Breast cancer serum proteomics: Sample processing and protein profiling by mass spectrometry.

Grassl, Julia

Breast Cancer Serum Proteomics

Sample Processing and Protein Profiling by Mass Spectrometry

by

Julia Grassl B.Sc. (Honours Genetics)

Thesis submitted to the University of Wales in fulfilment of the requirement for the degree of Doctor of Philosophy

School of Medicine
Swansea University
United Kingdom

Year of Submission 2007

Declaration

This work has not previously been accepted in substance for any degree and is not being currently submitted in candidature for any degree.

Signed .....(candidate)
Date ..... 19/08/07 .....

This thesis is the results of my own investigations, except where otherwise stated. No correction services were used.

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....(candidate)
Date ..... 19/08/07 .....

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ....(candidate)
Date ..... 19/08/07 .....

Summary

The aim of this project was to develop a method for discovery of biomarkers or a protein pattern, as a signature of breast cancer. Early detection of breast cancer is crucial to increase the survival rates of patients. Little was published about biomarker discovery from serum using mass spectrometry, so over the course of the project each factor of the methodology was analysed and optimized. It was shown that standardisation of sample preparation and handling is critical for any quantitative study. The presence of albumin and other highly abundant proteins in serum interferes with proteomic analysis and so depletion techniques were investigated. Centrifugal ultrafiltration was optimised and an extensive study showed it to be a robust and efficient method to enrich the LMW proteome for subsequent biomarker discovery in serum. SELDI-ToF and MALDI-ToF MS were compared for intact protein profiling for breast cancer. In contrast to SELDI-ToF, MALDI-ToF MS had been little tested for this purpose and therefore new software was developed for peak alignment enabling comparison of multiple spectra. LMW serum samples from 8 breast cancer and 8 control individuals were analysed in each experiment. Here we detected seven potential markers in total and gained initial peptide identifications for three markers. This study also tested the use of label-free quantitation using LC-MS on serum samples from breast cancer patients; one differentially-expressed peptide was discovered. The lack of a software tool for comparison of the resulting spectra limited the detection of further markers. The profiling results showed that the use of replicates all the way from starting with the initial serum sample through to data retrieval is crucial due to variation between the biological replicates, and also to reduce any variation occurring from sample preparation.

Acknowledgements

So many have supported and helped me throughout the last 3 ½ years, that it is difficult to name everyone. I would like to thank everyone in the School of Medicine and Biological Sciences, who has stood by me with advice and encouragement over the years, making me believe that I can do it.

I have to thank Prof Robert Leonard and Prof Gerry Thomas for taking me on as their PhD student and providing me with a project where I could develop myself and take the research into exciting and new directions.

I would like to thank Prof Gareth Brenton for becoming my supervisor so late into the project and providing me with the much needed support and encouragement to carry on writing and finish the thesis.

A very special thanks goes to Ed (Dr Dudley) for being super supportive, always interested, always there as a "supervisor" and as a friend. Thank you for putting up with my moods and rants at everything and everyone. I couldn't have done this without you.

A big thank you goes to Sarah (Dr Forbes-Robertson) for her patience, un-divided support and assistance in the lab and especially during the write-up. You have been brilliant.

I would like to thank Gorka for being my PhD buddy and keeping me company so many evenings when we were working late. Everyone in the lab and in the office, but especially Rachel, Claire and Regina, and Clare, Liz and Ally for making the days go by fast and the work fun.

I am equally thankful to Suzanne Williams for taking all my blood samples, and Marc, Emma and Karl for spinning so many of my serum samples.

Finally, most credit goes to my parents for getting me to where I am, always supporting and encouraging me. Especially my mum, believing that I could do it. The same thanks goes to John, who has always believed in me and put up with all the late nights and the random rants although it made so little sense.

The work has been carried out at the School of Medicine, University of Wales. This work was financially supported by a Molecular Oncology Research grant.

Abbreviations

2D-PAGE	2 dimensional polyacrylamide gel electrophoresis
ACN	acetonitrile
C.V.	coefficient of variance
CEA	carcinoembryonic antigen
CHAPS	3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate
CID	collision-induced dissociation
CV	cartridge volumes
Da	dalton
DCIS	ductal carcinoma in situ
deltaCn	delta correlation
DHB	2,5-dihydroxybenzoic acid
ESI-MS	electrospray ionisation mass spectrometry
FA	formic acid
FT	flow-through
HPLC	high performance liquid chromatography
ICAT	isotope-coded affinity tagging
IEF	isoelectric focusing
IgG	immunoglobulin G
IMAC-Cu2+	immobilized metal affinity capture coupled with copper
LC	liquid chromatography
LCIS	lobular carcinoma in situ
LCM	laser capture microdissection
LMW	low molecular weight
m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption/ionization
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MudPit	multidimensional protein identification technology
MV	markerview
MWCO	molecular weight cut-off
NF	normalization factor
OGP	octyl glucopyranoside

