bioinformatics 7 (2006): 403.

5. FTMS and TOF/TOF mass spectrometry in concert: Identifying peptides with high reliability using matrix prespotted MALDI target plates 67
Dekker, L.J., Burgers, P.C., Guzel, C., and Luider, T.M.
Journal of Chromatography B, 847 (2007) 62–64

Part C: Biomarker research CSF and serum

6. MALDI-TOF mass spectrometry analysis of cerebrospinal fluid tryptic peptide profiles to diagnose leptomeningeal metastases in breast cancer patients 75
Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P.,
Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M.,
and Sillevis Smitt, P. A.
Mol Cell Proteomics 4 (2005): 1341-1349.

7. Identification of leptomeningeal metastasis-related proteins in cerebrospinal fluid of patients with breast cancer by a combination of MALDI-TOF, MALDI-FTICR and nanoLC-FTICR mass spectrometry 95
Rompp, A., Dekker, L., Taban, I., Jenster, G., Boogerd, W., Bonfrer, H., Spengler, B.,
Heeren, R., Sillevis Smitt, P., and Luider, T. M.
Proteomics 7 (2007): 474-81.

8. Differential expression of protease activity in serum samples of prostate carcinoma patients with metastases. 111
Dekker, L. J., Burgers, P. C, Charif, H., van Rijswijk, A.L, Titulaer, M. K., Jenster, G., Bischoff, R., Bangma, C. H., and Luider, T. M.
Submitted

Part D: Conclusions and summary

9. Concluding remarks and future perspectives 133

10. Summary & samenvatting 145

Appendices

Acknowledgments / Dankwoord 156

List of publications 158

Curriculum vitae 159

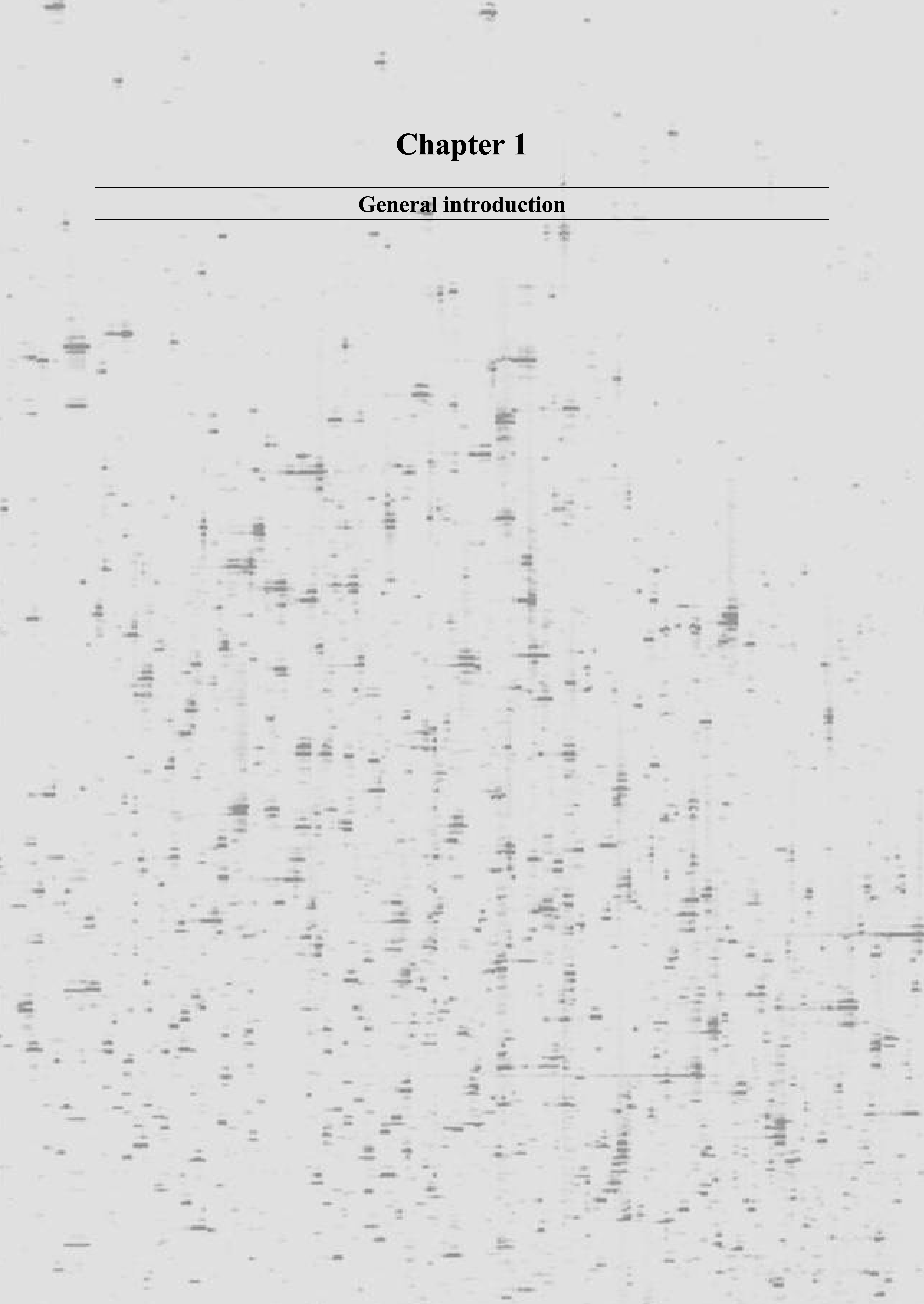
List of abbreviations

ACN	acetonitrile
ACL	access control list
ALS	amyotrophic lateral sclerosis
ASCII	American standard code for information interchange
AUC	area under the curve
CID	collision induced dissociation
CSF	cerebrospinal fluid
CSV	comma separated value
CV	coefficient of variance
DHB	2,5-dihydroxybenzoic acid
DW	dwell (time)
ECD	electron capture dissociation
ERD	entity relationship diagram
ERSPC	European Randomized study of Screening for Prostate Cancer
ESI	electro-spray ionization
FFT	fast Fourier transformation
FID	free induction decay
FTICR	Fourier transform ion cyclotron resonance
FTP	file transport protocol
FWHM	full width at half maximum
GUI	graphical user interface
HCCA	α-cyano-4-hydroxy-cinnamic acid
ICAT	isotope-coded affinity tag
ICP	inter cranial pressure
IP	internet Protocol
IRMPD	infrared multiphoton dissociation
iTRAQ	isotope tag for relative and absolute quantitation
JAR	Java archive
JDBC	Java database connectivity
JVM	Java virtual machine
LC	liquid chromatography
LIMS	laboratory information management system
LM	leptomeningeal metastasis

LP	lumbar puncture
MALDI-TOF	matrix-assisted laser desorption ionization time of flight
MRI	magnetic resonance imaging
MRM	multiple reaction monitoring
MS	mass spectrometry
MS/MS	tandem MS experiment
mzXML	mass over charge extensible markup language
nanoLC	nano-scale liquid chromatography
NTFS	Windows NT file system
ppm	parts per million
PSA	prostate specific antigen
ROC	receiver operating characteristic curve
S/N	signal to noise
SELDI-TOF	surface enhanced laser desorption ionization time of flight
SEM	standard error of the mean
SILAC	stable isotope labeling by amino acids in cell culture
SNAP	sophisticated numerical annotation procedure
SORI	sustained off-resonance irradiation
SQL	structure query language
SSL	secure socket program layer
SWT	standard widget toolkit
TCP	transmission control protocol
TFA	trifluoroacetic acid
TOF	time-of-flight
TRASH	thorough high-resolution analysis of spectra by horn
VEGF	vascular endothelial growth factor

Chapter 1

General introduction



1 General introduction

The ultimate aim of this thesis is to detect new biomarkers that will improve the accuracy of diagnosing leptomeningeal metastasis (LM) in breast cancer patients and of diagnosing prostate cancer. To reach this goal, we first developed new proteomics methods to reliably measure samples and analyze data. These newly developed methods were subsequently applied to the detection of biomarkers in the cerebrospinal fluid (CSF) from breast cancer patients with LM and in the serum from patients with metastatic prostate cancer.

1.1 Proteomics

Proteomics can be defined as the field of research that attempts to describe or explain biological phenomena in terms of qualitative and/or quantitative changes in various cells and extra cellular biological materials. The impulse for the development of proteomics was the availability of techniques that could transport proteins or peptides into the gas-phase, that is to say, into a solvent-free environment without breaking the protein or peptide apart. Coupled with mass spectrometry (MS), such techniques allow the accurate measurement of the masses of the intact proteins or peptides and of their sequence characteristic dissociation products (MS/MS).

In this thesis proteomics work is described that is mainly focused on biomarker discovery in body fluids. All work presented was performed within the framework of two projects namely an Erasmus MC Revolving fund (top-down) project (brain cancer and prostate cancer biomarker discovery) and an EU project P-MARK (prostate cancer biomarker discovery). At the onset of these projects, available proteomics methods did not yet allow reliable measurement of samples or analysis of data. For that reason we first present newly developed methods that are required to measure samples for biomarker research. These include the development of (i) a reproducible technique for peptide profiling of body fluid proteins; (ii) software for data storage and analyses; (iii) methods for depletion and fractionation, and (iv) identification methods.

1.2 Biomarkers in CSF from breast cancer patients with LM

Approximately 5% of patients with metastatic breast cancer will develop LM. The response to treatment largely depends upon early diagnosis. At present, LM in patients with breast cancer is diagnosed by magnetic resonance imaging (MRI) and cytological examination of the CSF. The sensitivity of MRI is approximately 75%. The sensitivity of CSF cytology increases from 75% after the first lumbar puncture to 90% after the third. False positive CSF cytology is rarely seen, resulting in an almost 100% specificity (1). Several biochemical markers of LM in CSF have been described, including lactate dehydrogenase iso-enzymes, carcinoembryonic antigen, beta-glucuronidase, and vascular endothelial growth factor (VEGF) (1-5). However, these markers are relatively nonspecific, resulting in false positive diagnoses and they are rarely used in clinical practice. In this thesis, we set out to discover new markers for LM, to facilitate early diagnosis and to gain insight into some aspects of the pathogenesis of LM.

1.3 Biomarkers in prostate cancer

The introduction of prostate specific antigen (PSA) for the detection of prostate cancer had an important influence on the management of this disease. As a result of the wide-spread application of PSA in asymptomatic men, a stage shift at diagnosis has occurred, resulting in a reduction of the number of men diagnosed with advanced prostate cancer (6). However, PSA lacks specificity, because in a number of other disorders including prostatitis and benign prostate hyperplasia elevated PSA levels are also observed. These false positive, high PSA levels will expose men to unnecessary treatments and the associated complications. In this thesis we attempt to find new biomarkers that may help to reduce the number of false positive diagnoses and to better predict the prognosis of the patient.

1.4 Scope of this thesis

This thesis can be divided into four parts. Part A contains an introduction to mass spectrometry based profiling of body fluids for biomarker research. Part B contains three chapters dealing with method development. In part C of this thesis, these methods are applied to the biomarker research of LM in breast cancer and to the biomarker research of prostate cancer. In part D, the findings of the thesis are summarized and discussed and future directions are indicated.

Chapter 2 gives a detailed overview of how mass spectrometry is used in the field of biomarker discovery, the current developments and the problems associated with this technique. In chapter 3, a newly developed method is described for tryptic peptide profiling of body fluids using matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS). An improvement in reproducibility is obtained by including replicate measurements in the analyses of the data. This improved reproducibility resulted in more reliable peptide profiles with a better chance of finding potential biomarkers. In chapter 4 we describe a newly developed database application for the analyses and storage of mass spectrometry data for biomarker research. The development of a software tool that is able to handle the huge amount of data that are generated with mass spectrometry profiling was necessary, because the required functionalities were not commercially available. This developed open source software is a modular application and generates output files of a standard format that can be read by commercial bioinformatics tools. The last chapter of part B describes a method for the identification of peptides in highly complex MALDI spectra using nanoLC-MALDI-TOF/TOF-MS and matrix-assisted laser desorption/ionization Fourier transform mass spectrometry (MALDI-FTMS) measurements. Direct identification procedures with MALDI-TOF/TOF-MS in complex samples are often unsuccessful and unreliable. By using a combination of two mass spectrometric techniques we were able to identify peptides in highly complex samples with a high degree of certainty. The nanoLC step provides a separation of the peptides that allows the sequencing of individual peptides. The subsequent MALDI-FTMS measurements provide a very accurate mass for the peptide of interest that can be used to confirm the identification. In part C of this thesis we apply the developed techniques to biomarker research. In chapter 6 we describe the results of a MALDI-TOF-MS tryptic profiling study on CSF samples of breast cancer patients with LM. Chapter 7 describes the identification of the peptides that were found differentially expressed in the profiling experiment described in chapter 6. The proteins have been identified with a combination of two FTMS mass spectrometry techniques, viz. MALDI and electrospray ionization (ESI) to obtain reliable identification. In chapter 8 another profiling study is described using serum of prostate cancer patients with metastases. This profiling study was performed with a fully automated magnetic bead purification followed by MALDI-TOF-MS. The differentially expressed proteins were identified by nanoLC-MALDI-TOF/TOF and confirmed by MALDI-

FTMS as described in chapter 5. In chapter 9 the results of the studies reported in this thesis are evaluated and an outlook for the future developments in biomarker research is presented.

References

1. DeAngelis, L. M., and Boutros, D. "Leptomeningeal metastasis," *Cancer Invest* 23 (2005): 145-54.
2. DeAngelis, L. M. "Current diagnosis and treatment of leptomeningeal metastasis," *J Neurooncol* 38 (1998): 245-52.
3. Herrlinger, U., Wiendl, H., Renninger, M., Forschler, H., Dichgans, J., and Weller, M. "Vascular endothelial growth factor (VEGF) in leptomeningeal metastasis: diagnostic and prognostic value," *Br J Cancer* 91 (2004): 219-24.
4. Twijnstra, A., van Zanten, A. P., Nooyen, W. J., and Ongerboer de Visser, B. W. "Sensitivity and specificity of single and combined tumour markers in the diagnosis of leptomeningeal metastasis from breast cancer," *J Neurol Neurosurg Psychiatry* 49 (1986): 1246-50.
5. van de Langerijt, B., Gijtenbeek, J. M., de Reus, H. P., Sweep, F. C., Geurts-Moespot, A., Hendriks, J. C., Kappelle, A. C., and Verbeek, M. M. "CSF levels of growth factors and plasminogen activators in leptomeningeal metastases," *Neurology* 67 (2006): 114-9.
6. Reynolds, M. A., Kastury, K., Groskopf, J., Schalken, J. A., and Rittenhouse, H. "Molecular markers for prostate cancer," *Cancer Lett* (2007).

Chapter 2

Peptide profiling in body fluids

Dekker, L. J., Burgers, P. C., Kros, J. M., Smitt, P. A., and Luider, T. M.

Partly adapted from:
Expert Rev Proteomics 3 (2006): 297-309

Abstract

The search for biomarkers is driven by the increasing clinical importance of early diagnosis. Reliable biomarkers can also assist in directing therapy and in monitoring disease activity and efficacy of treatment. In addition, the discovery of novel biomarkers might provide clues to the pathogenesis of a disease. The dynamic range of protein concentrations in body fluids exceeds 10 orders of magnitude. These huge differences in concentrations complicate the one-step detection of proteins with low expression levels. Since all classical biomarkers have low expression levels (e.g. PSA 2-4 µg/L; CA125: 20-35 U/ml) new developments with respect to identification and validation techniques of these low abundant proteins are required. In this chapter we will discuss the current status of profiling body fluids using mass spectrometry-based techniques, and the problems associated with it.

2.1 Introduction

Two main approaches are used for biomarker discovery: the targeted and profiling approaches (1). The targeted approach focuses on a selection of proteins that are known to be related to a disease. These proteins and combinations thereof can be examined as potential biomarkers. Alternatively, these proteins can be used experimentally as a bait to find related proteins that can subsequently be tested as new targets. By focusing on a limited set of proteins, the targeted approach reduces the complexity of the analyses dramatically (1). Unlike the targeted approach, the profiling approach does not use prior information on proteins of interest, thereby increasing the ability to find new unexpected candidate biomarkers (2, 3).

The dynamic range of protein concentrations in body fluids exceeds 10 orders of magnitude (1). These huge differences in concentrations complicate the one-step detection of proteins with low expression levels. Prefractionation is required to unmask these proteins, since all classical biomarkers have low expression levels (e.g., prostate-specific antigen: 1–4 µg/l; and CA125: 20–35 U/ml). This thesis will focus on the profiling approach to detect, identify, and validate biomarkers in cerebrospinal fluid (CSF) and serum using mass spectrometry based techniques.

2.2 Body fluids in biomarker discovery

Diverse techniques are used in biomarker discovery, including electrophoresis, chromatography (top-down or bottom-up), laser capture microdissection, MS and protein array technology. The biomarker search can be performed on a wide variety of biological materials, such as tissue, body fluids or *in vitro* cultured cells. The literature is focused mainly on serum and, for the study of neurological diseases, also on CSF. Other biofluids, including urine, saliva, tears and sweat, are also useful, although variation in these fluids can be relatively large compared with serum and CSF, which are under a stricter homeostasis control.

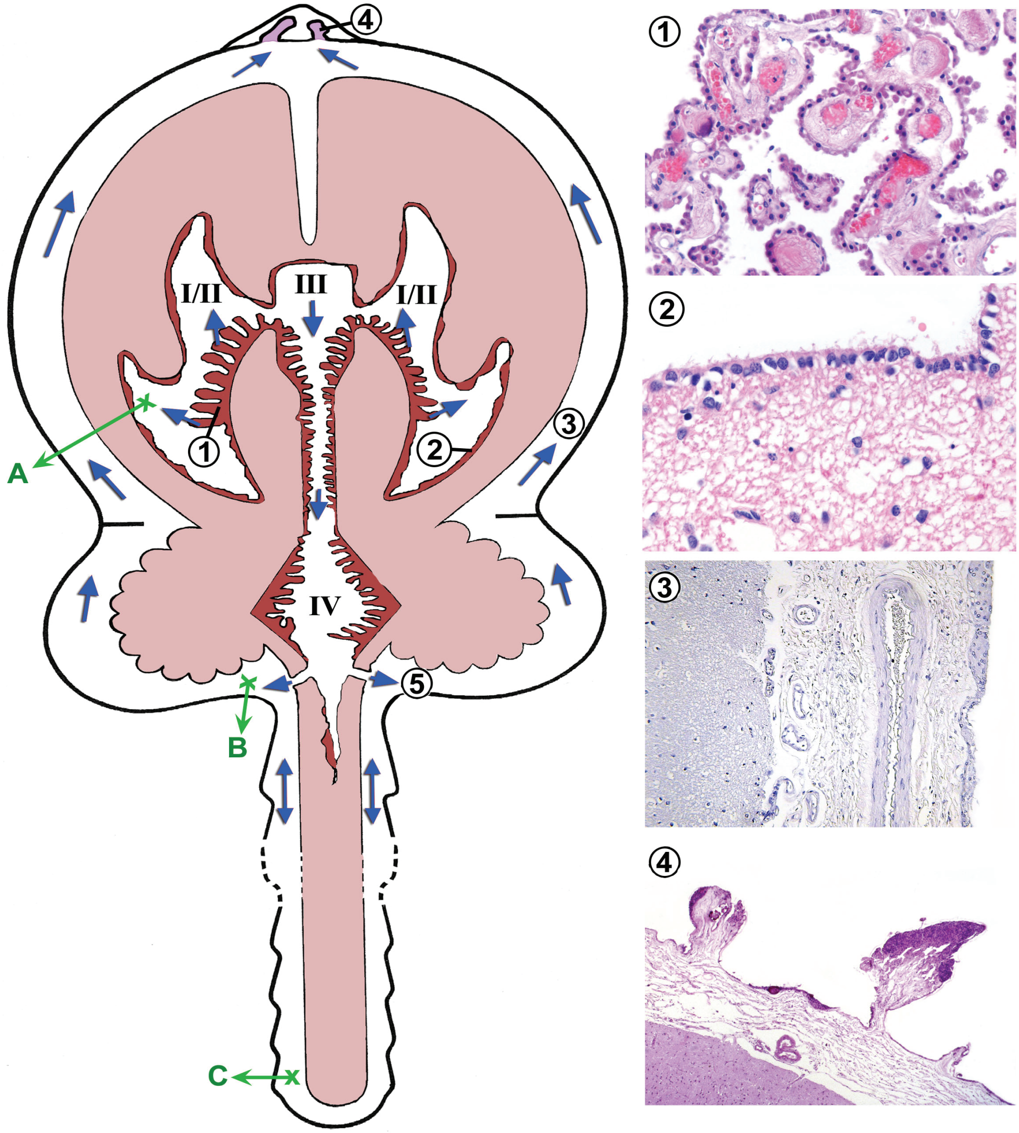


Figure 2.1 Scheme of CSF circulation. The pink areas represent brain tissue. The blue arrows indicate the flow direction of CSF from the cerebral ventricles to the subarachnoidal space. The CSF is produced mainly in the choroid plexuses (1) of the lateral (I and II), third (III) and fourth (IV) ventricles and to some extent also by the ependymal cells which line the ventricles (2) and the external surface of the brain parenchyma. The CSF reaches the subarachnoidal space (3) via a median aperture (foramen of Magendi) and two lateral recesses (foramina of Luschka) (5). CSF surrounds the brain and spinal cord and slowly moves to arachnoid villi (granulations of Pacchioni) (4) which represent the sites of absorption and deposition into the venous circulation. The total CSF protein concentration varies along the CSF pathway: A (256±59 mg/l), B(316±58 mg/l), C(420±55 mg/l)(4).

Tissue is very useful for biomarker discovery because the complexity and dynamic range of the tissue proteome are relatively low compared with the enormous dynamic range observed in serum and CSF. Albumin and antibodies comprise approximately 90% of the total protein concentration in serum and CSF (5). However, tissue from patients is not always easily accessible, especially in neurological disorders other than neoplasms. The resulting low number of tissue samples hampers reliable statistical analysis.

Serum, on the other hand, is easily accessible. CSF is generally obtained by lumbar puncture, a procedure that is more invasive than drawing blood. However, for many neurological disorders, archives of paired serum and CSF samples are collected and stored under uniform conditions with informed consent from the patients. These collections can be used for and are instrumental to biomarker research in clinical neurology. Standardization and an elaborate evaluation of how CSF should be prepared for different types of analysis (proteomics as well as metabolomics) is essential and has not yet been described in large detail in the literature. Standardization will increase the value of CSF biomarker research. Standardization should at least address the contamination of blood products, information about derangements of the blood–CSF barrier (albumin ratio), cell count, and work-up and storage conditions.

2.3 Biomarker discovery in serum and plasma

Serum and plasma are the most frequently used body fluids in the biomarker discovery. Easy accessibility of blood and routine sampling has resulted in large archives of serum and plasma samples of different disease types. These blood derived product have been used for a considerable time for all kind of clinical measurements resulting in clinical information for most samples e.g. number of blood cells, protein concentration. All this information can be used in combination with the proteomics data.

Blood has a lot of different functions: 1) transport of oxygen, nutrients, hormones and waste products, 2) regulation of pH, temperature and osmosis, 4) coagulation and 5) immunological processes. The most abundant proteins that are found in blood all play a role in these basic functions. Besides the most abundant proteins also tissue specific proteins can be found in blood as a result of excretion or cell death, and the general idea is that almost all proteins present in the body can be found in blood to some extent. Serum is obtained after centrifugation of coagulated blood and for this reason, proteins that are involved in coagulation of the blood should theoretically not longer be present. Plasma is obtained by centrifugation of blood, which has been treated with an anti-coagulation reagent (heparin, EDTA or citrate); for proteomics purposes EDTA plasma is most often used since the inference of EDTA with mass spectrometry is minimal. Uniform treatment and preparation is an important issue in biomarker research and standard protocols should be used for collection. Important issues are hemolysis, coagulation time, time before storage, storage time, type of collection tubes and the number of freeze-thaw cycles (6). Apart from the fact that serum or plasma is easy obtainable also much is known about the proteins in plasma and serum, which also could be an important factor to select serum or plasma for biomarker research. Since the HUPO project started, the number of proteins that have been identified has increased greatly and is still increasing (7). This information can be used to create specific databases, which make the identification of differentially expressed proteins easier.

2.4 Biomarker discovery in CSF

CSF is an interesting body fluid to search for biomarkers. CSF has several functions, including protection of the brain to large pressure differences, transport of active biological substances for normal maintenance of the brain and excretion of toxic and waste substances. CSF is produced by ultra filtration and active secretion, mainly in the ventricular choroid plexuses (Figure 2.1). Following passage through the foramina of Luschka and Magendie, the CSF circulates around the spinal cord and convexity of the brain, where bulk reabsorption takes place through the subarachnoid villi into the superior sagittal sinus and other venous structures. The amount of CSF produced is approximately 500 ml in 24 h (4). Due to the blood–brain and blood–CSF barriers, and their active transport systems, the concentrations of proteins and metabolites in the CSF can be quite different compared with serum. In general, the protein concentration in serum (60–80 g/l) is much higher than in CSF (0.2–0.5 g/l) (4). However, the ratio of concentrations of individual proteins can be very different. For instance, prostaglandin D synthase has a concentration that is approximately 30 times higher in CSF than in serum (5). By contrast, haptoglobin is approximately 1000 times less concentrated in CSF than in serum (5).

The exchange between the CSF compartment and the serum compartment is an active process that involves the blood–CSF or blood–brain barrier. The dynamics of both brain- and blood-derived proteins in CSF can be described with mathematical equations (8). These equations enable the calculation of the blood- and brain-derived fraction for each protein. When the blood–CSF barrier is disrupted, the change in CSF flow results in a change in CSF protein concentration. These changes can be calculated, to a certain extent, with these equations (9). This is relevant in CSF biomarker discovery, because the barrier is affected in a number of neurological diseases (e.g., meningitis, tumors and the Guillain–Barre syndrome). The relatively low total protein concentration in CSF and the fact that brain-specific proteins are passively or actively shed into CSF will result in a relatively high concentration of brain-specific proteins in the CSF, which is advantageous for tracing candidate biomarkers. For various diseases of the central nervous system, including glioma, Alzheimer's disease and amyotrophic lateral sclerosis, candidate biomarkers have been described that were first found in CSF (10-13).

2.5 Fractionation

The huge dynamic abundance range of proteins in serum and CSF (≥10 orders of magnitude (1)) and the extreme complexity of this material renders the detection of less abundant proteins nearly impossible using MS. This makes an enrichment and/or depletion of high-abundance proteins necessary for the detection and identification of proteins that are less abundant. The protein ratios and abundance levels in these depleted or enriched samples are influenced and, as a consequence, the reproducibility of these methods is an important concern that should be determined before applying these techniques on large sample groups. In standard analyses of serum samples without any further fractionation, approximately 200 proteins can be identified by nano-liquid chromatography (LC) MS methodology (7). This is only a very small fraction of the number of proteins that are expected to be present in the serum proteome. Therefore, it is unlikely that specific low-abundance markers will be detected without any prefractionation method (14). Currently, the high-abundance proteins can be removed reproducibly with reliable immunoaffinity columns (15) and by using lectin (16) and phosphospecific columns, as reviewed by Garcia and coworkers (17). Methods are available to automate these fractionation processes, which enable high-throughput measurements. Many of the abundant proteins have a transport function. This results in the binding of small molecules and low-molecular-weight peptides and proteins to these transport proteins. By depletion of abundant transport proteins (e.g., albumin) potential interesting molecules can be removed unintentionally. For this reason, the analysis of the fraction with the abundant proteins is also of interest (18, 19).

Different types of separation techniques can be used and combined. A gel-based or chromatographic step is often used to fractionate the proteins. The chromatographic techniques can also be combined with magnetic bead technology, which is ideally suited for automation. Of specific interest for the biomarker discovery field is the development of reverse-phase monolithic columns, because they offer a short analysis time, high resolution, compatibility with matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) MS, and the ability to separate both peptides and proteins (20, 21). Combinations of the aforementioned procedures have increased the number of identified proteins in serum to several thousands (7). Developments and improvement of separation technologies will enable more sensitive, accurate and high-throughput analyses of serum and CSF samples. These developments will be of crucial importance in the field of biomarker discovery, because they will enable the detection and identification of less abundant proteins.

2.6 Mass spectrometric techniques for peptide profiling

The majority of MS-based techniques for peptide profiling use two ionization techniques: MALDI and ESI. In most MALDI analyses, the mass of the analyte is determined by a time-of-flight (TOF) analyzer. However, more recently, MALDI has been coupled to a Fourier-transform (FT) mass spectrometer, which enables very accurate mass measurements. For quantification purposes, MALDI has also been coupled to a triple quadrupole mass analyzer. For ESI, various types of analyzers can be uses (e.g., TOF, quadrupole, triple quadrupole, ion trap and FTMS). Many of these instruments can perform tandem MS (MS/MS) experiments to determine the sequence of a peptide. In ESI, mostly multiple protonated peptides are found, whereas only singly protonated species are observed in MALDI. Thus, one peptide in ESI can produce several protonated species [MH_n^{n+}], whereas in MALDI, only the singly protonated peptide [MH^+] is observed. Thus, MALDI spectra are simpler than ESI spectra. In principal, surface-enhanced laser desorption/ionization (SELDI) is the same as MALDI; the only difference is that a chip with an active chromatographic surface is used to enable the binding of specific peptides and proteins from serum or other body fluids.

2.7 Profiling intact proteins and naturally occurring peptides

Petricoin and coworkers' paper on the use of SELDI in serum to identify ovarian cancer dramatically increased interest in profiling of proteins and peptides in body fluids (22). Since then, several studies of this SELDI approach appeared in the literature. From this literature, it was concluded that it is possible to discriminate various cancers, such as renal carcinoma, ovarian carcinoma and prostate cancer, from non-tumorous controls (22-26). This approach has also been applied to CSF for patients with Alzheimer's disease, dementia, amyotrophic lateral sclerosis and multiple sclerosis (27-32). The concept of the method developed by Petricoin and coworkers is based on a chip with a chromatographic surface, which enables binding of specific peptides and proteins from serum or other body fluids. The profiles of the bound proteins and peptides were measured with a mass spectrometer. Both patient and control sample groups were investigated. A comparative analysis of the obtained profiles enabled the distinction of the patient groups from control groups. The results of these initial SELDI studies appeared to be very promising. However, following validation using independent sets of samples and after re-analysis of the original data and identification of differential peaks in the mass spectra, the initial enthusiasm waned. It appears that this method suffers from a reproducibility problem (33-37). Indeed, the differences that have been identified to date are primarily due to the high-abundance proteins, viz., acute-phase proteins and inflammation-related proteins that are not considered to be disease-specific (38, 39).

Other protein profiling approaches that have been used include prefractionation by chromatographic beads in combination with measurement using high-end MS equipment (40). These approaches result in spectra with improved resolution compared with SELDI, and thus, more reliable profiling comparisons are possible. The surface of the beads is much larger than the surface of the SELDI chips, and therefore, prefractionation is more effective. It has also been demonstrated that the magnetic bead fractionation can be fully automated, which enables the high-throughput analysis of sample groups. The fractionation can also be performed on larger volumes, which is essential for further identification steps. The SELDI and magnetic bead methods are both based on the measurement of intact proteins or naturally occurring peptides in body fluids (40, 41). The drawback of both techniques is that the reproducibility of the measured intensities is relatively low for a biological test (coefficient of variance (CV): 15–30%) (23, 42-44) and even higher if all peptides or proteins detected are taken into account. The proteins and peptides identified demonstrate an intensity bias towards low-molecular-weight proteins or protein fragments.

Naturally occurring peptides in CSF are very attractive for biomarker research, since the CSF compartment acts as a sink for proteolytic products of large proteins, neuropeptides, growth factors and, possibly, also products of neurodegenerative processes. However, the number of studies that are focusing on naturally occurring peptides has been limited to date, probably due to sensitivity issues. The required volume of CSF for these analyses was at least 0.5 ml per analysis (45, 46). This is a large volume in comparison with the 3–5 ml of CSF that is routinely taken from patients for diagnostic purposes. With more recent analysis methods, the sensitivity is increased, thereby enabling the analysis of large groups of samples (47).

2.8 Tryptic peptide profiling

Another method to discover biomarkers is to analyze peptides of enzymatically digested proteins. This approach has several advantages. First, the enzymatic digestion of CSF proteins into peptides improves the resolution and sensitivity of the mass measurements by 1–2 orders of magnitude. For a conventional MALDI-TOF mass spectrometer, resolution of approximately 1500 for proteins with molecular weights in the range of 30,000–100,000 Da can be obtained. For peptides in the mass range of 1000–2000 Da measured in the reflectron mode, the resolution can be as high as 15,000. Also, the signal-to-noise ratio for the analysis of peptides is ten- to 100-times larger than that for measurements on proteins.

Second, since more peptides per protein can be found and analyzed, differentially expressed proteins can be identified more reliably. Furthermore, analyzing enzymatically obtained peptides enables the study of very large proteins (i.e., those not easily amenable to current mass spectrometric techniques). However, the very fact that one protein yields many peptides complicates the mass spectrum, and thus higher resolution is needed (e.g., FT-MS). Both MALDI and ESI can be used as ionization techniques to study the enzymatic digestion of complex protein mixtures. MALDI mass spectrometers and especially MALDI-TOF-MS have a high-throughput capacity, and thus are well suited for screening purposes (48).

The direct sequencing options of MALDI mass spectrometers on complex samples are still limited. The ESI ionization technique combined with online LC separation appears to be more amenable to sequencing. However, the throughput of this method is limited and thus, it is not quite as suitable as MALDI for the analyses of large sample groups (100–1000). In addition, ESI is less tolerant towards contaminations.

In 2003, Ramstrom and coworkers demonstrated that it was possible to use ESI to distinguish patients with amyotrophic lateral sclerosis from controls based on the tryptic peptide pattern of 400 µl CSF samples obtained with a nano-LC/FT-MS approach (49, 50). In this study, 12 patients with amyotrophic lateral sclerosis and ten matched controls were compared. From an average of 3700 detected peptides per sample, 165 peptides were differentially expressed in both groups. An attempt to perform a direct identification based on the exact masses of the peptides with a database search did not result in any significant matches. Indeed, the identification of especially low abundant proteins from peptide profiles remains difficult.

In another study, Ramstrom and coworkers demonstrated that they were able to detect approximately 6500 unique peptide masses in a single LC-FT-MS run of a tryptic digest of 32 µl of CSF. However, only 39 proteins could be identified from these 6500 peptide masses. Since then, the efficiency of identification by MS/MS has improved by better MS equipment and by the development of kits and columns that remove the most abundant proteins. However, even after applying these procedures only approximately 200–500 proteins can be identified in serum or plasma per analysis method (7, 51).

2.9 MALDI-FTMS peptide profiling

More recently, a new technique has been added to the armory for peptide profiling: MALDI-FTMS. This technique offers all the advantages one expects of FTMS such as excellent resolution and mass accuracy, but compared to ESI-FTMS, MALDI-FTMS allows for rapid sample throughput, and, because only singly charged ions are observed in MALDI (vide supra), spectral analysis is relatively straightforward. Also, the problem of ion suppression appears less serious in MALDI than in ESI. Compared to TOF-MS, FTMS offers some critical advantages besides its superior resolution. First, the problem of detector saturation that often occurs in MALDI-TOF and which makes it difficult to analyze and identify masses in the close neighborhood of a relatively intense peak, do not appear to occur in FTMS. A huge advantage of FTMS over TOF-MS is that chemical noise, which can be so annoying in TOF-MS, especially for weak peaks, appears to be absent in MALDI-FTMS, allowing the mass annotation of even very weak peaks. The combined advantages result in a technique that is able to analyze very complex samples with limited sample preparation and relatively large sensitivity.

2.10 Reproducibility

An important factor in the different procedures for peptide profiling is the reproducibility of the intensities of the mass peaks. In the literature, more and more analyses are described that assess the reproducibility of different MS-based techniques (36, 52-57). The reproducibility of measured intensities is relatively low in MALDI- and SELDI-TOF-MS compared with ESI-MS. For MALDI- and SELDI-TOF-MS, the CV for the peak intensities is in the range of 10–30%, depending on how the analysis was performed (all peaks or a peak selection) (23, 42-44, 58). The low reproducibility of peak intensities is caused by ion suppression, variation in the amount of matrix and variation in the crystal homogeneity. The latter depends on a number of factors, including contamination of the analyte and the ratio of matrix and analyte (59, 60). Knowledge of the reproducibility of a system or a technique is required to correctly design an experiment (61). The reproducibility of data is not only dependent on the instrument and the sample preparation, but sample storage and handling also have an influence. Many articles have been published on the effects of sample storage and handling, and as a result, optimal sample storage and handling protocols can be developed (6, 62). These protocols should lead to standardization of sample handling and storage, resulting in more reproducible and reliable data. The degree of reproducibility has a direct effect on the number of replicate measurements required per sample (and thus on the sample size) and on the statistics used for data analysis.

2.11 Quantification

The quantification of peptides or proteins is important for biomarker research, because the difference observed between patients and control samples is not necessarily a black and white phenomenon. In many cases, only a difference in concentration (i.e., intensity) is observed. Many different methods have been developed to use MS in a quantitative way. For example, relative quantification is possible using isotopic labeling (isotope-coded affinity tag (ICAT)) (63), isotope tag for relative and absolute quantitation (iTRAQ) and stable isotope labeling by amino acids in cell culture (SILAC) (64)), whereas other methods employ internal standards. A method for absolute quantification has appeared in the literature, which uses peak intensity and an index-based quantification. The latter is based on the number of sequenced peptides per protein in relation to the theoretical maximum number of peptides (65). Only the ICAT method has been applied to CSF by Zhang and coworkers (66, 67). The reproducibility of the quantitative aspect of the ICAT method is described in detail by Molloy and coworkers, and if all peaks are included in the analyses, the average CV is in the 20–30% range (57). In the publications of Zhang and coworkers the effects of aging and Alzheimer's disease on protein expression in the CSF are investigated.

The quantification by MS of metabolites and small molecules can be performed in an automated and high-throughput manner in a highly reproducible way (CV < 5%). This low CV value was obtained by exploiting selective capability of triple quadrupole mass analyzer. The first quadrupole is used to select an ion of interest. In the second quadrupole, the ion is fragmented into product ions and, in the third quadrupole, characteristic fragment ions of the precursor ion are selected and subsequently detected. This technique is referred to as multiple reaction monitoring and the total ion current of a characteristic fragment can be used for very precise relative quantification without the use of any internal standard or stable isotope. Recently, this method has also been applied to tryptic digests of plasma proteins. The CVs obtained were lower than 10% on average, and when a depletion procedure was applied, even better reproducibilities were obtained (5%) (68).

The aforementioned techniques work very well for the more abundant peptides in most cases. However, for low-abundance peptides and/or proteins the CVs are much larger. Thus, it is preferable to first perform a screening with a semiquantitative technique. In this way, a rough selection of potential biomarkers can be made, but it is possible that some small differences are missed. Sensitivity and reproducibility of the multiple reaction-monitoring technique are relatively high for drug analyses, and this technique will be a useful tool for peptide profiling and quantitative analysis in the near future if the sensitivity and reproducibility can be further improved.

2.12 Analysis

A number of steps are involved in the data processing of mass spectra, including data reduction, peak finding and clustering of data. The most important part of this process is the reduction and filtering of the raw data; however, there are some techniques available in which the analysis can be performed directly on the raw data without any preprocessing. Principal component analyses and artificial neural networks can be used to this end (69, 70). The latter analysis is not often used for large files (gigabyte files) because computer clusters or multiple processor servers are required to perform these types of complex analyses.

In all other cases, data reduction is applied by using only peak lists in the subsequent analysis. Rather than using entire spectra, a baseline subtraction and/or a smoothing is performed on the spectra prior to the peak detection. For peak detection, a peak-finding algorithm is used. The algorithm to be used depends largely on the type of mass spectrometer. For low-resolution data (e.g., obtained on a linear MALDI-TOF-MS), very simple peak-finding algorithms are appropriate and these only include a threshold value and window size. For higher resolution data, more advanced peak-finding algorithms are required, such as sophisticated numerical annotation procedure (SNAP) or thorough high-resolution analysis of spectra by horn (THRASH) (71). These algorithms include the picking of monoisotopic peaks only and its recognition as distinct from background signals by analysis of the isotopic distribution. A second step in the analysis is the alignment of the spectra. This can be performed using internal standards or omnipresent peaks in the profiles (56). For a discussion of the problems associated with spectral alignment and for a more detailed explanation, see (72).

The last processing step is to cluster the data. This is necessary because, in the statistical analysis (which is a comparative analysis), it is required that peaks are compared that either have exactly the same mass or are clustered in a group of masses that are seen as an entity in the statistical analysis. Peaks in a user-defined mass window are given the same mass and are subsequently clustered. The result of this analysis is a table in which the intensity (or presence or absence) of a peak is reported for each detected peak position. This table can be used as an import file into different statistical packages. The statistical analyses can be performed in two ways: uni- and multivariate.

In a univariate analysis, all peak positions are analyzed, resulting in independent p-values for each individual peak. The second type of analysis is the multivariate analysis, in which the data is used as an input rather than considering the individual peak positions. In this case, a combination of peaks that are not individually significant in the univariate analysis can, in combination with each other, have a good predictive value. Combinations of peaks can be used to create a predictive model that classifies unknown samples on the basis of their mass spectra. Different algorithms are used for this purpose, including decision trees, genetic algorithms and a unified maximum separability algorithm. Validation of a model is possible with an internal or external validation set. To perform an internal validation, the original data are used to test the performance of the model. An external validation includes the measurement of a new set of samples on which the original generated model is tested.