OP	breast cancer patient serum samples
PCA	principal component analysis
PSA	prostate specific antigen
Q-ToF	quadrupole time-of-flight
R2	correlation coefficient
RP	reverse phase
RT	retention time
S/N	signal-to-noise ratio
S1	sample set 1
S2	sample set 2
SA	sinapinic acid
SCX	strong cation exchange
SDS	sodium dodecyl sulphate
TFA	trifluoroacetic acid
TIC	total ion current
ToF	time-of-flight
UF	centrifugal ultrafiltration
V	control serum samples
VBA	visual basic for applications
WAX	weak anion exchange
WCX	weak cation exchange
Xcorr	cross correlation
XIC	extracted ion chromatogram
αCHCA	α-cyano-4-hydroxycinnamic acid

List of Figures

List of Tables

Contents

PUBLICATIONS AND CONFERENCE PAPERS

APPENDICES ..... on CD

CHAPTER 1

Introduction to Proteomics of Breast Cancer

1.1 Introduction to Breast Cancer

1.1.1 Staging and Histological Typing of Breast Cancer

Breast cancer is a heterogeneous disease, with a UK incidence rate of approximately 40,000 women in 2004 [1] and a mortality rate of 13,000 women in 2002 [2]. In its early stages, breast cancer is classified as either ductal carcinoma *in situ* (DCIS), arising from ductal epithelium, or lobular cancer *in situ* (LCIS), arising from the epithelium of the lobules. Due to the increasing use of screening mammography, these early, non-invasive cancers are more frequently diagnosed and now constitute 15-20% of all breast cancers [3]. DCIS, thought to be a direct precursor of invasive breast cancer, is highly curable by surgical removal. Even invasive cancer, that has infiltrated the basement membrane, can still be cured in over 90% of patients if detected at stage I (Figure 1.1). Therefore as in other cancers, early detection of breast cancer is vital. However, despite surgical removal of early stage tumours, some tumour cells may remain (micrometastases) and the cancer could recur locally or in distant organs. To reduce the risk of recurrence, adjuvant treatment is prescribed for the majority of patients in present clinical practice. To avoid unnecessary treatment and to help predict response to adjuvant therapies, markers to assist in decisions on further therapy are urgently required.

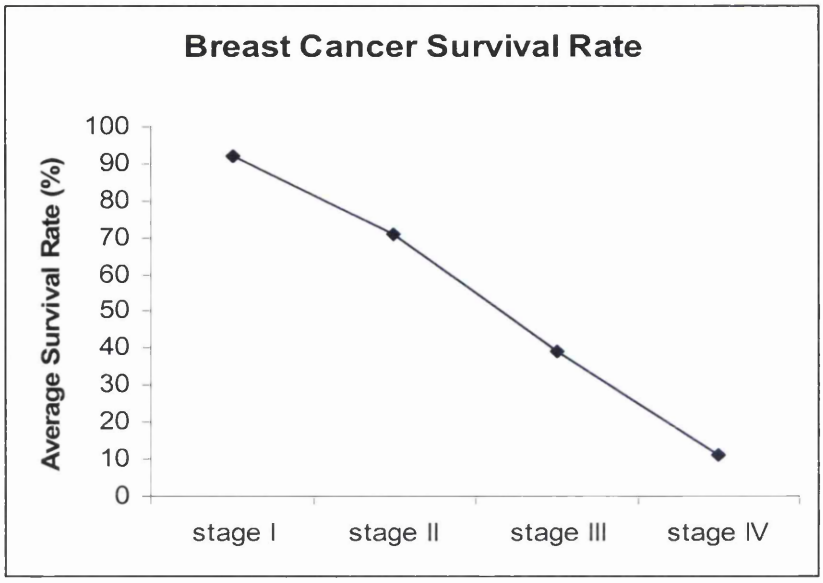


Figure 1.1: Average survival rate of breast cancer patients. Survival increases with early disease discovery and treatment. Data taken from Corporation Imaginis [3].