Classifiers often have problems with robustness. With an internal validation method, the performance of the model is often acceptable on the original data set. However, when tested with an external validation set, the performance is often considerably less (23). This demonstrates that the models are not presently very robust and that the quality of profiles is not sufficient for clinical applications (73). Thus, identification of the peptides or proteins of interest is crucial for the development of robust clinical assays.

2.13 Identification

In the profiling approach, peaks or peak patterns can be used to distinguish patients from controls without any knowledge of the identity of the peaks. This constitutes an advantage and increases the sensitivity of the method compared with methods in which each individual peak must be identified first. However, the identification of the peptides of interest is still crucial. The identification of peptides or proteins in complex samples, such as serum or CSF, almost always requires an enrichment or purification technique. The enriched or purified fraction can be analyzed on- or offline with a MS/MS approach.

New developments include the identification based on very accurate mass measurements (74). However, a direct identification of one peptide by accurate mass remains difficult, even with the mass accuracies of 0.5 ppm that can currently be obtained with most commercially available FT ion cyclotron resonance mass spectrometers.

Figure 2.2 illustrates that mass accuracies below 0.1 ppm are required to identify a peptide based on its accurate mass using the Swiss protein database. Also shown in this graph is the beneficial effect of using a specific homemade database (in this case for CSF). This specific database (containing information of approximately 356 CSF proteins) results in only one hit for most peptides over 1000 Da. Disadvantages of such specific databases are that peptides that are not present in this database cannot be identified. Furthermore, in most cases, these databases are not present or at best incomplete. If two or more accurate peptide masses per protein are obtained for a protein, a less accurate measurement is required to identify the protein from which the peptides are derived (~0.5 ppm; Figure 2.3).

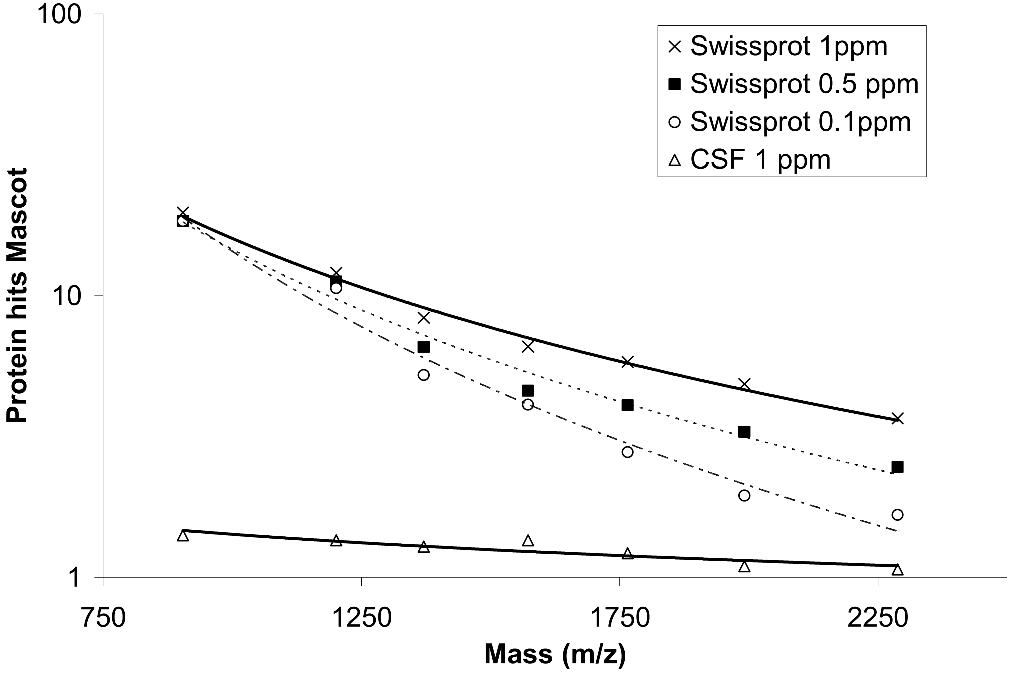


Figure 2.2: Database searches with parts per million and sub part per million mass accuracies. We investigated the effect of sub-ppm mass accuracies for database searches. For this plot we used all tryptic fragments of albumin in the mass range of 800-2,500 Da. For each individual mass a database search is performed in which no miss cleavages were allowed. This is performed under three different conditions against the Swissprot (human) database 0.1, 0.5 and 1 ppm mass accuracy and for one condition (1 ppm) against our in-house created CSF database (containing 360 human CSF proteins) In the plot is displayed the average number of proteins hits for each 250 Da mass interval.

The accurate mass tag and time approach developed by Pacific Northwest National Laboratories (WA, USA) is another method that uses accurate mass measurements to identify peptides. This approach combines high mass accuracy with the retention time of a peptide (75-77). Specific databases that contain as many peptides as possible with an accurate mass and retention time index are created. Thus, identification is based on two observable quantities; mass and time. The disadvantage of this approach is that a specific database must be created, which is time consuming. In addition, such databases will be incomplete and the identity of a protein must already be known. An advantage is that, when such a database is available, the identification of the peptides can be extremely sensitive because MS/MS is not required.

However for the identification of post-translational modifications, mutations and new proteins MS/MS experiments are still crucial. Thus, the further development of sensitive and accurate MS/MS methods is necessary. Accurate MS/MS measurements open new ways for sequencing based on the exact masses of the fragments (78). With this method, the entire sequence or parts thereof can be calculated very accurately without the use of a protein database.

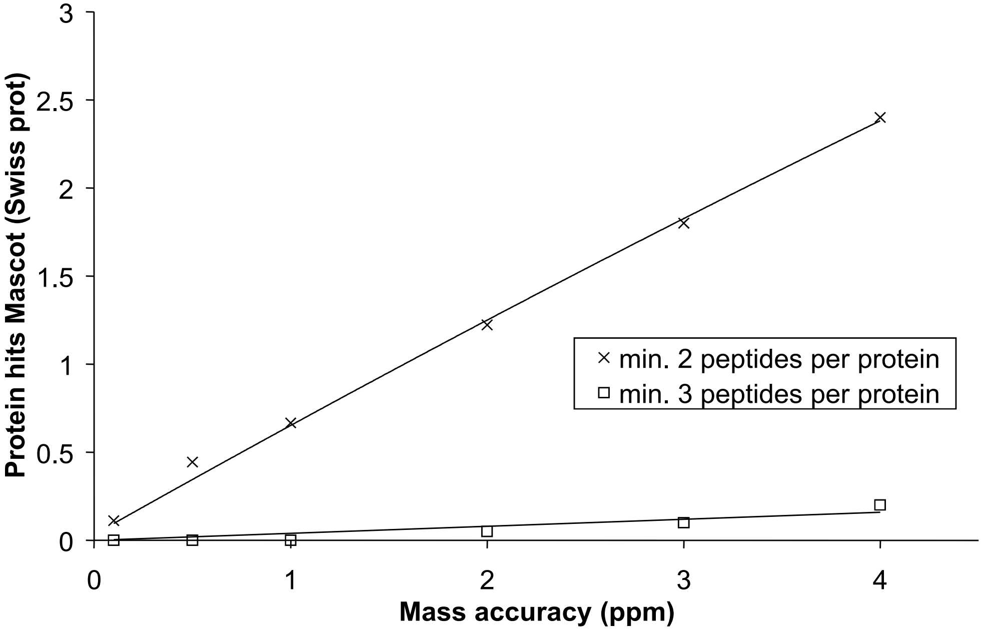


Figure 2.3: Background of database searches. From the random mascot database, 200 peptide were taken in 800-2,000 Da mass range. For different mass accuracies (0.1, 0.5, 1, 2, 3, 4 and 5 ppm) we performed a database search ten times, each time using 20 different peptides from the 200 random peptide masses against the human Swiss-Prot database with no miss cleavages included. For each mass accuracy used, the plot illustrates the average number of protein hits with a minimum of two and three peptides per protein hit.

The sensitivity of MS/MS on MALDI ions in FT-MS by collision-induced dissociation (CID) for complex peptide mixtures is still poor. Other fragmentation methods may be used that could lead to greater sensitivities, such as sustained off-resonance irradiation (SORI) CID, infrared multiphoton dissociation (IRMPD) or electron capture dissociation (ECD) fragmentation. All three of these methods can be performed in the measurement cell of the FT mass spectrometer, thereby circumventing the loss of fragments during transport to the cell. The aforementioned MS/MS methods have already been successfully implemented in ESI-FT-MS. However, the low speed with which MS/MS measurements can be performed in FT-MS still hinders the production of large numbers of high-quality MS/MS spectra; especially during LC/ESI experiments in which the time window to perform the MS/MS is limited. The recently introduced Orbitrap mass spectrometer is capable of high-resolution and high mass accuracy measurements, with mass accuracies as low as 0.5 ppm (including lock mass). In addition, the measurement time per scan is considerably shorter than in FT-MS. In 0.25 s, a MS/MS spectrum can be obtained compared with at least 1 s in FT-MS (79). Thus, this type of mass spectrometer is very useful for reliable and high-throughput identification in biomarker discovery.

2.14 Validation

After the identification of the differentially expressed proteins, a number of validation steps have to be performed. First, the results have to be confirmed on a new and often larger set of samples with the same technique. Often also additional control sets of other related diseases are included to determine the specificity of the potential marker. In addition, samples can be divided into subgroups based on the grade of the disease. In this way the sensitivity and specificity of the marker can be determined more precisely. The next step is to perform the validation of the differential expression of peptides and proteins with another technique and this is most often done by quantitative techniques (immunoassay). In addition, the tumor or disease tissue can be screened for the presence of the candidate biomarker. Sometimes post-mortem or surgical resection material is available to this end. In this way independent evidence can be obtained by e.g. immunohistochemistry or microarray whether a candidate biomarker is related to the target tissue or to a secondary phenomena e.g. inflammation or immune defense. The combination of proteomics and other independent technologies such as immunohistochemistry and microarray can help to validate biomarkers and to understand pathogenesis of a disease. A further external validation and a thorough clinical evaluation is then still required for the transition of a candidate biomarker to a clinically accepted biomarker.

References

1. Anderson, L. "Candidate-based proteomics in the search for biomarkers of cardiovascular disease," *J Physiol* 563 (2005): 23-60.
2. de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F., and Strouboulis, J. "Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice," *Proc Natl Acad Sci U S A* 100 (2003): 7480-5.
3. Pandey, A., and Mann, M. "Proteomics to study genes and genomes," *Nature* 405 (2000): 837-46.
4. Fishman, R. A. *Cerebrospinal fluid in diseases of the nervous system*. Philadephia: W.B. Saunders Company, 1992.
5. Felgenhauer, K. "Protein size and cerebrospinal fluid composition," *Klin Wochenschr* 52 (1974): 1158-64.
6. Rai, A. J., Gelfand, C. A., Haywood, B. C., Warunek, D. J., Yi, J., Schuchard, M. D., Mehigh, R. J., Cockrill, S. L., Scott, G. B., Tammen, H., Schulz-Knappe, P., Speicher, D. W., Vitzthum, F., Haab, B. B., Siest, G., and Chan, D. W. "HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples," *Proteomics* 5 (2005): 3262-77.
7. Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. "Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database," *Proteomics* 5 (2005): 3226-45.
8. Reiber, H. *CSF flow - its influence on CSF concentration of brain-derived and blood-derived proteins.* New York: Plenum, 1997.
9. Reiber, H. "Dynamics of brain-derived proteins in cerebrospinal fluid," *Clin Chim Acta* 310 (2001): 173-86.
10. Zheng, P. P., Luider, T. M., Pieters, R., Avezaat, C. J., van den Bent, M. J., Sillevis Smitt, P. A., and Kros, J. M. "Identification of tumor-related proteins by proteomic analysis of cerebrospinal fluid from patients with primary brain tumors," *J Neuropathol Exp Neurol* 62 (2003): 855-62.
11. Blennow, K., Vanmechelen, E., and Hampel, H. "CSF total tau, Abeta42 and phosphorylated tau protein as biomarkers for Alzheimer's disease," *Mol Neurobiol* 24 (2001): 87-97.
12. Blennow, K., Wallin, A., Agren, H., Spenger, C., Siegfried, J., and Vanmechelen, E. "Tau protein in cerebrospinal fluid: a biochemical marker for axonal degeneration in Alzheimer disease?," *Mol Chem Neuropathol* 26 (1995): 231-45.
13. Romeo, M. J., Espina, V., Lowenthal, M., Espina, B. H., Petricoin, E. F., 3rd, and Liotta, L. A. "CSF proteome: a protein repository for potential biomarker identification," *Expert Rev Proteomics* 2 (2005): 57-70.
14. Govorukhina, N. I., Keizer-Gunnink, A., van der Zee, A. G., de Jong, S., de Bruijn, H. W., and Bischoff, R. "Sample preparation of human serum for the analysis of

tumor markers. Comparison of different approaches for albumin and gamma-globulin depletion," *J Chromatogr A* 1009 (2003): 171-8.

15. Zolotarjova, N., Martosella, J., Nicol, G., Bailey, J., Boyes, B. E., and Barrett, W. C. "Differences among techniques for high-abundant protein depletion," *Proteomics* 5 (2005): 3304-13.

16. Fromell, K., Andersson, M., Elihn, K., and Caldwell, K. D. "Nanoparticle decorated surfaces with potential use in glycosylation analysis," *Colloids and Surfaces B-Biointerfaces* 46 (2005): 84-91.

17. Garcia, B. A., Shabanowitz, J., and Hunt, D. F. "Analysis of protein phosphorylation by mass spectrometry," *Methods* 35 (2005): 256-264.

18. Mehta, A. I., Ross, S., Lowenthal, M. S., Fusaro, V., Fishman, D. A., Petricoin, E. F., 3rd, and Liotta, L. A. "Biomarker amplification by serum carrier protein binding," *Dis Markers* 19 (2003): 1-10.

19. Lowenthal, M. S., Mehta, A. I., Frogale, K., Bandle, R. W., Araujo, R. P., Hood, B. L., Veenstra, T. D., Conrads, T. P., Goldsmith, P., Fishman, D., Petricoin, E. F., 3rd, and Liotta, L. A. "Analysis of albumin-associated peptides and proteins from ovarian cancer patients," *Clin Chem* 51 (2005): 1933-45.

20. Rieux, L., Lubda, D., Niederlander, H. A., Verpoorte, E., and Bischoff, R. "Fast, high-efficiency peptide separations on a 50-mum reversed-phase silica monolith in a nanoLC-MS set-up," *J Chromatogr A* (2006).

21. Wienkoop, S., Glinski, M., Tanaka, N., Tolstikov, V., Fiehn, O., and Weckwerth, W. "Linking protein fractionation with multidimensional monolithic reversed-phase peptide chromatography/mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant proteins," *Rapid Commun Mass Spectrom* 18 (2004): 643-50.

22. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet* 359 (2002): 572-7.

23. Rogers, M. A., Clarke, P., Noble, J., Munro, N. P., Paul, A., Selby, P. J., and Banks, R. E. "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility," *Cancer Res* 63 (2003): 6971-83.

24. Grizzle, W. E., Adam, B. L., Bigbee, W. L., Conrads, T. P., Carroll, C., Feng, Z., Izbicka, E., Jendoubi, M., Johnsey, D., Kagan, J., Leach, R. J., McCarthy, D. B., Semmes, O. J., Srivastava, S., Thompson, I. M., Thornquist, M. D., Verma, M., Zhang, Z., and Zou, Z. "Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer," *Dis Markers* 19 (2003): 185-95.

25. Malik, G., Ward, M. D., Gupta, S. K., Trosset, M. W., Grizzle, W. E., Adam, B. L., Diaz, J. I., and Semmes, O. J. "Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer," *Clin Cancer Res* 11 (2005): 1073-85.

26. Soltys, S. G., Shi, G., Tibshirani, R., Giaccia, A. J., Koong, A. C., and Le, Q. "The use of plasma SELDI-TOF MS proteomic patterns for detection of head and neck squamous cell cancers (HNSCC)," *Int J Radiat Oncol Biol Phys* 57 (2003): S202.

27. Lewczuk, P., Esselmann, H., Meyer, M., Wollscheid, V., Neumann, M., Otto, M., Maler, J. M., Ruther, E., Kornhuber, J., and Wiltfang, J. "The amyloid-beta (Abeta) peptide pattern in cerebrospinal fluid in Alzheimer's disease: evidence of a novel

carboxyterminally elongated Abeta peptide," *Rapid Commun Mass Spectrom* 17 (2003): 1291-6.

28. Lewczuk, P., Esselmann, H., Groemer, T. W., Bibl, M., Maler, J. M., Steinacker, P., Otto, M., Kornhuber, J., and Wiltfang, J. "Amyloid beta peptides in cerebrospinal fluid as profiled with surface enhanced laser desorption/ionization time-of-flight mass spectrometry: evidence of novel biomarkers in Alzheimer's disease," *Biol Psychiatry* 55 (2004): 524-30.

29. Ruetschi, U., Zetterberg, H., Podust, V. N., Gottfries, J., Li, S., Hviid Simonsen, A., McGuire, J., Karlsson, M., Rymo, L., Davies, H., Minthon, L., and Blennow, K. "Identification of CSF biomarkers for frontotemporal dementia using SELDI-TOF," *Exp Neurol* 196 (2005): 273-81.

30. Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., Hochstrasser, D. F., and Sanchez, J. C. "A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease," *Proteomics* 3 (2003): 1486-94.

31. Ranganathan, S., Williams, E., Ganchev, P., Gopalakrishnan, V., Lacomis, D., Urbinelli, L., Newhall, K., Cudkowicz, M. E., Brown, R. H., Jr., and Bowser, R. "Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis," *J Neurochem* (2005).

32. Irani, D. N., Anderson, C., Gundry, R., Cotter, R., Moore, S., Kerr, D. A., McArthur, J. C., Sacktor, N., Pardo, C. A., Jones, M., Calabresi, P. A., and Nath, A. "Cleavage of cystatin C in the cerebrospinal fluid of patients with multiple sclerosis," *Ann Neurol* 59 (2006): 237-47.

33. Diamandis, E. P. "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Mol Cell Proteomics* 3 (2004): 367-78.

34. gSorace, J. M., and Zhan, M. "A data review and re-assessment of ovarian cancer serum proteomic profiling," *BMC Bioinformatics* 4 (2003).

35. Diamandis, E. P. "Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems," *J Natl Cancer Inst* 96 (2004): 353-6.

36. Baggerly, K. A., Morris, J. S., and Coombes, K. R. "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics* 20 (2004): 777-85.

37. Karsan, A., Eigl, B. J., Flibotte, S., Gelmon, K., Switzer, P., Hassell, P., Harrison, D., Law, J., Hayes, M., Stillwell, M., Xiao, Z., Conrads, T. P., and Veenstra, T. "Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis," *Clin Chem* 51 (2005): 1525-8.

38. Diamandis, E. P. "Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics?," *Clin Chem* 49 (2003): 1272-5.

39. Diamandis, E. P. "Proteomic patterns in serum and identification of ovarian cancer," *Lancet* 360 (2002): 170; author reply 170-1.

40. Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., and Tempst, P. "Serum Peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry," *Anal Chem* 76 (2004): 1560-70.

41. Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J., and Tempst, P. "Correcting common errors in identifying cancer-specific serum peptide signatures," *J Proteome Res* 4 (2005): 1060-72.

42. Won, Y., Song, H. J., Kang, T. W., Kim, J. J., Han, B. D., and Lee, S. W. "Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons," *Proteomics* 3 (2003): 2310-6.

43. Schaub, S., Wilkins, J., Weiler, T., Sangster, K., Rush, D., and Nickerson, P. "Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry," *Kidney Int* 65 (2004): 323-32.

44. Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y., and Chan, D. W. "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clin Chem* 48 (2002): 1296-304.

45. Heine, G., Zucht, H. D., Schuhmann, M. U., Burger, K., Jurgens, M., Zumkeller, M., Schneekloth, C. G., Hampel, H., Schulz-Knappe, P., and Selle, H. "High-resolution peptide mapping of cerebrospinal fluid: a novel concept for diagnosis and research in central nervous system diseases," *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 782 (2002): 353-361.

46. Stark, M., Danielsson, O., Griffiths, W. J., Jornvall, H., and Johansson, J. "Peptide repertoire of human cerebrospinal fluid: novel proteolytic fragments of neuroendocrine proteins," *J Chromatogr B Biomed Sci Appl* 754 (2001): 357-67.

47. Selle, H., Lamerz, J., Buerger, K., Dessauer, A., Hager, K., Hampel, H., Karl, J., Kellmann, M., Lannfelt, L., Louhija, J., Riepe, M., Rollinger, W., Tumani, H., Schrader, M., and Zucht, H. D. "Identification of novel biomarker candidates by differential peptidomics analysis of cerebrospinal fluid in Alzheimer's disease," *Combinatorial Chemistry & High Throughput Screening* 8 (2005): 801-806.

48. Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M., and Sillevis Smitt, P. A. "MALDI-TOF Mass Spectrometry Analysis of Cerebrospinal Fluid Tryptic Peptide Profiles to Diagnose Leptomeningeal Metastases in Patients with Breast Cancer," *Mol Cell Proteomics* 4 (2005): 1341-1349.

49. Ramstrom, M., Ivonin, I., Johansson, A., Askmark, H., Markides, K. E., Zubarev, R., Hakansson, P., Aquilonius, S. M., and Bergquist, J. "Cerebrospinal fluid protein patterns in neurodegenerative disease revealed by liquid chromatography-Fourier transform ion cyclotron resonance mass spectrometry," *Proteomics* 4 (2004): 4010-8.

50. Ramstrom, M., Palmblad, M., Markides, K. E., Hakansson, P., and Bergquist, J. "Protein identification in cerebrospinal fluid using packed capillary liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry," *Proteomics* 3 (2003): 184-90.

51. Mann, M., *ASMS*. San Antonio, 2005.

52. Bischoff, R., and Luider, T. M. "Methodological advances in the discovery of protein and peptide disease markers," *J Chromatogr B Analyt Technol Biomed Life Sci* 803 (2004): 27-40.

53. Baggerly, K. A., Morris, J. S., Edmonson, S. R., and Coombes, K. R. "Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer," *J Natl Cancer Inst* 97 (2005): 307-9.

54. Anderle, M., Roy, S., Lin, H., Becker, C., and Joho, K. "Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum," *Bioinformatics* 20 (2004): 3575-3582.

55. Bons, J. A., de Boer, D., van Dieijen-Visser, M. P., and Wodzig, W. K. "Standardization of calibration and quality control using surface enhanced laser

desorption ionization-time of flight-mass spectrometry," *Clin Chim Acta* 366 (2006): 249-56.

56. Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M. "A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra," *Rapid Commun Mass Spectrom* 19 (2005): 865-870.
57. Molloy, M. P., Donohoe, S., Brzezinski, E. E., Kilby, G. W., Stevenson, T. I., Baker, J. D., Goodlett, D. R., and Gage, D. A. "Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling," *Proteomics* 5 (2005): 1204-8.
58. Hong, H., Dragan, Y., Epstein, J., Teitel, C., Chen, B., Xie, Q., Fang, H., Shi, L., Perkins, R., and Tong, W. "Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS)," *BMC Bioinformatics* 6 Suppl 2 (2005): S5.
59. Hoffmann, E., and Stroobant, V. *Mass spectrometry*. West Sussex, England: Wiley, 2002.
60. Bornsen, K. O. "Influence of salts, buffers, detergents, solvents, and matrices on MALDI-MS protein analysis in complex mixtures," *Methods Mol Biol* 146 (2000): 387-404.
61. White, C. N., Chan, D. W., and Zhang, Z. "Bioinformatics strategies for proteomic profiling," *Clin Biochem* 37 (2004): 636-41.
62. West-Nielsen, M., Hogdall, E. V., Marchiori, E., Hogdall, C. K., Schou, C., and Heegaard, N. H. "Sample handling for mass spectrometric proteomic investigations of human sera," *Anal Chem* 77 (2005): 5114-23.
63. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags," *Nat Biotechnol* 17 (1999): 994-9.
64. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics," *Mol Cell Proteomics* 1 (2002): 376-86.
65. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein," *Mol Cell Proteomics* 4 (2005): 1265-72.
66. Zhang, J., Goodlett, D. R., Quinn, J. F., Peskind, E., Kaye, J. A., Zhou, Y., Pan, C., Yi, E., Eng, J., Wang, Q., Aebersold, R. H., and Montine, T. J. "Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease," *J Alzheimers Dis* 7 (2005): 125-33.
67. Zhang, J., Goodlett, D. R., Peskind, E. R., Quinn, J. F., Zhou, Y., Wang, Q., Pan, C., Yi, E., Eng, J., Aebersold, R. H., and Montine, T. J. "Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid," *Neurobiol Aging* 26 (2005): 207-27.
68. Anderson, N. L., and Hunter, C. L. "Quantitative mass spectrometric MRM assays for major plasma proteins," *Mol Cell Proteomics* (2005).
69. Lee, K. R., Lin, X., and Park, D. C. "Megavariate data analysis of mass spectrometric proteomics data using latent variable pojection method," *Proteomics* 2 (2003): 1680-1686.

70. Ball, G., Mian, S., Holding, F., Allibone, R. O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I. O., Creaser, C., and Rees, R. C. "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers," *Bioinformatics* 18 (2002): 395-404.

71. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules," *J Am Soc Mass Spectrom* 11 (2000): 320-32.

72. Yu, W., Wu, B., Lin, N., Stone, K., Williams, K., and Zhao, H. "Detecting and aligning peaks in mass spectrometry data with applications to MALDI," *Comput Biol Chem* (2005).

73. Diamandis, E. P., and van der Merwe, D. E. "Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations," *Clin Cancer Res* 11 (2005): 963-5.

74. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., and Smith, R. D. "Utility of accurate mass tags for proteome-wide protein identification," *Anal Chem* 72 (2000): 3349-54.

75. Strittmatter, E. F., Ferguson, P. L., Tang, K., and Smith, R. D. "Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry," *J Am Soc Mass Spectrom* 14 (2003): 980-91.

76. Smith, R. D., Anderson, G. A., Lipton, M. S., Masselon, C., Pasa-Tolic, L., Shen, Y., and Udseth, H. R. "The use of accurate mass tags for high-throughput microbial proteomics," *Omics* 6 (2002): 61-90.

77. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. "An accurate mass tag strategy for quantitative and high-throughput proteome measurements," *Proteomics* 2 (2002): 513-23.

78. Spengler, B. "De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry," *J Am Soc Mass Spectrom* 15 (2004): 703-14.

79. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. "Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap," *Mol Cell Proteomics* 4 (2005): 2010-21.

Chapter 3

A new method to analyze MALDI-TOF peptide profiling mass spectra

Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M.

Rapid Commun Mass Spectrom 19 (2005): 865-870.

Abstract

In protein and peptide (mass spectrometry) profiling, the number of peaks, their masses and their intensities are important characteristics. Because of the relatively low reproducibility of peak intensities associated with complex samples in MALDI-TOF MS it is difficult to accurately assess the number of peaks and their intensities. In this study we evaluate these two characteristics for tryptic digest of CSF. We observed that the reproducibility of peak intensities was relatively poor (CV = 42%) and that additional normalization or spiking did not lead to a large improvement (CV = 30%). Moreover, at least seven mass spectra per sample were required to obtain a reliable peak list. An improvement of the sensitivity (eventually more peaks are detected) is observed if more replicates per sample are measured. We conclude that the reproducibility and sensitivity of peptide profiling can be significantly improved by a combination of measuring at least seven spectra per sample and a dichotomous scoring of the intensities. Our approach will aid the analysis of large numbers of mass spectra of patient samples in a reproducible way for the detection and validation of candidate biomarkers.

3.1 Introduction

Mass spectrometry is extensively used in biomarker research. This results in the development of methods for the analysis and comparison of large numbers of mass spectra. We describe here a new method for the analysis of tryptic peptide measurements in cerebrospinal fluid (CSF) by MALDI-TOF (Matrix-Assisted Laser Desorption Ionization Time of Flight Mass Spectrometry) that addresses the problem of the low reproducibility of intensities. The standard method of using peak intensities was compared with a new approach that uses peak frequencies. To this end, the reproducibility of peak intensities and of peak frequencies was determined.

The MALDI-TOF MS technique is characterized by a relatively high mass accuracy and precision. In contrast, the reproducibility of measured intensities is relatively low in MALDI-TOF and SELDI-TOF MS (Surface Enhanced Laser Desorption Ionization Time of Flight), compared to MS with electro spray ionization. For SELDI-TOF MS, the coefficient of variance (CV) for the peak intensities is in the range of 10-30% (1-4). The relatively low reproducibility of peak intensities is caused by ion suppression, variation in the amount of matrix, and variation in the crystallization of the matrix as a function of the analyte concentration or ratio. Crystallization depends on a number of factors including contamination of the analyte and the ratio of matrix and analyte (5, 6). If the intensities of peaks in the MALDI-TOF mass spectra of complex samples of peptides or proteins are to be used for analysis, it is essential that the reproducibility of these intensities be determined and used in the analysis.

Indeed, more and more publications appear that describe peptide profiling for the analysis of complex protein samples (7, 8). To our knowledge there are no previous reports on the reproducibility of MALDI-TOF measurements of these complex peptide mixtures. On the other hand, a few papers have reported the reproducibility of measurements of complex protein samples by SELDI-TOF (1, 2, 4, 9-11). However, the conclusions from these publications are limited for a number of reasons. First, the reproducibility of peak detection was calculated for a limited number of peaks (in most cases less than 10). Secondly, the reproducibility was determined for only one or two samples. Thirdly, the reproducibility was calculated from at most eight replicates. In addition, Baggerly et al. (10) also determined the reproducibility of observed protein peaks (how often a peak is present within a mass window of 0.05%). On comparing 24 replicates of a single sample, these authors detected a total of 702 protein peaks. Sixty-eight of these peaks were present in all 24 spectra while 245 peaks were present in only one spectrum. From these findings it can be concluded that, in protein profiling by SELDI-TOF, the reproducibility of both peak intensities and peak positions is low and quite variable. This systematic variability in sample handling and measuring has to be addressed in a statistically correct way.

Compared to protein profiling, peptide profiling offers an important advantage, the mass accuracy and resolution increases by at least an order of magnitude for peptides because they can be measured in the reflectron mode by a TOF MS. This makes peak detection more accurate and precise, and in addition the isotopic peaks can be used to distinguish noise from peptide peaks. Also it is possible to generate directly MS/MS spectra of the peptides of interest. For protein identification by the SELDI technique a protein of interest has to be first purified and digested, and subsequently it can be identified using a high-end mass spectrometer (7).

3.2 Methods

3.2.1 Samples

CSF samples were obtained from controls without a neurological disease or cancer. CSF samples were stored at –80° C.

3.2.2 Sample preparation of tryptic CSF peptides

From each sample, 20 µl CSF was taken and transferred to a 96-well low-binding plate (Nunc, USA). Twenty µl of 0.2% Rapigest (Waters, USA) in 50 mM ammonium bicarbonate was added to each well. The samples were incubated for 2 minutes at 37° C. Four µl of 0.1 µg/µl Gold grade trypsin (Promega, USA) in 5 mM Tris HCl was added to each well, and the 96-well low-binding plates were incubated at 37°C. After two hours of incubation, 2 µl of 500 mM HCl was added in order to obtain a final concentration of 30-50 mM HCl (pH < 2). Subsequently, the plates were incubated for 45 minutes at 37°C. A 96-well Zip C18 microtiter plate (Millipore, USA) was pre-wetted and washed twice with 200 µl acetonitrile per well. Maximum vacuum (20 inches of Hg) was applied to the plate using a vacuum manifold (Millipore). Three µl of acetonitrile was put on the C18 resin without vacuum to prevent it from drying. Each sample was mixed with 200 µl water (HPLC grade) containing trifluoroacetic acid (TFA, 0.1%). Subsequently, the samples were loaded on the washed and pre-wetted 96-well Zip C18 plate (Millipore, USA); 5 inches of Hg vacuum was applied. After the wells were cleared, they were washed twice with 100 µl of 0.1% aqueous TFA. Maximum vacuum was applied until all wells were empty. The samples were eluted in a new 96-well low-binding plate (Nunc, USA) with an elution volume of 15 µl of 50% acetonitrile/water (HPLC grade) containing 0.1% TFA; a pressure differential of 5 inches of Hg vacuum was used. After elution the samples were stored at 4° C in the 96-well plates covered with aluminum seals. All samples were spotted on a MALDI target (600/384 anchor chip with transponder plate, Bruker Daltonics). Two µl of elute was mixed with 10 µl matrix solution containing 2 mg of α-cyano-4-hydroxy-cinnamic acid (HCCA, Bruker Daltonic, Germany) saturated in 1 ml acetonitrile, using an ultrasonic water bath for 30 min. Samples were measured using an automated MALDI-TOF MS (Biflex III, Bruker Daltonic, Germany).

3.2.3 Measurements of spectra by MALDI-TOF MS

MALDI-TOF mass spectra were obtained using a Bruker Biflex III (Bruker Daltonic, Germany) instrument operated in the reflectron mode, using the Bruker standard method for peptide measurements, (default file Bruker, Daltonics, Germany “1-2kD positive” with the mass range changed to 300-3000 Da). Ions were generated by a nitrogen laser (337 nm) and were accelerated to 19 keV. In all experiments the deflection high voltage (“matrix suppressor”) was set at 400 Da. For automation, the following settings were used: an initial laser power of 20% and a maximum of 35%. For a spectrum to be accepted the most intense peak above 750 Da had to have a signal-to-noise ratio of at least 5 and a minimum resolution (FWHM) of 5000. After every 30 laser shots the summed spectrum was checked for these criteria. If the summed spectrum did not meet these criteria, it was rejected. If 13 summed spectra of 30 shots each met the criteria, these were combined and saved; when 50 summed spectra of 30 shots were rejected, the measurement of that spot was ended and the next spot was measured.

3.2.4 Analysis of spectra

The raw spectra files were converted to a general file format for analysis; this was done with a java script. The raw data files were rewritten as standard ASCII files in which the intensities of channel numbers (flight-time windows) of the spectra are displayed. First we developed a peak detection algorithm in R (http://www.r-project.org). In this algorithm a "validated" peak is defined by a) the intensity of the peak ought to be above a predefined threshold and b) the intensity of the peak ought to be the highest in a given window. This peak finding algorithm was tested on a small set of spectra with different settings for the threshold and the mass window. The settings for the peak finding process were chosen such that they resulted in a peak list most resembling the manually assigned peaks. A mass window of 0.5 Da and a threshold of 98.5% (the intensity at that mass window must belong to the 1.5% highest intensity values of the spectrum) were chosen. A quadratic fit with a number of internal calibrants was needed to translate the channel numbers (flight times) into masses; this process also aligns the spectra and makes the peak comparison possible. For this alignment/conversion step, five omnipresent albumin peaks (m/z 960.5631, 1000.6043, 1149.6156, 1511.8433, and 2045.0959) were used. The accurate masses of these albumin peaks were obtained by performing an *in silico* tryptic digest of the human albumin amino acid sequence with MS-digest (http://prospector.ucsf.edu/ucsfhtml4.0/msdigest.htm). During the process of alignment and conversion, the quality of the spectra was also checked. If two or more of the omnipresent albumin peaks were not detected, the spectrum was not used in the further analysis. This peak finding algorithm with the above mentioned settings was used to detect the peaks in all the spectra.

3.2.5 CSF peptide measurements

In a first set of two samples the measurements and analysis were repeated 36 times to study the reproducibility of the validated peaks. Next, a second set of 10 samples, including the two samples mentioned above, were measured and analyzed 12 times to study the reproducibility of peak frequency and peak intensity. Two different types of normalization were used to improve the reproducibility, namely, normalization using total ion current and normalization by a spiked peptide. For the latter type of normalization, two CSF samples were mixed with a synthetic peptide $P_{14}R$ (PPPPPPPPPPPPPPPR) (Sigma Aldrich, USA), to obtain a final concentration of 25 fmol per spot. The procedures of sample preparation, measurement and analysis were identical to those used for the other samples.

We tested the influence of the addition of CSF peptides on the quantitative measurement of a single peptide. Various amounts of the synthetic peptide $P_{14}R$ (25, 50, 100, 200 and 400 fmol) were used to create a calibration curve in the absence and presence of CSF peptides. Each of the different concentrations was spotted and measured 8 times. The analysis of the data was performed using Flex analysis 2.0 software (Bruker Daltonics, Germany). This software package allows the calculation of peak area and peak intensity.

The influence of diluting a CSF sample on the number of validated peaks was tested. Three CSF samples were serially diluted (1x, 2x and 4x) in 50 % acetonitrile/water (HPLC grade) with 0.1% TFA, and measured in 18 replicates. The numbers of validated peak positions for the different dilutions were compared using a Kruskall-Wallis test in SPSS software.

3.3 Results

Tryptic digests of two CSF samples were prepared and measured 36 times each. In figure 3.1 the number of validated peaks, as defined above, is plotted against the number of replicate spectra. Four conditions were tested, i.e., number of validated peaks present in at least 3%, 25%, 50% or 75% of the MS spectra. The numbers of validated peaks present in at least 3% of the spectra do not reach a maximum, indicating a significant contribution from noise. However, the numbers of validated peaks detected in 25, 50 and 75% of spectra stabilizes in the range from 7-12 spectra. This finding indicates that analysis of at least 7 spectra per sample is necessary to obtain the maximum possible number of validated peptide peaks.

We then examined the reproducibility of the numbers of validated peaks detected in a larger panel of ten samples, each analyzed twelve-fold. The average number of total detected validated peaks per sample was 828.8 ± 71.4 (CV 11%). In the 75%, 50% and 25% frequency categories, 153 (CV 9%), 226 (CV 6%) and 328 (CV 2%) peaks are detected on average, respectively (see figure 3.2). If only one replicate per sample was analyzed the average number of validated peaks was 314 (CV 11%).

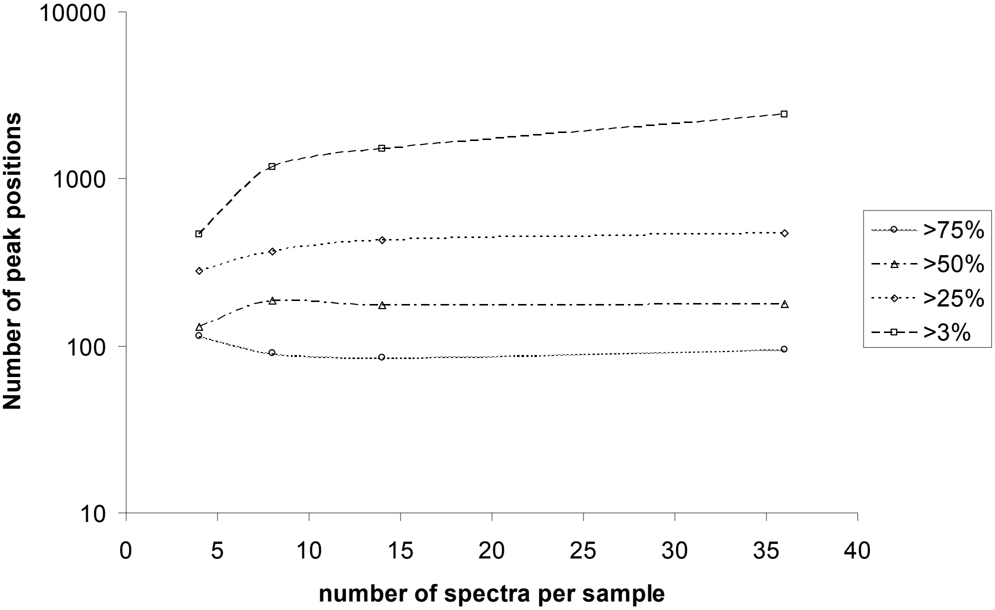


Figure 3.1: The number of validated peaks versus the number of spectra per sample. Tryptic digests of two CSF samples were prepared and measured 36 times each. The number of validated peaks is plotted against the number of replicate spectra for four different frequency categories defined in terms of what percentage of the 36 spectra contain the validated peaks. On the x-axis the number of spectra measured per sample is displayed. The y-axis indicates on a logarithmic scale the average number of validated peaks for the measurements of the two digests.

For the same set of ten samples we determined the CV of the detected peak intensities in 10 CSF samples. Each sample was measured 12-fold, and a normalization on total ion current was applied for each spectrum. The average CV (over the 10 samples) of the peak intensities of peaks present in 100% of the spectra of a CSF sample was 32% (range 19 – 42 %) (see figure. 3.3), with an average number of 83 peaks per sample. The average CV of peak intensities for peaks that are present in more than 50% of the spectra was 29% (range 20-40%) with an average number of 226 peaks. If no normalization on total ion currents was applied, the average CV of the intensities of peaks present in 100% of the spectra increased from 32% to 42%. Also, a normalization on the peak intensity for a spiked peptide with a known concentration was tested for two samples; this resulted in an average CV of 43% after normalization compared to a CV of 65% when no normalization was applied.

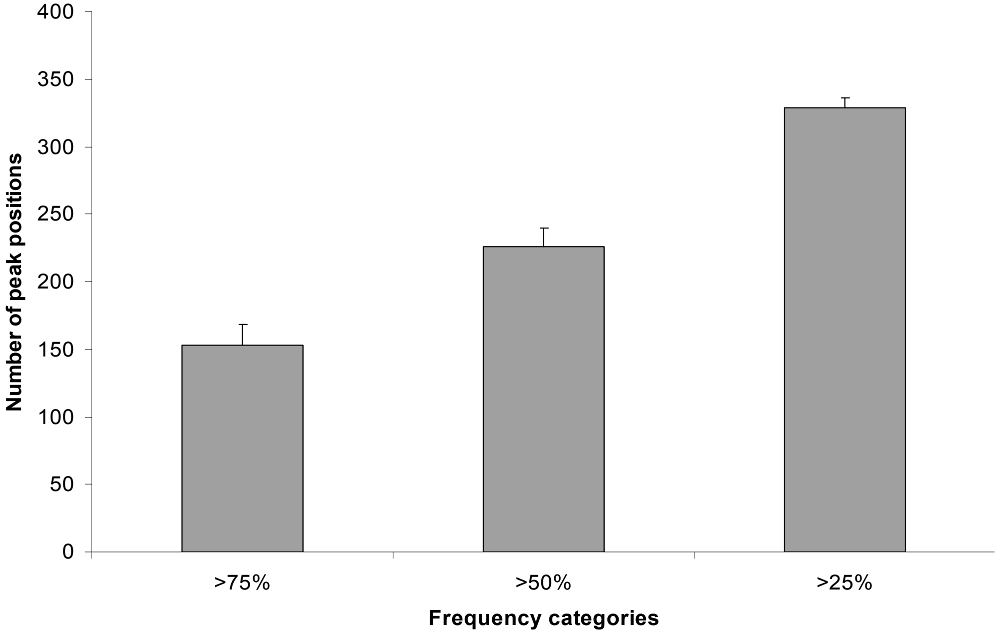


Figure 3.2: The reproducibility of peak frequencies. The reproducibility of the numbers of peaks detected in a larger panel of ten samples, for each of which 12 spectra were obtained, is investigated. The numbers of validated peaks for three frequency categories are shown: the x-axis shows 3 categories defined in terms of the percentage of the 12 spectra per sample the validated peaks are observed. They-axis indicates the numbers of validated peaks, evaluated as the averages over the 10 samples; the error bars indicate the standard deviations evaluated over the 10 samples.

To test the influence of the addition of a complex mixture on the quantification of a single peptide, a serial dilution of a synthetic peptide was measured in both the absence and presence of a tryptic digest. Eight spectra were obtained for each dilution, and the average intensities of the spiked peptide peak over these eight measurements were plotted against spiked concentration. This resulted in a calibration curve with an R^2 value of 0.99 for the synthetic peptide alone, but for the peptide in combination with the CSF tryptic digest a linear correlation between intensity and concentration was no longer observed. For both curves the variation in intensity of the eight replicates at each peptide concentration was relatively large (CV of 48% on average).

A serial dilution (1x, 2x, 4x) of CSF digests was analyzed to obtain 8 spectra each, to detect the influence of the concentration of the sample on the number of validated peptide peaks. This resulted in no significant differences (p=0.88, Kruskal-Wallis) in numbers of peptide peaks for the different dilutions.

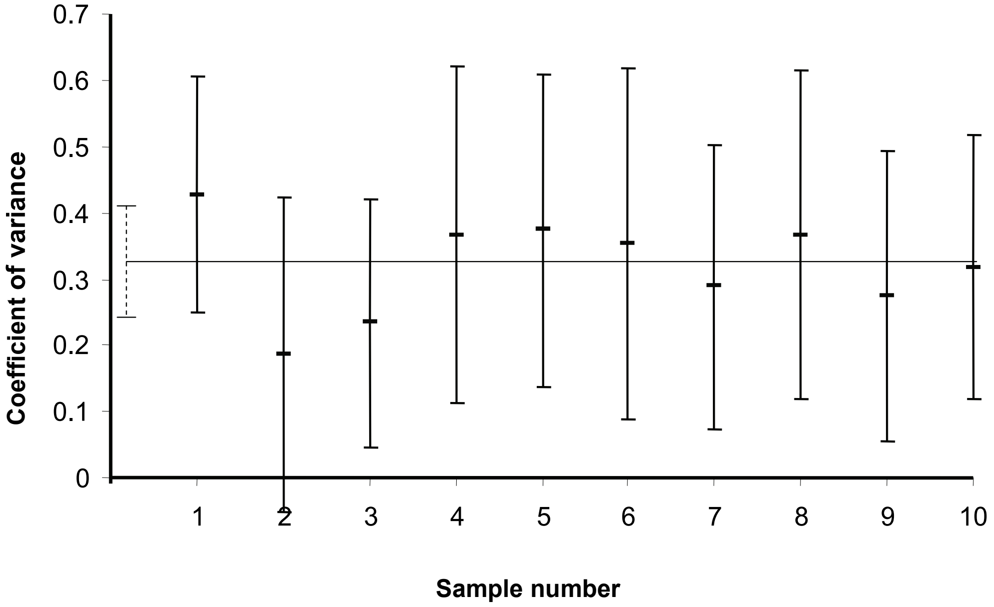


Figure 3.3: Reproducibility of peak intensities. The reproducibility of peak intensities of peaks detected in 100% of the replicate spectra of a sample is investigated. For each of the ten samples 12 replicates spectra are analyzed. The x-axis indicates the (arbitrary) sample number, on the y-axis the average CV of the peak intensities is displayed. The error bars indicate the standard deviation of the different CVs. The horizontal line indicates the average CV for the ten samples and the dashed vertical bar on the line shows the standard deviation of the CV of the ten samples.

3.4 Discussion

This paper reports the development of a method for the analysis of tryptic peptide profiles (see figure 3.4). This new method is based on a new characteristic of mass spectra, the frequency with which a peak is present (as judged by objective criteria) in a number of replicate spectra. This new characteristic improves the reproducibility and the sensitivity of the detected peaks (see Table 2.1).

Table 2.1: Improved reproducibility and sensitivity when new characteristic is used.

	Classical	New
Intensity	32% CV	NA*
Sensitivity (number of peaks)	314	829
Reproducibility of the number of peaks	11% CV	2% CV**

*Not applicable

**For peaks that are present in at least 25% of 12 replicates (328)

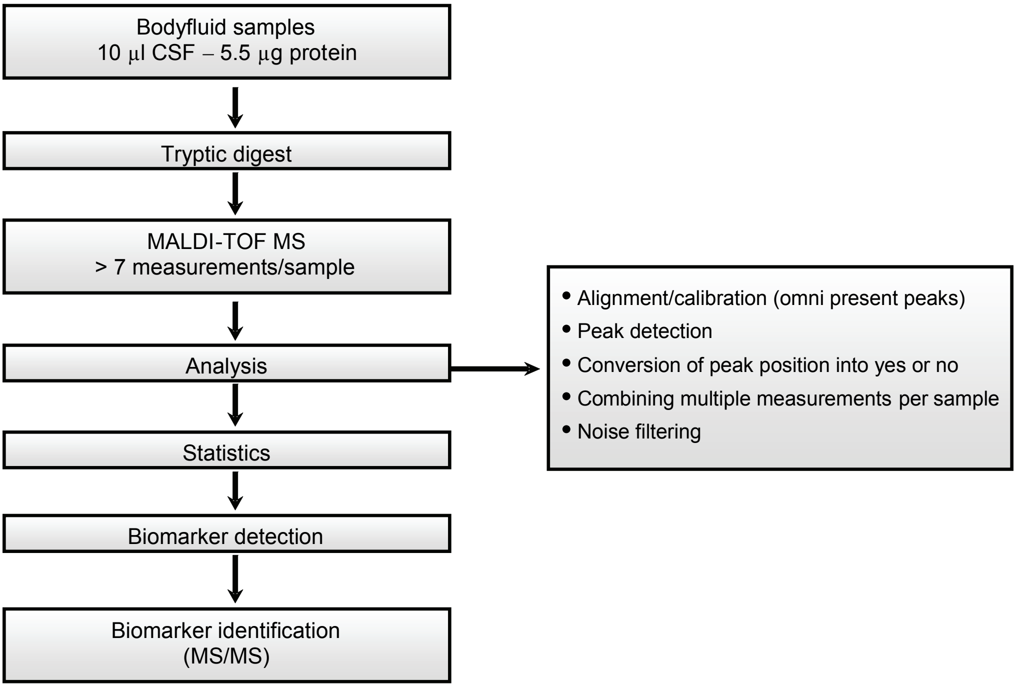


Figure 3.4: Flow chart of a new method to analyze peptide profiles for biomarker research.

The standard method used in protein profiling, viz., comparison of peak intensities, has several disadvantages. The main disadvantage is the relatively low reproducibility of peak intensities with MALDI-TOF MS. Our results show that the CVs of peak intensity for peptide present in 100% of the spectra of a sample are of the same order of magnitude as for protein profiling, i.e., on average 42%, ranging from 6%-120%. An increase in reproducibility was obtained by normalization using total ion current (a decrease of CV of approximately 10% to

32%). Normalization on a spiked peptide with a known concentration did not result in a further improvement of the reproducibility of peak intensities of CSF peptides. This is probably due to the differences in the ionization process for different peptides. These data show that the CV of peak intensity, even for the most abundant peptides with a normalization on total ion current, is still not lower than 30%.

Also there is no longer a linear correlation between intensity and concentration of a specific peptide measured in the presence of a CSF digest although such a linear correlation was obtained for the peptide in the absence of the CSF peptides. This indicates that quantitative peptide analysis is less straightforward for the discovery of biomarkers, for which large numbers of patient samples have to be analyzed.

The presence or absence of a peak is quite variable if multiple measurements of a single sample are compared. This variability in MALDI-TOF spectra of complex samples is mainly due to ion suppression (10). Only a few peaks are always present. To address the variability more than one measurement per sample is needed to obtain a reliable profile of the validated peaks for any sample. The number of validated peaks per sample depends on the number of replicate spectra examined. We obtained an optimum when 7 or more replicates per sample were studied. With more than 7 replicates per sample an increase in the number of validated peaks is still observed, but the peaks that are added after 7 replicates are mainly peaks present in less than 25% of the spectra. Since these peaks are not very reliable they should not be included in the analysis. We conclude that to obtain a peak list with the maximum number of "reliable" peaks, at least seven spectra should be combined and evaluated with respect to the category requiring that peaks deemed reliable, must be validated for >25% of the spectra. The reproducibility for the number of peaks present in at least 25% of 12 replicates of a sample is relatively high (2 % see Table 2.1).

We have also shown that the reproducibility of the peak intensity for peptide profiling is low compared to that of the number of validated peaks. A quantitative comparison of samples by MALDI-TOF MS must be addressed with caution. A better method to analyze and measure large peptide profiling experiments would be to use only the presence or absence of a peak (8) and to combine the information from multiple measurements of a sample. This new method (see figure 3.4) is not only more sensitive (more peaks are detected) but also the reliability of measured m/z values improves. The use of peptide digests instead of proteins for biomarker discovery offers the additional advantage of the possibility of direct identification of the peptides of interest by MS/MS (7). Modern MALDI TOF/TOF mass spectrometers can handle large numbers of samples for such analyses.

Whole CSF or pre-fractionated CSF proteins are attractive for peptide or protein profiling due to the relatively low total protein concentration compared to serum (12, 13). However, biological variation in body fluids and variations in measurements (not related to biomarkers) can not be neglected in the discovery of biomarkers. The present method for analysis of MALDI-TOF mass spectra addresses the variations in body fluids and in measurement using a MALDI TOF mass spectrometer. This will help to analyze large numbers of mass spectra in a more reproducible and reliable way.

References

1. Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y., and Chan, D. W. "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clin Chem* 48 (2002): 1296-304.
2. Rogers, M. A., Clarke, P., Noble, J., Munro, N. P., Paul, A., Selby, P. J., and Banks, R. E. "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility," *Cancer Res* 63 (2003): 6971-83.
3. Schaub, S., Wilkins, J., Weiler, T., Sangster, K., Rush, D., and Nickerson, P. "Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry," *Kidney Int* 65 (2004): 323-32.
4. Won, Y., Song, H. J., Kang, T. W., Kim, J. J., Han, B. D., and Lee, S. W. "Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons," *Proteomics* 3 (2003): 2310-6.
5. Bornsen, K. O. "Influence of salts, buffers, detergents, solvents, and matrices on MALDI-MS protein analysis in complex mixtures," *Methods Mol Biol* 146 (2000): 387-404.
6. Hoffmann, E., and Stroobant, V. *Mass spectrometry*. West Sussex, England: Wiley, 2002.
7. Koomen, J. M., Zhao, H., Li, D., Abbruzzese, J., Baggerly, K., and Kobayashi, R. "Diagnostic protein discovery using proteolytic peptide targeting and identification," *Rapid Commun Mass Spectrom* 18 (2004): 2537-48.
8. Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., and Tempst, P. "Serum Peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry," *Anal Chem* 76 (2004): 1560-70.
9. Baggerly, K. A., Morris, J. S., and Coombes, K. R. "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics* 20 (2004): 777-85.
10. Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C., and Coombes, K. R. "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics* 3 (2003): 1667-72.

11. Coombes, K. R., Fritsche, H. A., Jr., Clarke, C., Chen, J. N., Baggerly, K. A., Morris, J. S., Xiao, L. C., Hung, M. C., and Kuerer, H. M. "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization," *Clin Chem* 49 (2003): 1615-23.

12. Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., Hochstrasser, D. F., and Sanchez, J. C. "A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease," *Proteomics* 3 (2003): 1486-94.

13. Mannes, A. J., Martin, B. M., Yang, H. Y., Keller, J. M., Lewin, S., Gaiser, R. R., and Iadarola, M. J. "Cystatin C as a cerebrospinal fluid biomarker for pain in humans," *Pain* 102 (2003): 251-6.

Chapter 4

A database application for pre-processing, storage and comparison of mass spectra

Titulaer, M.K., Siccama, I., Dekker, L.J., van Rijswijk, A.L.,
Heeren, R.M., Sillevis Smitt, P.A., and Luider, T.M.

BMC Bioinformatics 7 (2006): 403.

Abstract

Statistical comparison of peptide profiles in biomarker discovery requires fast, user-friendly software for high-throughput data analysis. Important features are flexibility in changing input variables and statistical analysis of peptides that are differentially expressed between patient and control groups. In addition, integration of mass spectrometry data with the results of other experiments, such as microarray analysis, and information from other databases requires a central storage of the profile matrix, where protein id's can be added to peptide masses of interest. A new database application is presented, to detect and identify significantly differentially expressed peptides in peptide profiles obtained from body fluids of patient and control groups. The presented modular software is capable of central storage of mass spectra and results in fast analysis. The database application is capable to distinguish patient Matrix Assisted Laser Desorption Ionization (MALDI-TOF) peptide profiles from control groups using large size datasets. The modular architecture of the application makes it possible to handle also large sized data from MS/MS and Fourier Transform Ion Cyclotron Resonance (FT-ICR) mass spectrometry experiments.

4.1 Introduction

In mass spectrometry (MS), analysis of mass spectra is possible with various software packages. In general these software applications work fine for the analysis of individual spectra, but lack the ability to compare very large number of spectra and address differences in (peptide) profile masses to certain groups, such as patient and control groups. Therefore, it is necessary to have fast, user-friendly software for high throughput data pre-processing, flexibility in changing input variables and statistical tools to analyze the peptides that are significantly differentially expressed between the patient and control groups. Statistical calculations are performed within seconds to at most several hours. To the best of our knowledge the only open source project that is capable of peptide profiling with raw MS fid (free induction decay) files (Bruker Daltonics, Germany) is the RProteomics 3-tier architecture of the Cancer Biomedical Informatics Grid, presented in a concurrent versions system (cabigcvs.nci.nih.gov). In the RProteomics project, the main development language is R and the application has a web interface.

This paper describes an application where MS data preprocessing is expanded with a kind of Laboratory Information Management Systems (LIMS). It requires no grid architecture, can even be installed on a stand-alone computer, and due to local file interfaces can easily be integrated with commercial statistical software packages visualization applications, such as Spotfire^{TM} (www.spotfire.com) and Omniviz^{TM}. The presented software architecture is capable of central storage of mass spectra and analysis results. A central database holds all meta-data. Meta-data consist of the origin of the measured samples, experiments performed on different mass spectrometers and allocation of samples to different groups. Meta-data can also link the experimental results to clinical information. Information from the database can be retrieved with Structured Query Language (SQL) and can be linked to other databases on common keys, such as patient code. In this study, the application is built in fast Java code, which provides an excellent Graphical User Interface (GUI), and statistic R routines are called if needed. In addition, the protein origin of the significant peptide masses can be identified by comparing the centrally stored peptide masses of interest with those calculated from the human mass spectrometry protein sequence database (for example MSDB) or by mass spectrometry assisted sequencing. Both identification techniques use the Mascot™ search engine (www.matrixscience.com).

4.2 Methods

4.2.1 Software architecture, packages and interfacing

The MS analysis software architecture consists of 4 pillars, a GUI written in Java^{TM} (java.sun.com), a MySQL^{TM} database (www.mysql.com), which contains all metadata, such as experiment numbers and sample codes, and a FTP (File Transport Protocol) server to store all raw MS fid files and processed data and fourth R. The software package R is used for statistical calculations (www.r-project.org). Figure 4.1 gives a schematic overview of the architecture. The Java software components are developed and tested on the Eclipse^{TM} platform (www.eclipse.org). The raw MS fid files can manually be selected by the Java GUI on the client and stored on a central FTP server. For calculations, the Java client retrieves the information in these files again. After processing of the data, the results of analysis are transported to the FTP server again. The FTP file storage is installed on a central server, and the information can be retrieved by different Java client workstations. However, for testing, the FTP service and MySQL database are both installed on the client workstation, with hostname *localhost*. Special Java archives (Jar's) have to be in the Java Virtual Machine's class path. The edtftpj-1.4.8.jar (www.enterprisedt.com) provides an interface for programming the standard FTP commands in Java. The Java Database Connectivity (JDBC) driver mysql-connector-Java-3.1.6-bin.jar (www.mysql.com) gives an interface for SQL database access (1). In this way, a communication between the Java client and the MySQL database or FTP service is established. There are several ways to set up an interface between Java and the statistical software package R (2). Java's Runtime.exec() command is used in the database application. The advantage of applying this method is that it requires no other adaptations than a default installation of R. Lemkin et al. (3) implemented the method in the Micro Array Explorer project (maexplorer.sourceforge.net). The Runtime.exec() command in Java can execute a Windows^{TM} cmd.exe (command interpreter) batch file. The batch file, Rterm.bat, subsequently starts an Rterm^{TM} process. The Rterm process has a file-based communication with Java (Figure 4.1). The Java client generates all R scripts and R input files. The name and path of the input and output files are defined in the generated R script. Java waits until Rterm has finished the job, and reads the output file(s). The Java application warns if Rterm is not installed in the default installation path on the client workstation.

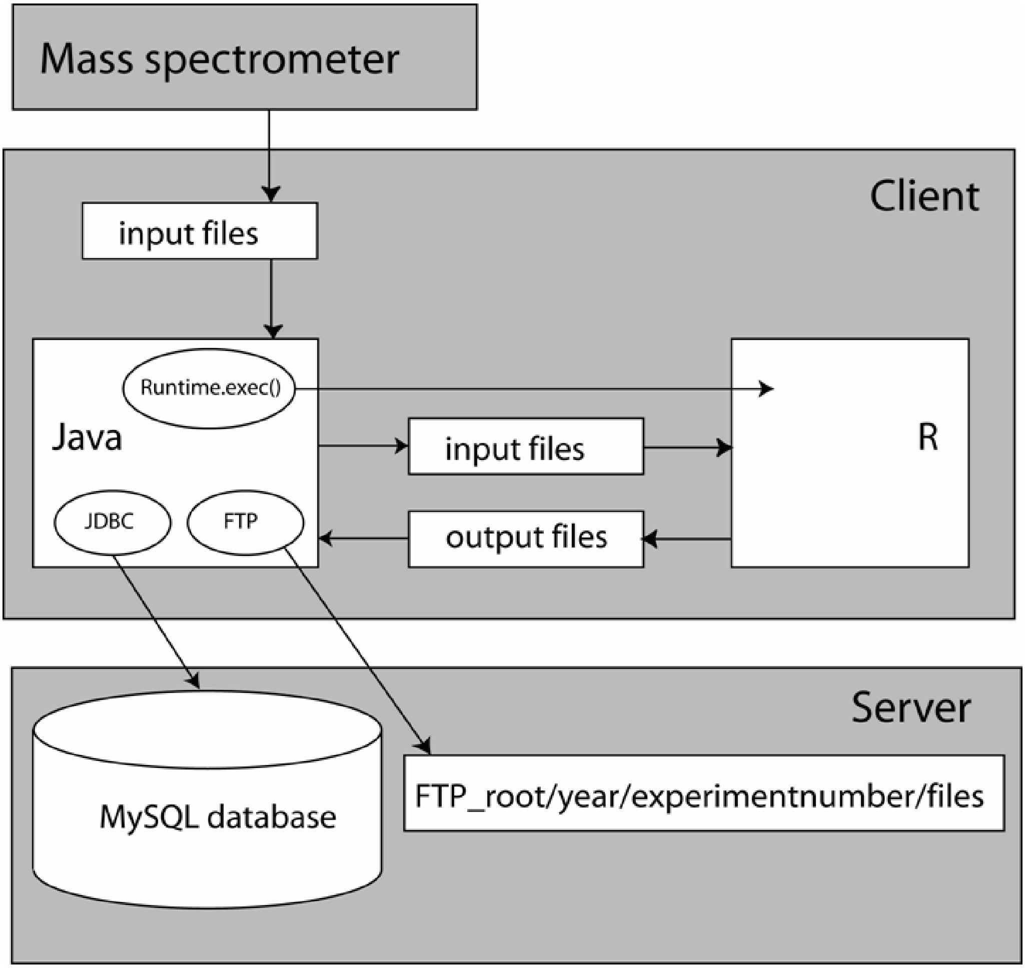


Figure 4.1: System architecture. The system architecture consists of a client JAVA code for fast processing of data while a MySQL database on the server contains all the MS metadata. An FTP service puts all the raw files and processed data on the server and client R is used for statistical analysis.

4.2.2 Database design

The database is kept at minimum size. The ERD distinguishes two sets of tables or entities. One set contains records with metadata of the MS measurements, namely equipment, experiment, result, sample, group, person, material and origin. The other set consists of system tables. The records of the table result contain pointers to the MS files on the FTP server. These pointers are the filenames in the fields of these records, which also hold information about

MALDI target plate spot positions. Each sample generates one or more mass spectra. Therefore, records in the table result keep the foreign key of the sample records. The database application selects the replicate spectra of each sample in order of the ascending *resultid* value of the result table. The reversed selection of replicate spectra is also studied by changing the order in descending *resultid* value of the result table. The samples have to be allocated to a certain group; control, breast cancer, or breast cancer with LM. The foreign key of the table group in the table result achieves indirectly a link between sample and group. There is no direct link between the table sample and group. In this way, samples can be allocated to different groups in different experiments. This gives more flexibility to the application, and avoids storage of redundant sample information. Information about the origin of a sample, for example *lumbar puncture*, can be stored in a table, as well as information about the material, *CSF*. A patient-id in records of the table person can link the MS results with other clinical data. A second set of system supporting tables are named *systemcode, systemcodeitem, itemvalue*, and *unit*. The mass spectra can be internally calibrated, the masses for internal calibration are stored in records of the table *itemvalue*. A series of calibration masses can be named and stored in a record of the table *systemcodeitem*. The table *systemcode* offers the possibility to store more series of internal calibration masses.

4.2.3 GUI components and functions

The software architecture contains the following GUI components and functions: 1) Import of the MS files from the (local) file system and transport of these files to the FTP server; 2) search and selection of table records; 3) a screen to update or insert the records; 4) allocation of the samples in different groups; 5) creating the profile matrix; and 6) performing the Wilcoxon-Mann-Whitney rank sum test on matrix values (4-6). The GUI to select and import MS files to the FTP server is based on the Java's JFileChooser Class (1). JFileChooser is a member of the Swing^{TM} library for the GUI design. Most GUI components were built with this toolkit to keep the same look and feel throughout the application (except for the ugly JTextField), though SWT is getting more and more popular for these purposes, like the (SWT based) Eclipse IDE for development. One or more file(s) or even complete directories can be selected, and all files including subdirectories are transported to the FTP server location. The combination of file type and the type of instrument determines how the data in the files should be processed. File types that can be imported into the system are at present binary fid and text files in ASCII format (American Standard Code for Information Interchange). This can be extended with any other file type. If the file type is fid, Bruker related acqu and acqus files, containing the calibration constants are also transported to the FTP server. The calibration constants have a totally different meaning for data measured with the TOF or FT-ICR technique. When the mass spectra are imported into the system, result records are created in the database. Each record of the table result refers to each mass spectrum, which is measured for a certain sample. For statistical analysis of the data, these result records have to be linked to samples and samples have to be placed in groups. The allocation module achieves this by constructing a link between the records of the result and sample table, and the result and group_ table, respectively. The field filename in a record of the result table holds the spot position on the anchor chip, because it is part of the filename. Records in the table sample and table group_ hold the sample and group codes in their table fields. Table maintenance screens can add additional sample information, such as person, material, and origin. The matrix of number of occurrences of mass peaks in replicates of different samples allocated to different groups is created in another module. Three different matrices are produced simultaneously, one with the number of occurrences of masses in replicate spectra of different samples, a binary table with number of occurrences of masses above a specific threshold, and finally a matrix with the mean intensity of the present peaks in the mass spectra replicates. The matrices of all samples are stored in Comma Separated Value (CSV) format on the FTP server and in the local document root. The total matrix can be visualized by importing the table in the statistical package Spotfire. R's Wilcoxon-Mann-Whitney rank sum test is performed for each matrix peptide, based on the numbers of peptide mass occurrences per sample in different groups. The Wilcoxon-Mann-Whitney test discriminates the peptide masses between the groups with a probability value (p-value). The frequency distribution of the calculated p-values of the peptide masses in the matrix is presented in a histogram. A separate Wilcoxon-Mann-Whitney GUI generates this histogram and creates a list of the masses with corresponding p-values. In this screen, the test can be performed on matrices generated in different experiments and between different groups. The results of the Wilcoxon-Mann-Whitney rank sum test on a matrix are stored in a file with CSV format. The p-values of all peptide masses, as well positive (+) as negative (-) expressed between the groups are listed in this file. The file is stored on the FTP server and in the local document directory.

4.2.4 Algorithms

Calibration constants

A small storage size of the files on the FTP server is guaranteed, due to the fid format of MS spectra, a byte array of 92000 channel intensities. The TOF, time, can be calculated from the MS channel number, i, in the fid files by

$$time_i = DELAY + (i \cdot DW) \quad i = 1,2,...,92000 \tag{1}$$

The values of the constants DW (dwell time) and DELAY are stored in the acqus and acqu files, which are also transported to the FTP server. Other important values are those of the ML1, ML2 and ML3 calibration constants in the acqus files, which are used to calculate the peptide masses from the TOF. Theoretically, the square root of the mass over charge, is proportional with the TOF time.

$$0 = A \cdot \left(\sqrt{\frac{m}{z}_i} \right)^2 + B \cdot \sqrt{\frac{m}{z}_i} + C(time_i) \tag{2}$$

Therefore, the value of constant B is about 40.000 times larger than the value of constant A, where, $A = ML3$, $B = \sqrt{\frac{10^{12}}{ML1}}$ and $C(time_i) = (ML2 - time_i)$

The mass over charge is $$\frac{m}{z}_i = \left(\frac{-B + \sqrt{B^2 - 4 \cdot A \cdot C(time_i)}}{2A} \right)^2 \tag{3}$$

Peak finding

A peak list consists of mass over charge (m/z), channel number i, and intensity. It is constructed from the data in the raw fid files. A histogram of the number of channels with a specific intensity can be constructed. The integral under the distribution curve represents the amount of 92000 instrument channels. From this distribution curve, the R quantile function calculates an intensity threshold, where the probability is 98 % to find channels with a lower intensity. The effect of changing R quantile percentages between 97 and 99 % in the create matrix GUI is examined. The MS peaks are expected to be in the channels numbers, i, with intensity higher than this threshold, namely in the range of the 3 % highest intensities. The peak finding algorithm determines the highest channel intensity within a certain mass over charge (m/z) window, for example 0.5 Da at both sides. A second condition is that this local maximum intensity must be above the quantile threshold intensity. Noise spectra do not contain real peaks with a high intensity flanks. As a consequence, many noise peaks are above the quantile threshold. Peak lists with too many peak masses above an arbitrary number of 450 fall off, because a large part of these peak positions are probably noise peaks.

Internal calibration

Internal calibration is necessary to align all the spectra in the matrix. There are several methods reported to align mass spectra datasets. The alignment algorithms of Wong et al. (7) and Jeffries (8) have in common that they use special reference masses or peaks between the spectra. Wong et al. have developed an algorithm written in C^{++} where spectral data points are added or deleted in regions with a low intensity, in order to a shift peaks. This algorithm has a slight effect on the shape of the peaks. However, the signals in MS are presented by peaks and not by the regions of minimal intensity. Jeffries compares peaks lists generated from mass spectra. He uses R's smooth spline function to correct measured masses with help of reference calibrate masses. A smooth spline function, f_λ, is drawn through the ratio of measured over real mass on the y-axis against the measured mass of the calibrate peaks on the x-axis, which results in a factor close to 1. Division of the measured masses by the calculated function f_λ interpolates all data points. Theoretically, a cubic spline function needs to pass through all of the calibrate data points. This results in a lot of curvature. A smooth spline is a compromise; where the function may deviate from calibrate data points within a certain limit, due to a factor λ, which diminishes the amount of slope. The amount of slope is expressed by the integrating the square of the second derivative of the spline function (8). Another alignment algorithm assumes no knowledge of peaks in common (9, 10). This method considers the shape of the spectra, and aims to minimize the phase differences between the spectra. This process is named dynamic time warping. It is however easier to calibrate the channel numbers of the MALDI-TOF equipment against known masses, since the square root of mass over charge is theoretically proportional to the time. This dependency can be fit with a polynomial function. The masses in the peak list are internally calibrated, using the at least 4 of the 5 omnipresent albumin masses. The channel numbers in the peak list, with corresponding masses, which are the closest with a window of 0.5 Da to one of the albumin masses, are determined. Peak lists without the required number of albumin masses fall off. The channel numbers, i, and corresponding albumin masses, $\frac{m}{z}_i$, are fit in a second-degree polynomial function

$$\frac{m}{z}_i = a[1] \cdot i^2 + a[2] \cdot i + a[0] \tag{4}$$

The coefficients, a[0], a[1], and a[2] are calculated with R's linear model (lm) function where $y = \frac{m}{z}_i$, $x = i$, and a is the array of a[0], a[1], and a[2]

$$ft3 \leftarrow lm(y \sim I(x\hat{}2) + x) \tag{5}$$

$$a \leftarrow coeff\,(ft3) \tag{6}$$

All peptide masses in the peak list are recalculated, using these coefficients and the polynomial function.

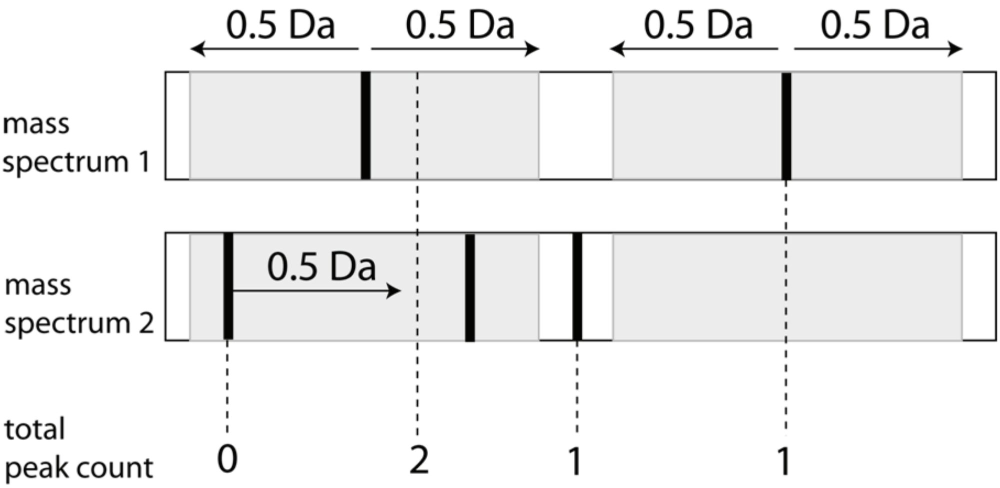


Figure 4.2: Data reduction. Data reduction by finding peak maxima and combining the measured peptide masses in the different spectra. Occurrences of masses with a window of 0.5 Da are summed, and the average value of the mass is calculated (dashed line). The occurrence of only one peptide in the second spectrum is summed in the mass window of the first spectrum. The second peak of spectrum 2 and not the first one is combined with the first peak of spectrum 1, since it has the closest distance in mass (Da) with the first peak of spectrum 1. Not previously registered masses in the first spectrum are added to the list. The clustering continues iteratively through all mass spectra of the samples in the matrix.

Data reduction

Last step is the creation of the profile matrix, which consists of peptide masses in all spectra of the samples in the columns against the occurrences in replicate spectra of the samples in the rows. The matrix is the input file for the Wilcox-Mann-Whitney test, but can also be input for other statistical packages, like Spotfire. The matrix is stored on the FTP server, as well as in the local document directory. Figure 4.2 schematically shows the clustering of two spectra in the matrix. Within a mass window of 0.5 Da at both sides of a peptide mass in the first spectrum, the occurrence of at least one peptide mass in the second spectrum is investigated, closest in distance to the peak in the first spectrum. If that is the case, the average mass of both peptide masses is calculated, and the number of occurrences in both spectra summed. For each mass spectrum, only one peptide occurrence, 1 or 0, is summed for each mass window. If a peptide mass is present in the second spectrum, but not in the first spectrum it is added to themass list. The average intensity of the present masses is also stored in a separate matrix. The clustering continues iteratively though all spectra of the samples in the matrix. All averages are calculated at the end of the clustering routine. In the database application, the clustering of selected spectra is in the order of ascending *groupid*, ascending *sampleid*, and ascending *resultid* values in records of the table result, which are pointers to the files on the FTP server. The effect of reversed clustering is studied by changing the order in descending *groupid*, descending *sampleid*, and descending *resultid* values of the table result.

4.3 Discussion

The database application can clearly distinguish the MALDI-TOF peptide profiles between different patient and control groups. It can determine differences in the frequency and intensities of peptide masses in spectra from both groups. A strong feature of the here described architecture is that it can process different MS file formats, such as peak lists, MALDI-TOF and FT-ICR binary files from various manufactures in the same manner. More important are speed and memory usage by the client workstation. Peptide profile matrices have to be created in reasonable time. When dealing with large quantity of data, the Java application will easily run into out of memory errors with default settings of the JVM. Very important to use limit and offset strategies in MySQL queries to fetch no more than a buffered amount of 5000 table records each time when displaying them in the GUI. A specific MALDI-TOF MS matrix of 111 samples and 1949 masses has 216339 matrix fields and a CSV file size of 444 Kbytes. Three matrices, peptide mass occurrences, intensity, and binary of this size can be simultaneously built in the Java Virtual Machine's (JVM) allocated memory. However, a typical FTMS matrix with 374 samples and 10651 discriminated masses has an 18 times larger number of 3983474 matrix fields and an 18 times larger CSV file size of 7.9 Mbytes. It is impossible to build three matrices of this size simultaneously in the Java's memory space. These files have to be built in the user document root as a FileOutputStream and transported to the FTP server.

More advanced techniques such as Fourier transform ion cyclotron resonance (FT-ICR) MS and offline nano LC-MALDI (Liquid Chromatography) in combination with FT-ICR measure accurate masses in the 0.5 to 1 ppm range. Furthermore, the higher resolution of FT-

ICR MS prevents the clustering of peaks of different peptides. These techniques allow the identification of proteins from peptide masses by either peptide mapping or peptide sequencing. The database application can be adapted to handle the mass spectra of these experiments due to its modular architecture. The type of equipment, in combination with type of imported spectra will determine the handling of raw data, such as calibration and peak finding algorithms. In order to transform the spectra from the time domain to the frequency domain (11), an extra Fast Fourier Transformation (FFT) step to handle raw data of FT-ICR experiments is constructed. The peptide masses can subsequently be calculated from the cyclotron frequency. It is also possible to apply a de-isotope algorithm on the peptide masses due to the higher resolution and mass accuracy of FT-ICR. Peak centroiding will be implemented, which calculates the real mass of the peak maximum, weighted by the intensity of the points surrounding the local maximum.

In conclusion, a new software architecture is presented which can analyze high throughput MS data from MALDI-TOF MS and MALDI-FTMS measurements in a efficient way. Results of the analysis are stored in a centralized relational database and FTP server. Meta data of the experiment and samples can be stored as well, and can be used to link the results to clinical data or data from other types of experiments. The database application generates a matrix with the frequency of masses in replicate spectra from different samples, a binary table with the frequency of masses above a specific threshold, and a matrix with the mean intensity of the present peaks in the mass spectra replicates. The matrix, which is stored on the FTP server and in the local document directory, can be imported in statistical packages or in (commercial) analysis software such as Spotfire. Statistical analysis of two test datasets by the Wilcoxon-Mann-Whitney test in R clearly distinguishes the peptide-profiles of patient body fluids from those of controls. Finally, the modular architecture of the application makes it possible to also handle data from FT-ICR experiments or other MS devices.

References

1. Deitel, H., and Deitel, P. . *Java how to program* New Jersey: Pearson - Prentice Hall, 2005.
2. Zschunke, M., Nieselt, K., and Dietzsch, J. . "Connecting R to Mayday, Chapter 2: Calling R from within Java (www.zbit.uni-tuebingen.de)," *Studienarbeit Bioinformatik*, 2004.
3. Lemkin, P. F., Thornwall, G., Alvord, W. G., Lubomirski M., and Sundaram, S. . "Extending MicroArray Explorer with R Language Scripts (http://maexplorer.sourceforge.net)," *Frederick bioinformatics forum* 2003.

4. Mann, H. B., and Whitney, D. R. . "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Stat.* 18 (1947): 50-60.
5. Siegel, S., and Castellan, N.J. . *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill Book Co, 1988.
6. Wilcoxon, F. "Individual comparisons by ranking methods," *Biometrics Bull.* 1 (1945): 80-83.
7. Wong, J. W. H., Cagney, G., and Cartwright, H.M. . "SpecAlign—processing and alignment of mass spectra datasets," *Bioinformatics* (2005): 2088-2090.
8. Jeffries, N. "Algorithms for alignment of mass spectrometry proteomic data," *Bioinformatics* (2005).
9. Lin, S. M., Haney, R.P., Campa, M.J., Fitzgerald, M.C., Patz, Jr.E.F "Characterizing phase variations in MALDI-TOF data and correcting them by peak alignment," *Cancer Informatics* 1 (2005): 32-40.
10. Ramsay, J. O., and Li, X. . "Curve registration," *J. Roy. Stat. Soc., Ser. B* 60 (1998): 351-363.
11. Press, W. H., Flannery, B.P., Teukolsky, S.A., and Vetterling W.T. . *Numerical recipes in C: the art of scientific computing, Chapter 12: Fast Fourier Transform* Cambridge: Cambridge University Press, 1992.

Chapter 5

FTMS and TOF/TOF mass spectrometry in concert: Identifying peptides with high reliability using matrix prespotted MALDI target plates

Dekker, L.J., Burgers, P.C., Guzel, C., and Luider, T.M.

Journal of Chromatography B, 847 (2007) 62–64

Abstract

In this chapter we describe a combination of the mass spectrometric techniques MALDI–TOF/TOF and MALDI–FTMS to identify proteins in complex samples using prespotted MALDI target plates. By this procedure accurate FTMS mass measurements and TOF/TOF data are obtained from the same spot. We have found that this combination of techniques leads to more reliable identification of peptides.

5.1 Introduction

MALDI-TOF mass spectrometry is an attractive technique for peptide profiling for reasons of its sensitivity, reliability, and high-throughput capability (1). However, the extreme complexity of peptide mixtures derived from proteins as well as the large dynamic range of the abundances of proteins in body fluids, tissue and cell lysates makes it all but impossible to detect all tryptic peptides from a sole mass spectrum. In addition, TOF peptide profiles are often difficult to interpret, mainly because a monoisotopic peak of one peptide may overlap with an isotopic peak of another, a feature the peak-picking algorithm may not detect. Higher mass resolution and mass accuracy as supplied for example by FTMS may alleviate some of these drawbacks. These matters are exemplified in Figure 5.1, which shows a partial MALDI-TOF (panel A) and MALDI-FTMS (panel B) mass spectrum of a trypsinized (albumin depleted) serum sample: the monoisotopic peaks of the two peptides can be easily identified from the FTMS mass spectrum and the exact masses are then obtained, as indicated. Figure 5.1 illustrates another major advantage of MALDI-FTMS over MALDI-TOF, namely that the FTMS spectra, and in contrast to the MALDI-TOF spectra, hardly contain signals often referred to as "chemical noise" (2). Thus the signal to noise ratio in FTMS is much larger (134 for FT) than that for TOF data (14 for TOF). Using FTMS, mass measurement with low ppm accuracies are now routinely obtained. Using judiciously chosen calibration procedures accuracies below 0.5 ppm can be obtained on a Bruker Daltonics Apex Q 9.4 Tesla instrument. In addition the FTMS technique offers sensitivity in the femtomole range in complex samples (3).

However, the identification of peptides from a mass measurement alone requires a mass accuracy below 0.1 ppm which is not yet possible with most FTMS mass spectrometers (4). Also the sensitivity and selectivity of MS/MS experiments on MALDI ions in FTMS by collision-induced dissociation (CID) for complex peptide mixtures is still not on a par with multiply charged peptides generated by electrospray ionization.