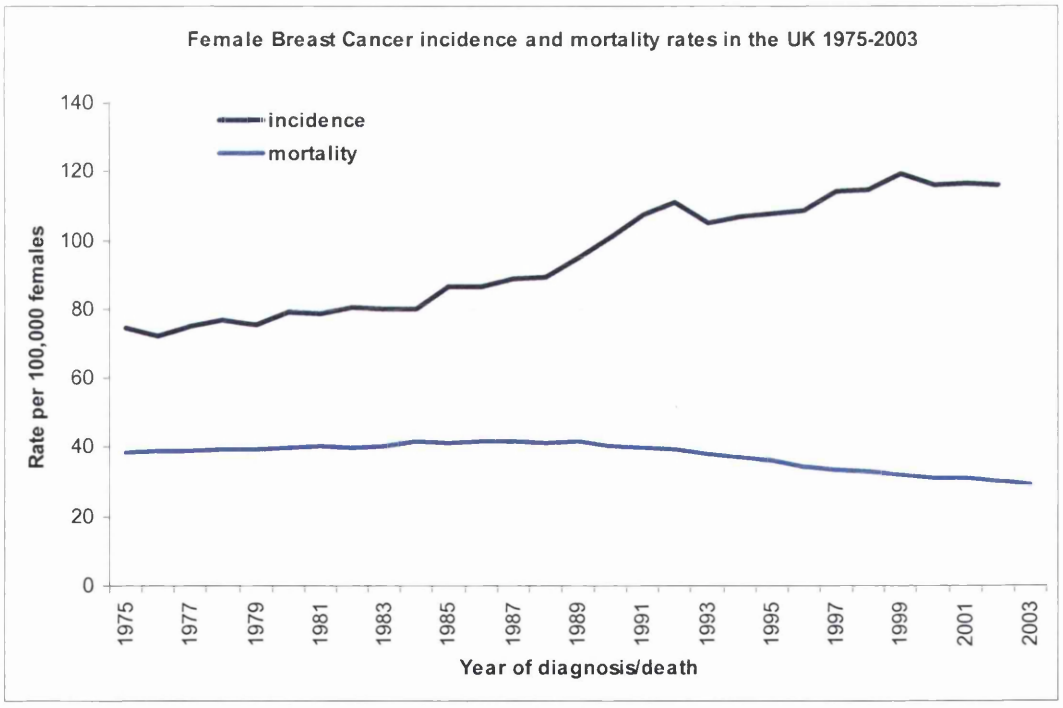


Figure 1.2: Female breast cancer incidence and mortality trends in the UK. The incidence of breast cancer cases has increased over the past 20 years however due to better treatment the mortality has decreased. Data taken from Cancer Research UK [4].

Recognised prognostic factors for breast cancer include histological subtype, tumour size, grade, lymphovascular invasion and axillary node metastases and these allow classification of the tumour (Table 1.1 and Table 1.2). However within each stage of disease there are differences in survival so new and better prognostic indictors are needed. The aim of this study was to identify potential prognostic markers in serum in metastatic breast cancer patients. These markers may enable earlier detection of recurrence of breast cancer disease or define patients at high risk of metastases.

Table 1.1: TNM Classification of breast cancer, according to the National Cancer Institute [5].

	T - Primary Tumour	
TX	Primary tumour cannot be assessed	
T0	No evidence of primary tumour	
Tis	Non-infiltrating intraductal carcinoma	
	Tis (DCIS)	Ductal carcinoma in situ
	Tis (LCIS)	Lobular carcinoma in situ
	Tis (Paget's)	Paget's disease of the nipple with no tumour.
T1	Tumour ≤2.0 cm in greatest dimension	
T2	Tumour >2.0 cm but ≤5.0 cm in greatest dimension	
T3	Tumour >5.0 cm in greatest dimension	
T4	Tumour of any size with direct extension to chest wall or skin	
	N - Regional lymph nodes	
NX	Regional lymph nodes cannot be assessed (e.g., previously removed)	
N0	No regional lymph node metastasis	
N1	Metastasis to movable ipsilateral axillary lymph node(s)	
N2	Metastasis to fixed ipsilateral axillary lymph node(s)	
N3	Metastasis in ipsilateral internal mammary lymph node(s)	
	pN - Pathologic classification	
pNX	Regional lymph nodes cannot be assessed (e.g., not removed for pathologic study or previously removed)	
pN0	No regional lymph node metastasis histologically	
pN1	Metastasis in 1 to 3 axillary lymph nodes, and/or in internal mammary nodes	
pN2	Metastasis in 4 to 9 axillary lymph nodes, or internal mammary lymph nodes	
pN3	Metastasis in ≥10 axillary lymph nodes	
	M - Distant metastasis	
MX	Presence of distant metastasis cannot be assessed	
M0	No distant metastasis	
M1	Distant metastasis	

Table 1.2: Breast cancer stage grouping, according to the National Cancer Institute [5].