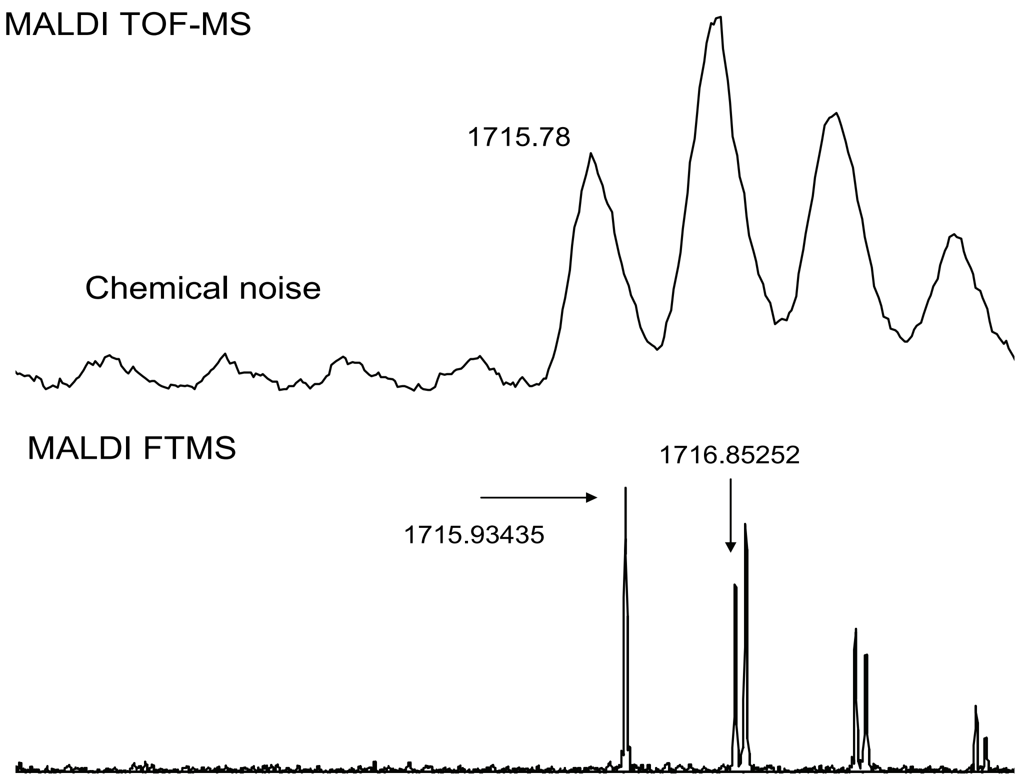


Figure 5.1: Partial MALDI-TOF and MALDI-FT mass spectra of a tryptic digest of proteins of an albumin depleted serum sample. Monoisotopic peaks of the two peptides can be easily identified from the FTMS mass spectrum, because of the superior resolution compared to the MALDI-TOF measurement. In addition the chemical noise, present in the MALDI-TOF spectrum, is absent in the MALDI-FT mass spectrum.

5.2 Methods

We have therefore developed a simple method for the identification of peptides in complex mixtures whereby the high mass accuracy and resolution of MALDI-FTMS is exploited for directing and confirmation purposes. First, from MALDI-FTMS peptide profiling experiments peaks are identified that show a significant difference in expression between the control and patient group. Next, a fractionation was performed using a C18 Pep Map column (75 µm i.d. x 150 mm, 3µm, Dionex, Sunnyvale, CA, USA) on a nanoscale liquid chromatography system (nanoLC) (Dionex, Sunnyvale, CA, USA). Five µl of the sample was loaded onto the trap column (300 µm i.d. x 5mm, 5µm, Dionex, Sunnyvale, CA, USA). Fractionation was performed using a 130 minute gradient from 0% to 76% of acetonitrile (ACN), (solution A (100% H_2O, 0.05% trifluoroacetic acid (TFA)) and solution B (80% ACN, 20% H_2O and 0.04% TFA); 0 to 15 min, 0% solution B, 15.1 min 15%, 75 min 40%, 90 min 70%, 90.1-100 min 95%, 100.1 min 0% and 130 min 0%). Fifteen second fractions were spotted automatically onto a commercially available prespotted MALDI plate containing 384 spots (Bruker Daltonics, USA) covered with α- cyano-4-hydroxycinnamic acid (HCCA) matrix, using a robotic system (Probot Micro Fraction Collector, Dionex, Sunnyvale, CA, USA). To each fraction, 1 µl water was added. Finally, salts were removed by washing the pre-spotted plate for 5 seconds with a 10mM $(NH)_4H_2PO_4$ solution in 0.1% TFA/water solution. The spots were subsequently measured by automated MALDI-TOF/TOF (Ultraflex, Bruker Daltonics, Germany using WARLP-LC software (Bruker Daltonics, Germany)). By this procedure MS spectra of each individual spot was obtained and subsequently MS/MS experiments are performed on each peptide. The best spots for performing the MS/MS measurements were determined automatically by the WARLP-LC software.

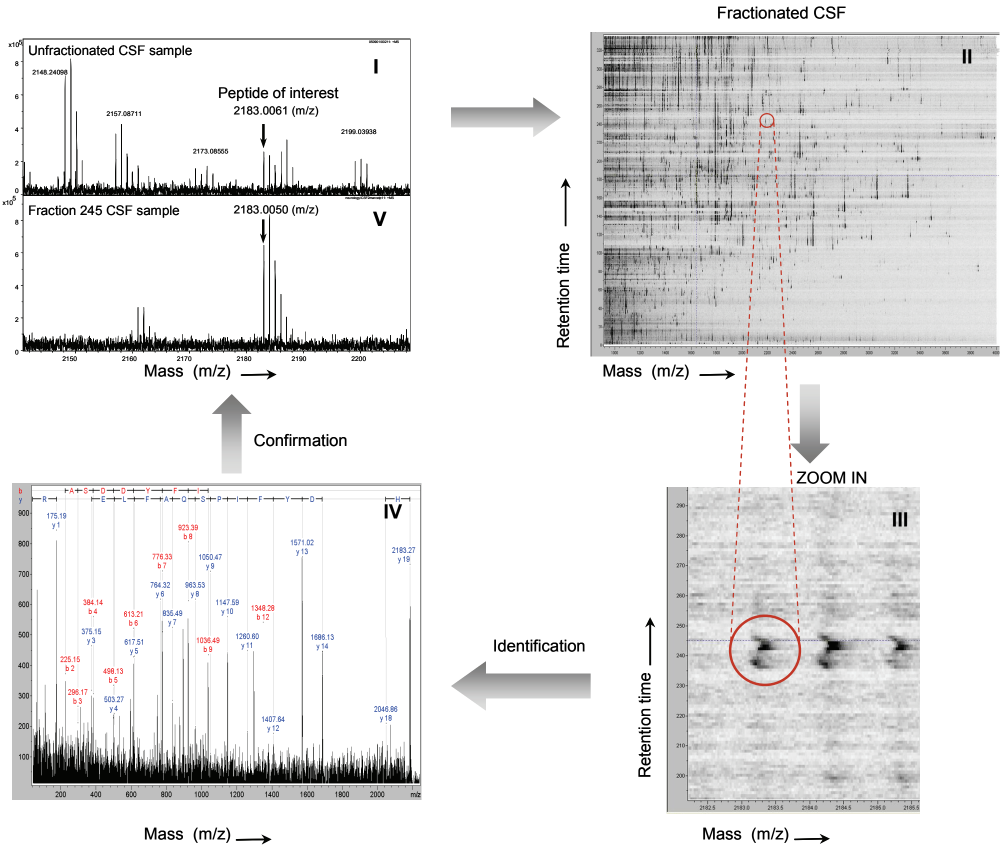


Figure 5.2: Visualization of the identification method for differential peptides from a MALDI-FTMS profiling study. In section I of this figure a zoom-in of a MALDI-FTMS spectrum is displayed in which a peak that showed in a series of measured samples a significant difference in expression between the control and patient group. It can be seen that the peptide of interest is in an area of the mass spectrum that contains many peaks. This decreases the chances of successfully performing a direct identification with MALDI-TOF. In section II a 2 dimensional view of a LC MALDI-TOF run is shown of the sample that is measured for the spectrum in section I. The tryptic digested CSF sample is analyzed on a nano-LC system with a reverse phase C18 pepmap column. Each 15 seconds the eluting fraction is spotted on a prespotted anchorchip plate containing 384 spots. Subsequently, the entire plate is measured in the MALDI-TOF. In section III a zoom in of the 2-dimensional view is displayed in which the peptide of interest again can be seen. Section IV shows the MS/MS measurement of the peptide of interest; the software automatically decides what the best fraction is to perform MS/MS for this peak (parameters that are used for this are signal to noise and the presence of neighboring peaks). After the identification is performed this can be confirmed by re-measuring the spot of interest with MALDI FTMS. This is displayed in section V, the exact mass measured is compared to the original peptide of interest to be sure that it is the same peptide and also to the calculated mass to confirm the identification.

5.3 Results and discussion

The idea was to use this prespotted plate also for MALDI-FTMS experiments to confirm, by exact mass measurement, the identity of the differentially expressed peptides. Unfortunately on our Bruker Apex Q 9.4 Tesla FTMS, the matrix α-cyano-4-hydroxy-cinnamic acid (HCCA) produces only very weak signals in MALDI-FTMS experiments. However, when after the TOF/TOF data have been acquired, the spot of interest is covered with 0.5 µL of a DHB solution (10 mg/mL 2,5-dihydroxy benzoic acid in 0.1% TFA) and then introduced into the FTMS, intense spectra are obtained allowing FTMS experiments on the same spot as was used to acquire the TOF/TOF data. The workflow of this procedure is shown in Figure 5.2. Confirmation proceeds via two paths: The exact mass measured in the final step is compared to the calculated mass of the identified peptide and to the exact mass measured in the peptide profiling experiment.

An example of our procedure is presented in Figure 5.3 which shows the full MALDI-FTMS mass spectrum of a nanoLC fraction of digested CSF. This spectrum, following the workflow of Figure 5.2, was obtained after the TOF/TOF measurements. This spectrum contained as many peaks as the MALDI-FTMS mass spectrum of the original sample (i.e. prior to LC separation). In the inset is given a small part of the spectrum of the original sample (panel A) and of the above fraction (panel B). It can be seen that even almost undetectable signals in the original spectrum can, after LC separation, be measured with great precision.

Finally, the prespotted plate with the deposited peptide mixtures can be stored in a dark environment at room temperature for at least one month without significant loss of signal intensity, allowing remeasurement when required at a later stage. We have applied our procedure successfully to various projects and the results will be published in separate papers (5, 6).

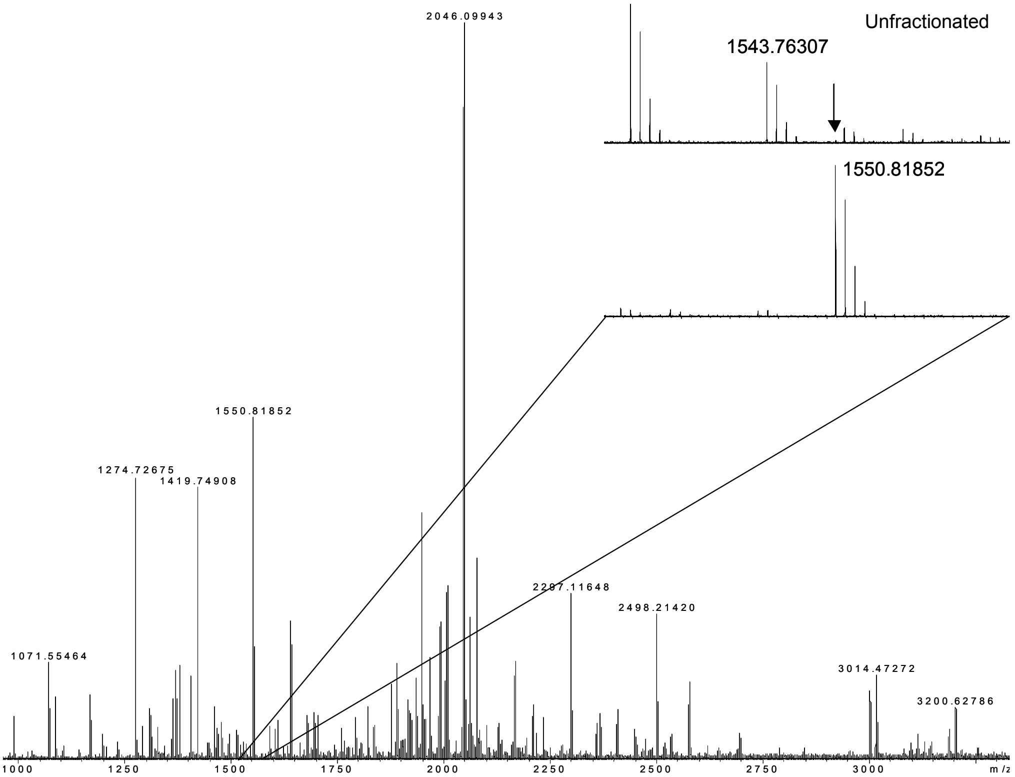


Figure 5.3: MALDI-FT mass spectrum of a nanoLC fraction of a frozen section of trypsinized placenta material. This fraction contains as many peaks as the unfractionated material. The upper part of the inset shows part of the mass spectrum obtained from the unfractionated sample. Even the very small peak indicated by the arrow gives after separation an intense signal.

References

1. Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M. "A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra," *Rapid Commun Mass Spectrom* 19 (2005): 865-870.
2. Krutchinsky, A. N., and Chait, B. T. "On the nature of the chemical noise in MALDI mass spectra," *J Am Soc Mass Spectrom* 13 (2002): 129-34.
3. Shen, Y., Tolic, N., Masselon, C., Pasa-Tolic, L., Camp, D. G., 2nd, Hixson, K. K., Zhao, R., Anderson, G. A., and Smith, R. D. "Ultrasensitive proteomics using high-efficiency on-line micro-SPE-nanoLC-nanoESI MS and MS/MS," *Anal Chem* 76 (2004): 144-54.
4. Dekker, L. J., Burgers, P. C., Kros, J. M., Smitt, P. A., and Luider, T. M. "Peptide profiling of cerebrospinal fluid by mass spectrometry," *Expert Rev Proteomics* 3 (2006): 297-309.

5. Mustafa, D. A., Bergers, P. C., Dekker, L. J., Charif, H., Titulaer, M., Sillevis Smitt, P. A., Luider, T. M., and Kros, J. M. "Identification of glioma neovascularisation-related proteins by using MALDI-FTMS and nano-LC fractionation to microdissected tumor vessels," *Mol Cell Proteomics* (2007).

6. de Groot, C. J. M., Guzel, C., Steegers-Theunissen, R. P. M., de Maat, M., Derkx, P., Roes, E., Heeren, R., Luider, T. M., and Steegers, E. A. P. "Specific peptides identified by mass spectrometry in placental tissue from pregnancies complicated by early onset preeclampsia attained by laser capture dissection," *Proteomics Clin. Appl.* 1 (2007): 325-335.

Chapter 6

MALDI-TOF mass spectrometry analysis of cerebrospinal fluid tryptic peptide profiles to diagnose leptomeningeal metastases in breast cancer patients

Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M., and Sillevis Smitt, P. A.

Mol Cell Proteomics 4 (2005): 1341-1349.

Abstract

Leptomeningeal metastasis (LM) is a devastating complication occurring in 5% of breast cancer patients. Early diagnosis and initiation of treatment are essential to prevent neurological deterioration. However, early diagnosis of LM remains challenging because 25% of CSF samples test false negative at first cytological examination. We developed a new, mass spectrometry based method to investigate the protein expression patterns present in the CSF from breast cancer patients with and without LM.

CSF samples from 106 patients with active breast cancer (54 with LM and 52 without LM) and 45 controls were digested with trypsin. The resulting peptides were measured by matrix-assisted laser desorption ionization - time of flight mass spectrometry (MALDI-TOF MS). Then, the mass spectra were analyzed and compared between patient groups using newly developed bioinformatics tools.

A total of 895 possible peak positions was detected and 164 of these peaks discriminated between the patient groups (Kruskal-Wallis, p <0.01). The discriminatory masses were clustered and a classifier was built to distinguish breast cancer patients with and without LM. After bootstrap validation, the classifier had a maximum accuracy of 77% with a sensitivity of 79% and a specificity of 76%.

Direct MALDI-TOF analysis of tryptic digests of CSF gives reproducible peptide profiles that can assist in diagnosing LM in breast cancer patients. The same method can be used to develop diagnostic assays for other neurological disorders.

6.1 Introduction

Leptomeningeal metastases (LM) arise when tumor cells metastasize to the cerebrospinal fluid (CSF). The flow of CSF results in widespread dissemination of the tumor cells along the surface of the central nervous system, causing symptoms by invading the brain, spinal cord, cranial nerves and nerve roots (1).

One of the tumors most frequently associated with LM is breast cancer. During the course of their disease approximately 5% of metastatic breast cancer patients will develop symptoms caused by LM. Response to therapy of this debilitating complication depends upon early treatment. However, diagnosis of LM remains challenging because 25% of samples tested are false negative at the first cytological examination of the CSF, probably due to sampling error (1).

Recently, protein expression profiling of body fluids from cancer patients has become a valuable tool for obtaining information on the state of protein circuits inside tumor cells and outside the cells at the host-tumor interface (2, 3). In serum and CSF, low molecular weight proteins and peptides that are related to this altered micro-environmental ‘cancerous’ state can be detected.

We studied the differential tryptic peptide profiles in the CSF from breast cancer patients with and without LM and controls. Studying CSF has several advantages over studying serum. Firstly, tumor cells in LM patients are located in the CSF and in the leptomeninges that are surrounded by CSF. Prior to their transport into serum, tumor-related proteins will therefore first be shed into the CSF. Secondly, the normal protein concentration of CSF is 100-400 fold lower than in serum (4). This results in a significant over-representation of LM-related proteins in CSF compared to serum. The identification of protein profiles specific for LM may be helpful in diagnosing patients with clinical suspicion of LM and negative cytology. In addition, such proteins may reveal cellular mechanisms relevant to the biology of LM.

With the advent of mass spectrometry into the field of clinical proteomics, the comparison of large numbers of proteins in complex biological samples such as serum and CSF has become feasible (3 , 5). Up to now, the most commonly used instrument is the SELDI-TOF MS (Surface enhanced laser desorption ionization time of flight mass spectrometry). Using SELDI-TOF MS analysis of various body fluids, discriminatory protein expression profiles have been identified in various diseases (3, 6). However, SELDI-TOF MS does not allow a direct identification of the discriminatory proteins and suffers from low reproducibility and accuracy (7-10). To improve the reproducibility and accuracy and to find better ways to identify relevant discriminatory proteins (11), we first digested our samples with trypsin and analyzed the resulting peptide mixtures by MALDI-TOF MS (matrix assisted laser desorption ionization time of flight mass spectrometry) (12). The reproducibility of this type of analysis has been described elsewhere (13).

We analysed CSF samples from 106 patients with active breast cancer, 54 of whom had LM, and 57 controls. Tryptic peptide mixtures were measured by MALDI-TOF MS and analyzed using a newly designed bio-informatics tool. We could identify unique peptide patterns that discriminated the LM patients from the other breast cancer patients and from controls.

6.2 Methods

6.2.1 Patient selection

Using clinical databases and CSF banks, we retrospectively identified all breast cancer patients with available CSF samples collected in the last seven years in four participating institutions (Erasmus MC, Netherlands Cancer Institute, UMC Nijmegen and Innsbruck Medical University). We only included patients with advanced breast cancer defined as metastatic or locally progressive disease. Patients with positive cytology or with a compatible neurological syndrome and diagnostic MRI were considered to have LM (Group I, n=54). When cytology was negative and when clinical follow-up was incompatible with LM, patients were classified as having advanced breast cancer without LM (Group II, n=52). Controls (Group III, n=45) were patients who were not known to have cancer and who did not suffer any known neurological disease (collected at Erasmus MC, Netherlands Cancer Institute, and LUMC). We noted the date of diagnosis of breast cancer, the date of lumbar puncture and the date of last follow-up or death. The symptoms at the time of lumbar puncture were scored as either central, cranial nerve involvement, radicular, compatible with raised intracranial pressure, or other. The following CSF parameters were noted: total protein concentration, white cell count and cytology (positive or negative). MRI and CT scans were scored as either positive, suggestive or negative for LM. Nodular or focal linear meningeal enhancement was considered diagnostic of LM and communicating hydrocephalus suggestive of LM. All samples had been routinely centrifuged to discard cellular elements prior to storage at –80° C.

6.2.2 Sample preparation and measurement of samples

All samples were blinded and analyzed in random order. From each sample, 20 µl CSF was put into a 96-well plate, and 20 µl of 0.2% Rapigest (Waters, USA) in 50 mM ammoniumbicarbonate buffer was added to each well. The samples were incubated for 2 minutes at 37° C. Four µl of 0.1 µg/µl 3 mM Tris HCl gold grade trypsin (Promega, USA) was added to each well, and the 96-well plates were incubated at 37°C. After two hours of incubation, 4 µl of 500 mM HCl was added in order to obtain a final concentration of 30-50 mM HCl (pH < 2). The 96-well plates were then incubated again for 45 minutes. A 96-well zip C18 microtiter plate (Millipore, USA) was pre-wetted and washed twice with 200 µl acetonitrile per well. Full vacuum was applied to the plate using a vacuum manifold (Millipore, USA). Three µl of acetonitrile was put on the C18 resin without vacuum to prevent it from drying. Each sample was mixed with 200 µl water HPLC grade / TFA 0.1%. Subsequently the samples were loaded on the washed and pre-wetted 96-well zip C18 plate (Millipore); a pressure differential of 5 inches of Hg vacuum was used. After the wells had been cleared, the wells were washed twice with 100 µl 0.1% TFA. Full vacuum was applied until all wells were empty. The samples were eluted in a new 96-well plate with an elution volume of 15 µl 50% acetonitrile/water HPLC grade 0.1% TFA; a pressure differential of 5 inches of Hg vacuum was used. After elution, the samples were stored at 4° C in the 96-well plates covered with aluminum seals. All samples were spotted on a MALDI target (600/384 anchor chip with transponder plate; Bruker Daltonics, Germany) in three-fold. In order to do so, two µl of elute was mixed with 10 µl matrix solution (2 mg HCCA (α-Cyano-4 hydroxy-cinnamic acid, Bruker Daltonics, Germany) in 1 ml acetonitrile 30 min using an ultrasonic bath). Afterwards, samples were automatically measured on a MALDI-TOF MS (Biflex III, Bruker Daltonics, Germany). The digestion step was repeated twice for each sample, the purified peptides were spotted in three-fold, and all the spots were measured in three-fold. This resulted in 18 spectra for each sample. The standard method for peptide measurements on the MALDI-TOF MS was used (default file Bruker “1-2kD positive” with the measurement range changed to 300-3000 Da). For the automated measurements, the following settings were used: an initial laser power of 20% and a maximum of 35%. The highest peak above the 750

Da had to have a signal to noise ratio of at least 5 and a minimum resolution of 5000. Every 30 laser shots, the sum spectrum was checked for these criteria. If the sum spectrum did not meet these criteria, it was rejected. If 13 sum spectra from 30 shots met the criteria, these were combined and saved; when 50 sum spectra from 30 shots were rejected, the measurement of that spot was then ended and the next spot was measured.

6.2.3 Analysis of spectra

First, the raw binary data files were converted to ASCII files containing the measured intensities for all channel indices of the spectra. We then developed a peak detection algorithm in the statistical language R (http://www.r-project.org). The definition of a peak (or local maximum) in this algorithm states that the intensity of the peak position has to be above a predefined threshold and has to be the highest intensity value in a surrounding mass window. This peak finding algorithm was tested on a small set of spectra with different settings for the threshold and the mass window. The settings for the peak finding were chosen such that the resulting peak list most resembled peaks that would be manually assigned, this optimizing the trade-off between signal sensitivity and noise detection. We chose a percentile threshold of 98.5% (the intensity of the position must belong to the 1.5% highest intensity values of the spectrum) and a mass window of 0.5 Dalton. A quadratic fit with a number of internal calibrants was used to calibrate the channel numbers to masses. For this mass calibration, five omnipresent albumin peaks (960.5631, 1000.6043, 1149.6156, 1511.8433, 2045.0959 m/z) were used. The accurate mass of these albumin peaks was obtained by performing an *in silico* “tryptic digest” on the human albumin amino acid sequence with MS-digest (http://prospector.ucsf.edu/ucsfhtml4.0/msdigest.htm).

During the process of alignment and conversion, the quality of the spectra was checked as follows. If two or more of the omnipresent albumin peaks were not detected, the spectrum was not used in the further analysis. The peak finding algorithm was then used to create a list of peak positions for each individual spectrum. These peak lists were combined by comparing the lists one by one. If peak positions were present in a mass window of 0.5 Dalton in both spectra, these peak positions were combined. The combined peak list was then compared with a new spectrum until all peak lists had been combined. The latter peak list was used to create a matrix displaying the frequency of each peak position for each sample. Peak positions that were present in less than 5% of the spectra were deleted from the matrix to reduce the number of noise peaks. The matrix created in this way was used for statistical analysis of the data. Using a univariate analysis in R a p-value was determined for every peak position. When comparing more than two groups, we used the Kruskal-Wallis test and when comparing two groups we used the Wilcoxon-Mann-Whitney test.

To investigate whether differences in the total CSF protein concentration of the samples affected the performance of the MALDI-TOF, we first used the Bio-Rad DC protein assay (Biorad, USA) to determine the protein concentration of all the CSF samples. We then calculated the sum of albumin peaks detected in all seven spectra of each sample (excluding the albumin peaks that had been used for calibration). Using GraphPad Prism version 4.0 software (GraphPad Software, USA, 2002), we compared the total protein concentration and the sum of the albumin peaks of the three groups. To test for statistically significant differences between the three groups, we used one-way ANOVA followed by Bonferroni's multiple comparison test. All tests were two-sided, and $p<0.05$ was considered statistically significant. Also the correlation between peak frequency and protein concentration was calculated for each individual peak position. A histogram of all correlation coefficients was created and the distribution was compared with a normal distribution using the Kolmogorov Smirnov (SPSS Software, USA, 2001).

All peak positions with a frequency that was two times higher in group I than in the control groups II and III were selected. These peak values were submitted to the Mascot search engine (Matrix Science, London, UK) to search the MSDB human database using a 100 ppm tolerance.

6.2.4 Building a predictive model

A supervised multivariate analysis method was used to determine whether sample groups I and II could be separated on the basis of their peak positions. For each patient, 7 mass spectra were used, and the peak positions of each of those 7 spectra were combined. Therefore the number of times that a peak was present varied between 0 and 7. To reduce noise, a minimum of 2 peaks was required to determine whether a peak was present (≥2, 1) in a sample or not (<2, 0) allowing the formation of a binary data matrix. The required frequency is kept low to minimize loss of signal. To reduce the number of variables, a clustering was performed that combined peaks of similar behaviour. Peptide peaks that often occurred simultaneously were grouped into the same cluster using a hierarchical clustering algorithm. The distance between each possible pair of peptide peaks was determined with the Manhattan distance measure, i.e. the sum of the absolute differences for all patients. The number of clusters was set at 50. With 50 clusters, isotope peaks were generally grouped into the same cluster. The clusters generated represented groups of peptide peaks that might, at least in part, be derived from the same protein or proteins. The clustering of the masses made it possible to compose a new data matrix. Each matrix cell contained the number of peaks present for a certain patient relative to the total number of peaks in a particular cluster. In other words, each cell in the new matrix defined the proportion of peptides that was present in a cluster for a certain patient. To further reduce the complexity of the data we set a threshold for the presence of a cluster to obtain a binary data matrix. Using the clustered, binary variables we constructed a non-linear predictive model that separated group I from group II. In the model thus generated, a maximum of 8 clusters was allowed and only those clusters with an area under the curve (AUC) greater than 0.62 were considered. Genetic programming was used to search for the model with the highest AUC (14). To obtain an unbiased estimate of the predictive accuracy of the model, we utilized the bootstrapping method (15, 16). Bootstrap data sets were created by randomly selecting patients with replacements from the original data set. As an extra precaution, the clustering step was included in the bootstrapping process as well. Hundred bootstrapped matrices were created from the original matrix, by resampling with replacement. The clustering was repeated for each of these resampled matrices and a predictive model was constructed. The AUC of each model was measured on the bootstrap data set as well as on the original data set. The average difference between the performance on the bootstrap data set and the performance on the original data set provided a correction factor which gave an estimate of the bias of our model development process. Finally, we developed a model on the original data and corrected its AUC with the correction factor, producing a conservative estimate of the performance of the model.

6.3 Results

6.3.1 Patients

Clinical information, CSF data and imaging results are summarized in Table 5.1. Forty-six percent of breast cancer patients with LM (Group I) presented more than one neurological symptom while the majority of breast cancer patients (73%) in group II presented a single symptom at the time of lumbar puncture (LP). All patients in Group I had LM while diagnoses in Group II included bone metastases (n=15), tension headache (n=5), metabolic encephalopathy (n=3), carpal tunnel syndrome (n=3), brain metastases (n=2), migraine (n=2), DIC (n=2), dural metastasis (n=2), lumbago (n=2), and psychiatric (n=2). All of the following conditions were diagnosed, each in one patient: jugular vein thrombosis, herniated disk, dementia, anaplastic astrocytoma, whiplash injury, polyneuropathy, and syncope. In two patients, no cause for the symptoms was found.

The cytological examination of the CSF revealed tumor cells in all Group I patients and in none of the group II patients. The CSF white cell count was increased (> 4 cells / µl) in 56% of Group I and 0% of Group II patients. The total protein concentration was higher than the institutional upper limit in 85% of Group I and in 30% of group II patients. Imaging studies were obtained in 39 out of 41 Group I patients, 35 MRI and 4 CT-scans. Imaging results were positive for LM in 23 (2 CT-scans) patients, suggestive in 4 and negative in 10 patients (2 CT-scans). MRI results were available in 26 patients in group II. In 22 patients, the MRI was considered negative, suggestive in 3 and positive for LM in one patient. The patient with a false positive MRI had a single non-symptomatic dural enhancing lesion at Th9; the CSF examination was three times normal.

Table 5.1. Clinical information on advanced breast cancer patients with (Group I) and without leptomeningeal metastasis (Group II) and controls (Group III).

	Group I	*Group II*	*Group III*
Number of patients	41	46	43
Age at LP (mean ± SD)	52± 10	52± 13	38± 21
Time from cancer diagnosis (months) (mean ± SD)	62± 65	69± 59	
Time from LP to last follow-up or death (months) (mean ± SD)	6± 10	19± 19	
Alive at last follow-up	1 (3%)	7 (16%)	
Number of symptoms			
One symptom	22 (54%)	33 (73%)	
Two symptoms	12 (29%)	12 (27%)	
Three symptoms	7 (17%)	-	
Symptoms			
Central	28 (68%)	14 (31%)	
Cranial Nerve	11 (27%)	10 (22%)	
Raised ICP	16 (39%)	5 (11%)	
Radicular	12 (29%)	8 (18%)	
Other	-	20 (44%)	
CSF			
White cell count /µl (mean ± SD)	21.2 ± 34.4	0.8 ± 0.8	NA
Increased white cell count (>4/µl	23 (56%)	0 (0%)	
Positive cytology	41 (100%)	0 (0%)	0 (0%)
Protein concentration > norm institute	35 (85%)	14 (30%)	0 (0%)
Glucose (mmol/l, mean ± SD)	2.0 ± 1.3	3.8 ± 0.9	NA
Imaging (MRI/CT)	35/4*	26	
Positive	23 (56%)	1 (2%)	
Negative	12 (29%)	22 (48%)	
Suggestive	4 (10%)	3 (7%)	
NA	2 (5%)	20 (43%)	

* Two of four CT-scans were positive for LM while the other two were negative.

6.3.2 Peak detection

We detected an average of 350 peaks per spectrum (95% CI 250 - 450). Spectra with more than 450 detectable peaks were excluded from the analysis because these spectra consisted mainly of noise peaks (peaks without an isotopic distribution). After alignment and quality control, the number of good-quality spectra per sample was counted and samples with fewer than 7 good quality spectra were discarded. The number of 7 spectra is based on an earlier reproducibility study (13). At this threshold, the remaining number of cases per patient group was 41 in group I, 46 in group II, and 43 in group III (Table 5.2). Of the samples with more than seven spectra, only the first seven spectra were used. The combined peak list of all these spectra contained 2006 possible peak positions. After noise reduction, the matrix created from this list contained 895 peak positions.

Table 5.2: Number of patient per group before an after quality control of spectra.

	Group I	*Group II*	*Group III*
Before quality control	54	54	45
After quality control	41	46	43

6.3.3 Univariate analysis

For each peak, the significance of difference in distribution over groups I-III was tested with the Kruskal Wallis test and a p-value was calculated. We detected 323 peak positions with $p<0.05$ and 172 with $p<0.01$, indicating that a considerable number of peaks correlated with diagnosis groups. In Figure 6.1 A, the number of peaks (frequency) is presented for each p-value interval. On the same data, we performed a cross validation by randomly assigning a group number to each CSF sample and then repeating the Kruskal Wallis test. This scrambling procedure was repeated 10,000 times. The mean frequency of peaks per p-value interval is presented by the blue line in Figure 6.1A. The flat distribution of the p-value histogram indicates that there is no correlation between peak positions and groups after scrambling. The p-value histogram of the actual experiment is clearly skewed to lower p-values and is significantly different from the histogram after scrambling (Figure 6.1A).

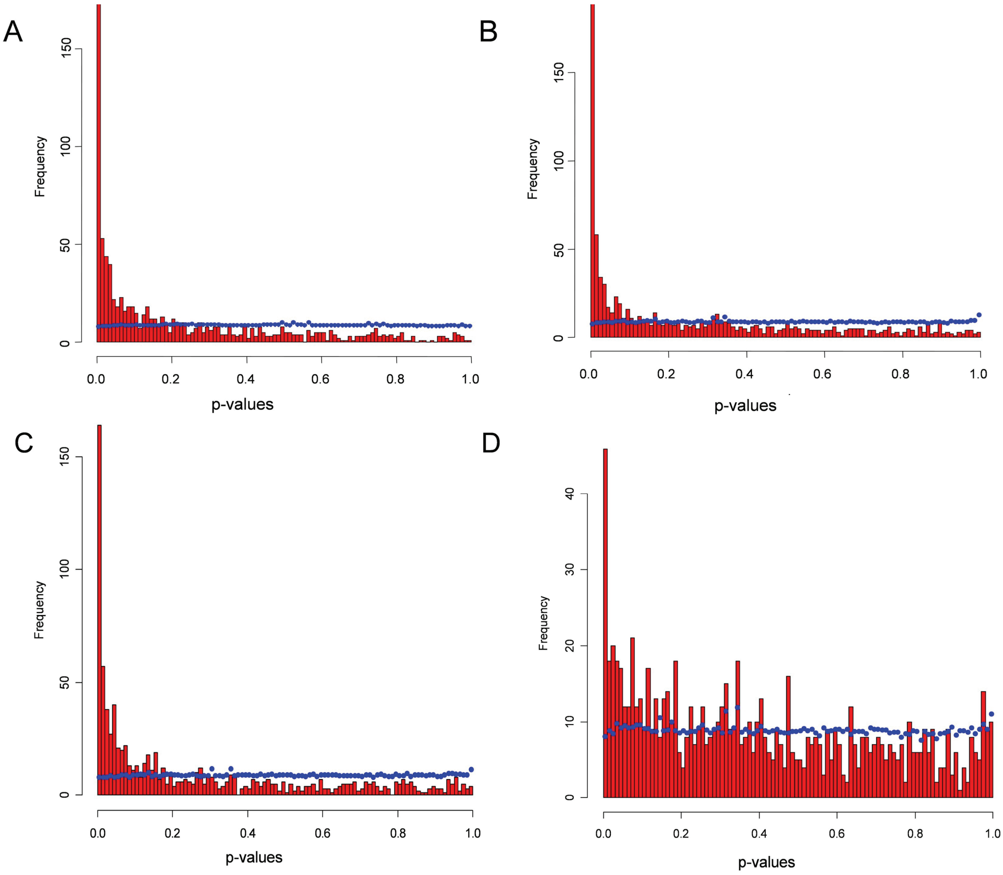


Figure 6.1: A significant number of peptides is differentially expressed between breast cancer patients with (Group I) and without LM (Group II) and normal controls (Group III). The figure shows histograms of p-values where the height of each bar denotes the number of peptide peaks while the horizontal base corresponds to the p-value interval (interval size 0.01).The blue line represents the histogram of p-values after cross validation. The height of the blueline shows the average number of peptide peaks after 10,000 scrambling procedures (see result section). A: Histogram of p-values after comparing groups I, II and III using the Kruskal-Wallis test. The distribution is clearly different from the random distribution (blue line) and skewed to the left indicating a high number of peptides that discriminate between the three groups (low p-value). B-D: Histogram of p-values for peak positions comparing two groups with the Wilcoxon-Mann- Whitney test. Groups I and III are compared in panel B, Groups I and II in panel C and Groups II and III in panel D. In blue the distribution of a re-sampled situation is displayed. The re-sampling is repeated 10,000 times.

We then compared the groups pair wise with a Wilcoxon-Mann-Whitney test. The highest number of peak positions with a significant p-value was found when groups I and III were compared (p<0.05: 329 peak positions; p<0.01: 190 peak positions) (Figure 6.1B). When groups I and II were compared, the p-value histogram was significantly different from a random distribution (p<0.05: 326 peak positions; p<0.01: 164 peak positions) (Figure 6.1C). When groups II and III were compared, fewer peak positions with a significant p-value were detected (p<0.05: 119 peak positions; p<0.01: 46 peak positions) (Figure 6.1D). As a cross-check, the correlation between the different institutes and the peak occurrences were compared. The histogram of the p-values did not differ significantly from a random distribution.

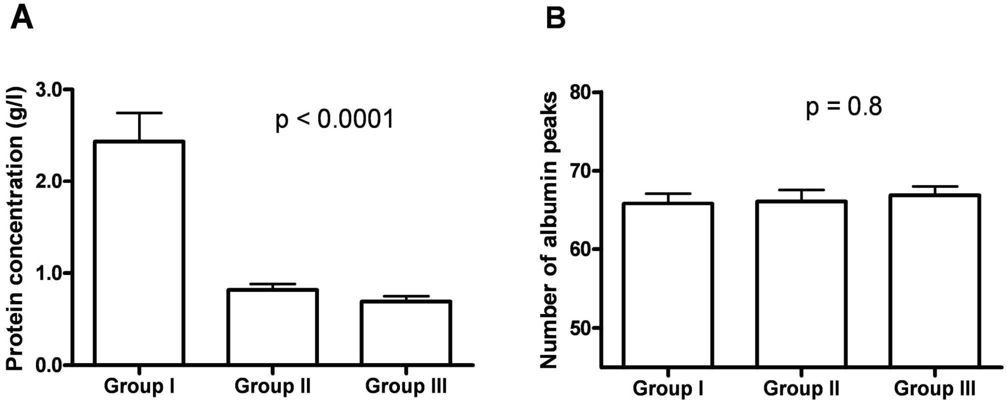


Figure 6.2: Differences in protein concentration do not affect MALDI-TOF analysis of CSF samples A: The mean protein concentration differed significantly between the 3 groups (ANOVA p<0.0001). Differences were significant between Groups I and II (Bonferroni multiple comparison test, p<0.0001), Groups I and III (p<0.0001) but not between Groups II and III (P>0.05). B: The mean number of albumin peaks per group does not differ significantly between the 3 groups (ANOVA, p = 0.8). Bars denote standard error of the mean (SEM).

The CSF total protein concentration of the samples differed significantly between Groups I-III (ANOVA, p< 0.001; Figure 6.2A). However, the sum of albumin peaks detected per sample did not differ between the groups (ANOVA, p = 0.8; Figure 6.2B). For all individual peak positions, the correlation between the peak frequency and the protein concentration was calculated and plotted in a histogram (Figure 6.3). The distribution of the correlation coefficients did not significantly differ from a normal distribution (Kolmogorov

Smirnov test, p=0.35). The constant number of albumin peaks and the lack of effect of protein concentration on the peak frequency indicated that differences in total protein concentration had not significantly affected MALDI-TOF performance.

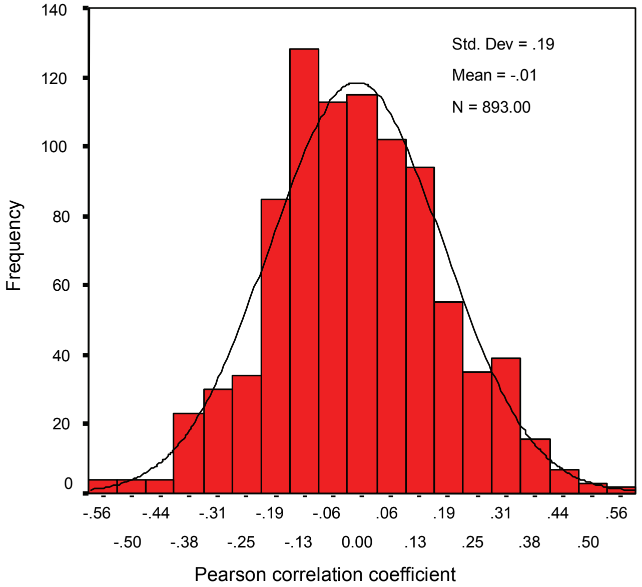


Figure 6.3: Correlation coefficients of protein content and peak frequency. The histogram shows the distribution of correlation coefficients for protein content and peak frequency for each peak position. The histogram does not differ significantly from a normal distribution with a mean value of zero (Kolmogorov Smirnov test, p=0.35).

From the matrix file, we made a selection of all peak positions that were up-regulated two times in group I compared to both control groups II and III. This list contained 52 peak positions that were used in a mascot database search against the MSDB human database. Five of the 52 peptides matched to apolipoprotein A1 (28 kDa). These five matching peptides had an average mass error of 19 ppm compared to the MSDB database.

6.3.4 Clustering analysis

A clustering on the masses was performed on a matrix in which the samples had been sorted on group number. This clustering resulted in the detection of group-specific clusters (Figure 6.4). In Figure 6.4A, a zoom in of the dendrogram is displayed that shows peak positions that have higher frequencies in breast cancer patients without LM (group II) and normal controls (group III) than in breast cancer patients with LM (group I). In panel B peak positions with a higher frequency in breast cancer patients with LM (group I) than in groups II and III are shown.

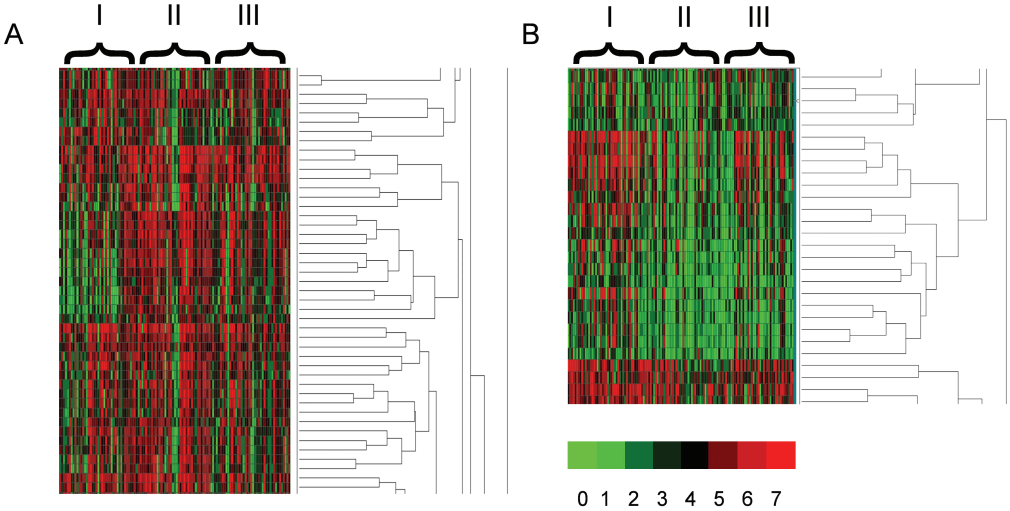


Figure 6.4: Unsupervised clustering demonstrates peptide peaks that differentiate breast cancer patients with and without LM. The figure illustrates close-ups of an unsupervised clustering dendrogram. On the y-axis, the clustered masses are displayed, while the x axis represents the samples ordered by group. Colors represent the frequency of a peak position in the seven spectra of a sample (0-3 spectra, increasing intensity green; 4 spectra, black; 5-7 spectra, increasing intensity red; see legend figure. Panel A illustrates peak positions that are less frequently expressed in CSF from breast cancer patients with LM (group I) than in the CSF from breast cancer patients without LM (group II) and in control CSF (group III). Panel B illustrates peak positions that are more frequently expressed in Group I than in Groups II and III.

6.3.5 Predictive model

Clustering of the masses resulted in a reduction of the variables from 895 peaks to 50 clusters. All albumin peaks were assigned to the same cluster (a large cluster containing 164 peaks). Table 5.2 lists 10 clusters ordered by their univariate area under the receiver operating characteristic curve (ROC) , indicating the highest discriminatory value. As predicted, isotopic peaks are grouped together in the same cluster (Table 5.3). The AUC can vary between 0.5 and 1 (0.5 for a random prediction and a value of 1 for a optimal prediction). The cluster with the highest AUC consisted of 10 peaks, all of which had p-values <0.01. When adjoining isotope peaks were combined, four distinctive peptide peaks in this cluster remained. The model with the highest predictive value, as selected with genetic programming, used six clusters (Table 5.3). The AUC achieved by this model on the original data set was 0.936**;** the bootstrap-corrected AUC was 0.852. The maximum accuracy that could be achieved after bootstrapping was 77%. At this cut-off point the corrected sensitivity was 79% and the corrected specificity was 76%.

Table 5.3: Ten clusters with the highest predictive value

Cluster	*AUC*	*Number of peaks*	*Peak masses (p-value)*
1*	0.74	10	1195.60 (0.0053) 1196.61 (0.0053) 1250.65 (0.047) 1285.61 (0.00023) 1478.73 (16e-05) 1479.74(0.00037) 1480.75 (0.00038) 1629.78 (7.5e-6) 1630.82 (0.00033) 1631.86 (3.4e-07)
2	0.69	8	777.26 (0.00033) 968.54 (0.0012) 1114.59 (0.13) 1395.70 (0.0099) 1753.79 (0.0052) 1754.81 (0.0052) 1787.84 (4.9e-05) 1912.00 (2e-05)
3*	0.67	17	395.21 (0.016) 418.23 (0.079) 426.29 (0.053) 427.32 (0.0759) 519.24 (0.028) 520.28 (0.0740) 522.23 (0.047) 522.96 (0.14) 525.10 (0.014) 526.14 (0.012) 533.26 (0.12) 534.26 (0.012) 544.20 (0.0290) 550.94 (0.520) 552.04 (0.068) 567.13 (0.03) 687.30 (0.031)
4	0.65	4	1191.62 (0.038) 1216.70 (0.110) 1433.74 (0.00012) 1632.79 (0.002)
5*	0.65	5	1366.76 (0.12) 1367.77 (0.018) 1499.77 (0.033) 1648.76 (0.18) 1746.78 (0.0025)
6	0.64	10	1086.59 (0.00076) 1190.62 (0.049) 1201.68 (0.061) 1255.56 (0.052) 1288.66 (0.19) 1302.68 (0.0092) 1344.67 (0.0081) 1345.73 (0.016) 1346.73 (0.39) 1430.75 (0.160)
7	0.64	2	1349.71 (0.00130) 1350.71 (0.0023)
8*	0.63	15	650.06 (0.01) 656.03 (4.5e-05) 665.97 (0.0083) 671.99 (0.00047) 825.04 (0.00052) 841.03 (9.8e-05) 843.34 (0.0052) 965.45 (0.056) 1187.61 (0.61) 1248.62 (0.0011) 1661.80 (0.077) 1785.80 (6.6e-06) 1786.83 (1.7e-05) 1910.01 (0.0024) 1911.01 (8.3e-05)
9*	0.63	3	1034.54 (0.28) 1538.71 (0.074) 1731.77 (0.0044)
10*	0.63	4	1837.97 (0.67) 1838.92 (0.033) 1857.98 (0.27) 1877.02 (0.19)

* Clusters that are used in the predictive model

6.4 Discussion

We studied the value of proteomic profiling in the diagnosis of LM in breast cancer patients. Following digestion of CSF proteins with trypsin, we analysed the resulting peptides with MALDI-TOF. In a univariate analysis, we detected many peptides that were differentially expressed between breast cancer patients with and without LM and normal controls. We then built a predictive model to diagnose LM in breast cancer patients. After validation by bootstrapping, the model achieved a sensitivity of 79% and specificity of 76%. In current clinical practice, diagnostic tests for LM include CSF cytology and gadolinium (Gd) enhanced MRI (1, 17, 18). The sensitivity of CSF cytology (75%) and Gd MRI (76%) are comparable to our predictive model (17). The specificity of MRI (77%) is also similar but the specificity of CSF cytology is much higher (100%) (17). We conclude that our test may be useful to support the diagnosis of LM in breast cancer patients. Importantly, the test only requires 20 µl of CSF and can therefore easily be combined with cytological examination of the CSF.

Following the original highly intriguing report that the serum proteome profile can be used for the early detection of ovarian cancer (3), many researchers have applied the SELDI-TOF technology to detect proteome profiles specific for other forms of cancer and non-malignant disease (6, 19). However, criticism has focused on the low reproducibility of the SELDI-TOF analytical tool (7, 9, 10, 20-25). Models based on SELDI-TOF protein profiling data generally performed poorly upon external validation in time (26). This lack of reproducibility may be due to variation in chip batches, mass spectrometers, sample stability, the low reproducibility of peak height and the low number of measurements per sample (7, 20, 27). We believe that our model is less affected by these variations for several reasons: Firstly, the sample preparation is simple, fully automated and does not require chips or fractionations. Secondly, we did not include the height of the peaks in the model because quantitative measurements of peak heights with both the MALDI and SELDI methods are poorly reproducible (28). In addition, we have carefully determined prior to analysis the number of replicates per sample that provided the optimal reproducibility (13). The number of replicates that we used (18) was much higher than in other studies. Thirdly, the predictors that we used in our model were clusters of peaks and not single peaks improving the robustness of the model. Changes in one peak position of a cluster, used as a predictor, will not have a dramatic effect on the performance of the predictive value of the entire cluster. In the future, the reliability of the method can be further improved by linking multiple peptide peaks to a single protein.

The direct identification of peptides from complex samples remains difficult, because of the complexity of tryptic digests of body fluids. A direct MS/MS identification of the peptides using MALDI TOF/TOF is not possible due to the presence of multiple peptides, even in small mass windows. Although off line nano LC-MALDI could solve this problem, we believe that the best method to identify peptides in complex mixtures is Fourier transform MS in which the exact mass of the peptide of interest is obtained. In most cases, the detection of multiple peptides derived from a single protein will allow identification of the protein. Our database search on the up-regulate peptides has demonstrated the feasibility of this approach for apolipoprotein A1. The up-regulation of different forms of apolipoprotein has been observed before in different SELDI-TOF studies (29-31). We are currently performing Fourier transform MS to identify the other up-regulated peptides as well.

Confounding factors in the present study could be differences in sample collection and storage between institutes, differences in total protein concentration and white cell count between the groups, reproducibility of the method and patient selection bias. The number of peptides, differentially expressed between the institutes, was identical to a chance distribution, excluding potential biases introduced by differences in sample handling. The white cell count and protein concentration in CSF from patients with LM were increased compared to both breast cancer patients without LM and normal controls. All samples were routinely centrifuged after lumbar puncture making contamination of the supernatant with cellular debris unlikely. The elevated protein concentration in the CSF from LM patients is well-known and due to dysfunction of the blood-barrier (4) resulting in an increase of high abundant serum proteins in the CSF. A normalization on protein concentration could have been used to compensate for this difference. However, this implies that less CSF should be used from LM samples. This would result in a lower amount of CSF specific proteins compared to the control samples. In our opinion this would result in a bias between the three sample groups. To investigate the potential confounding effect of differences in total protein concentration, we calculated the average number of tryptic peptide digests derived from albumin in each group. The number of albumin derived peptide peaks did not differ between the three groups. In addition, no significant negative or positive correlation between the number of peaks and the protein concentration could be detected. This provided strong evidence that the differences in protein concentration had not interfered with the analysis.

All breast cancer patients in the present study had signs or symptoms compatible with LM, which led to the performance of a lumbar puncture. All patients in group I, diagnosed with LM, had positive cytology. All patients in group II had negative cytology combined with clinical follow-up indicating an alternative diagnosis. At this stage we did not include a group of patients with 'false-negative' CSF cytology as indicated by MRI and/or clinical follow-up. It will be particularly interesting to investigate the performance of this proteomic based test also in these patients, preferably in a prospective manner.

We conclude that MALDI-TOF analysis of tryptic peptide digests derived from the CSF of breast cancer patients can support the diagnosis of LM. We expect that the use of more accurate and sensitive measurements by Fourier-transform mass spectrometry will further improve the identification of disease-specific patterns and markers from body fluids in the near future.

References

1. DeAngelis, L. M. "Current diagnosis and treatment of leptomeningeal metastasis," *J Neurooncol* 38 (1998): 245-52.
2. Zheng, P. P., Luider, T. M., Pieters, R., Avezaat, C. J., van den Bent, M. J., Sillevis Smitt, P. A., and Kros, J. M. "Identification of tumor-related proteins by proteomic analysis of cerebrospinal fluid from patients with primary brain tumors," *J Neuropathol Exp Neurol* 62 (2003): 855-62.
3. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet* 359 (2002): 572-7.
4. Fishman, R. A. *Cerebrospinal fluid in diseases of the nervous system*. Philadephia: W.B. Saunders Company, 1992.
5. Pusch, W., Flocco, M. T., Leung, S.-M., and Thiele, H. "Mass-spectrometry based clinical proteomics," *Pharmacogenomics* 4 (2003): 463-476.
6. Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., Hochstrasser, D. F., and Sanchez, J. C. "A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease," *Proteomics* 3 (2003): 1486-94.
7. Diamandis, E. P. "Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics?," *Clin Chem* 49 (2003): 1272-5.
8. Diamandis, E. P., and van der Merwe, D. E. "Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations," *Clin Cancer Res* 11 (2005): 963-5.
9. Diamandis, E. P. "Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems," *J Natl Cancer Inst* 96 (2004): 353-6.
10. Diamandis, E. P. "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Mol Cell Proteomics* 3 (2004): 367-78.
11. Koomen, J. M., Zhao, H., Li, D., Abbruzzese, J., Baggerly, K., and Kobayashi, R. "Diagnostic protein discovery using proteolytic peptide targeting and identification," *Rapid Commun Mass Spectrom* 18 (2004): 2537-48.
12. Ramstrom, M., Palmblad, M., Markides, K. E., Hakansson, P., and Bergquist, J. "Protein identification in cerebrospinal fluid using packed capillary liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry," *Proteomics* 3 (2003): 184-90.
13. Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M. "A new method to analyze matrix-assisted laser desorption/ionization

time-of-flight peptide profiling mass spectra," *Rapid Commun Mass Spectrom* 19 (2005): 865-870.

14. Koza, J. R. *Genetic programming*: MIT press, 1992.
15. Efron, B., and Tibshirani, R. "Bootstrap methods for standard errors, confidence intervals, and other measures of accuracy," *Statistical science* 1 (1986): 54-77.
16. Efron, B., and Tibshirani, R. *An introduction to the bootstrap*. New York: Chapman and Hall, 1993.
17. Straathof, C. S., de Bruin, H. G., Dippel, D. W., and Vecht, C. J. "The diagnostic accuracy of magnetic resonance imaging and cerebrospinal fluid cytology in leptomeningeal metastasis," *J Neurol* 246 (1999): 810-4.
18. Freilich, R. J., Krol, G., and DeAngelis, L. M. "Neuroimaging and cerebrospinal fluid cytology in the diagnosis of leptomeningeal metastasis," *Ann Neurol* 38 (1995): 51-7.
19. Soltys, S. G., Shi, G., Tibshirani, R., Giaccia, A. J., Koong, A. C., and Le, Q. "The use of plasma SELDI-TOF MS proteomic patterns for detection of head and neck squamous cell cancers (HNSCC)," *Int J Radiat Oncol Biol Phys* 57 (2003): S202.
20. Baggerly, K. A., Morris, J. S., and Coombes, K. R. "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics* 20 (2004): 777-85.
21. Check, E. "Proteomics and cancer: running before we can walk?," *Nature* 429 (2004): 496-7.
22. White, C. N., Chan, D. W., and Zhang, Z. "Bioinformatics strategies for proteomic profiling," *Clin Biochem* 37 (2004): 636-41.
23. Diamandis, E. P. "How are we going to discover new cancer biomarkers? A proteomic approach for bladder cancer," *Clin Chem* 50 (2004): 793-5.
24. Diamandis, E. P. "Proteomic patterns to identify ovarian cancer: 3 years on," *Expert Rev Mol Diagn* 4 (2004): 575-7.
25. Diamandis, E. P. "Identification of serum amyloid a protein as a potentially useful biomarker for nasopharyngeal carcinoma," *Clin Cancer Res* 10 (2004): 5293; author reply 5293-4.
26. Rogers, M. A., Clarke, P., Noble, J., Munro, N. P., Paul, A., Selby, P. J., and Banks, R. E. "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility," *Cancer Res* 63 (2003): 6971-83.
27. Qu, Y., Adam, B.-L., Yasui, Y., and Ward, M. D. "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostrate cancer from noncancer patients," *Clinical chemistry* 48 (2002): 1835-1843.
28. Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., and Tempst, P. "Serum Peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry," *Anal Chem* 76 (2004): 1560-70.
29. Heike, Y., Hosokawa, M., Osumi, S., Fujii, D., Aogi, K., Takigawa, N., Ida, M., Tajiri, H., Eguchi, K., Shiwa, M., Wakatabe, R., Arikuni, H., Takaue, Y., and Takashima, S. "Identification of serum proteins related to adverse effects induced by docetaxel infusion from protein expression profiles of serum using SELDI ProteinChip system," *Anticancer Res* 25 (2005): 1197-203.

30. Malik, G., Ward, M. D., Gupta, S. K., Trosset, M. W., Grizzle, W. E., Adam, B. L., Diaz, J. I., and Semmes, O. J. "Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer," *Clin Cancer Res* 11 (2005): 1073-85.

31. Allard, L., Lescuyer, P., Burgess, J., Leung, K. Y., Ward, M., Walter, N., Burkhard, P. R., Corthals, G., Hochstrasser, D. F., and Sanchez, J. C. "ApoC-I and ApoC-III as potential plasmatic markers to distinguish between ischemic and hemorrhagic stroke," *Proteomics* 4 (2004): 2242-51.

Chapter 7

Identification of leptomeningeal metastasis-related proteins in cerebrospinal fluid of patients with breast cancer by a combination of MALDI-TOF, MALDI-FTICR and nanoLC-FTICR mass spectrometry

Rompp, A. [1], Dekker, L. [1], Taban, I., Jenster, G., Boogerd, W., Bonfrer, H., Spengler, B., Heeren, R., Sillevis Smitt, P., and Luider, T. M.

Proteomics 7 (2007): 474-81

[1]First two authors contributed equally to this manuscript.

Abstract

Leptomeningeal metastasis is a devastating complication occurring in 5% of breast cancer patients. However, the current 'gold standard' of diagnosis, namely microscopic examination of the cerebrospinal fluid, is false-negative in 25% of patients at the first lumbar puncture. In a previous study, we analyzed a set of 151 CSF samples (tryptic digests) by MALDI-TOF and detected peptide masses which were differentially expressed in breast cancer patients with LM. In the present study, we obtain for a limited number of samples exact masses for these peptides by MALDI-FTICR MS measurements. Identification of these peptides was performed by electrospray FTICR MS after separation by nano-scale liquid chromatography. The database results were confirmed by targeted high mass accuracy measurements of the fragment ions in the FTICR cell. The combination of automated high-throughput MALDI-TOF measurements and analysis by FTICR MS leads to the identification of seventeen peptides corresponding to nine proteins. These include proteins that are operative in host-disease interaction, inflammation and immune defense (serotransferrin, alpha 1-antichymotrypsin, hemopexin, haptoglobin and transthyretin). Several of these proteins have been mentioned in the literature in relation to cancer. The identified proteins alpha1-antichymotrypsin and apolipoprotein E have been described in relation to Alzheimer's disease and brain cancer.

7.1 Introduction

Approximately 5% of all breast cancer patients will develop leptomeningeal metastasis (LM) during the course of their disease. Currently, the diagnosis is performed by cytological examination of the CSF ('gold standard') with or without gadolinium enhanced MRI. However, the false negative rate of both methods is approximately 25% at first examination. In an earlier study (1), we have shown that peptide profiling by MALDI-TOF provides a method for the detection and characterize clusters of tryptic peptides that correlate with the presence of LM in breast cancer patients. Our analyses of tryptic peptide profiles of CSF samples from 106 patients with active breast cancer (54 with LM and 52 without LM) and 45 control patients by MALDI-TOF showed that 190 peaks, of the in total 895 detected peaks, were significantly differentially expressed between sample groups. In this report, we focus on the identification of the peptides and proteins that are differentially expressed in the CSF from breast cancer patients with LM.

The complexity of tryptic peptide mixtures of CSF makes it almost impossible to perform direct MS/MS identification by MALDI-TOF/TOF. Sample pre-fractionation, higher mass resolution and mass accuracy are needed for such samples. Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) offers a range of features that are ideally suited for the identification of biological macromolecules in complex mixtures (2, 3). These features include a mass resolution in excess of 1,000,000, the ability to routinely achieve measurements with low or sub-ppm mass accuracy and a high sensitivity. However, the high mass resolution is of limited help if too many peptides in a wide concentration range have to be measured. Low abundant peptides will be suppressed by the main components in the mixture. Therefore an additional separation step prior to mass spectrometric detection should be added. In this study an online combination of nano-scale liquid chromatography (nanoLC) and FTICR MS interfaced by a nanospray ionization source was used for the identification of the differentially expressed peptides. The identification of biomarkers requires a statistically sound basis and this is achieved by the analysis of large numbers of samples. However, it is still very time-consuming to analyze hundreds of samples by high resolution mass spectrometry. Therefore we decided to use only a limited set of samples for these measurements; the selection of these samples was based on the statistical analysis of MALDI-TOF data. Because of the relative low mass accuracy of TOF mass measurement, the linking of the peptides identified by nanoLC FTICR MS to the original MALDI-TOF data is not straight-forward. For this reason, we first performed accurate mass measurements by MALDI-FTICR MS to obtain sub-ppm mass accuracies for the significantly differential peptides observed in the original study (1). We used the accurate masses obtained by MALDI-FTICR MS, and the original MALDI-TOF data to link and identify the differentially expressed peptides by nanoLC-FTICR MS. Screening by high-throughput automated MALDI-TOF mass spectrometry combined with more complex and time consuming nanoLC-FTICR MS for identification purposes allows the identification of differentially expressed peptides in large numbers of CSF samples on a statistically sound basis.

7.2 Methods

7.2.1 Samples

CSF samples used in this study are described in more detail in Dekker et al 2005 (1). Briefly, patients with positive cytology or with a compatible neurological syndrome and diagnostic MRI were considered to have LM. When cytology was negative and when clinical follow-up was incompatible with LM, patients were classified as having advanced breast cancer without LM. Controls were patients who were not known to have cancer and who did not suffer from any known neurological disease. Hence, our study comprises three groups: breast carcinoma with LM, breast carcinoma without LM and control. Twelve samples (4 of each group) from the original set of 151 CSF samples (1) were chosen for the high resolution FTICR MS measurements.

7.2.2 Sample preparation

Of each CSF sample, 50 µl were tryptically digested according to the protocol described previously (1). The tryptic digest was divided into two equal volumes and stored at -20ºC. Additionally, the six most abundant proteins were removed from four samples (two breast carcinoma / two breast carcinoma with LM) by a multiple affinity removal system HPLC column (Agilent Technologies, USA). The column contains polyclonal antibodies to human albumin, transferrin, haptoglobin, alpha1-antitrypsin, IgG, and IgA. Fifty microliters of each CSF sample were diluted fivefold according to the manufacturer's protocol. The flow-through fractions from the injections were collected and enzymatically digested. For digestion 0.1 µg/µl gold grade trypsin (Promega, USA) in 3 mM Tris-HCl was added at a 1:10 (v/v) ratio and the samples were subsequently incubated overnight at 37 ºC. Digested protein samples were stored at -20 ºC until measurements were performed.

7.2.3 MALDI-FTICR MS

One microliter of the tryptic digest (corresponding to 0.5 µl CSF) was spotted onto an anchorchip target plate (600/384 anchorchip with transponder plate; Bruker Daltonik GmbH, Bremen, Germany). Before the spots dried, one microliter of 10 mg dihydroxy-benzoic acid (DHB) matrix (Bruker Daltonics GmbH, Bremen, Germany) dissolved in 1 ml 0.1% TFA water was added and the spots were allowed to dry at room temperature. All spots were measured with a MALDI FTICR MS (Apex Q 9.4 tesla equipped with a combi-source, Bruker Daltonics, USA). For each scan the ions generated by 10 laser shots were accumulated in the storage hexapole. For each sample 100 scans were combined to obtain one mass spectrum.

All MALDI-FTICR MS spectra were analyzed by Data Analyses software package (Bruker Daltonics, USA version 3.4 build 150) and each spectrum was internally calibrated with respect to the most intense albumin peaks in the sample. This ensured a mass accuracy of less than 1 ppm. Subsequently, peak picking is performed with the SNAP 2 algorithm with a signal to noise threshold of 4 and a quality factor of 0.9. The generated peak lists of all samples are clustered with a mass tolerance of 1 ppm using an Excel macro to generate a combined peak list.

7.2.4 NanoLC-FTICR MS

Separation was performed on a nanoscale liquid chromatography system (nanoLC) (LC Packings, Amsterdam, Netherlands) with a 60 min gradient (10%-35% acetonitrile/H_20). The injection volume was 1 µl corresponding to 0.5 µl CSF. After pre-concentration on a trap column (1 mm x 300 µm I.D.) the peptides were separated on a C18 PepMap column (150 mm x 75 µm I.D) at 200 nL/min (LC Packings, Amsterdam, The Netherlands). A UV dector (214 nm) was used to monitor the separation. The nanoLC was coupled to the mass spectrometer by a nanospray source. The separated peptides were detected by a linear ion trap Fourier transform ion cyclotron resonance (LT-FTICR) mass spectrometer (*Finnigan LTQ FT*, Thermo Electron, Bremen, Germany). All samples were measured in the selected ion monitoring (SIM) method in order to identify proteins in a database search. In this method, the peptide masses are measured in a survey scan with reduced resolution (R= 50,000). In the next step, selected precursor masses are detected with a lower ion population in the ICR cell in order to increase mass accuracy. Fragment ion masses are measured in the linear ion trap to speed up the analysis. Only precursor ions that corresponded to the list of interesting peptide masses (as determined by the MALDI-TOF measurements) were selected for fragmentation. The mass tolerance of the inclusion list for this selection was 0.5 Da.

Peptides and their corresponding proteins were identified with SEQUEST using the Bioworks software (Version 3.2, Thermo Electron, San Jose, USA). As database was used the human database downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The mass accuracy for the database search was set to 2 ppm for the precursor ions. Only peptides with ion probabilities of less than 0.001 were considered for further analysis. In order to validate the database search results, the identified peptides were further analyzed in targeted MS/MS experiments: in the experiment the fragment ion masses were also measured in the FTICR part of the *Finnigan LTQ FT* to obtain highly accurate masses for the fragment ions. The fragment ion masses were manually compared to the calculated masses of the peptides identified in the database search. The mass accuracy of these measurements was below 2 ppm. A peptide search with 2 ppm mass accuracy for the fragment ions is not possible with the current version of the Bioworks software.

7.2.5 Data analyses

From the original MALDI-TOF peak list described by Dekker et al. 2005 (1) a list of masses of significantly differentially expressed peptides was extracted. This list contained all the peaks that showed a significant difference ($p < 0.01$) in a Kruskall-Wallis test between the three sample groups. All the masses of the peptides that were observed by MALDI-FTICR MS were compared with the MALDI-TOF peaks. If masses matched within a mass tolerance of 50 ppm, the accurate mass was linked to the original MALDI-TOF mass. All the linked MALDI-FTICR MS masses (accuracy < 1ppm) were then compared with the combined peak list of all ESI-FTICR identifications with a 2 ppm mass tolerance. If a linked mass matched within a 2 ppm mass window, the mass was considered as identified.

An *in silico* digestion using the peptide mass tool (http://www.expasy.ch/cgi-bin/peptide-mass.pl) was performed on the sequences of the identified proteins. The calculated masses of all these peptides were compared with the list of MALDI-FTMS masses that were also present in the MALDI-TOF analyses. If masses matched within a mass tolerance of 1 ppm these peptides were considered as additional detected peptides of the identified proteins.

For patients in the group with breast cancer and LM, the original data was also used to test whether a significant difference exists in peak occurrence for the identified peaks with respect to patients showing low and high total protein concentration in the CSF. This group contained a total of 41 patients and in 35 of these patients the measured protein concentration was above the normal range (1). Based on this criterion two new groups were created, viz. patients with a protein concentration higher than the normal range (n=35) or patients within the normal range (n=6). A reduced matrix containing the occurrence of the identified differentially expressed peptide peaks was created for these two groups. For each individual identified differential peptide peak a p-value was calculated using a Wilcoxon test (SPSS 13.0, 2004, USA).

7.3 Results

The following analysis is based on peptide peaks in the mass range between 800 and 2500 Da. These peaks were detectable by all mass spectrometers used in this study. The MALDI-TOF profiling experiments of 151 CSF samples resulted in 654 peptide peaks in this mass range of which 165 were differentially expressed. Isotopic clusters could only be partially resolved for these measurements. The number of monoisotopic peaks was estimated to be about 300. The MALDI-FTICR MS measurements of the twelve samples resulted in the detection of 850 unique monoisotopic peaks (S/N≥4). The measurements performed with nanoLC electrospray FTICR MS resulted in the detection of 8621 monoisotopic peaks. The MALDI-TOF and MALDI-FTICR peak lists were compared with a mass window of 50 ppm. This resulted in an overlap of 293 monoisotopic peptide masses (Figure 7.1). This indicated that the vast majority of monoisotopic MALDI-TOF peaks were also detected by MALDI-FTICR. The comparison of peptide masses detected by MALDI-FTICR and nanoLC-FTICR measurements with a tolerance of 2 ppm resulted in an overlap of 331 peaks (39% of MALDI-FTICR masses) (Figure 7.2).

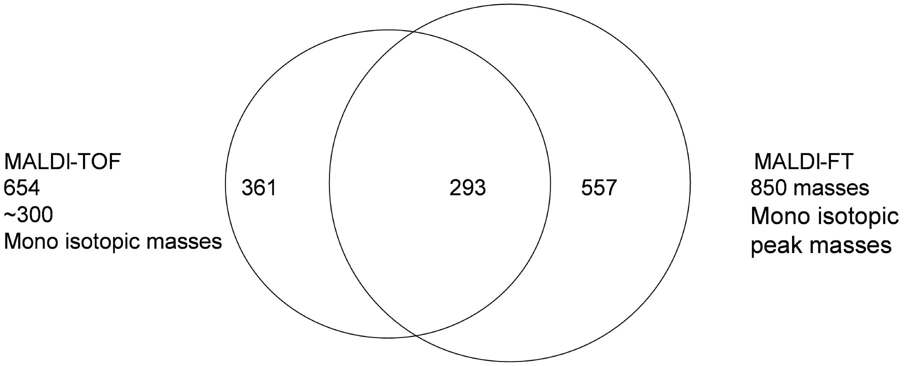


Figure 7.1: Overlap between peak masses of MALDI measurements. The peaks of 151 tryptic digested CSF samples measured by MALDI-TOF compared with the MALDI-FTICR measurements of 12 samples (from the set of 151) with a 50 ppm mass tolerance. For MALDI-TOF measurements all isotopic peaks are included; for MALDI-FT only monoisotopic peaks.

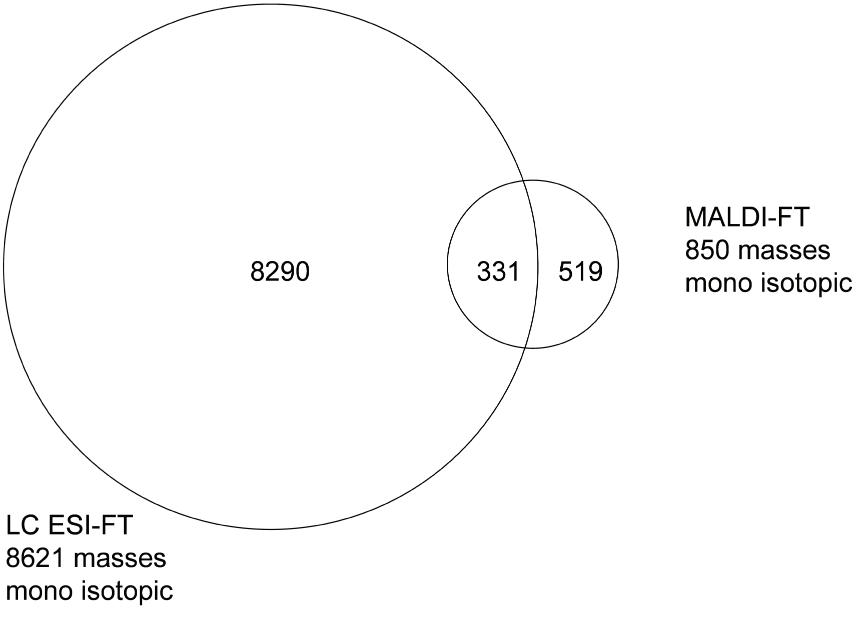


Figure 7.2: Overlap between peak masses of MALDI-FTICR MS and nanoLC-FTICR MS measurements. Twelve identical tryptic digested CSF samples are measured and analysed by both techniques and compared with a mass tolerance of 2 ppm.

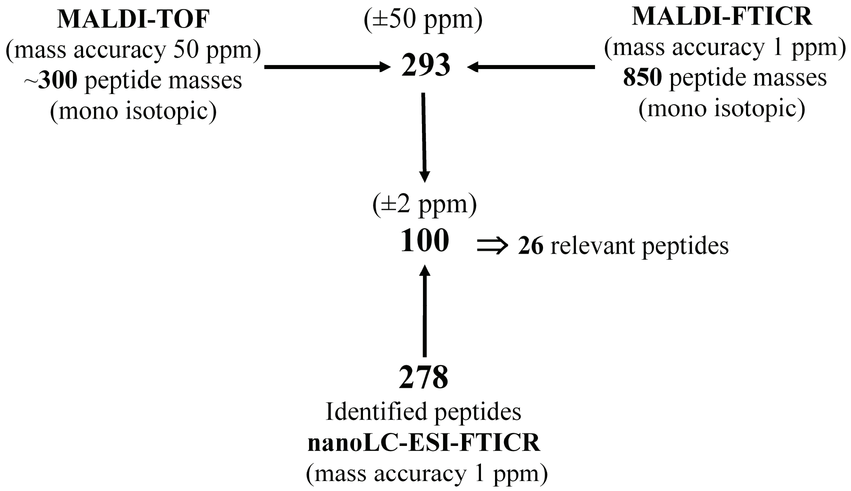


Figure 7.3: Linking/identification procedure of MALDI-TOF peaks. MALDI-TOF peak masses are linked to accurate masses from MALDI-FTICR measurements (50 ppm mass tolerance in a mass range of 800-2500 Da). All linked accurate masses are compared with the identified masses of the nanoLC-FTICR MS measurements with a mass tolerance of 2 ppm. Twenty-six of the 100 identified peptides showed a significant p-value in the MALDI-TOF analysis.

The nanoLC-FTICR measurements resulted in the identification of 278 peptides corresponding to 115 proteins. Forty-seven of these proteins were identified by the presence of at least two peptides. It is important to note that these measurements were specifically focused on the identification of the peptides already observed in MALDI-TOF experiments and not on the identification of as many proteins as possible. The mass tolerance of the inclusion list (containing the peak masses of the MALDI-TOF measurements) for MS/MS measurements was ± 0.5 Da. Therefore peptides in this mass range were also fragmented and (partially) identified, even if they were not included in the MALDI-TOF list. The linking of peptides identified by nanoLC-FTICR MS to peak masses detected in the MALDI-TOF measurements is illustrated in Figure 7.3. One hundred of the 278 identified peptides were also detected in both the MALDI-TOF and the MALDI FTICR measurements and 26 of these showed p-values of less than 0.01, i.e. they were differentially expressed between the three sample groups. A few peptides were exclusively identified in the CSF samples that were depleted of the most abundant proteins (see methods). The peptide masses 1773.76494 Da (alpha1-antichymotrypsin), 1301.64844 Da (apolipoprotein A1) and 842.50943 (putative protein) could only be identified in these samples.

Table 7.1: Identified peptide masses with a significant p-value (<0.01) in MALDI-TOF profiling experiment. The column FT-MSMS confirmation indicates if the measured fragment ions match with those of the identified peptide within a mass tolerance of 2 ppm.

Theoretical Mass (MH+)	Protein	Accession number	Peptide sequence	Measured m/z	Charge	MH+ (Measured)	Mass accuracy ppm	FT-MS/MS confirmation
1632.82231	Alpha1-antichymotrypsin	1340142	K.MEEVEAMLLPETLK.R	816.91498	2	1632.82268	-0.23	
1773.76494*	Alpha1-antichymotrypsin	1340142	K.WEMPFDPQDTHQSR.F	591.92657	3	1773.76516	-0.12	
1283.5725	Apolipoprotein A-I	2914178	K.WQEEMELYR.Q	642.29004	2	1283.5728	-0.24	x
1301.64844*	Apolipoprotein A-I	2914178	R.THLAPYSDELR.Q	651.32737	2	1301.64746	0.75	x
1497.80198	Apolipoprotein E	178851	R.AATVGSLAGQPLQER.A	749.4043	2	1497.80132	0.44	
1730.8443	Apolipoprotein E	178851	K.SELEEQLTPVAEETR.A	865.92596	2	1730.84464	-0.2	x
968.55236	Apolipoprotein E	178851	R.LGPLVEQGR.V	484.77921	2	968.551143	1.26	x
1203.63681	Haptoglobin	1620396	K.VTSIQDWVQK.T	602.32245	2	1203.63762	-0.68	x
1837.88677	Hemopexin	1708182	K.SGAQATWTELPWPHEK.V	613.30096	3	1837.88833	-0.85	x
1785.8766	Prostaglandin D2 synthase	189772	-.APEAQVSVQPNFQQDK.F	893.44287	2	1785.87846	-1.04	x
1909.95418	Prostaglandin D2 synthase	189772	K.AQGFTEDTIVFLPQTDK.C	637.32227	3	1909.95226	1.01	x
842.50943*	Putative [Homo sapiens]	553734	K.GITLSVRP.-	421.75793	2	842.508583	1.01	x
1366.75899	Transthyretin	4558178	R.GSPAINVAVHVFR.K	456.25845	3	1366.7608	-1.32	x
1195.55243	Serotransferrin	4389230	K.DSGFQMNQLR.G	598.27936	2	1195.55144	0.83	
1478.73481	Serotransferrin	4389230	K.MYLGYEYVTAIR.N	739.87146	2	1478.73564	-0.56	x
1577.81446	Serotransferrin	4389230	R.TAGWNIPMGLLYNK.I	789.41028	2	1577.81328	0.75	x
1629.8159	Serotransferrin	4389230	K.EDPQTFYYAVAVVK.K	815.41089	2	1629.8145	0.86	x

* Peptides exclusively identified in the depleted samples

In order to validate the identification a manual check of all identified masses in the MALDI-FTICR spectra was performed. This check included the quality of the peak and the presence of high intensity isotopic peaks. The presence of an isotopic peak close to the identified peak could mean that the differential peak in the MALDI-TOF analyses is not the identified peak but an isotopic peak. That is to say the identified peak in the MALDI-FTICR spectra lies close to an isotope peak of another peptide which would not be apparent from the low resolution TOF data. This reduced the number of identified peptides from 26 to 17. These 17 identified peptides are listed in Table 1 with the sequence information and accuracy of the measurement of the parent ion obtained from the ESI-FTICR MS.

The database identification of the peptides was validated by MS/MS measurements in the FTICR cell of the *Finnigan LTQ FT*. An example of fragment ions measured with high mass accuracy is shown in Figure 7.4. The masses of the fragment ions were compared to those generated by the peptide identified in the database search. In case of the peptide of m/z 739.87146 (serrotransferin) the fragment ion masses matched within 1 ppm in most cases. Peptides were only included in the list of identified peptides/proteins if these masses matched within an accuracy of less than 2 ppm. These criteria were met for 13 peptides as indicated in the last column in Table 1. Given the high mass accuracy of these measurements the probability of a false hit was greatly reduced by this additional step.

In silico digestion of all identified proteins resulted in a list of possible peptides. The calculated masses of these peptides were compared to the MALDI-FTMS masses that were also present in the original MALDI-TOF data set. This resulted in the detection of 20 additional peptides belonging to the identified proteins that matched within 1 ppm of the calculated mass. The nanoLC-FTICR MS measurements were also searched for additional identified peptides that match the proteins, which were differentially expressed. The number of these additional peptides for each protein is shown in Table 7.2.

Table 7.2: Functional information of identified peaks that were differentially expressed in MALDI-TOF analyses

Protein name	**Up/down regulation in LM**	**Function**	**Acute phase reactant**	**Number of identified peptides***	**Additional peptides**	
					MALDI**	**ESI-FTMS**
Alpha1-antichymotrypsin	↑	Protease inhibitor	x	2	4	6
Apolipoprotein A-1	↑	Transport		2	3	11
Apolipoprotein E	↓	Transport		3	5	6
Haptoglobin	↑	Transport	x	1	3	7
Hemopexin	↑	Transport	x	1	2	2
Prostaglandin D2 synthase	↓	Metabolism/transport		2	1	4
Putative	↓	Unknown		1	0	1
Serotransferrin	↑	Transport	x	4	2	8
Transthyretin	↓	Transport	x	1	0	2

*Detected by all three MS methods
**Peptides that have been detected with both MALDI methods

From all the identified differentially expressed peptides, the peak frequencies in the original MALDI-TOF data of LM patients with an elevated protein concentration (higher than normal level n=35) and a normal protein concentration (n=6) were compared (1). This was done in order to test whether the disruption or dysfunction of the blood-CSF barrier has an effect on the expression of these differentially expressed peptides. A Wilcoxon test indicated that for none of these peptides a significant difference between the high and low protein concentration group could be observed.