Stage Groupings			
Stage 0	Tis	N0	M0
Stage 1	T1	N0	M0
Stage 2a	T0-T1	N1	M0
	T2	N0	M0
Stage 2b	T2	N1	M0
	T3	N0	M0
Stage 3a	T0-T2	N2	M0
	T3	N1-N2	M0
Stage 3b	T4	Any N	M0
Stage 3c	Any T	N3	M0
Stage 4	Any T	Any N	M1

Table 1.3: World Health Organization Classification of Carcinoma of the Breast [6]

Pathological classifications
Noninvasive carcinoma
Ductal carcinoma *in situ*
Lobular carcinoma *in situ*
Invasive carcinoma
Invasive ductal carcinoma
Invasive lobular carcinoma
Mucinous carcinoma
Medullary carcinoma
Papillary carcinoma
Tubular carcinoma
Adenoid cystic carcinoma
Secretory (juvenile) carcinoma
Apocrine carcinoma
Carcinoma with metaplasia (metaplastic carcinoma)
Inflammatory carcinoma
Other (specify)
Paget's disease of the nipple

1.1.2 Diagnosis and Management

To date, the most commonly used technology for early diagnosis of breast cancer is mammography. Although tumours detected by screening are significantly smaller than those presenting clinically, specificity and sensitivity of the method can be improved. Mammography fails to identify about 10% of cases and also gives a certain number of false positive diagnosis [7]. Other imaging techniques (e.g. ultrasound and MRI) have been developed to detect small tumour masses; however, the techniques still suffers from the lack of sensitivity to detect small numbers of cells. Therefore a large amount of research has focused on looking for biomarkers to detect disease in tissue and in serum. A biomarker may be a specific physical trait used to measure or indicate the effects or progress of a disease, illness, or condition [8]. However, the major role of current blood markers has been the diagnosis and monitoring of metastatic disease, where elevation of established blood tumour markers is correlated with the extent of the metastatic breast cancer. In fact, tumour marker measurements are now used, if not routinely, as a complementary test in the diagnosis of symptomatic metastases [9]. To advance early diagnosis and treatment of breast cancer, more reliable markers need to be found.

1.1.3 Tumour Markers of Breast Cancer

A large number of tumour markers have been proposed over the years for breast cancer, some of which are described below. However, to date, these markers have only been used for detection of metastatic disease and/or assessment of treatment response, and have shown little utility for diagnostic purposes.

The most widely-used clinical tumour markers are serum levels of MUC-1 family of mucin glycoproteins (e.g. CA15.3, CA27.29) and onco-foetal proteins such as carcinoembryonic antigen (CEA). The clinical data on the current use of markers is controversial, with some studies showing a correlation of CA15.3 levels in serum with disease progression, as high levels of CA15.3 occur in stage IV disease with ~90% specificity [10-14]. In combination, CEA, c-erb-2 and CA15.3 have shown a sensitivity of 88% in patients with recurrence of breast cancer [15]. However, the

American Society of Clinical Oncology, as well as the European Group on Tumour Markers, chose a cautious policy and state that present data regarding both CA15.3 and CEA are insufficient to recommend their routine use in the diagnosis of recurrent breast cancer follow-up [16, 17]. Therefore, at present, in the absence of any established alternative, further evaluation of these markers and new ones is required, to provide a much-needed clinical tool to assess response to metastatic cancer therapy.

The growth factor encoding gene, c-erb-2/HER2/*neu*, is overexpressed in 20-30% of breast cancer patients [18-20] and elevated protein levels have been found in the serum of 29% of patients diagnosed with breast carcinoma [15]. The extracellular domain of HER-2/*neu* is shed from breast cancer cells into the circulation and measurable by immunoassay. Serum HER-2/*neu* receptor protein levels have successfully predicted the presence and progression of HER-2/*neu* -positive breast cancer. Collected published studies revealed that more than 80% of patients showed a significant correlation between serum HER-2/*neu*-protein levels and either disease recurrence, metastasis, shortened survival or predicting response to chemotherapy and hormone therapy [17, 21-24].

Detection of specific antibodies against cancer cells might allow very early detection of tumours, before any markers have been released by the tumour, in the same way we can diagnose infectious disease based on the humoral immune response at early stages. Auto-immunity against cancer proteins, such as p53, heatshock protein 90, c-erb-2/HER2/*neu*, and mucin-related antigens has been used for detection and monitoring of breast cancer [22, 25, 26]. The presence of p53 has been found in 15% of breast cancer patients and is associated with poor prognosis [27]. However, these techniques are still being evaluated clinically.

One of the emerging techniques to improve biomarker detection is proteomics. Proteomics is the analysis of the proteome: all proteins in an organism, tissue or body fluid expressed at the time of interest. The proteome may differ between two samples depending on the disease state (e.g. cancer or non-cancer) and this difference may give information on what proteins are involved or influenced by tumour formation or progression. Proteomics is therefore a critical technique in finding new markers to detect tumours in breast cancer.