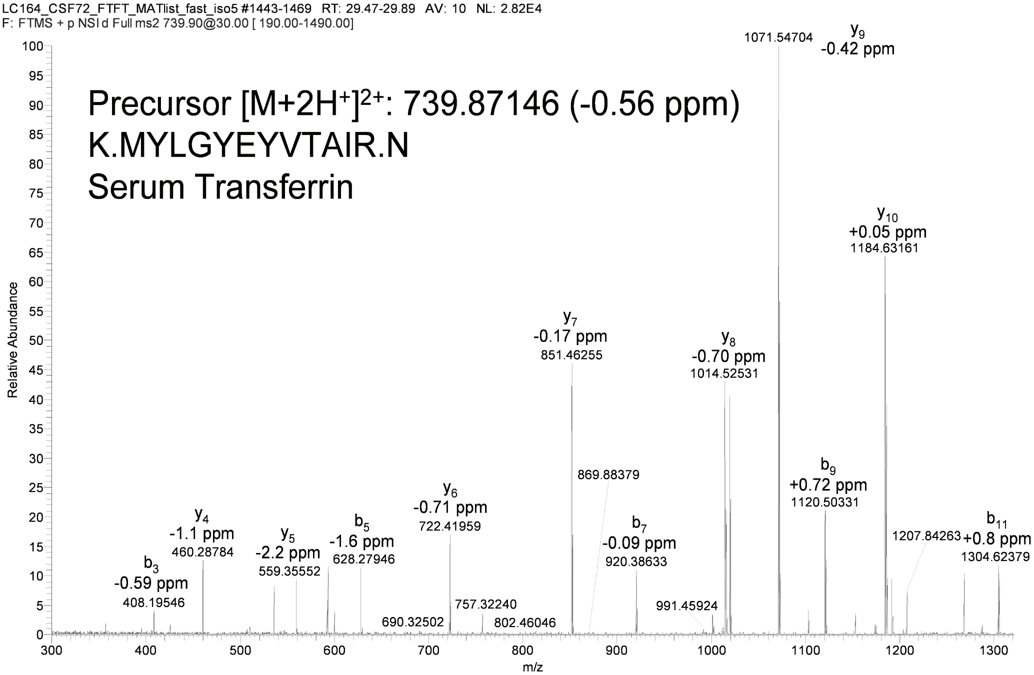


Figure 7.4: An example of a manual examination of an MS/MS spectrum measured in the FTICR cell of the Finnigan LTQ FT.

7.4 Discussion

It has already been shown that in spite of different ionization mechanisms associated with ESI and MALDI there is substantial overlap between both techniques (4). As stated earlier, in this paper we exploit the overlap between ESI and MALDI to identify peptides that have previously been detected in a MALDI-TOF peptide profiling experiment (1). Integration of both FTICR techniques applied to exactly the same samples and using the same mass range increases the reliability of identification. In our experiment the overlap between MALDI and ESI is 39%. Considering this, the number of peptides that have been identified is relatively high. Out of all 654 (~300 monoisotopic) masses detected by MALDI-TOF in a comparable mass range with both FTICR techniques, 100 peptides could be identified.

In a first comparison 27 masses of the 165 (~80 monoisotopic) differentially expressed peptides could be assigned to proteins. A further manual check of the MALDI-FTMS resulted in a reduction of this number to 17. The improved resolution of the FTMS data was used as a quality check on the MALDI-TOF data. For some masses multiple peaks were observed in the

50 ppm mass window of the MALDI-TOF peak and for this reason the peak could not be assigned with confidence to a single protein. In some cases the signal detected in MALDI-TOF was not the monoisotopic peak and was therefore discarded from the list. By performing an *in silico* tryptic digestion of the identified proteins and by matching the calculated masses to the masses measured in the FTMS, 20 extra peptides belonging to the identified proteins could be found in the MALDI-TOF and in the MALDI-FTMS data. This, in combination with the identified peaks, resulted in an improved assignment of peptides to the proteins.

We assumed that peptide peaks that were detected in the MALDI-FTICR MS and nanoLC-FTICR MS measurements with the same accurate mass (+/- 1 ppm) correspond to the same peptide. A random match of two different peptides in the MALDI and ESI measurement is quite unlikely at this low mass tolerance. For the mass 1366.75899 Da ($[M+H]^+$) there are only 3 different tryptic peptides that could derive from human proteins. This number was derived from a MS-Tag search in ProteinProspector (assuming a mass tolerance of +/- 1 ppm, no modifications, SwissProt database). It is important to keep in mind that this number refers to all human proteins, most of these proteins will not necessarily be present in the cerebrospinal fluid.

The analysis of the depleted samples resulted in the identification of three additional peptides from the list of significant peaks of the MALDI-TOF analysis. The reason for this limited increase in identifications is that in the large scale MALDI-TOF screening no depletion is performed so the chances of picking up low abundant proteins were limited. However, the depletion leads to better quality MS/MS spectra in several cases, facilitating the identification of peptides. However, to avoid the necessity of using large amounts of CSF, miniaturization of the depletion technique is a prerequisite for its routine application to CSF analysis in the near future.

Most of the identified differentially expressed proteins are common and known serum and CSF proteins that have a function in transport (Table 7.2). Serotransferrin, hemopexin (5) and haptoglobin are part of the iron and heme transport to the liver, degradation of hemoglobin and the re-uptake of iron. Apolipoprotein A1 participates in the reverse transport of cholesterol from tissue to the liver. Apolipoprotein E and transthyretin transport lipids and thyroxine, respectively, from the bloodstream to the brain. Alpha1-antichymotrypsin functions as a protease inhibitor. Changes in serum levels of alpha1-antichymotrypsin, hemopexin, haptoglobin (up-regulation), transthyretin and serotransferrin (down regulation) are described in reaction to an acute phase response, e.g. host-disease interaction, immune response and/or inflammation (6-8).

Many of the identified differentially expressed proteins have been described as potential markers in different types of cancers such as ovarian cancer (apolipoprotein a-1, serotransferrin, transthyretin), hepatocelluar carcinoma (apolipoprotein A1), pancreatic cancer (apolipoprotein E, alpha-1-antichymotrypsin), glioblastoma (apolipoprotein E) and renal cancer (haptoglobin) (9-12).

Disruption of the blood-CSF barrier in patients with LM is a common observation. This disruption is also indicated by the increased total protein concentration in 85% of patients in the LM group. However, the disruption of the blood-CSF barrier is only partial (13). This disruption of the blood-CSF barrier can explain the elevated levels of serum proteins in CSF of LM patients. However, some proteins such as immunoglobulins are not elevated. An increase in concentration for all proteins would be expected if a complete disruption in the blood-CSF barrier would occur (14). This is not the case as also supported by the original MALDI-TOF data: no significant difference in the occurrence of the identified peptides is observed between patients with low and high protein concentration. This indicates that the disruption of the blood-CSF barrier has only a limited effect on the up- or down-regulation of these differentially expressed proteins. The effects observed on the serum proteins are probably a combination of an acute phase response to the tumor metastasis and the partial disruption of the blood-CSF barrier.

The decrease of prostaglandin synthase D2 and the unchanged level of cystatin-c in the CSF of patients with a higher CSF/serum albumin concentration (as seen in our study) is an indication of inflammatory meningeal processes (14). Both of these proteins are highly brain specific and are synthesized by meningeal cells. In case of blood-CSF barrier dysfunction, the flow rate of the CSF is reduced which results in a higher concentration of prostaglandin synthase D2 and of cystatin-c in the CSF (14). In case of an inflammation of the meninges the production of both of these proteins is decreased and no elevation of these proteins in CSF is observed.

Changes in alpha-1-antichymotrypsin and apolipoprotein E expression have been reported in Alzheimer's disease. The regulation that we observe, down-regulation for apolipoprotein E and up-regulation for alpha-1-antichymotrypsin, is observed also by Licastro and coworkers in Alzheimer's disease and in aging of the brain in a mouse model (15). They propose an inhibiting effect of apolipoprotein E on Alpha-1-antichymotrypsin expression. Both proteins have also been detected in brain tumor cells by immunohistochemistry (16, 17).

In conclusion, by integration of three mass spectrometry techniques we were able to identify CSF proteins determined as being significantly differentially expressed in a large scale MALDI-TOF study. This combination of techniques used is in general well-suited for biomarker research, the MALDI-TOF screening is a fast technique allowing the analyses of relatively large numbers of samples, which is a requirement for reliable statistical analyses, the FTMS measurements are very precise limiting the chances of false positive identifications. All identified peptides belong to abundant CSF proteins. Based on the literature there are indications that the identified proteins are not necessarily related to a disruption of the blood-CSF barrier, but probably have a function in acute-phase response and other brain specific processes. This is in agreement with our results that show that the disruption of the blood-CSF barrier has only a limited effect on the up- or down-regulation of the identified proteins.

References

1. Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M., and Sillevis Smitt, P. A. "MALDI-TOF Mass Spectrometry Analysis of Cerebrospinal Fluid Tryptic Peptide Profiles to Diagnose Leptomeningeal Metastases in Patients with Breast Cancer," *Mol Cell Proteomics* 4 (2005): 1341-1349.
2. Rompp, A., Taban, I. M., Mihalca, R., Duursma, M. C., Mize, T. H., McDonnel, L. A., and Heeren, R. M. "Examples of Fourier transform ion cyclotron resonance mass spectrometry developments: from ion physics to remote access biochemical mass spectrometry," *Eur J Mass Spectrom (Chichester, Eng)* 11 (2005): 443-56.
3. Qian, W. J., Camp, D. G., 2nd, and Smith, R. D. "High-throughput proteomics using Fourier transform ion cyclotron resonance mass spectrometry," *Expert Rev Proteomics* 1 (2004): 87-95.
4. Bodnar, W. M., Blackburn, R. K., Krise, J. M., and Moseley, M. A. "Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage," *J Am Soc Mass Spectrom* 14 (2003): 971-9.
5. Tolosano, E., and Altruda, F. "Hemopexin: structure, function, and regulation," *DNA Cell Biol* 21 (2002): 297-306.
6. Ingenbleek, Y., and Young, V. "Transthyretin (prealbumin) in health and disease: nutritional implications," *Annu Rev Nutr* 14 (1994): 495-533.
7. Gruys, E., Toussaint, M. J., Niewold, T. A., and Koopmans, S. J. "Acute phase reaction and acute phase proteins," *J Zhejiang Univ Sci B* 6 (2005): 1045-56.
8. Ceciliani, F., Giordano, A., and Spagnolo, V. "The systemic reaction during inflammation: the acute-phase proteins," *Protein Pept Lett* 9 (2002): 211-23.
9. Kozak, K. R., Su, F., Whitelegge, J. P., Faull, K., Reddy, S., and Farias-Eisner, R. "Characterization of serum biomarkers for detection of early stage ovarian cancer," *Proteomics* 5 (2005): 4589-96.

10. Steel, L. F., Shumpert, D., Trotter, M., Seeholzer, S. H., Evans, A. A., London, W. T., Dwek, R., and Block, T. M. "A strategy for the comparative analysis of serum proteomes for the discovery of biomarkers for hepatocellular carcinoma," *Proteomics* 3 (2003): 601-9.
11. Yu, K. H., Rustgi, A. K., and Blair, I. A. "Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry," *J Proteome Res* 4 (2005): 1742-51.
12. Tolson, J., Bogumil, R., Brunst, E., Beck, H., Elsner, R., Humeny, A., Kratzin, H., Deeg, M., Kuczyk, M., Mueller, G. A., Mueller, C. A., and Flad, T. "Serum protein profiling by SELDI mass spectrometry: detection of multiple variants of serum amyloid alpha in renal cancer patients," *Lab Invest* 84 (2004): 845-56.
13. Taillibert, S., Laigle-Donadey, F., Chodkiewicz, C., Sanson, M., Hoang-Xuan, K., and Delattre, J. Y. "Leptomeningeal metastases from solid malignancy: a review," *J Neurooncol* 75 (2005): 85-99.
14. Reiber, H. "Dynamics of brain-derived proteins in cerebrospinal fluid," *Clin Chim Acta* 310 (2001): 173-86.
15. Licastro, F., Campbell, I. L., Kincaid, C., Veinbergs, I., Van Uden, E., Rockenstein, E., Mallory, M., Gilbert, J. R., and Masliah, E. "A role for apoE in regulating the levels of alpha-1-antichymotrypsin in the aging mouse brain and in Alzheimer's disease," *Am J Pathol* 155 (1999): 869-75.
16. Ikeyama, Y., Orita, T., Nishizaki, T., Aoki, H., and Ito, H. "[Immunohistochemical expression of alpha-1-antitrypsin in human gliomas]," *No Shinkei Geka* 19 (1991): 1047-51.
17. Nicoll, J. A., Zunarelli, E., Rampling, R., Murray, L. S., Papanastassiou, V., and Stewart, J. "Involvement of apolipoprotein E in glioblastoma: immunohistochemistry and clinical outcome," *Neuroreport* 14 (2003): 1923-6.

Chapter 8

Differential expression of protease activity in serum samples of prostate carcinoma patients with metastases.

Dekker, L. J., Burgers, P. C., Charif, H., van Rijswijk, A.L., Titulaer, M. K.,
Jenster, G., Bischoff, R., Bangma, C. H., and Luider, T. M.

Submitted

Abstract

Prostate cancer is the most frequently diagnosed form of cancer in men over fifty. New markers for diagnostic and prognostic purposes are needed to, if necessary, initiate treatment in an earlier stage and to optimize treatment for the individual patient. In this manuscript we present results from mass spectrometric peptide profiling experiments aimed at discovering potential markers for prostate cancer. MALDI-TOF profiling experiments were performed on tryptic digests of serum samples (obtained by the European Randomized study of Screening for Prostate Cancer) of prostate cancer patients with metastases (n=27) and controls (n=30) after purification with surface-active magnetic beads. This resulted in the detection of eight repeatedly observed differentially expressed peptides, which were then identified by nanoLC-MALDI-TOF/TOF and confirmed by MALDI-FTMS exact mass measurements. All differentially expressed peptides are derived from two homologous parts of human serum albumin; two of the eight peptides were tryptic and six non-tryptic. The presence of the non-tryptic fragments indicates that a proteolysis is not mediated by trypsin. Since the non-tryptic fragments were found at significantly higher levels in control samples compared to metastases samples, it is hypothesized that a specific inhibition of the proteolytic process is in effect in the serum of prostate cancer patients. Experiments using synthetic peptides showed that this proteolytic activity occurs ex vivo and is sequence specific. Importantly, the observed prostate carcinoma related inhibition of the proteolysis was reproduced ex vivo using synthetic peptides.

8.1 Introduction

Prostate cancer is the most commonly diagnosed cancer in men over the age of 50 (1, 2). The current methods for diagnosing prostate cancer are screening for elevated prostate-specific antigen (PSA) levels, digital rectal examination, transrectal ultrasound imaging and needle biopsy of the prostate. The diagnosis of prostate cancer on the basis of PSA levels alone lacks specificity; in combination with other tests sensitivity and specificity increases but still remain limited and variable (3). In a variety of other conditions including prostatitis, benign prostate hyperplasia and non-cancerous neoplasia also elevated PSA levels are observed. The PSA concentration in serum is used as a trigger to perform further diagnostic tests. The concentration in serum that is used to initiate these further diagnostics varies between 1-4 ng/ml and is subject to debate. In early detection programs like the European Randomized study of Screening for Prostate Cancer (ERSPC) study a serum concentration of 3 ng/ml is used (4, 5). In this study 75 % of the patients with a PSA level higher than 3 ng/ml are false positive. Evidently, there is a need for additional biomarkers with higher specificity than PSA for the detection of prostate cancer. These new markers could be used in combination with PSA levels to reduce the number of false positive diagnosed patients. Also, prognostic markers are important to distinguish patients that will develop a more aggressive and metastatic form of prostate cancer from patients with clinically insignificant prostate cancer. In case of clinically insignificant prostate cancer the progression of the disease can be monitored and no further treatment is required (active surveillance). The number of patients that undergo an unnecessary prostatectomy that can result in incontinence or impotence can thereby be decreased. We selected for this study two sample groups, a control group of men with a low PSA value also during follow-up and a group of men with metastatic prostate cancer and a high PSA level. These two groups allow also the finding of markers that relate to metastasis, which could be of prognostic use.

Currently it is possible to perform protein or peptide profiling on large sample sets using different forms of mass spectrometry. Several studies have been published in which disease-specific protein profiles are detected with high sensitivity and specificity (6-10). Most of the identified differentially expressed proteins are high abundant serum proteins or fragments thereof that are related to stress, infection, hormonal modulation and acute phase reactions (11). The possible clinical applicability of such proteins or break-down products of proteins has been discussed (12). In addition, some recent studies have shown that protease activities can be highly disease-specific for different types of cancers (13). In a previous study we have shown that tryptic peptide profiling is a powerful technique to detect differentially expressed peptides (14). In our current study we describe the combination of tryptic peptide profiling of serum proteins with a fully automated magnetic bead purification procedure (15). Using this technology we identified a novel prostate carcinoma-related inhibition of proteolyses in serum that can be monitored by experiments using specific synthetic peptides.

8.2 Methods

8.2.1 Samples

All samples were obtained via the ERSPC biobank (Erasmus MC) and are collected under uniform conditions. In total 57 serum samples were used, 30 control samples and 27 samples of prostate cancer patients with metastasis and a PSA level > 4 ng/ml. The obtained serum samples were thawed on ice once, aliquoted and immediately refrozen.

8.2.2 Depletion

Depletion was performed on an immunoaffinity column (4.6 x 50 mm: Agilent, Santa Clara, USA) according to the recommendations of the manufacturer. Briefly, 25 µL serum was diluted to 125 µL with loading buffer and spin-filtered (0.22µm) for 20 min at 13,000 rpm in an Eppendorf centrifuge at 4°C. Seventy-five µL of each sample was loaded onto the column using an autosampler cooled to 4°C. Depletion was performed at room temperature on an AKTA FPLC system (GE-Healthcare, Chalfont St. Giles, UK) using the following program: 10 minutes at 100% eluent A at 0.25 mL/min; 3.5 minutes at 100% eluent B at 1 mL/min; 5.5 minutes at 100% eluent A at 1 mL/min. Fraction collection was started automatically when a threshold of 25 mAU at 280 nm was exceeded and the collected fractions were snap frozen in liquid nitrogen within 5 minutes after collection and stored at –80°C (16).

8.2.3 Serum sample profiling

Pre-fractionation of tryptic digests of the full and depleted serum samples were performed using hydrophobic interaction MB-HIC 18 ClinProt magnetic beads, (Bruker Daltonik, Leipzig, Germany). Immediately before tryptic digestion, samples were thawed to assure that each sample was treated equally in terms of freeze-and-thaw cycles and incubation/storing time at room temperature. Six µl of each serum sample was diluted 10 times in milliQ to obtain a total volume of 60 µl. To each sample 6 µl of 1% Rapigest (Waters, Milford, USA) dissolved in 50 mM ammoniumbicarbonate was added. Samples were incubated for 2 minutes at 37°C and subsequently 7.0 µl of 0.1 µg/µl gold grade trypsin (Promega, Madison, USA) in 3 mM Tris-HCL were added and incubated overnight at 37°C. After overnight incubation, 8 µl of 500 mM HCl were added in order to obtain a final concentration of 50 mM HCl (pH<2), and incubated further for 45 minutes at 37°C to hydrolyze the Rapigest. After this preparation procedure, samples were used for magnetic bead fractionation.

All magnetic bead preparations were performed according to the manufacturer's instructions. Beads and sample were incubated in binding buffer for several minutes at room temperature. The beads were separated from the supernatant using a magnetic separation device. After washing, the bound peptides were eluted by 0.1% TFA/50% ACN/50% water.

The eluted peptides were spotted in four-fold onto an MTP AnchorChip 600/384 target plate using α-cyano-4-hydroxycinnamic acid (HCCA) (Bruker Daltonik, Bremen, Germany) as matrix according to the manufacturer's recommendations. MALDI-TOF mass spectra were obtained in the automated AutoXecute mode using an Ultraflex TOF/TOF instrument equipped with a nitrogen laser (Bruker Daltonik, Bremen, Germany), operated in reflectron mode in the mass range of 800-4000 Da.

For all samples, the magnetic bead-based sample fractionation and MALDI-TOF target preparation were performed on a fully automated robotic platform (ClinProt Robot, Bruker Daltonik, Leipzig, Germany). This was done to enable high throughput and to assure optimal reproducibility. The whole experiment was performed in threefold with intervals of at least one month allowing a reliable assessment of the reproducibility and to only select differentially expressed peaks that are robust over time (present in 2 out of 3 separate experiments).

The magnetic bead fractions of two samples (control and metastasis) were re-measured by MALDI FTMS. In addition for these two samples also the depleted serum and albumin fraction of the depletion was measured. One µl of eluted sample was spotted onto an anchorchip target plate (600/384 anchorchip with transponder plate; Bruker Daltonik GmbH, Bremen, Germany). Before the spots were dried, one µl of 2,5-dihydroxy-benzoic acid (DHB) matrix (Bruker Daltonics GmbH, Bremen, Germany), 10 mg/ml in 0.1% TFA water was added and the spots were allowed to dry at ambient temperature. All spots were measured with an Apex Q 9.4 Tesla MALDI FTMS equipped with a combi-source (Bruker Daltonics, Billerica, USA). For each measurement 100 scans were summed up and for each scan ions generated by 10 laser shots were accumulated.

8.2.4 Reproducibility

Intra- and inter-experimental reproducibility of the peak intensity was calculated. using the ClinProtools software package 2.0 build 365 (Bruker Daltonics, Bremen, Germany). This software tool was used for alignment, normalization and peak detection. A signal to noise level of >2 was used for peak picking, the peak list generated in this way contained an average peak intensity and standard deviation for each peak position. The average intensities and standard deviations were used to calculate the coefficient of variance (CV) for each individual peak position for ten samples (five controls, five metastases). The intra-experimental CV was calculated for all masses present in all four replicate measurements within one experiment. The inter-experimental CV was calculated based on the peak positions present in the twelve measurements of one sample over the three experiments.

8.2.5 Statistical analysis of differences in peptide profiles

The differences in intensities of peptide masses in MALDI-TOF mass spectra of patients and a control group were compared. The differences were statistically analyzed with a database application developed by Titulaer et al. (17). A new version of the database application was used, which was adapted for peak picking of masses above a signal to noise threshold of 4, as described in Horn et al. (18). Spectra from the three MS experiments performed with a time intervals of one month were compared. Each experiment contained four replicate spectra of 27 samples of prostate cancer patients and 30 samples of a control group. From the, in total, four replicate spectra of each sample, three where randomly used to create a profile matrix of the mean intensity of each peptide mass for each sample (the same exclusion criteria were used as previously described (14)). Each spectrum was internally calibrated based on at least four masses present of in five intense albumin-derived peptides (masses: 1296.7045, 1511.8427, 1623.7875, 1639.9377 and 2045.0953 Da, denoted as the p_mark calibration list). Spectra with less then four of the five albumin masses were rejected and not used in the further analyses. Peptide masses present in at least four spectra of all samples were clustered within a mass window of 0.5 Da. If a mass was not present in a certain spectrum, the background intensity at that point was taken for correct statistical comparison. The Wilcoxon-Mann-Whitney test was performed on each peptide mass in the matrix, comparing the metastatic prostate cancer group and control group of experiment 1, the metastatic prostate cancer and the control group of experiment 2, the prostate cancer and control group of experiment 3, respectively. In a second statistical comparison all spectra of experiment 1, 2 and 3 were combined, which gave twelve replicate spectra for each sample. From these replicates a number of eight spectra where combined to create a profile matrix of the mean intensity of each peptide mass for each sample (which allows a maximum of four rejected spectra per sample, if less than four spectra were rejected eight spectra were selected at random). Again peptide masses present in at least 4 spectra of all samples were clustered within a mass window of 0.5 Da. All differentially expressed peaks with a p-value < 0.01 (Wilcoxon-Mann-Whitney test) in at least two of the three experiments and present with a p-value < 0.01 in the combined experiment were used as candidates for identification.

8.2.6 MALDI FTMS data analyses

All MALDI-FTMS spectra were analyzed by Data Analyses (Bruker Daltonics, USA version 3.4 build 169) and each spectrum was internally calibrated with respect to the most intense albumin peaks in the sample (960.5631, 1000.6043, 1149.6156, 1511.8427, 2045.0953 m/z). Subsequently, peak picking was performed with the SNAP 2 algorithm with a signal to noise threshold of 4 and a quality factor of 0.9.

8.2.7 Identification by NanoLC MALDI-TOF/TOF

Fractionation was performed using a monolithic column (200 µm i.d. Dionex, Sunnyvale, CA, USA) on a nanoscale liquid chromatography system (nanoLC) (Dionex, Sunnyvale, CA, USA). The samples in which the differentially expressed peaks had the highest intensity were pooled for identification. From this sample, 5 µl was loaded onto a trap column (250 µm i.d. x 5 mm, Dionex, Sunnyvale, CA, USA). Fractionation was performed using a 50 minute gradient from 0% to 64% of acetonitrile, (solution A (100% H_2O, 0.05% TFA) and solution B (80% ACN, 20% H_2O and 0.04% TFA); 0 to 3 min, 0% solution B, 35 min 45%, 35.1 min 80%, 38 min 80%, 38-50 min 0% with a flow of 2 µl/min. Ten second fractions were spotted automatically onto a commercially available prespotted MALDI plate containing 384 spots (Bruker Daltonics, USA) of α-cyano-4-hydroxycinnamic acid (HCCA) matrix, using a robotic system (Probot Micro Fraction Collector, Dionex, Sunnyvale, CA, USA). To each fraction, 0.33 µl water was added. Finally, we used a 10 mM $(NH)_4H_2PO_4$ in 0.1% TFA/water solution to wash the pre-spotted plate for 5 seconds to remove salts. The plate was subsequently measured by automated MALDI-TOF/TOF-MS (Ultraflex, Bruker Daltonics, Germany) using WARP-LC software (Bruker Daltonics, Germany) which obtained MS spectra of each individual spot and subsequently performed MS/MS measurement for selected precursor masses. The precursor masses for performing MS/MS measurements were determined automatically by the WARLP-LC software based on signal intensity and presence of interfering peaks. The MS/MS data resulting from the WARP-LC measurements were searched against the Swiss Prot database using the Mascot search engine with a 100 ppm mass tolerance for the parent ion and a 0.6 Da mass tolerance for fragments. All identified differentially expressed peptide peaks were confirmed by comparing the accurate mass of the MALDI-FTMS measurement with the calculated mass for the identified sequence within a mass window of 2 ppm. For some peaks it was not necessary to perform an LC separation and these peaks have been identified by direct MS/MS measurements using the same equipment as described above.

8.2.8 Proteolytic assays using synthetic peptides

Synthetic peptides with the sequences listed in table 8.1 and having a purity of >98% were purchased from Pepscan, Lelystad, The Netherlands. Control serum samples undiluted, 10-, 100- and 1000-fold diluted were used, respectively. One µl of the diluted serum sample was incubated with 10 µl of 10 pmol/µl synthetic peptide in water for 30 minutes at 37˚C and subsequently 0.5 µl was spotted with 0.5 µl of DHB on an anchorchip plate and measured by MALDI-FTMS. In addition, a time series was performed: 0.5 min, 1 min, 2 min, 5 min, 10 min, 20 min, 40 min, 80 min, 120 min and overnight incubation at 37˚ C, respectively. For these experiments 100-fold diluted serum was used to obtain a similar serum concentration as in the profiling experiment. These samples were also spotted and measured by MALDI-FTMS. Also, negative controls were included in which the peptides were replaced by water or in which the serum was replaced by water. Control samples (n=5) and prostate cancer with metastasis samples (n=5) were diluted 100-fold and subsequently 1 µl of the diluted serum was added to 10 µl of 10 pmol/µl of the synthetic peptide (RHPYFYAPELLFFAK) followed by incubation for two, four and fifteen hours. The resulting mixtures were measured by MALDI-FTMS. In an additional experiment, 50 µl of the 100-fold diluted serum samples were first incubated for two hours with 50 µl of 0.1 % Rapigest (Waters, USA) dissolved in 50 mM ammoniumbicarbonate and 10 µl of 0.1 µg/µl gold grade trypsin in 3 mM Tris HCL. Subsequently, one µl of this solution was added to 10 µl of 10 pmol/µl of the synthetic peptide

RHPYFYAPELLFFAK and measured by MALDI-FTMS. This experiment was performed independent in three fold. In addition, similar experiments were performed in which the incubation time of the serum with trypsin was increased to four hours. The linear relation between the amount of sample and the intensities in MALDI-FTMS measurements was ascertained by measuring a dilution series of peptide RHPDYSVVLLLR (1467.8430 Da) in the concentrations of 0.5, 1, 2, 5 and 10 pmol/µl spiked with 2 pmol/µl of peptide RHPYFYAPELLFFAK (1898.9951 Da) in the presence of a 100-fold diluted serum sample. MALDI-FTMS measurements have been performed as described above.

Table 8.1: Synthetic peptides purchased at >98% purity

Sequence	Description	Mass MH+
RHPYFYAPELLFFAK	Sequence 1	1898.9951
RHPYFYAPELLFFAKRYKA	Extended C-terminus sequence 1	2417.2917
ARRHPYFYAPELLFFAK	Extended N-terminus sequence 1	2126.1334
KAFFLLEPAYFYPHR	Sequence 1 reverse	1898.9951
RHPDYSVVLLLR	Sequence 2	1467.8430
ARRHPDYSVVLLLR	Extended C-terminus sequence 2	1694.9812
RHPDYSVVLLLRLAKT	Extended N-terminus sequence 2	1881.1068
RLLLVVSYDPHR	Sequence 2 reverse	1467.8430

8.3 Results

8.3.1 Peptide profiling

A set of 57 tryptically digested serum samples (control n=30, prostate carcinoma with metastasis n=27) were used for a tryptic peptide profiling experiment and the resulting spectra were analyzed and statistically compared. The reproducibility of peak intensity, intra- and inter-experimental, were calculated, and the average CV's for the peptides present in all spectra were 30% and 45%, respectively. In total 2410 possible peak positions were detected in the analyses of the three combined experiments. Of these 2410 peak positions 94 were significantly differentially expressed with $p<0.01$ as determined by the Wilcoxon-Mann-Whitney test. In Figure 8.1, the number of peaks for each p-value interval is presented. On the same data, we performed a cross validation by randomly assigning a group number to each serum sample and then repeating the Wilcoxon-Mann-Whitney test. This scrambling procedure was repeated 1,000 times. The average frequency of possible background peaks per p-value interval is presented by the red line in Figure 8.1. The relative flat distribution of the p-value histogram indicates that there is no correlation between peak positions and groups after scrambling. The p-value histogram of the actual experiment is clearly skewed to lower p-values and is significantly different from the histogram after scrambling (Figure 8.1). All experiments were also analyzed separately and all significant peaks from these analyses were combined; peaks that were present in at least two experiments were extracted from the list. It was found that 22 differentially expressed peak positions were present in at least two experiments. If presence in all three experiments was set as a condition, this number drops to 18. We used only peaks that were present at significantly differential levels in two of three for the identifications experiments to minimize background and unreliable differences. The 22 peak positions corresponded to 8 mono-isotopic masses (Table 8.2). For mass 1249.6 Da a graphical presentation is shown in Figure 8.2 indicating a significant difference in intensity between the control and the metastatic cancer samples.

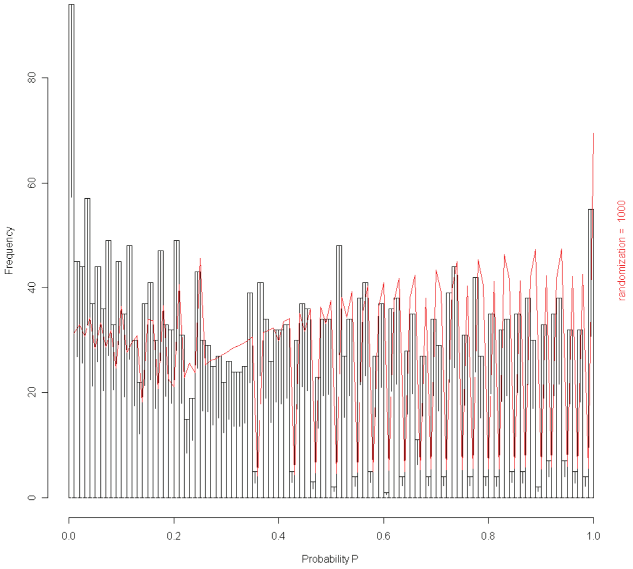


Figure 8.1: A significant number of peptides are differentially expressed between prostate cancer patients with metastasis and control patients The figure shows a histogram of p-values where the height of each bar denotes the number of peptide peaks while the horizontal base corresponds to the p-value interval (interval size 0.01), the part of the bar in white represent the number of peaks with a higher intensity in the control and the part in black the peaks with a higher intensity in the prostate cancer with metastasis. The red line represents the histogram of p-values after cross validation. The height of the red line shows the average number of peptide peaks after 1,000 scrambling procedures. The distribution is clearly different from the random distribution (red line) and skewed to the left indicating a high number of peptides that discriminate between the two groups (low p-value). See also color figures (page 166).

8.3.2 Identification of differentially expressed peptides

The 8 mono-isotopic peaks that were significantly differently expressed in at least 2 experiments were identified by MALDI-TOF/TOF-MS and nanoLC MALDI-TOF/TOF-MS experiments and identifications were confirmed by performing exact mass measurements of the parent ion with MALDI-FTMS (Table 8.3) (19). Surprisingly, this resulted in the identification of eight peptides from human serum albumin (HSA), two of them being tryptic and six semi-tryptic (N-terminal tryptic and C-terminal non-tryptic). The semi-tryptic peptides were derived from two specific amino acid sequences (Figure 8.3a & b). We also checked the MALDI-FTMS mass spectra for additional HSA fragments also derived from these two amino acid sequences; if masses matched within 2 ppm they were included in Figure 8.3. Also for most of these peptides MS/MS data were obtained as indicated in Table 8.3. All selected fragments in Figure 8.3 have significant p-values (p<0.01) in the Wilcoxon-Mann-Whitney test of the combined data. The two tryptic fragments show an overlap in sequence as can be seen from Figure 8.3c. The average peak intensity and standard deviations of the identified peptides are displayed in Table 8.3. In Figure 8.4 the patient/control ratios for the average intensities of the identified peptides are plotted. In this figure one can observe that the intensities of the tryptic fragments are higher in prostate carcinoma patients with metastasis compared to the controls. Conversely, the semi-tryptic fragments are significantly higher in the control samples. In the spectra acquired from both fractions of the depleted samples these specific albumin fragments were not present and the identified peptides were also not observed as naturally occurring peptides in full serum.

Table 8.2: Results of statistical analyses

Masses	p-values	number of independent experiments with p-value <0.01	Protein	sequence	up/down regulation in metastasis
1155.592	1.16E-04	3	Human serum albumin	HPDYSVVLLL	down
1156.588	3.82E-04	3			
1249.603	4.96E-05	2	Human serum albumin	HPYFYAPELL	down
1250.614	1.16E-04	3			
1293.628	8.07E-04	2	Human serum albumin	HPYFYAPEL	down
1405.715	8.75E-05	3	Human serum albumin	RHPYFYAPELL	down
1406.721	1.01E-04	3			
1407.727	1.16E-04	3			
1408.728	2.59E-04	3			
1552.796	0.001842	3	Human serum albumin	RHPYFYAPELLF	down
1553.800	0.001463	3			
1554.825	0.002063	2			
1770.910	0.005471	3	Human serum albumin	RHPYFYAPELLFFA	down
1771.916	0.007431	3			
1772.926	0.009998	2			
1899.006	3.36E-04	3	Human serum albumin	RHPYFYAPELLFFAK	up
1900.009	7.14E-04	3			
1901.009	7.14E-04	3			
1902.009	0.002581	3			
2055.097	4.34E-04	3	Human serum albumin	RHPYFYAPELLFFAKR	up
2056.091	5.58E-04	3			
2057.106	0.001463	3			

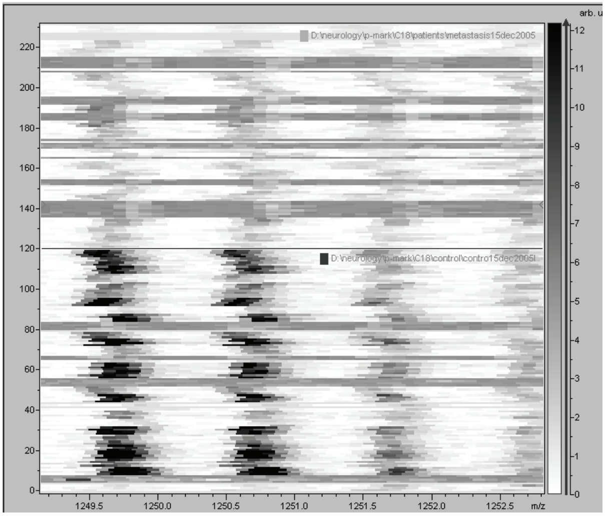


Figure 8.2: A clear difference in intensity between control samples and prostate cancer patients with metastasis for mass 1249 Da. A zoom in of mass 1249 Da in a gel view representation of all measured spectra.

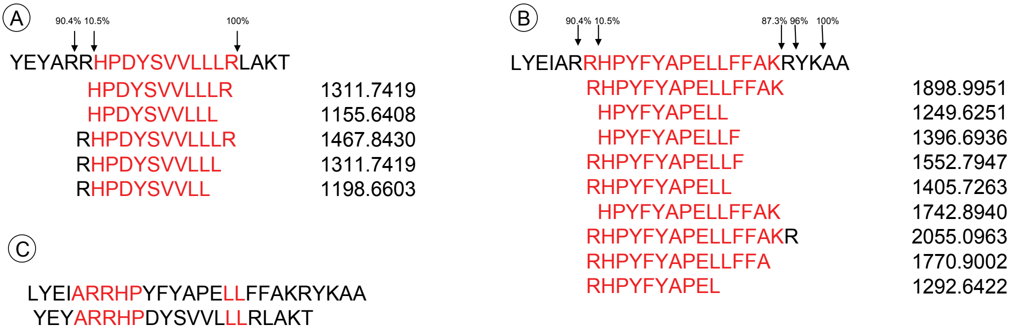
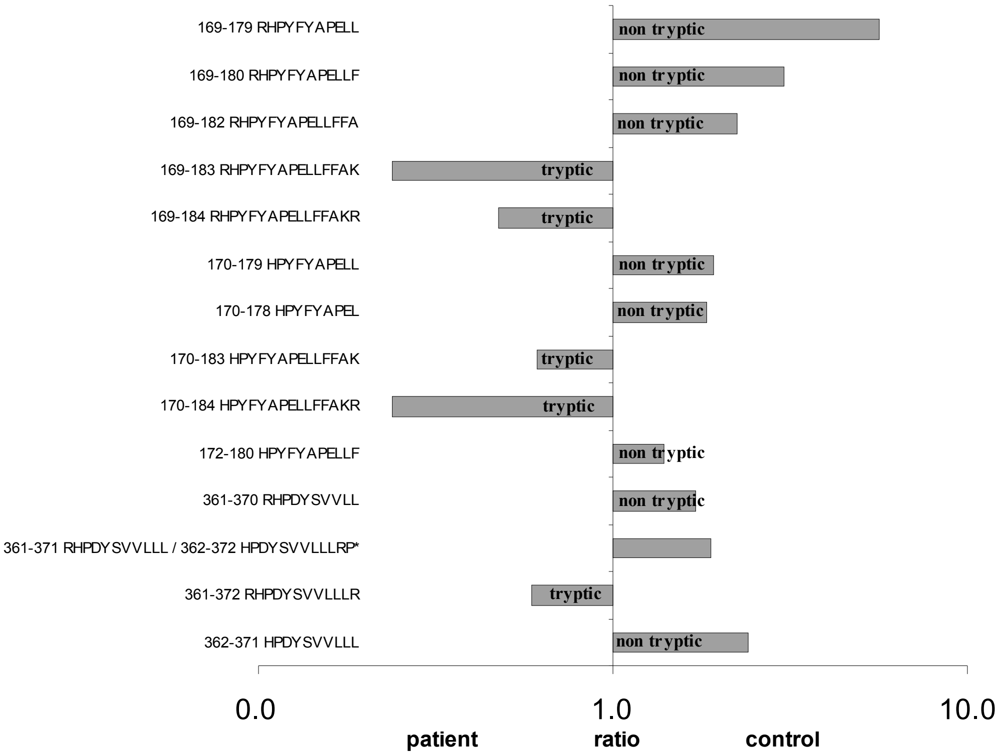


Figure 8.3: Identified differentially expressed tryptic peptides. In panel a and b two amino acid sequences that are part of the HSA sequence are shown. The arrows indicate cleavage sites for albumin and the percentage above the arrow the specificity of cleavages. All identified peptides that derive from those sequences are shown with their calculated masses (MH$^+$). In panel c we show the homology between the two amino acid sequences.

Table 8.3: Information of identifications and peptides

observed mass MALDI-TOF	observed mass MALDI-FTMS	calculated mass	mass accuracy (ppm)	MS/MS	miss cleavages	sequence	peak intensity control	stdev control	peak intensity patients	stdev patient	ratio control/patient
1405.64	1405.7257	1405.7263	-0.4	*	1	169-179 RHPYFYAPELL	5968	5358	1055	768	5.7
1552.70	1552.7943	1552.7947	-0.3	*	1	169-180 RHPYFYAPELLF	6286	5423	2076	2303	3.0
1770.79	1770.9049	1770.9002	2.7	*	1	169-182 RHPYFYAPELLFFA	6091	4400	2715	1328	2.2
1898.88	1898.9944	1898.9952	-0.4	*	1	169-183 RHPYFYAPELLFFAK	3212	5075	13530	16217	0.2
2055.96	2055.0919	2055.0963	-2.1	*	1	169-184 RHPYFYAPELLFFAKR	679	359	1426	852	0.5
1249.57	1249.6244	1249.6251	-0.6	*	0	170-179 HPYFYAPELL	1118	444	579	207	1.9
1292.60	1292.642	1292.6422	-0.2	*	1	169-178 RHPYFYAPEL	1118	574	607	211	1.8
1742.79	1742.8932	1742.894	-0.5	*	0	170-183 HPYFYAPELLFFAK	739	340	1207	780	0.6
1898.88	1898.9944	1898.9952	-0.4	*	1	170-184 HPYFYAPELLFFAKR	3212	5075	13530	16217	0.2
1396.66	1396.6938	1396.6936	0.1	na	0	172-180 HPYFYAPELLF	828	293	592	209	1.4
1198.59	1198.6591	1198.6603	-1.0	*	1	361-370 RHPDYSVVLL	1346	1311	788	752	1.7
1311.66	1311.7419	1311.7419	0.0	*	1	361-371 RHPDYSVVLLL	17947	14679	9490	5886	1.9
1467.74	1467.8435	1467.8431	0.3	*	1	361-372 RHPDYSVVLLLR	3612	3190	6119	4544	0.6
1155.58	1155.6405	1155.6408	-0.3	*	0	362-371 HPDYSVVLLL	1504	993	627	205	2.4
1311.66	1311.7419	1311.7419	0.0	*	0	362-372 HPDYSVVLLLR	17947	14679	9490	5886	1.9



*Peptides have same mass and cannot be distinguished

Figure 8.4: A difference in ratio (control/patients) of peptide intensities between tryptic and semi-tryptic fragments of albumin. On the x-axis the ratio (control/patients) of peak intensities is shown on a logarithmic scale. On the y-axis the sequence of the peptide is indicated.