1.1.4 The Breast Cancer Proteome

Despite numerous efforts to find a marker to diagnose breast cancer, similar to prostate specific antigen (PSA) for prostate cancer, no such protein has been found. Hence, it could be hypothesised that the reason for not finding a single reliable biomarker, or pattern to detect early breast cancer so far, is due to the heterogeneity of the disease. The mammary gland is a cellular ecosystem in which each cell type is constantly proliferating and differentiating (Figure 1.3), particularly epithelial cells, due to hormonal and growth factor influences [28]. Consequently, most breast cancers are carcinomas (malignant epithelial tumours). Furthermore, numerous different classifications of breast cancer in the epithelium are found depending on their origin and histology. These can be divided into two classes; *in situ* carcinomas of either ducal or lobular origin, which do not invade through the basement membrane; and *invasive* carcinomas, where the basement membrane is damaged or destroyed permitting cancer cells to invade surrounding tissue, leading to metastasis [29]. Invasive carcinomas are further classified according to the differences in their histological appearances. In addition, other cells/tissues, such as blood cells, blood vessels and fibroblasts, are usually found within a tumour and cause even greater cellular heterogeneity. This makes proteomic analysis of biopsies very problematic. Furthermore, local inflammation (common in breast cancers) and hormonal influences cause a constantly developing environment and make serological proteomics additionally challenging. It is important to realise that the blood proteome is constantly changing as a consequence of the pathophysiology of the system, subtracting from or modifying the circulating proteome. These disease-related differences might be the result of proteins being over-expressed and/or abnormally shed and added to the serum proteome. They may also arise from a specific process of protein clipping, degradation and/or proteolysis as a consequence of the disease process itself, or may be removed from the proteome due to abnormal proteolytic degradation pathway activation [30].

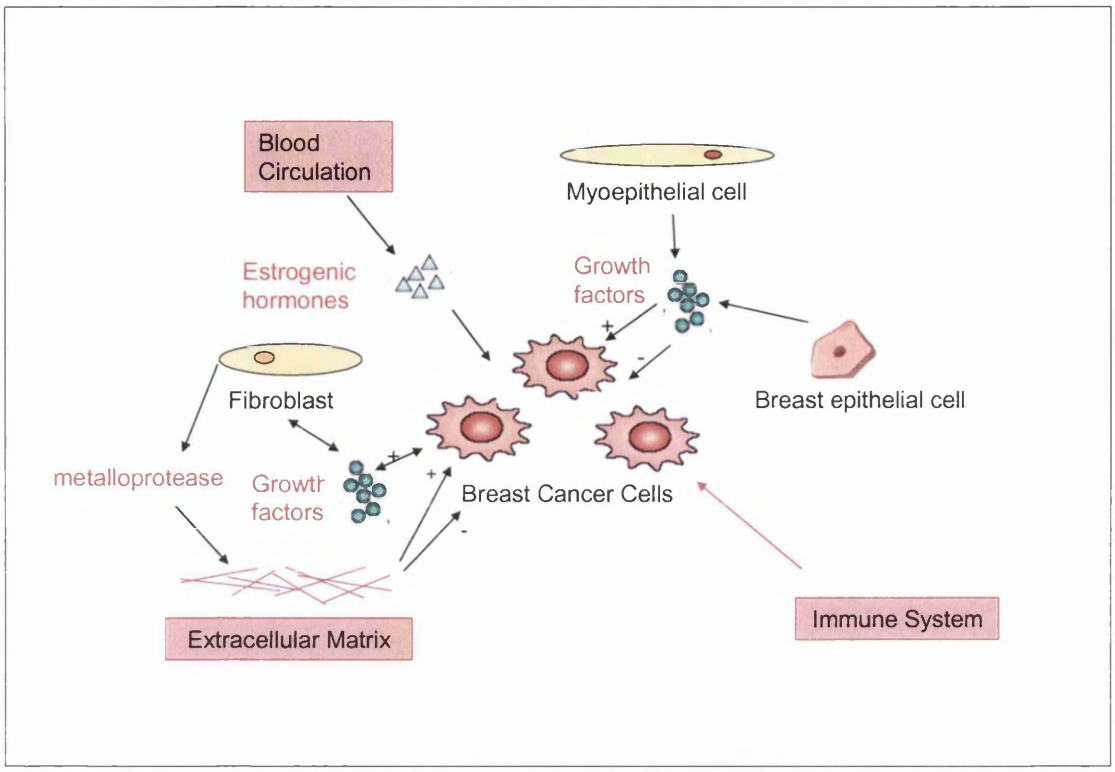


Figure 1.3: Breast cancer cells are constantly subjected to cellular interactions through hormones and growth factors. A + indicates a stimulation and a – inhibition of breast cancer cell growth.

1.1.5 Biomarker Discovery in the Literature

There are two ultimate goals for proteomic analysis of breast cancer: to identify new marker candidates for diagnosis and profiling disease, and to gain a greater understanding of the mechanism of cancer and the signalling pathways that initiate and lead to progression of breast tumours [28]. Surveying the entire proteome of a sample rather than searching for a specific protein may provide a greater chance of identifying markers or a diagnostic pattern. The core of proteomics involves the comparison of two clinically/biologically different samples (e.g. cancer vs. non cancer), in the case of breast cancer this could be tissue, cells or body fluids such as plasma, serum or nipple aspirate.

The first reported study of proteomics appears to have been conducted in 1974 [31] comparing clinical samples using 2-dimensional polyacrylamide gel electrophoresis (2D-PAGE) from serum, although, no proteins were identified. Mass spectrometry was not used for proteomic studies until 1993, where down-regulation of tropomyosin 1, 2 and 3 in mammary carcinomas was linked to breast neoplasia [32, 33]. Wulfkuhle J [34] identified 57 differentially expressed proteins between normal ductal/lobular units compared to ductal carcinoma tissue samples, using 2D-PAGE. Even though all markers were confirmed by immunohistochemistry in tissue samples, the technique lacks reproducibility due to lack of standardisation and the heterogeneity of the biological material. Hence, the data are not yet clinically relevant for diagnosis, treatment choice or prognosis. More recent examples of the use of 2D-PAGE include a study by Luo [35] where 25 proteins were found to be differentially expressed comparing infiltration ductal carcinoma tissue with normal breast tissue. Most biomarkers are quantitative and no protein has yet been found to be exclusively present or absent in breast cancer versus non-cancer tissue. Nevertheless one example where breast cancer proteomics has already identified a potential marker, is the molecular chaperone 14-3-3 sigma [36-39]. Using 2D-PAGE and matrix assisted laser desorption/ionization time-of-flight (MALDI-ToF) MS, 14-3-3 sigma was shown to be down-regulated in primary breast carcinomas and MCF-7 and MDA-MB-231 cell lines [39]. This has been confirmed by mRNA studies as an early event in breast cancer carcinogenesis [40]. There is also evidence that a loss of 14-3-3 sigma expression through epigenetic silencing or p53 mutations may lead to cancer formation. Thus 14-3-3 sigma may have more than one role in cancer formation [41]. However a more recent study, looking at the levels of expression of 14-3-3 sigma in primary breast tumours, using a proteomic approach complemented by IHC analysis showed that the loss of 14-3-3 sigma protein is not a frequent event in breast cancer [42]. To address the problem of heterogeneity of cells in the tumour, laser capture microdissection (LCM) has been used [43-46], although the technique requires a minimum of 100,000 cells for 2D-PAGE [43]. Also the decision on what is considered normal breast tissue in the constantly proliferating environment can be challenging.

Thus, many studies have used cell culture as a starting point to create a homogeneous and controlled cellular environment, to identify biomarkers that are later sought in tissue biopsies. Westley [47] identified a 46 kDa glycoprotein, secreted from breast cancer cells when induced by oestrogen, as the protease cathepsin D. Later it was

found that normal and cancerous epithelial cells produce different subgroups of keratin [48-50]. Furthermore, from cell culture two new breast cancer prognosis markers were found to be proteinase inhibitors TIMP1 and PAI-1 [51].