8.3.3 Synthetic peptides assay shows proteolytic effect in serum

Synthetic peptides were purchased with the same sequence as the two HSA sequences that show proteolysis (RHPDYSVVLLLR and RHPYFYAPELLFFAK) in the control serum samples. In addition, two peptide sequences with an extended C-terminus and an extended N-terminus of the aforementioned peptides were purchased to determine which part of the sequence is important to observe this proteolytic activity. For background assessment, peptides with the reverse sequence were also purchased (Table 8.1). In an initial experiment 10 pmol of each peptide was incubated with a 100 times diluted control serum sample for two hours. For both peptides proteolysis is observed to 13 % and 54.5 % of the original peptide, respectively. The observed proteolysis pattern is similar to that of the profiling experiment (Figure 8.5). In the negative control experiment no proteolysis of the synthetic peptide was observed. For the reverse sequence of the peptide only a small amount of peptide is degraded < 5% (this cannot be determined more precisely because of impurities in the synthetic peptides with the same exact mass as the expected fragments). For the peptides with the extended C-terminus, a similar low amount of proteolysis is observed. Peptides with the extended N-terminus showed a clear increase in proteolysis to 23% and 93%, respectively. For the peptide with sequence RHPYFYAPELLFFAK a time series and serum dilution series were performed. With an increasing serum concentration; a higher percentage of proteolysis was observed. For a 1000-, 100-, 10- and 0-fold dilution of the serum, the percentage of the degraded peptide is 1.5, 19.2, 92.2 and 95 %, respectively, after half an hour incubation at 37 ˚C. For the time series a similar effect is observed, i.e. more proteolysis at longer incubation times (Figure 8.6). Note that the proteolytic products are not present in the purchased synthetic peptide nor are they formed by trypsin. The measurements of a dilution series of peptide RHPDYSVVLLLR (1467.8430 Da) spiked with RHPYFYAPELLFFAK (1898.9951 Da) in the presence of serum resulted in a linear correlation between the concentration and the relative intensity of peptide RHPDYSVVLLLR (R^2=0.993).

Between control serum samples (n=5) and serum samples of patients with prostate cancer and metastasis (n=5) no significant difference in the proteolysis of peptide RHPYFYAPELLFFAK is observed (p=0.775 Wilcoxon-Mann-Whitney) when 100-fold diluted serum is incubated for two, four and fifteen hours. To exactly replicate the experimental conditions of the serum profiling experiment we first incubated the 100-fold diluted serum samples with a trypsin solution for two hours and subsequently added the synthetic peptide followed by incubation for another two hours. In this setup, a significant difference in ratio between the synthetic peptide and the proteolysis products was observed (p=0.005 Wilcoxon-Mann-Whitney) between prostate cancer and control samples (Figure 8.7). The experiment was performed in three-fold resulting in similar differences in ratio for control serum samples and prostate cancer metastases serum samples. If the serum samples were incubated with trypsin solution for four hours prior to the addition of the synthetic peptide no proteolysis was observed anymore. Incubation of the synthetic peptides with serum for two, four and fifteen hours after two hours pre-incubation with trypsin, were evaluated. For two and four hours, similar results were obtained but after an incubation period of 15 hours the difference between controls and prostate cancer samples was not longer significant.

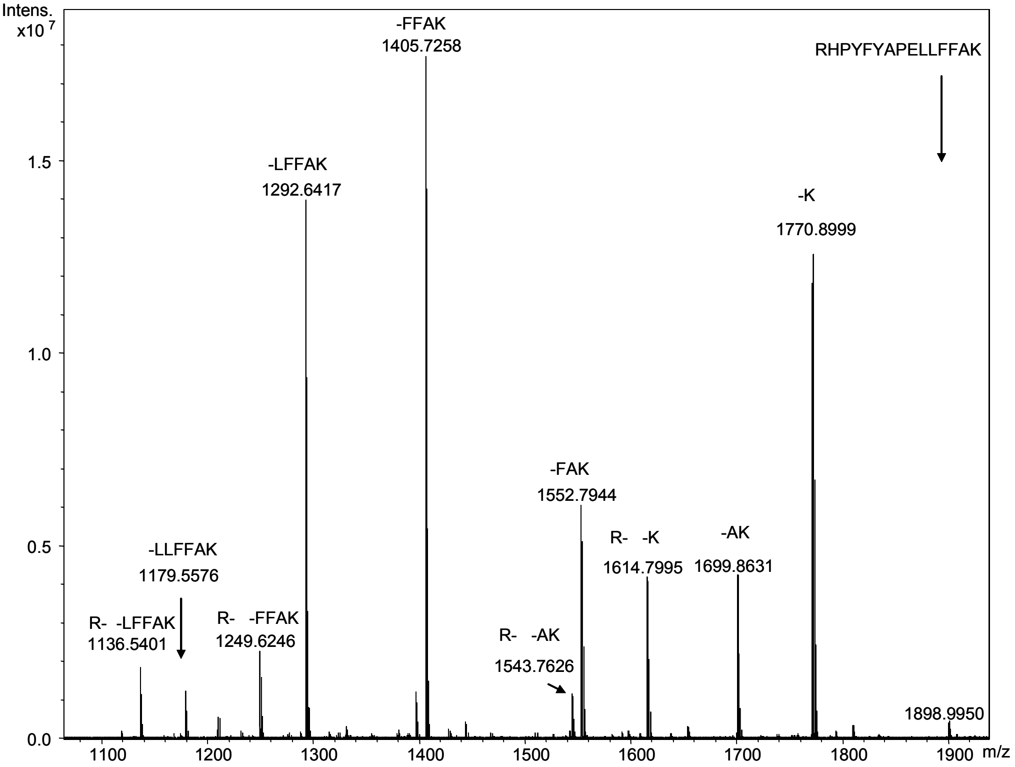


Figure 8.5: Proteolytic degradation of synthetic peptide RHPYFYAPELLFFAK. The synthetic peptide is incubated overnight at 37 ˚C with 100 fold diluted serum and subsequently measured by MALDI-FTMS. The resulting fragmentation pattern is similar to the fragmentation observed in the profiling experiment.

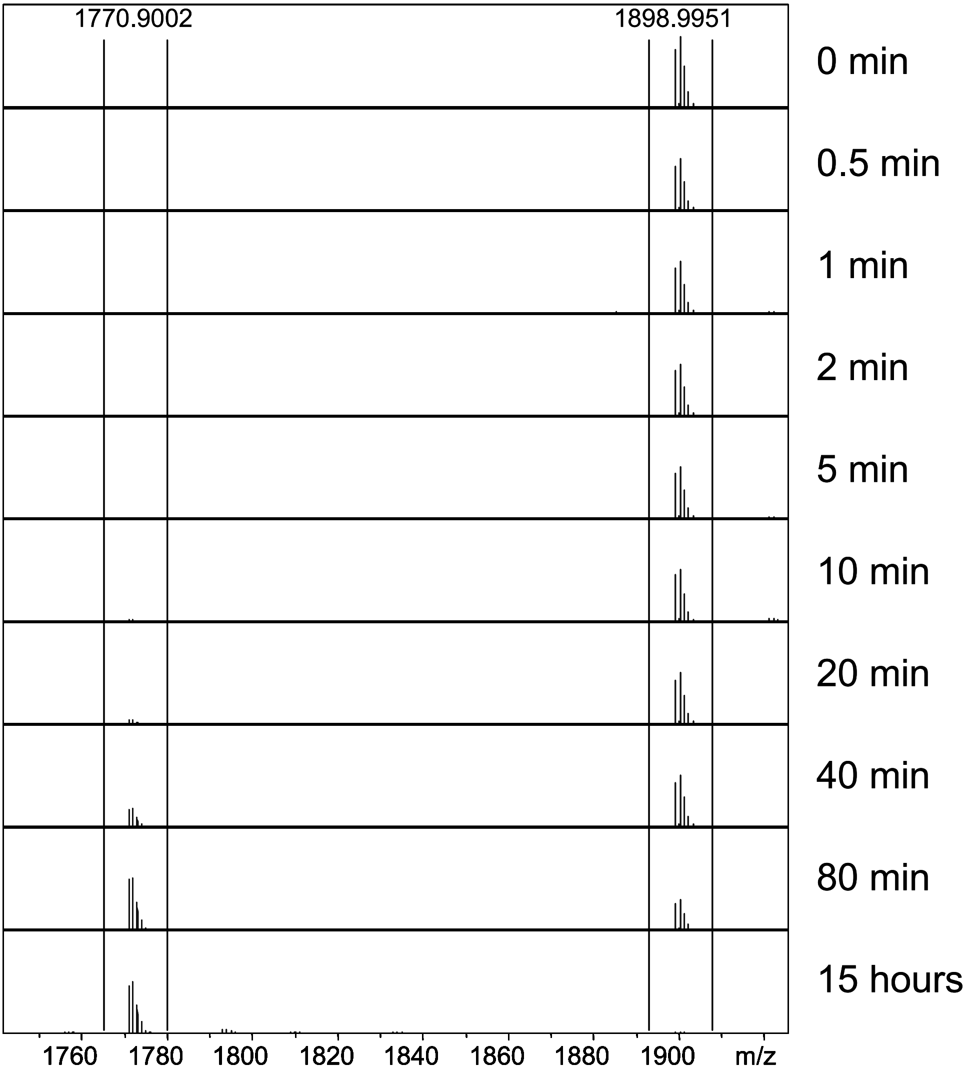


Figure 8.6: Proteolytic degradation of synthetic peptide RHPYFYAPELLFFAK is time dependent. The synthetic peptide is incubated with 100 fold diluted serum at 37˚ C and subsequently measured by MALDI-FTMS at different time points. In time a clear decrease of mass 1898 Da (RHPYFYAPELLFFAK) and a clear increase of mass 1770 Da (RHPYFYAPELLFFA, one of the fragments) can be observed in the mass spectra.

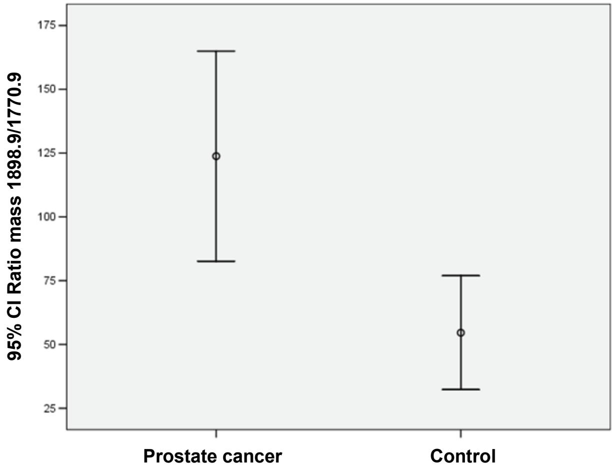


Figure 8.7: Differential proteolysis of synthetic peptide observed between control serum samples and prostate cancer serum samples. In this plot the confidence interval of the ratio between the synthetic peptide with mass 1898.9 Da and the proteolytic product with mass 1770.9 for five tryptic digests of control serum samples and five tryptic digest of prostate cancer serum samples incubated with synthetic peptide (RHPYFYAPELLFFAK) for two hours is shown.

8.4 Discussion

The aim of this study was to find additional markers for prostate cancer. The clinical serum marker commonly used for the detection of prostate cancer is PSA. The diagnosis of prostate cancer on basis of PSA levels lacks specificity (3). Additional markers could decrease the number of false positive diagnoses. In addition the performed experiments are also aimed at the discovery of prognostic markers. The prognostic capabilities of PSA in low concentration range <4 ng/ml are limited. If prognostic markers were available the decision to perform a prostatectomy could be based on the aggressiveness of the cancer. This could prevent men with clinically insignificant prostate cancer from undergoing an unnecessary prostatectomy with all the possible complications.

Mass spectrometric profiling and identification of body fluid proteins can assist in discovering novel biomarkers for diagnoses and understanding the pathogenesis of a disease. Body fluids like serum are easily accessible and only a limited amount is necessary for profiling and identification. Blood samples are already routinely taken from patients and controls in a uniform way, resulting in large archives of samples with follow-up. In this study we performed magnetic bead sample preparation on tryptic digests of serum samples followed by a MALDI-TOF screening. This resulted in a number of significantly differentially expressed peptide peaks. The number of differentially expressed peaks is higher than when a randomized analysis of the data is performed. The relatively poor reproducibility of peak intensities in MALDI-TOF-MS has been acknowledged and discussed (20). In an earlier study we showed that to overcome this problem multiple technical replicate measurements per sample are necessary (21). In this study we did not only analyze multiple replicates per sample but we also repeated the overall experiment three times with intervals of one month to show that the differentially expressed peptides in this study are stable over time.

All identified differentially expressed peptides are tryptic and semi tryptic fragments of two homologous amino acid sequences in the HSA protein. In control serum we find mainly C-terminal non-tryptic proteolytic degraded fragments from these two HSA sequences, while in serum of prostate cancer patients with metastasis we find the intact tryptic fragments (Figure 8.4). A large number of abundant serum proteins are active in proteolysis or in the inhibition of proteolyses. This proteolytic process can occur both *in vivo* or *ex vivo* or in a multi step process in which the first steps occur *in vivo* and the final steps *ex vivo* as proposed by Villanueva et al. (11, 13). The large number of proteolytic enzymes and inhibitors present in blood and the proposed multi step proteolysis will result in vast numbers of possible fragments. Proteolysis of proteins and peptides is a common phenomenon especially during cancer metastasis (22-24). During metastasis proteolytic enzymes are released in the blood stream and can possibly also result in an increased secretion of “specific” inhibitors (25). The variation in enzymes and specific inhibitors makes it possible that disease-specific protein/peptide profiles can be observed by MS techniques. The formation or absence of such specific proteolytic fragments in cancer patients can possibly be used for diagnostic and prognostic tests and may give clues about the cancer-related proteases and inhibitors that play a role in metastasis. The fragmentation products that were found in this study have to our knowledge not been described previously. Possibly, the described observations do not occur *in vivo* since naturally occurring peptides in serum samples of prostate cancer and control patients have been studied (13, 26, 27), showing that the indicated peptide fragments are not originally present in serum. This can also be concluded from the results of the measurements of tryptic-digested, albumin-depleted serum samples (abundant and depleted fraction) in which none of the fragments were detected. The substrate of this enzymatic reaction is possibly created by tryptic digestion of HSA, explaining the absence of these fragments in untreated full serum.

The observed proteolytic fragmentation is specific for two homologous amino acid sequences in HSA. These two peptide sequences show a large overlap in sequence with kinetensin (IARRHPYFL), a signaling peptide (28). Kinetensin is a known substrate for the metalloprotease meprin B. However, it has a single cleavage site and does not show any exopeptidase activity (29) as observed in our experiments.

We tested the proteolytic activity in serum by the addition of synthetic peptides to serum with the same sequence as the two observed peptides derived from HSA. This resulted in a clear proteolysis pattern of the peptides, similar to the patterns observed in the profiling experiment. We were able to show that the activity of the proteolytic degradation process is dependent on the sequence. The extent of proteolysis is decreased or even absent when the peptide sequences are in reverse order or when the C-terminal part of the peptide is extended. If the N-terminal part of the homologous sequence ARRHP is present, the activity is increased indicating that the homologous part between both HSA peptides is of importance to observe this proteolytic activity. Also we showed that the proteolytic degradation activity was time- and serum concentration-dependent as expected for an enzymatic process. We hypothesized that an inhibitor of this specific proteolytic process is formed or activated during the tryptic digestion process. To test this hypothesis we incubated the serum with a trypsin solution prior to adding the synthetic peptide. This resulted in a significant decrease in proteolytic products in the prostate cancer serum samples, supporting the hypothesis that specific inhibitors of a proteolytic process are formed during the digestion process. In serum of prostate cancer patients this process is significantly different compared to control serum samples. The time of tryptic digestion of the serum is crucial to observe this effect. When the time is extended to 4 hours, no proteolytic degradation was observed, probably because the protease is completely digested by trypsin. If the time is shortened no inhibition is observed because the inhibitor is not yet formed.

In conclusion, this profiling experiment resulted in a number of peptides that are significantly differentially expressed between control serum samples and serum samples of prostate cancer patients with metastases. These differentially expressed peptides all relate to a proteolytic degradation process of two specific HSA sequences. We were able to reproduce this proteolytic process with an assay using synthetic peptides with the same sequences as the peptides from HSA. Also, we have shown that the proteolytic activity is related to the sequence of these peptides and that an inhibition of this process takes place in tryptic digests of prostate cancer serum samples. Peptide profiling on serum samples is biased towards the high abundant proteins. Low abundant but very active proteases and protease inhibitors are present in serum and can be expressed at significantly different levels in cancer patients compared to control patients. By using synthetic peptide assays in combination with mass spectrometry, these proteolytic enzymes and inhibitors can be detected and eventually identified. These measurements of the protease activity in serum might be useful as a diagnostic or prognostic tool.

References

1. Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., and Thun, M. J. "Cancer statistics, 2006," *CA Cancer J Clin* 56 (2006): 106-30.
2. Selley, S., Donovan, J., Faulkner, A., Coast, J., and Gillatt, D. "Diagnosis, management and screening of early localised prostate cancer," *Health Technol Assess* 1 (1997): i, 1-96.
3. Reynolds, M. A., Kastury, K., Groskopf, J., Schalken, J. A., and Rittenhouse, H. "Molecular markers for prostate cancer," *Cancer Lett* (2007).
4. Schroder, F. H., Roobol-Bouts, M., Vis, A. N., van der Kwast, T., and Kranse, R. "Prostate-specific antigen-based early detection of prostate cancer - Validation of screening without rectal examination," *Urology* 57 (2001): 83-90.
5. Schroder, F. H., Damhuis, R. A. M., Kirkels, W. J., DeKoning, H. J., Kranse, R., Nijs, H. G. T., and Blijenberg, B. G. "European randomized study of screening for prostate cancer - The Rotterdam pilot studies," *International Journal of Cancer* 65 (1996): 145-151.
6. Kozak, K. R., Su, F., Whitelegge, J. P., Faull, K., Reddy, S., and Farias-Eisner, R. "Characterization of serum biomarkers for detection of early stage ovarian cancer," *Proteomics* 5 (2005): 4589-96.
7. Sidransky, D., Irizarry, R., Califano, J. A., and Li, X. "Serum protein MALDI profiling to distinguish upper aerodigestive tract cancer patients from control subjects," *Journal of the national cancer institute* 95 (2003): 1711-1717.
8. Petricoin, E. F., 3rd, Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C. B., Levine, P. J., Linehan, W. M., Emmert-Buck, M. R., Steinberg, S. M., Kohn, E. C., and Liotta, L. A. "Serum proteomic patterns for detection of prostate cancer," *J Natl Cancer Inst* 94 (2002): 1576-8.
9. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet* 359 (2002): 572-7.
10. Villanueva, J., Martorella, A. J., Lawlor, K., Philip, J., Fleisher, M., Robbins, R. J., and Tempst, P. "Serum Peptidome Patterns That Distinguish Metastatic Thyroid Carcinoma from Cancer-free Controls Are Unbiased by Gender and Age," *Mol Cell Proteomics* 5 (2006): 1840-52.

11. Koomen, J. M., Shih, L. N., Coombes, K. R., Li, D., Xiao, L. C., Fidler, I. J., Abbruzzese, J. L., and Kobayashi, R. "Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins," *Clin Cancer Res* 11 (2005): 1110-8.

12. Liotta, L. A., Ferrari, M., and Petricoin, E. "Clinical proteomics: written in blood," *Nature* 425 (2003): 905.

13. Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A., Erdjument-Bromage, H., Olshen, A. B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., Sanchez-Carbayo, M., Holland, E. C., Cordon-Cardo, C., Scher, H. I., and Tempst, P. "Differential exoprotease activities confer tumor-specific serum peptidome patterns," *J Clin Invest* 116 (2006): 271-84.

14. Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M., and Sillevis Smitt, P. A. "MALDI-TOF Mass Spectrometry Analysis of Cerebrospinal Fluid Tryptic Peptide Profiles to Diagnose Leptomeningeal Metastases in Patients with Breast Cancer," *Mol Cell Proteomics* 4 (2005): 1341-1349.

15. Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J., and Tempst, P. "Correcting common errors in identifying cancer-specific serum peptide signatures," *J Proteome Res* 4 (2005): 1060-72.

16. Dekker, L. J., Bosman, J., Burgers, P. C., van Rijswijk, A., Freije, R., Luider, T., and Bischoff, R. "Depletion of high-abundance proteins from serum by immunoaffinity chromatography: A MALDI-FT-MS study," *J Chromatogr B Analyt Technol Biomed Life Sci* (2006).

17. Titulaer, M. K., Siccama, I., Dekker, L. J., van Rijswijk, A. L., Heeren, R. M., Sillevis Smitt, P. A., and Luider, T. M. "A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls," *BMC Bioinformatics* 7 (2006): 403.

18. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules," *J Am Soc Mass Spectrom* 11 (2000): 320-32.

19. Dekker, L. J., Burgers, P. C., Guzel, C., and Luider, T. M. "FTMS and TOF/TOF mass spectrometry in concert: Identifying peptides with high reliability using matrix prespotted MALDI target plates," *J Chromatogr B Analyt Technol Biomed Life Sci* (2006).

20. Baggerly, K. A., Morris, J. S., and Coombes, K. R. "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics* 20 (2004): 777-85.

21. Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M. "A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra," *Rapid Commun Mass Spectrom* 19 (2005): 865-870.

22. Koblinski, J. E., Ahram, M., and Sloane, B. F. "Unraveling the role of proteases in cancer," *Clin Chim Acta* 291 (2000): 113-35.

23. Kim, J., Yu, W., Kovalski, K., and Ossowski, L. "Requirement for specific proteases in cancer cell intravasation as revealed by a novel semiquantitative PCR-based assay," *Cell* 94 (1998): 353-62.

24. Chambers, A. F., and Matrisian, L. M. "Changing views of the role of matrix metalloproteinases in metastasis," *J Natl Cancer Inst* 89 (1997): 1260-70.

25. Sloane, B. F., Rozhin, J., Moin, K., Ziegler, G., Fong, D., and Muschel, R. J. "Cysteine endopeptidases and their inhibitors in malignant progression of rat embryo fibroblasts," *Biol Chem Hoppe Seyler* 373 (1992): 589-94.

26. Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C., and Tempst, P. "Serum Peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry," *Anal Chem* 76 (2004): 1560-70.

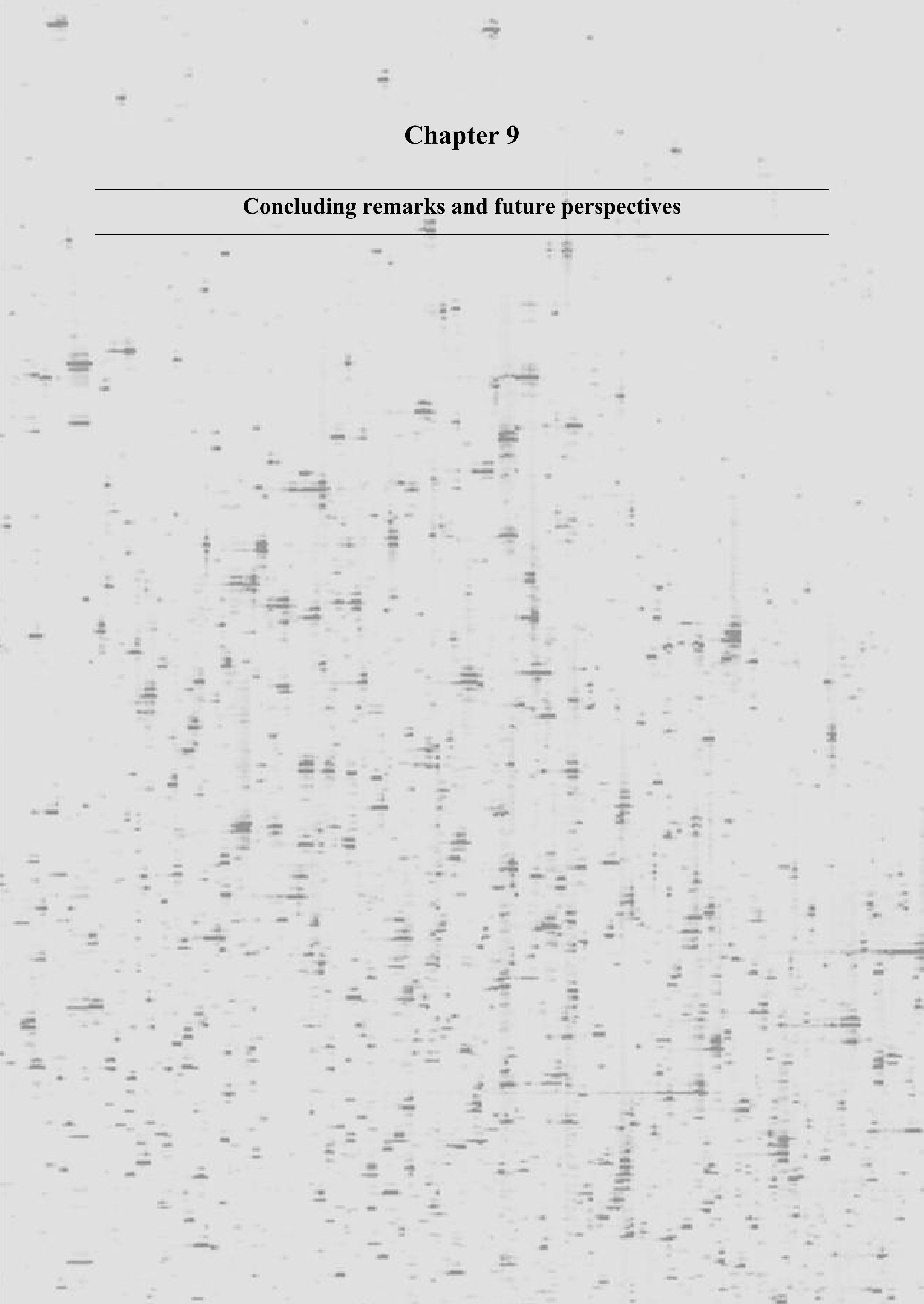
27. Koomen, J. M., Li, D., Xiao, L. C., Liu, T. C., Coombes, K. R., Abbruzzese, J., and Kobayashi, R. "Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery," *J Proteome Res* 4 (2005): 972-81.

28. Mogard, M. H., Kobayashi, R., Chen, C. F., Lee, T. D., Reeve, J. R., Jr., Shively, J. E., and Walsh, J. H. "The amino acid sequence of kinetensin, a novel peptide isolated from pepsin-treated human plasma: homology with human serum albumin, neurotensin and angiotensin," *Biochem Biophys Res Commun* 136 (1986): 983-8.

29. Bertenshaw, G. P., Turk, B. E., Hubbard, S. J., Matters, G. L., Bylander, J. E., Crisman, J. M., Cantley, L. C., and Bond, J. S. "Marked differences between metalloproteases meprin A and B in substrate and peptide bond specificity," *J Biol Chem* 276 (2001): 13248-55.

Chapter 9

Concluding remarks and future perspectives



9.1 Introduction

During the four years of this Erasmus MC Revolving fund (top-down) project, the proteomics field has developed and matured. The equipment currently available is superior in terms of robustness, sensitivity, mass accuracy and resolution compared to the equipment of four years ago (1, 2). Also the standards for performing experiments and analyzing data have been raised (3, 4). In this chapter we will discuss the developed methods, the biomarker discovery in leptomeningeal metastasis (LM) in breast cancer and in prostate cancer, new developments in biomarker research and the impact of these new developments on biomarker research in LM and prostate cancer.

9.2 Developed methods for peptide profiling

The detection of biomarkers in body fluids is challenged by the large biological and technical variation. Measures to overcome variations include the analysis of relatively large numbers of samples (> 100), pooling of samples and carefully controlling the experimental conditions. Major drawbacks of pooling samples include the loss of individual sample information resulting in less reliable quantitative information that may be biased towards an outlier. To measure large sample groups, a fast and reliable screening technique is required (5). We developed a technique that allows analysis of large numbers (>100) of samples. The combination of tryptic digestion and profiling with matrix-assisted laser desorption-ionization time of flight (MALDI-TOF) MS is a sensitive and rapid method that can be fully automated.

Technical variation can be compensated by analyzing more samples, by technological advances and by improving the data analyses. The reproducibility of MALDI-TOF in terms of peak intensity is limited (CV = 30%); however the high-throughput capability of the MALDI-TOF technique (a strong aspect of this type of mass spectrometry) allows the measurement of large numbers of samples. We compensated for the relatively poor reproducibility of the MALDI-TOF technique by measuring all samples in eighteen-fold and by using an all-or-nothing analysis approach i.e., the presence and absence of peaks. As shown in Chapter 2, this resulted in a considerable improvement in the reproducibility of data. The data obtained in this way can only be used in a semi-quantitative way. New developments such as matrix-assisted laser desorption ionization Fourier transform mass spectrometry (MALDI-FTMS) (as discussed later), can help to overcome this problem.

An additional important issue is sample handling: during sample handling differences among samples can be introduced leading to unwanted variations. To minimize these technical variations, samples (both patients and controls) have to be handled and stored under uniform conditions (6).

The measurement of large numbers of samples also requires a thorough statistical evaluation of the data. To this end we developed a database application dedicated to storage and analysis of mass spectrometry data. The modular structure of the application allows easy introduction of additional functionalities. Using this software we were able to extract information from the mass spectra and to perform statistical analyses on the data.

Statistical analyses of the data lead to the detection of differentially expressed peptide peaks. The next step is the identification of these differentially expressed peptides. For identification purposes a small number of samples is analyzed in much more depth. Combinations of MALDI related screening procedures and identification by advanced nanoLC ESI or MALDI tandem mass spectrometry are used to identify peptides that were detected in the MALDI high-throughput screening. A drawback of this method is that electrospray (ESI) ionizes a different chemical class of peptides than MALDI. Fortunately, in our hands the overlap between these two ionization methods is still quite large (approximately 30-50%) despite their complementary nature. Another disadvantage is the requirement of two expensive MS machines that generate two data streams that have to be integrated to come to a successful identification.

9.3 Biomarker discovery

9.3.1 LM in breast cancer

After implementation of our newly developed methods, subsequent experiments resulted in the identification of proteins that are differentially expressed between breast cancer patients with and without LM. The differentially expressed proteins that we have identified in the cerebrospinal fluid (CSF) of breast cancer patients with LM are probably indirect effects of the disease (acute phase response). The identification of this class of proteins clearly indicates the reliability of the method since such inflammation-related proteins are frequently up-regulated in cancer patients (7-10). The clinical importance of these proteins is not clear, because the specificity of these processes is limited. In addition, some brain/CSF specific proteins were identified such as prostaglandin synthase D2, cystatin-c and apolipoprotein E. The observed changes in apolipoprotein E and alpha-1-antichymotrypsin have been associated with Alzheimer’s disease and brain cancer (11-13). The profiling results have been used to build a classifier to distinguish breast cancer patients with and without LM. After bootstrap validation, the classifier had a maximum accuracy of 77% with a sensitivity of 79% and a specificity of 76%. The sensitivity that is obtained is similar to the sensitivity of CSF cytology and magnetic resonance imaging (MRI), the currently used diagnostic tools (14). The sensitivity and specificity of other CSF markers of LM is in the same range (beta-glucuronidase, beta2-microglobulin, carcinoembryonic antigen, lactate dehydrogenase isoenzymes and vascular endothelial growth factor (VEGF) (14-18)). Used as a single marker, none of these markers will add to the currently used diagnosis tools (17). In addition, all these markers will give false positive test results, because of their limited specificity. For this reason they only have an additional clinical value when combinations of markers are used together with CSF cytology and MRI.

In our study, we did not detect any of the traditional CSF markers for LM, indicating that the applied profiling technique lacks the sensitivity to detect and identify low abundant proteins. The detection of low abundant proteins is generally hampered by the presence of high abundant proteins. An additional complication is the increase in high abundant proteins (e.g. albumin) in the CSF from LM patients. This also limits the possibility of detecting differences in proteins that are present in lower concentrations. The fact that so many proteins are differentially expressed in CSF from LM patients makes it interesting to check whether combinations of these proteins may add to the specificity of the available markers.

9.3.2 Prostate cancer

In the serum samples from metastatic prostate cancer patients and controls, we identified differentially expressed non-tryptic breakdown products of albumin. We were able to show that the differential non-tryptic fragments were formed ex-vivo in serum and that this process is inhibited in the prostate cancer samples. The clinical applicability of such proteins or breakdown products of proteins has been envisioned (19, 20). Recent studies have shown that breakdown products can be highly specific for different disease types and are probably a result of the differential expression of a proteolytic enzyme or inhibitor (21). We have shown that synthetic peptide assays can be used to quantify the activity of a proteolytic enzyme and its inhibition in serum. Using synthetic peptide assays, enzymes that are low abundant but have high activities can be detected using mass spectrometry. This concept can be generalized by using peptide libraries to screen serum samples for differences in activity of known proteases. The simplicity of the measurement and the specificity of the readout system (the mass spectrometer) make this technique suited for large screening studies.

The main problem in both studies is that with the techniques used we were not able to directly detect or identify low abundant proteins. The detection of prostate specific antigen (PSA) is unlikely with a concentration in the ng/ml range and a heavy glycosylation of the protein. We estimated, based on the identified proteins, that a dynamic range of 2 to 3 orders of magnitude is covered by this technique. For this reason, further improvements are required to detect and identify also proteins that are in the range of the more classical markers in CSF and serum.

9.4 Technological developments in biomarker discovery

9.4.1 Dynamic range

One of the main problems in biomarker research, as also indicated by the experiments described in this thesis, is the limited dynamic range that can be covered. Only the most abundant proteins in samples can be detected if no fractionation is applied. The current mass spectrometry based techniques, in combination with the most advanced separation techniques that can currently be used routinely, can cover a dynamic range of about six orders of magnitude. In studies in which mass spectrometry methods that are less common and are more time consuming are used, a dynamic range of over seven orders of magnitude was reported (22). New developments in separation sciences and mass spectrometry will further extend this window to even lower abundant proteins (1, 23-25). Examples of new developments are new column materials (e.g. monolithic columns), immuno affinity depletion, chip technology and ion mobility mass spectrometry. In ion mobility mass spectrometry an extra step is included before entering the normal mass spectrometer. In this step the ions are transferred under the influence of a weak electric field through an inert buffer gas. The time for each ion to travel through this region filled with buffer gas is related to the mobility of the ion. The mobility of the ion depends on the strength of the electric field and the drift velocity of the ion. The measurements of ion mobiltities are very precise and vary only by 1-2%. The mobility of an ion depends on the structure of the ion: more compact structures will have higher mobilities. Large effects of this technique have already been reported on the dynamic range that can be covered for analyses of complex peptide mixtures (25-27).

9.4.2 Tissue analyses

Not only developments in equipment will help to detect lower abundant proteins, but also method changes can help to accomplish this. If the disease primarily affects tissue, the (relative) concentration of disease-specific proteins is probably higher in the tissue or tissue fluid compared to blood. Therefore, the chance of discovering new biomarkers is probably higher when studying diseased tissue, if it is available, rather than serum or plasma. Laser capture microdissection can be used to study disease tissue more specifically. The increased sensitivity of MS equipment makes it possible to study peptide profiles of the low numbers of cells that are obtained by laser capture microdissection (28, 29). This technique also allows for a comparison of the control and disease tissue from the same patient. The first discovery phase can be performed on a relatively small number of samples (control versus disease). The results of a pre-screening in tissue sections can be used to perform a more focused discovery in serum. In serum, the proteins that were found differentially expressed in tissue can be targeted. The number of candidate proteins can be further extended with proteins that are selected based on literature or that have a relation with the proteins that were found differentially expressed in tissue.

9.4.3 Reproducibility and quantification

We have shown that MALDI-TOF-MS can be used for peptide profiling. However, there are some problems associated with this technique, such as low reproducibility and limited possibilities to quantify. MALDI-FTMS results in a better reproducibility (CV=9%) and accuracy compared to MALDI-TOF-MS resulting in the detection of smaller differences in concentration. The measurement time per analysis on the MALDI-FTMS is longer compared to the MALDI-TOF-MS but because of the better reproducibility the number of replicate measurements can be decreased and large sample sets can be measured by MALDI-FTMS as well. Also, the quantitative capability of MALDI-FTMS in complex mixtures is much better compared to MALDI-TOF-MS. This is demonstrated by the almost ideal linear relationship between peak intensity and concentration of spiked synthetic peptides in complex samples (R^2=0.993, see chapter 7).

Developments in MALDI-FTMS are particularly interesting: since in the latest commercial FT mass spectrometers data acquisition is automated, resolutions of over 600,000 at mass 400 can be realized, sensitivity for peptides is at the amol level and a relatively large dynamic range can be achieved (4 orders of magnitude).

9.4.4 Identification

Another bottleneck in biomarker research is that differentially expressed peptides often remain unidentified. To further increase the sensitivity of identification, different sequencing methods can be used, e.g infrared multiphoton dissociation (IRMPD) and sustained off resonance irradiation collision induced dissociation (SORI/CID) options in MALDI FTMS can lead to the direct identification of peptides of interest. In the nanoLC-electrospray additional Electron Capture Dissociation (ECD) sequencing can be applied. Combinations of these techniques can considerably increase the number of identified peptides. New commercial mass spectrometers in which MS/MS spectra can be obtained with high mass accuracy can be used for compositional based sequencing (30). Also, improvements in identification can be obtained by construction of tailor-made or tissue specific databases. The advantage of these databases is that background (finding by chance) can be minimized considerably (see Figure 1.2). The downside of using databases is that if the protein is not present in the database, one will not be able to identify it. In the HUPO database a total of 9504 proteins are annotated of which 3020 proteins have been identified with two or more peptides (the Core Dataset) (31). A publicly available CSF database was until recently not available. For this reason we have constructed a home-made database with 356 CSF proteins that are described in literature. It is clear that these databases are not complete and efforts ought to be undertaken to obtain a higher coverage of all CSF and blood proteins in these databases. Pan et al. (32) identified a total of 2594 proteins in CSF that are now available in a publicly accessible database.

9.4.5 Validation

Biomarker research often results in long lists of candidate proteins. A validation of all proteins with antibody-based approaches is time-consuming and only feasible when antibodies are available. The multiple reaction monitoring (MRM) technique is a nice alternative to perform a pre-validation of proteins that were found in a first discovery round. It offers the possibility to quantify a large number of known peptides in a single run. In addition, the dynamic range of this technique is much larger than that of any other mass spectrometric technique. The

MRM technique has already been applied for years to the quantitative analysis of small molecules in body fluids. This technique can now also be used to quantify peptides and their precursor proteins in body fluids (33). In one run, several hundreds of peptides can be quantified, allowing analysis of relatively large numbers of samples.

The performance of MRM measurements is less affected by the enormous dynamic range of proteins in body fluids. However, a one-step detection of low abundant proteins is still not possible. A combination of depletion procedures and MRM leads to a better performance as shown by lower CV values and lower detection limits (33). Using beads to extract the tryptic peptides of interest from the digested serum sample can further decrease the CV values and detection limit. Antibodies specific for peptides of interest can be coupled to the beads for this purpose. Antibodies against peptides are easier to generate compared to antibodies raised against entire proteins. In addition, an absolute specificity of the antibody is not required, because the mass spectrometer will add to the specificity

A next step in the validation procedure is to obtain evidence that a candidate biomarker is expressed in the target tissue and not in the control tissue. For this purpose traditional immunohistochemical methods can be used. In the future, an alternative may be mass spectrometry based imaging, a technique that is only recently used in biomarker discovery (34). MS imaging has already been successfully applied in the fields of cancer research, neuroscience and pharmacology (35). MS based imaging techniques have greatly improved over the recent years in terms of sensitivity, spatial resolution and speed, making it more suitable for biomarker research. The unique ability of this technique is that it can detect numerous different compounds at the same time in tissue without losing the spatial distribution. This makes it possible to not only detect a difference in expression between tissue types but also to determine the exact location (spatial resolution <10 μm), of these differences (36). Furthermore, recent advances in MS imaging with respect to the direct identification of proteins in tissue (37), a further increase in measurement speed and sensitivity and the use of high resolution mass spectrometry (38) will undoubtedly result in an increased use of this technique in biomarker research and finally in pathology.

9.5 Effects of technological developments

For the biomarker research of LM in breast cancer patients the clinical application of identified markers could not be tested at the moment the experiments were performed. An external confirmation and further validation of the identified differentially expressed proteins was impossible, because not for all identified proteins ELISA kits were available and also for many patients not enough CSF was present. New developments in the technique can help to overcome these problems. The multiple reaction monitoring (MRM) technique using a triple quadrupole mass spectrometer can be used for this purpose. This type of mass spectrometry was until recently only available equipped with an ESI source; however, a triple quadrupole mass spectrometer with a MALDI source is under development and the first results look promising in our hands. It has been shown that this mass spectrometer can be used for quantification of peptides (39) allowing a confirmation and a relative quantification of the CSF proteins of interest. To validate the earlier findings externally we are currently collecting a new set of CSF samples from breast cancer patients with LM.