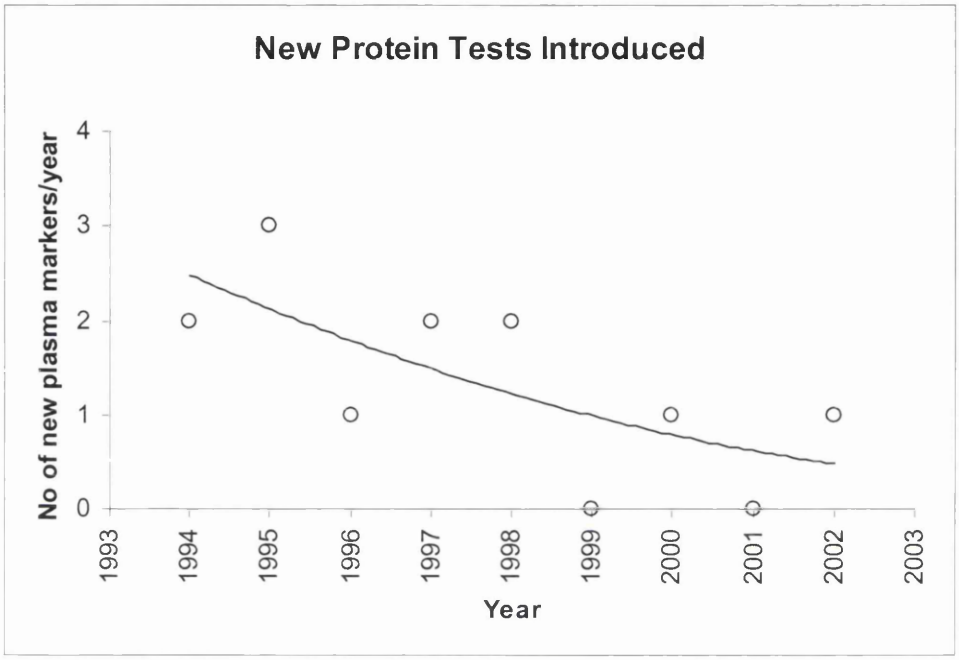


Figure 1.4: The declining rate of introduction of new protein tests data taken from Anderson and Anderson [52]

The rate at which new biomarkers have been discovered has declined dramatically over the last 10 years (Figure 1.4). Petricoin has described this by saying "The low hanging fruit have all been picked." Most of the estimated 10,000 serum proteins have not yet been detected by proteomic analysis, so finding a single biomarker for breast cancer is like searching for a needle in a haystack. Despite the numerous efforts to find markers for disease in serum [53-56], to date only potential serum biomarkers have been reported. It is even possible that a single informative biomarker to diagnose breast cancer does not exist but we ought to be looking for protein patterns [57]. This may be due to the fact that cancer is not an infectious disease, hence not a result of a foreign body entering the blood stream [58]. It may be more likely that cancer produces a pattern of changes, as a consequence of abnormal cellular mechanisms and the way by which the rest of the body reacts to change. Thus, even if the specific

protein pattern comprises products that are distant to the actual disease, they can retain specificity for the disease because this process can arise from a specific type of biomarker amplification. Importantly, recent findings indicate that the tumour–host microenvironment can generate cascades of enzymatic cleavage, shedding and sharing of growth factors. This interface could therefore be a source of low molecular weight biomarkers that are ultimately shed and amplified into the serum macroenvironment to bind with carrier proteins, enabling early disease detection and risk, severity and response assessment [57].

Recently surface-enhanced laser desorption/ionisation time-of-flight (SELDI-ToF) MS analysis has become a popular tool for protein pattern profiling: here the identification of the biomarker is secondary if a protein pattern can be used for cancer diagnosis. This technology was especially designed for biomarker discovery and possesses high sensitivity. The literature lists many examples of SELDI-ToF MS application in breast cancer [59-63]. In a study analysing 169 serum samples of stage 0-III breast cancer, healthy volunteers and benign breast cancer, SELDI-ToF analysis was able to identify 3 biomarkers (molecular weight: 4.3 kDa, 8.1 kDa and 8.9 kDa) [62]. These proteins could successfully distinguish between stage 0-I breast cancer from control sera and could again be used to identify stage II-III cancers. The biomarkers had an overall sensitivity of 85% for breast cancer with a specificity of 91%. Used in combination the sensitivity was increased to 93% and a specificity of 91%. Because of the multifactorial nature of breast cancer a combination of biomarkers may be beneficial. More recently two of these three biomarkers were validated in a completely independent study [64].

In a different study, using MALDI-ToF MS for identification of cell membrane proteins associated with breast cancer, three novel potential biomarkers (named breast cancer membrane protein BCMP11, BCMP84, and BCMP101) were discovered in the cell membrane [65]. With thorough validation, these could become diagnostic or used in cancer therapy through their membrane receptors. Similarly, candidate markers were discovered by 2D-PAGE in laser capture microscope cells from ER-negative, HER2/neu-positive tumour cells, which could be the driving force of more aggressive tumour proliferation [44, 46]. One of these, cytokeratin-19, a marker that has been use in immunohistochemistry in the past, has been found to be overexpressed in HER-2/*neu* positive tumours along with a number of other candidate proteins. This may be

an important finding to bring proteomics closer to finding markers that could become candidates in clinical settings.