In addition to these experiments, a second discovery round will be performed using more advanced separation techniques and mass spectrometry. For pre-fractionation we are developing a completely automated protein separation using monolithic-based columns. This type of columns minimizes analysis time and increases reproducibility. Protein separation helps to reduce the suppressing effect of the high abundant proteins resulting in a lower threshold of detection and a higher sensitivity of the measurements. This will lead to the detection of lower abundant proteins. Following the aforementioned reasoning, the MALDI-FTMS would be the mass spectrometer of choice for measuring and quantifying these samples.

In the prostate cancer work, the number of samples that are going to be analyzed in a follow-up experiment will be increased to 50 per group. Also, the number of groups will be increased by including different stages of the disease to allow for the detection of earlier protein/peptide markers. To enhance the sensitivity of the measurements we are planning to remove the high abundant proteins from the serum samples. We have already shown that we were able to remove six of the most abundant proteins from a large set (n>200) of serum samples in a reproducible way by immuno-affinity chromatography (40). Currently we extend this depletion procedure to include the twenty most abundant proteins. This, in combination with MALDI-FTMS measurements, will further increase the reproducibility and sensitivity of peptide profiling. We are currently also testing the possibility to further exploit and increase the mass accuracy of MALDI-FTMS measurements. We hypothesize that by using the information of a large number of mass measurements of different patients, the accuracy of single masses measured can be further increased using information of all mass spectra. This procedure will result for each peptide in an independent accuracy, which can be used in a database search or for confirmation during sequencing.

The work presented in this thesis clearly shows that peptide profiling is a technique with a large potential in the field of biomarker discovery. New advances in technology and analyses methods will elevate this technique to a higher level and will eventually lead to the discovery of clinically applicable biomarkers and knowledge about molecular mechanisms in cancer.

References

1. Qian, W. J., Jacobs, J. M., Liu, T., Camp, D. G., 2nd, and Smith, R. D. "Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications," *Mol Cell Proteomics* 5 (2006): 1727-44.
2. Veenstra, T. D., Conrads, T. P., Hood, B. L., Avellino, A. M., Ellenbogen, R. G., and Morrison, R. S. "Biomarkers: mining the biofluid proteome," *Mol Cell Proteomics* 4 (2005): 409-18.
3. Bradshaw, R. A., Burlingame, A. L., Carr, S., and Aebersold, R. "Reporting protein identification data: the next generation of guidelines," *Mol Cell Proteomics* 5 (2006): 787-8.
4. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. "The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data," *Mol Cell Proteomics* 3 (2004): 531-3.
5. McComb, M. E., Perlman, D. H., Huang, H., and Costello, C. E. "Evaluation of an on-target sample preparation system for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in conjunction with normal-flow peptide high-performance liquid chromatography for peptide mass fingerprint analyses," *Rapid Commun Mass Spectrom* 21 (2007): 44-58.
6. West-Norager, M., Kelstrup, C. D., Schou, C., Hogdall, E. V., Hogdall, C. K., and Heegaard, N. H. "Unravelling in vitro variables of major importance for the outcome of mass spectrometry-based serum proteomics," *J Chromatogr B Analyt Technol Biomed Life Sci* 847 (2007): 30-7.
7. Kozak, K. R., Su, F., Whitelegge, J. P., Faull, K., Reddy, S., and Farias-Eisner, R. "Characterization of serum biomarkers for detection of early stage ovarian cancer," *Proteomics* 5 (2005): 4589-96.
8. Steel, L. F., Shumpert, D., Trotter, M., Seeholzer, S. H., Evans, A. A., London, W. T., Dwek, R., and Block, T. M. "A strategy for the comparative analysis of serum proteomes for the discovery of biomarkers for hepatocellular carcinoma," *Proteomics* 3 (2003): 601-9.
9. Yu, K. H., Rustgi, A. K., and Blair, I. A. "Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry," *J Proteome Res* 4 (2005): 1742-51.

10. Tolson, J., Bogumil, R., Brunst, E., Beck, H., Elsner, R., Humeny, A., Kratzin, H., Deeg, M., Kuczyk, M., Mueller, G. A., Mueller, C. A., and Flad, T. "Serum protein profiling by SELDI mass spectrometry: detection of multiple variants of serum amyloid alpha in renal cancer patients," *Lab Invest* 84 (2004): 845-56.
11. Ikeyama, Y., Orita, T., Nishizaki, T., Aoki, H., and Ito, H. "[Immunohistochemical expression of alpha-1-antitrypsin in human gliomas]," *No Shinkei Geka* 19 (1991): 1047-51.
12. Nicoll, J. A., Zunarelli, E., Rampling, R., Murray, L. S., Papanastassiou, V., and Stewart, J. "Involvement of apolipoprotein E in glioblastoma: immunohistochemistry and clinical outcome," *Neuroreport* 14 (2003): 1923-6.
13. Licastro, F., Campbell, I. L., Kincaid, C., Veinbergs, I., Van Uden, E., Rockenstein, E., Mallory, M., Gilbert, J. R., and Masliah, E. "A role for apoE in regulating the levels of alpha-1-antichymotrypsin in the aging mouse brain and in Alzheimer's disease," *Am J Pathol* 155 (1999): 869-75.
14. DeAngelis, L. M. "Current diagnosis and treatment of leptomeningeal metastasis," *J Neurooncol* 38 (1998): 245-52.
15. Twijnstra, A., van Zanten, A. P., Nooyen, W. J., and Ongerboer de Visser, B. W. "Sensitivity and specificity of single and combined tumour markers in the diagnosis of leptomeningeal metastasis from breast cancer," *J Neurol Neurosurg Psychiatry* 49 (1986): 1246-50.
16. Herrlinger, U., Wiendl, H., Renninger, M., Forschler, H., Dichgans, J., and Weller, M. "Vascular endothelial growth factor (VEGF) in leptomeningeal metastasis: diagnostic and prognostic value," *Br J Cancer* 91 (2004): 219-24.
17. van de Langerijt, B., Gijtenbeek, J. M., de Reus, H. P., Sweep, F. C., Geurts-Moespot, A., Hendriks, J. C., Kappelle, A. C., and Verbeek, M. M. "CSF levels of growth factors and plasminogen activators in leptomeningeal metastases," *Neurology* 67 (2006): 114-9.
18. DeAngelis, L. M., and Boutros, D. "Leptomeningeal metastasis," *Cancer Invest* 23 (2005): 145-54.
19. Liotta, L. A., Ferrari, M., and Petricoin, E. "Clinical proteomics: written in blood," *Nature* 425 (2003): 905.
20. Veenstra, T. D., Prieto, D. A., and Conrads, T. P. "Proteomic patterns for early cancer detection," *Drug Discov Today* 9 (2004): 889-97.
21. Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A., Erdjument-Bromage, H., Olshen, A. B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., Sanchez-Carbayo, M., Holland, E. C., Cordon-Cardo, C., Scher, H. I., and Tempst, P. "Differential exoprotease activities confer tumor-specific serum peptidome patterns," *J Clin Invest* 116 (2006): 271-84.
22. Liu, T., Qian, W. J., Gritsenko, M. A., Xiao, W., Moldawer, L. L., Kaushal, A., Monroe, M. E., Varnum, S. M., Moore, R. J., Purvine, S. O., Maier, R. V., Davis, R. W., Tompkins, R. G., Camp, D. G., 2nd, and Smith, R. D. "High dynamic range characterization of the trauma patient plasma proteome," *Mol Cell Proteomics* 5 (2006): 1899-913.
23. Lee, H.-J., Lee, E.-Y., Kwon, M.-S., and Paik, Y.-K. "Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics," *Current Opinion in Chemical Biology* 10 (2006): 42-49.
24. Lee, H. J., Lee, E. Y., Kwon, M. S., and Paik, Y. K. "Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics," *Curr Opin Chem Biol* 10 (2006): 42-9.

25. Valentine, S. J., Plasencia, M. D., Liu, X., Krishnan, M., Naylor, S., Udseth, H. R., Smith, R. D., and Clemmer, D. E. "Toward plasma proteome profiling with ion mobility-mass spectrometry," *J Proteome Res* 5 (2006): 2977-84.
26. Valentine, S. J., Koeniger, S. L., and Clemmer, D. E. "A split-field drift tube for separation and efficient fragmentation of biomolecular ions," *Anal Chem* 75 (2003): 6202-8.
27. Valentine, S. J., Liu, X., Plasencia, M. D., Hilderbrand, A. E., Kurulugama, R. T., Koeniger, S. L., and Clemmer, D. E. "Developing liquid chromatography ion mobility mass spectometry techniques," *Expert Rev Proteomics* 2 (2005): 553-65.
28. de Groot, C. J., Steegers-Theunissen, R. P., Guzel, C., Steegers, E. A., and Luider, T. M. "Peptide patterns of laser dissected human trophoblasts analyzed by matrix-assisted laser desorption/ionisation-time of flight mass spectrometry," *Proteomics* 5 (2005): 597-607.
29. Ball, H. J., and Hunt, N. H. "Needle in a haystack: microdissecting the proteome of a tissue," *Amino Acids* 27 (2004): 1-7.
30. Spengler, B. "De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry," *J Am Soc Mass Spectrom* 15 (2004): 703-14.
31. "http://www.bioinformatics.med.umich.edu/app/hupo/ppp/."
32. Pan, S., Zhu, D., Quinn, J. F., Peskind, E. R., Montine, T. J., Lin, B., Goodlett, D. R., Taylor, G., Eng, J., and Zhang, J. "A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry," *Proteomics* 7 (2007): 469-73.
33. Anderson, N. L., and Hunter, C. L. "Quantitative mass spectrometric MRM assays for major plasma proteins," *Mol Cell Proteomics* (2005).
34. Reyzer, M. L., and Caprioli, R. M. "MALDI mass spectrometry for direct tissue analysis: a new tool for biomarker discovery," *J Proteome Res* 4 (2005): 1138-42.
35. Reyzer, M. L., and Caprioli, R. M. "MALDI-MS-based imaging of small molecules and proteins in tissues," *Curr Opin Chem Biol* 11 (2007): 29-35.
36. McDonnell, L. A., Piersma, S. R., MaartenAltelaar, A. F., Mize, T. H., Luxembourg, S. L., Verhaert, P. D., van Minnen, J., and Heeren, R. M. "Subcellular imaging mass spectrometry of brain tissue," *J Mass Spectrom* 40 (2005): 160-8.
37. Groseclose, M. R., Andersson, M., Hardesty, W. M., and Caprioli, R. M. "Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry," *J Mass Spectrom* 42 (2007): 254-262.
38. Taban, I. M., Altelaar, A. F. M., van der Burgt, Y. E. M., McDonnell, L. A., Heeren, R. M. A., Fuchser, J., and Baykut, G. "Imaging of Peptides in the Rat Brain Using MALDI-FTICR Mass Spectrometry," *Journal of the American Society for Mass Spectrometry* 18 (2007): 145-151.
39. Melanson, J. E., Avery, S. L., and Pinto, D. M. "High-coverage quantitative proteomics using amine-specific isotopic labeling," *Proteomics* 6 (2006): 4466-74.
40. Dekker, L. J., Bosman, J., Burgers, P. C., van Rijswijk, A., Freije, R., Luider, T., and Bischoff, R. "Depletion of high-abundance proteins from serum by immunoaffinity chromatography: A MALDI-FT-MS study," *J Chromatogr B Analyt Technol Biomed Life Sci* 847 (2007): 65-9.

Chapter 10

Summary & Samenvatting

Summary

In this thesis we present newly developed methods for biomarker discovery. We applied these methods to discover biomarkers of leptomeningeal metastasis (LM) in the cerebrospinal fluid (CSF) from breast cancer patients and in serum from patients with prostate cancer. Early diagnosis of LM remains challenging because 25% of CSF samples test false negative at first cytological examination. In addition, the sensitivity and specificity of magnetic resonance imaging (MRI) in diagnosing LM in solid tumors are approximately 75%. Early diagnosis and initiation of treatment of LM are essential to prevent neurological deterioration. We therefore set out to find biomarkers to increase the accuracy of early diagnostic testing in LM.

For prostate cancer a biomarker in the form of prostate specific antigen (PSA) is available. The specificity of this marker is however rather poor, in screening studies like the European Randomized study of Screening for Prostate Cancer (ERSPC) 75% of men with an elevated PSA level > 3 ng/ml are false positive. Resulting in a group of men that undergo unnecessary treatment with a risk of complications including impotence and incontinence. New markers in addition to PSA could help to reduce the large number of false positive diagnosed patients and to predict the course of the disease.

The dynamic range of protein concentrations in body fluids exceeds 10 orders of magnitude. These huge differences in concentrations complicate the detection of proteins with low expression levels. Since all classical biomarkers have low expression levels (e.g. PSA 1-4 ng/ml; CA125: 20-35 U/ml) new developments with respect to identification and validation techniques of these low abundant proteins are required. In chapter 2 a detailed overview is given of the application of mass spectrometry to biomarker discovery, the current developments and the problems associated with this technique.

In the second part of the thesis new developments for biomarker analyses are presented. In chapter 3 the reproducibility and set-up of protein and peptide (mass spectrometry) profiling experiments are discussed. The number of peaks, their masses and their intensities are important characteristics in mass spectrometry. Because of the relatively low reproducibility of peak intensities associated with complex samples in matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) it is difficult to accurately assess the number of peaks and their intensities. We evaluated these two characteristics for tryptic digests of CSF. We observed that the reproducibility of peak intensities was relatively poor (CV = 42%) and that additional normalization or spiking did

not lead to a large improvement (CV = 30%). Moreover, at least seven mass spectra per sample were required to obtain a reliable peak list. An improvement of the sensitivity (eventually more peaks are detected) is observed if more replicates per sample are measured. We conclude that the reproducibility and sensitivity of peptide profiling can be significantly improved by a combination of measuring at least seven spectra per sample and a dichotomous scoring of the intensities.

In chapter 4 we present a newly developed database application for the storage and analysis of mass spectrometry data. Statistical comparison of peptide profiles requires fast, user-friendly software for high-throughput data analysis. Important features are flexibility in changing input variables and in statistical analysis of peptides that are differentially expressed between patient and control groups. In addition, integration of the mass spectrometry data with the results of other experiments, such as microarray analysis, and with information from other databases requires a central storage of the profile matrix, where protein names can be added to peptide masses of interest. A new database application is presented to detect and identify significantly differentially expressed peptides in peptide profiles obtained from body fluids samples of patient and control groups. The presented modular software is capable of central storage of mass spectra and results in fast analysis. The database application is capable of distinguishing patient MALDI-TOF peptide profiles from control groups using large datasets. The modular architecture of the application allows handling of the large data sets from MS/MS- and Fourier Transform (FT) mass spectrometry experiments.

In chapter 5 a new method for identification of proteins in complex samples is presented. MALDI-TOF mass spectrometry is an attractive technique for peptide profiling because of its sensitivity, reliability, and high-throughput capability. However, the complexity of peptide mixtures derived from proteins as well as the large dynamic range of protein concentrations in body fluids and tissue and cell lysates hamper the detection of all tryptic peptides from a sole mass spectrum. In addition, TOF peptide profiles are often difficult to interpret, mainly because a monoisotopic peak of one peptide may overlap with an isotopic peak of another, a feature the peak-picking algorithm may not detect. This drawback may be partially alleviated by higher mass resolution and mass accuracy as provided by e.g. FTMS. We have developed a simple method for the identification of peptides in complex mixtures whereby the high mass accuracy and resolution of MALDI-FTMS is exploited for directing and confirmation purposes.

In the third part of the thesis we present a peptide profiling study to detect and identify biomarkers of leptomeningeal metastases (LM) in breast cancer (chapter 6 and 7) and of prostate cancer (chapter 8). We investigated the protein expression patterns present in the CSF from breast cancer patients with and without LM. CSF samples from 106 patients with active breast cancer (54 with LM and 52 without LM) and 45 controls were digested with trypsin. The resulting peptides were measured by MALDI-TOF MS. Next, the mass spectra were analyzed and compared between patient groups using newly developed bioinformatics tools. A total of 895 possible peak positions were detected and 164 of these peaks discriminated between the patient groups (Kruskal-Wallis, $p < 0.01$). The discriminatory masses were clustered and a classifier was built to distinguish breast cancer patients with and without LM. After bootstrap validation, the classifier had a maximum accuracy of 77% with a sensitivity of 79% and a specificity of 76%.

In chapter 7, we determined the exact masses of these significant peptides in a limited number of samples by MALDI-FTMS. Identification of these peptides was performed by electrospray FTMS after separation by nano-scale liquid chromatography. The database results were confirmed by targeted high mass accuracy measurements of the fragment ions in the FT cell. The combination of automated high-throughput MALDI-TOF measurements and analysis by FTMS leads to the identification of seventeen peptides corresponding to nine proteins. These include proteins that are operative in host-disease interaction, inflammation and immune defence (serotransferrin, alpha 1-antichymotrypsin, hemopexin, haptoglobin and transthyretin). Several of these proteins have been mentioned in the literature in relation to cancer. The identified proteins alpha 1-antichymotrypsin and apolipoprotein E have been described in relation to Alzheimer’s disease and brain cancer. Direct MALDI-TOF analysis of tryptic digests of CSF gives reproducible peptide profiles that can assist in diagnosing LM in breast cancer patients. The same method can be used to develop diagnostic assays for other neurological disorders.

In chapter 8 we present results from mass spectrometric peptide profiling experiments aimed at discovering potential markers for prostate cancer. MALDI-TOF profiling experiments were performed on tryptic digests of serum samples (obtained by the ERSPC) from patients with metastatic prostate cancer (n=27) and controls (n=30) after purification with surface-active magnetic beads. This resulted in the detection of eight repeatedly observed differentially expressed peptides, which were then identified by nanoLC-MALDI-TOF/TOF and confirmed by MALDI-FTMS exact mass measurements. All differentially expressed peptides are derived from two homologous parts of human serum albumin; two of the eight peptides were tryptic and six were semi-tryptic. The presence of the semi-tryptic fragments indicates that a proteolytic process operates other than by trypsin. Since the semi-tryptic fragments are significantly more abundant in control patient samples compared to the metastases samples, we hypothesize that a specific inhibition of the proteolytic process is in effect in the serum of metastatic prostate cancer patients. Experiments using synthetic peptides showed that this proteolytic activity occurs ex vivo and that it is sequence specific.

The developed methods for peptide profiling are successfully implemented in biomarker research as shown by the two aforementioned examples: LM in patients with breast cancer and metastatic prostate cancer. Proteins or specific cleavage product of proteins could be detected from which a clear distinction can be made between patient and control groups. Future developments in equipment and methodology will create new possibilities to find markers that are more directly related to the disease. An example of developing methodology is the combination of microdissected tissue and mass spectrometry. Also the development of more advanced high throughput separation techniques and more sensitive MS equipment will help to achieve this goal.

Samenvatting

In dit proefschrift worden nieuwe methodes voor het ontdekken van biologische markers gepresenteerd. Deze nieuwe methodes zijn gebruikt voor het vinden van biomarkers voor leptomingeale metastases (LM) in hersenvloeistof van borstkanker patiënten en in serum van prostaatkanker patiënten.

Een vroegtijdige diagnose van LM blijft moeilijk aangezien 25 % van de cytologische onderzoeken van het hersenvocht resulteert in een vals negatieve uitslag. Tevens is de sensitiviteit en specificiteit van magnetic resonance imaging (MRI) voor de diagnose van LM voor vaste tumoren maar ongeveer 75%. Een vroege diagnose en vroegtijdige start van de behandeling is van essentieel belang om neurologische achteruitgang te voorkomen. Het is daarom van belang om nieuwe biomarkers te vinden voor een vroegtijdige diagnose van LM.

Voor prostaatkanker is een biomarker aanwezig in de vorm van prostate specific antigen (PSA). De specificiteit van deze marker is echter laag, in screening studies zoals de European Randomized study for Screening for Prostate Cancer (ERSPC) is 75% van de mannen met verhoogd PSA niveau (>3ng/ml) vals positief. Dit resulteert in een groep mannen die een onnodige behandeling ondergaat met de risico's op eventuele complicaties zoals impotentie en incontinentie. Nieuwe biomarkers zouden in aanvulling op PSA gebruikt kunnen worden om het grote aantal vals positieve uitslagen te reduceren en om een betere voorspelling te kunnen doen van het ziekteverloop.

Het dynamisch bereik van eiwitconcentraties in lichaamsvloeistoffen is meer dan 10 ordes van grootte. Deze grote verschillen in concentratie bemoeilijken de detectie van eiwitten met lage expressie niveaus. Aangezien alle klassieke biomarkers op lage niveaus tot expressie komen is het van belang dat nieuwe technieken ontwikkeld worden die het mogelijk maken deze eiwitten te identificeren en te valideren. In hoofdstuk 2 wordt een gedetailleerd overzicht gegeven van het gebruik van massaspectrometrie in de biomarker discovery, nieuwe ontwikkelingen op dit gebied en de problemen die hiermee geassocieerd zijn.

In het tweede gedeelte van dit proefschrift worden nieuw ontwikkelde methodes voor biomarker onderzoek gepresenteerd. In hoofdstuk 3 word de reproduceerbaarheid en de opzet van eiwit en peptide (massaspectometrie) experimenten bediscussiëerd. Het aantal pieken, hun massa's en de bijbehorende intensiteiten zijn belangrijke karakteristieken in de massaspectrometrie. De relatief lage reproduceerbaarheid van piekintensiteiten in complexe samples met matrix-assisted laser desorption/ionization time of flight massaspectrometrie

(MALDI-TOF MS) maakt het ingewikkeld om op een accurate manier het aantal pieken en hun intensiteit in een massaspectrum te bepalen. In dit hoofdstuk evalueren we deze twee karateristieken voor tryptische digesten van hersenvloeistof. De reproduceerbaarheid van de intensiteiten is relatief laag (CV = 42%); het gebruik van normalisatie of het spiken van het monster resulteert niet in een grote verbetering (CV = 30%). Om tot een betrouwbare pieklijst te komen waren minimaal 7 metingen per monster nodig. Een verbetering in de sensitiviteit (meer pieken worden gedetecteerd) wordt eveneens verkregen door meer metingen uit te voeren. Hieruit concludeerden we dat de reproduceerbaarheid en de gevoeligheid van peptide profilering significant kan worden verbeterd door een combinatie van minstens 7 metingen per monster en het scoren van de intensiteit op basis van aan- of afwezigheid.

In hoofdstuk 4 wordt een nieuw ontwikkelde database-applicatie voor het bewaren en analyseren van massaspectrometrie data gepresenteerd. Voor een grootschalige statistische vergelijking van peptide profielen is snelle en gebruiksvriendelijke software nodig. Een belangrijk kenmerk van software voor deze toepassingen is flexibiliteit. Het is van belang dat data variabelen gemakkelijk kunnen worden aangepast en dat verschillende vormen van statistiek kunnen worden toegepast. Een ander belangrijk punt is de mogelijkheid tot integratie van verschillende data stromen, zodat het mogelijk is om data vanuit verschillende experimenten te combineren b.v. microarray en proteomics data. Ook het koppelen van de data met data uit andere databases is van belang; dit vereist een centrale opslag. Een nieuwe database applicatie is ontwikkeld die het mogelijk maakt massaspectra van patiënten en controle monsters te vergelijken en de significant verschillend tot expressie komende peptides te detecteren en uiteindelijk het geïdentificeerde eiwit hieraan te koppelen. De modulaire opbouw van de applicatie maakt het mogelijk ook data van andere massaspectrometers te analyseren zoals van Fourier Transform massaspectrometrie (FTMS) apparatuur.

In Hoofdstuk 5 wordt een identificatiemethode gepresenteerd om eiwitten te identificeren in complexe mengsels. MALDI-TOF is een aantrekkelijke techniek voor peptide profilering vanwege de hoge sensitiviteit, betrouwbaarheid en snelheid. De complexiteit van peptide mengsels afkomstig van eiwitten en het grote verschil in concentraties van eiwitten in lichaamsvloeistoffen en cellysaten maakt het onmogelijk alle peptides te detecteren in een enkel massaspectrum. Tevens zijn dit soort spectra moeilijk te interpreteren vanwege overlappende isotooppatronen. Dit kan ten dele worden opgelost door gebruik te maken van massaspectrometrie apparatuur met een hogere resolutie zoals FTMS. De door ons ontwikkelde identificatiemethode voor peptides in complexe monsters maakt gebruik van de hoge massa-accuratesse en resolutie van FTMS metingen voor identificatie en bevestiging.

In het derde gedeelte van dit proefschrift worden de eerder beschreven methodes toegepast voor het detecteren en identificeren van biomarkers voor LM in borstkanker patiënten (hoofdstuk 6 en 7) en voor prostaatkanker (hoofdstuk 8). We hebben de eiwitexpressie patronen van hersenvocht monsters van patiënten met en zonder LM onderzocht. Hersenvocht van 106 patiënten met actieve borstkanker (54 met LM en 52 zonder LM) en 45 controle monsters zijn voor dit doel gedigesteerd met trypsine. De resulterende peptides zijn gemeten met een MALDI-TOF-MS. Vervolgens zijn de massaspectra geanalyseerd en zijn de verschillende groepen statistisch vergeleken met nieuw ontwikkelde bio-informatica software. In totaal zijn 895 pieken gedetecteerd, hiervan bleken er 164 te discrimineren tussen de patiëntengroepen (Kruskal-Wallis, $p<0.01$). De discriminerende massa's zijn geclusterd en een voorspellend model is gemaakt om borstkanker patiënten met en zonder LM te kunnen onderscheiden. Na een bootstrap validatie van het model werd een maximale accuratesse van 77% met een sensitiviteit van 79% en een specificiteit van 76% behaald.

In hoofdstuk 7 beschrijven we de identificatie van de discrimineerde pieken door middel van accurate massametingen van een beperkt aantal monsters. Voor de identificatie van de peptiden is gebruik gemaakt van electrospray FTMS in combinatie met nano vloeistof chromatografie. De database resultaten zijn bevestigd door van alle geïdentificeerde peptiden hoge massa accuratesse MS/MS spectra te meten in de FT cel. De combinatie van geautomatiseerd snelle MALDI-TOF metingen en de analyse met FTMS hebben er toe geleid dat 17 peptiden afkomstig van negen eiwitten geïdentificeerd konden worden. Dit zijn eiwitten die actief zijn in ontstekingen en immunologische reacties (serotransferrine, alpha 1-antichymotrypsine, hemopexine, haptoglobuline en transthyretin). Een aantal van deze eiwitten zijn in de literatuur genoemd in relatie tot kanker. De geïdentificeerde eiwitten alpha 1-antichymotrypsine en apolipprotein E zijn beschreven in relatie met Alzheimer en hersenkanker. Het blijkt dus mogelijk met een directe MALDI-TOF analyse van tryptische digesten van hersenvocht reproduceerbare peptide profielen te genereren die mogelijk van nut kunnen zijn in de diagnose van LM in borstkanker patiënten. Deze methode kan ook gebruikt worden om diagnostieke methodes te ontwikkelen voor andere neurologische afwijkingen.

In hoofdstuk 8 presenteren we de resultaten van een profileringexperiment met als doel nieuwe markers voor prostaatkanker te ontdekken. MALDI-TOF profileringexperimenten zijn gedaan met tryptische digesten van serum monsters (verkregen van de ERSPC) van 27 patiënten met gemetastaseerde prostaatkanker en 30 controles die eerst zijn opgezuiverd met oppervlakte-actieve magnetische bolletjes. Dit resulteert in de detectie van 8 differentiëel tot expressie komende pieken. Deze pieken zijn geïdentificeerd met nanoLC-MALDI-TOF/TOF en de identificaties zijn bevestigd met exacte massa bepalingen met MALDI-FTMS. Alle peptiden die differentiëel tot expressie komen zijn afkomstig van twee homologe sequenties van humaan serum albumine; twee van de 8 peptiden zijn tryptisch en zes zijn semi-tryptisch. De aanwezigheid van de semi-tryptische fragmenten is een indicatie dat een proteolitisch proces actief is. Aangezien de semi-tryptische fragmenten significant hoger tot expressie komen in de controle monsters, hypothetiseren we dat een specifieke inhibitie van dit proteolitisch proces optreedt in de serum monsters van prostaatkanker patiënten. Experimenten met synthetische peptides hebben aangetoond dat dit proteolytische effect *ex vivo* optreedt en sequentiespecifiek is.

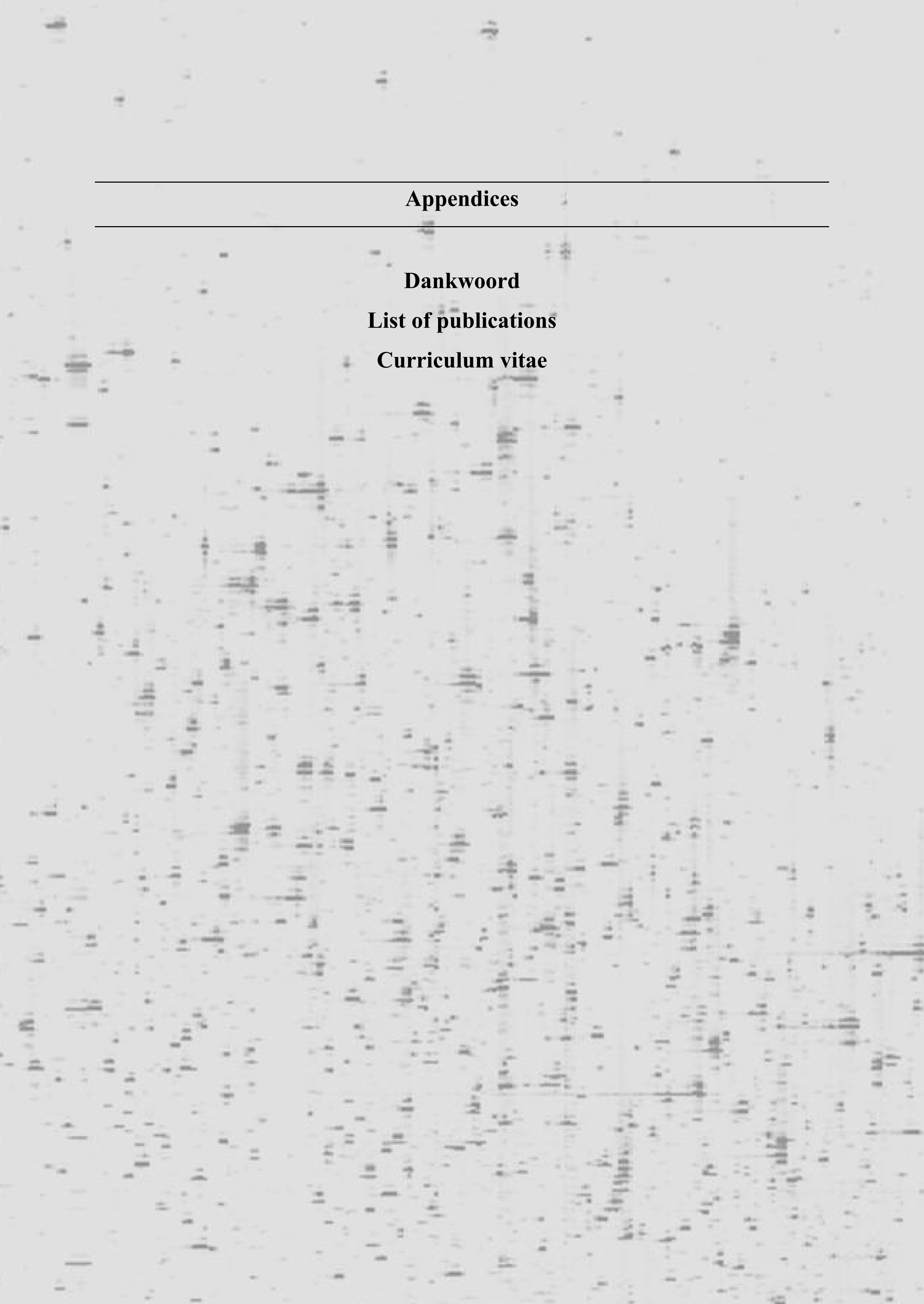
De ontwikkelde methodes voor peptide profilering zijn succesvol geïmplementeerd in het biomarker onderzoek zoals blijkt uit de twee bovengenoemde voorbeelden: LM in patiënten met borstkanker en gemetastaseerd prostaatkanker. Eiwitten of specifieke afbraak producten van eiwitten konden gebruikt worden om onderscheid te maken tussen de patiënt en controle groepen. Ontwikkelingen in apparatuur en methodologie zullen het mogelijk maken om ook markers te vinden die direct gerelateerd zijn aan de ziekte. Een voorbeeld van een ontwikkeling is de combinatie van microdissectie van weefsel en massa spectrometrie. Ook ontwikkelingen in scheidingstechnieken en gevoeligere MS apparatuur zullen hier toe bijdrage.

Appendices

Dankwoord

List of publications

Curriculum vitae



Dankwoord

Na een promotietraject van vier jaar is dit boekje het resultaat van mijn inspanningen. Al het voorgaande in dit proefschrift geeft een duidelijk beeld van hoe het mij op wetenschappelijk gebied de afgelopen vier jaar is vergaan. Het wordt echter niet duidelijk wat deze periode van vier jaar persoonlijk met mij heeft gedaan. Buiten dat ik veel heb geleerd, heb ik ook heel veel plezier gehad in het werk. Dit is niet alleen te danken aan het feit dat wetenschap gewoon heel leuk is maar vooral ook aan de fijne en inspirerende werkomgeving waarbinnen dit promotieonderzoek heeft plaatsgevonden. In dit dankwoord zou ik graag iedereen die daar aan heeft bijgedragen van harte willen bedanken. Natuurlijk wil ik ook nog even een aantal personen bij naam noemen, omdat zij in het bijzonder hebben bijgedragen.

In de eerste plaats wil ik Theo Luider bedanken. Hij heeft er in de vier jaar van deze promotie voor gezorgd dat we van een groepje van drie mensen die proteomics werk deden zijn uitgegroeid naar een complete klinische proteomics groep. We werken er nu met meer dan tien mensen en dit aantal neemt nog steeds toe, ook het arsenaal aan apparatuur dat tot onze beschikking staat blijft zich uitbreiden. Dit heeft er voor gezorgd dat we mijn experimenten maar ook die van de gehele klinische proteomics groep naar een hoger niveau konden tillen. Naast dat Theo de mogelijkheden voor mij creëerde om mijn onderzoek te kunnen doen heeft hij me ook altijd zeer prettig begeleid. Theo is in staat altijd overal op een positieve manier tegen aan te kijken en zelfs experimenten die ik als mislukt beschouwde toch nog op een positieve manier te interpreteren. Hartelijk dank hiervoor en ik hoop op een verdere succesvolle samenwerking in de komende jaren.

Mijn andere begeleider Guido Jenster en promotoren Peter Sillevis Smitt en Chris Bangma wil ik ook bedanken voor alle inspirerende en plezierige werkbesprekingen die we hebben gehad en de hulp die jullie hebben geboden bij het schrijven van de artikelen en het tot stand komen van dit proefschrift.

Iemand die mij zeer in het bijzonder de laatste 2,5 jaar van mijn promotie heeft geholpen is Peter Burgers. Beste Peter, naast dat jij altijd zorgt dat we wel ergens om kunnen lachen op de werkvloer, heb je ook met je enorme kennis en ervaring er toe bijgedragen dat mijn promotie een succes is geworden. Dankzij jouw hulp en aanwijzingen is het grootste gedeelte van mijn artikelen op dit moment al gepubliceerd. Ook was het voor mij zeer plezierig om onze kennis te combineren: mijn ervaring met biochemisch labwerk en jouw ervaring op meer hardcore chemie- en massaspectrometrie gebied. Hartelijk dank hiervoor en ik hoop dat we nog een aantal plezierig jaren zullen samenwerken.

Door de jaren heen heb ook altijd hulp gehad van een of meerdere analisten, zonder hen hadden een aantal dagen met veel experimenteel werk er heel anders uitgezien. In het bijzonder wil ik Hans bedanken die mij wegwijs maakte op het lab en mij bij mijn eerste experimenten ontzettend heeft geholpen. De rest van de analisten, Eric, Halima, Coşkun dank voor jullie hulp en de plezierige samenwerking. Mijn collega AIO's en postdoc's Marcel, Jeroen, Linda en Arzu wil ik ook bedanken voor de gezellige werksfeer, de leuke congres bezoeken en de bruikbare adviezen.

Ook alle externe samenwerkingen zijn voor het tot stand komen van dit proefschrift van essentieel belang geweest. Ik wil dan ook allen hiervoor van harte bedanken: de NKI groep van Dr. Hans Bonfrer, aan de universiteit van Groningen de groep van Prof. Rainer Bischoff, de AMOLF groep van Prof. Ron Heeren, Dr. Andreas Römpp van de universiteit van Giessen en Dr. Yuri van der Burgt LUMC.

Een laatste dankwoord is voor mijn vrienden en familie, zij hebben er voor gezorgd dat ik het werk in mijn vrije tijd van mij af kon zetten, zodat ik elke dag weer met nieuwe moed er tegenaan kon. In het bijzonder geldt dit voor Lineke die mij gedurende deze periode altijd gesteund heeft en mij van de broodnodige afleiding heeft voorzien.

List of publications

1. Dekker, L. J., Boogerd, W., Stockhammer, G., Dalebout, J. C., Siccama, I., Zheng, P., Bonfrer, J. M., Verschuuren, J. J., Jenster, G., Verbeek, M. M., Luider, T. M., and Sillevis Smitt, P. A. "MALDI-TOF Mass Spectrometry Analysis of Cerebrospinal Fluid Tryptic Peptide Profiles to Diagnose Leptomeningeal Metastases in Patients with Breast Cancer," *Mol Cell Proteomics* 4 (2005): 1341-1349.
2. Dekker, L. J., Bosman, J., Burgers, P. C., van Rijswijk, A., Freije, R., Luider, T., and Bischoff, R. "Depletion of high-abundance proteins from serum by immunoaffinity chromatography: A MALDI-FT-MS study," *J Chromatogr B Analyt Technol Biomed Life Sci* 847 (2007): 65-9.
3. Dekker, L. J., Burgers, P. C., Guzel, C., and Luider, T. M. "FTMS and TOF/TOF mass spectrometry in concert: identifying peptides with high reliability using matrix prespotted MALDI target plates," *J Chromatogr B Analyt Technol Biomed Life Sci* 847 (2007): 62-4.
4. Dekker, L. J., Burgers, P. C., Kros, J. M., Smitt, P. A., and Luider, T. M. "Peptide profiling of cerebrospinal fluid by mass spectrometry," *Expert Rev Proteomics* 3 (2006): 297-309.
5. Dekker, L. J., Dalebout, J. C., Siccama, I., Jenster, G., Sillevis Smitt, P. A., and Luider, T. M. "A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra," *Rapid Commun Mass Spectrom* 19 (2005): 865-870.
6. Mustafa, D. A., Burgers, P. C., Dekker, L. J., Charif, H., Titulaer, M., Sillevis Smitt, P. A., Luider, T. M., and Kros, J. M. "Identification of glioma neovascularisation-related proteins by using MALDI-FTMS and nano-LC fractionation to microdissected tumor vessels," *Mol Cell Proteomics* (2007).
7. Rompp, A., Dekker, L., Taban, I., Jenster, G., Boogerd, W., Bonfrer, H., Spengler, B., Heeren, R., Smitt, P. S., and Luider, T. M. "Identification of leptomeningeal metastasis-related proteins in cerebrospinal fluid of patients with breast cancer by a combination of MALDI-TOF, MALDI-FTICR and nanoLC-FTICR MS," *Proteomics* 7 (2007): 474-81.
8. Titulaer, M. K., Siccama, I., Dekker, L. J., van Rijswijk, A. L., Heeren, R. M., Sillevis Smitt, P. A., and Luider, T. M. "A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls," *BMC Bioinformatics* 7 (2006): 403.

Curriculum vitae

Lennard Dekker werd geboren op 28 maart 1979 te Dirksland. Zijn HAVO diploma behaalde hij aan de Prins Maurits scholengemeenschap te Middelharnis in 1997. In 2001 rondde hij zijn studie af op de Hogere Laboratorium School in Etten-Leur met als specialisatie biochemie en proefdierkunde. Zijn afstudeerstage voor deze opleiding doorliep hij op de afdeling Farmacologie van Organon, waar hij onderzoek deed naar de expressie van osteoporose-gerelateerde genen onder supervisie van Dr. W. Dokter. Na deze stage bleef hij nog enkele maanden werken als analist op deze afdeling. In september 2001 startte hij het HLO instroomprogramma van de opleiding biologie aan de Universiteit van Leiden. Tijdens deze opleiding heeft hij diverse stages doorlopen bij de onderzoeksgroep dierecologie van de Universiteit van Leiden. In mei 2003 rondde hij deze opleiding succesvol af en startte direct aansluitend met een promotieonderzoek op het Erasmus MC. Dit promotie-onderzoek werd uitgevoerd op de afdelingen Neurologie (hoofd Prof. Dr. P.A.E. Sillevis Smitt) en Urologie (Prof. Dr. C.H. Bangma) onder supervisie van Dr. T.M. Luider en Dr. G. Jenster. Gedurende dit promotie-onderzoek heeft hij verschillende cursussen en trainingen voor diverse massaspectrometrische technieken gevolgd. Op dit moment is Lennard Dekker nog steeds werkzaam op het Erasmus MC waar hij als post-doc onderzoek doet op de afdeling neurologie naar intracellulaire concentraties van HIV medicatie met behulp van nieuwe massaspectrometrie methoden in klinische monsters.