1.2 Introduction to Serum Proteomics

1.2.1 Serum Proteins

Serum is a very complex mixture of proteins, and serum proteins from cancer patients may reflect the pathological state of the tumour as well as the body's response, and therefore provide earlier detection on top of staging of cancer [52, 66]. Blood serum hosts most major categories of protein which have been shed from many organs into the blood stream, including extracellular and cellular proteins. It has even been proposed to contain all human proteins, as well as proteins from other organisms such as bacteria, viruses or fungi [67]. The dynamic range of proteins in serum or plasma is believed to be one of the greatest of any biological system (Figure 1.5), ranging from < pg/ml level such as prostate-specific antigen, to high abundance proteins such as albumin or immunoglobulins in the > mg/ml levels [68]. There also appears to be a correlation between protein abundance and their biological classification or source (Figure 1.5). Sample handling and preparation is critical for proteome research. A plasma sample is prepared when blood is drawn in the presence of anticoagulants (EDTA, sodium citrate of heparin) and the red blood cells are removed by centrifugation. A serum sample on the other hand is obtained in the absence of anticoagulants and the blood is left to clot before centrifugation and removal of the red blood cells. The composition of serum and plasma greatly differs and the debate on whether to use serum or plasma for proteomics analysis is ongoing [69].

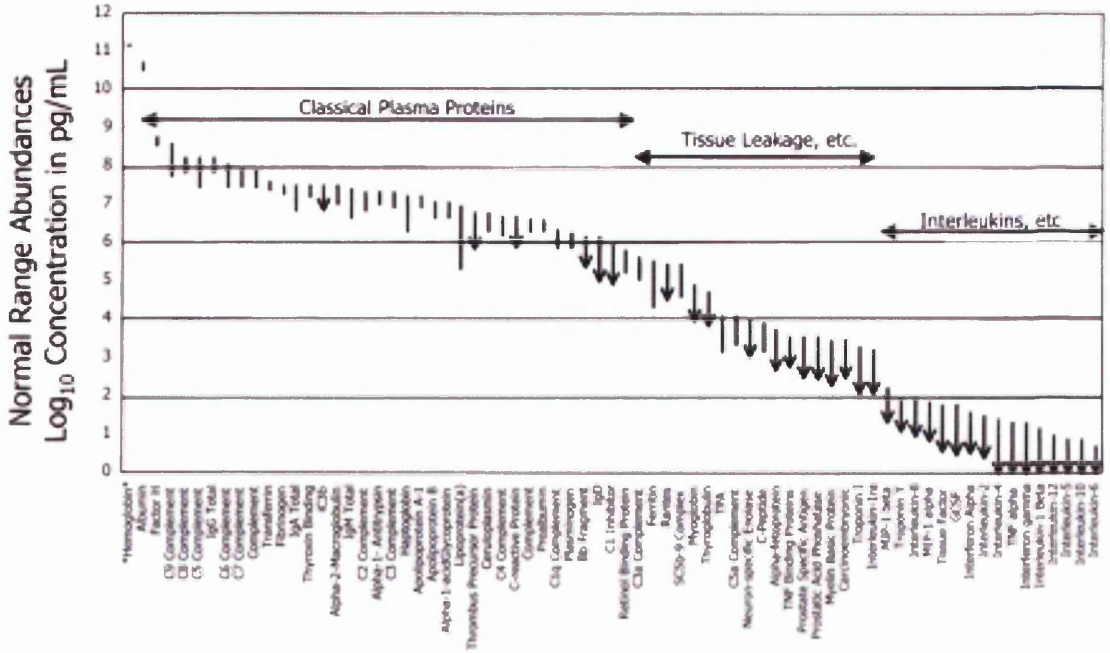


Figure 1.5: Reference intervals for 70 protein analytes in plasma. Abundance is plotted on a log scale spanning 12 orders of magnitude. Where only an upper limit is quoted, the lower end of the interval line shows an arrowhead. The classical plasma proteins are clustered to the left (high abundance), the tissue leakage markers (e.g. enzymes and troponins) are clustered in the centre, and cytokines are clustered to the right (low abundance). Haemoglobin is included (far left) for comparison [52].

1.2.2 Serum Complexity

Serum protein concentration ranges around 60-80 mg/ml, however 90% of the content is made up of only 10 proteins [70] (Figure 1.6). Human serum albumin accounts for almost 50% of the whole serum proteome, of the remaining 10%, further 12 high abundance proteins make up for 9%, of which most are well characterised (Figure 1.6). The final 1% of the serum proteome is composed of proteins that may be of clinical or biological interest (Figure 1.7). Unfortunately, the dynamic range of protein abundance in serum leaves complete characterisation of this proteome nearly impossible with current analytical methods. The high abundance proteins such as human serum albumin immunoglobulin G, antitrypsin, IgA, transferrin, and haptoglobin [71] mask the detection of the remaining proteome during 2D-PAGE and mass spectrometry. This is a particular problem because low abundance proteins and

peptides may have higher accuracy than traditional biomarkers for cancer detection [72].

Instead the proteome has to be broken down and analysed in smaller bits to recover a maximum number of proteins, with particular focus on the poorly characterised low abundance proteins. Serum and its component low abundance proteins therefore have an immense diagnostic potential.

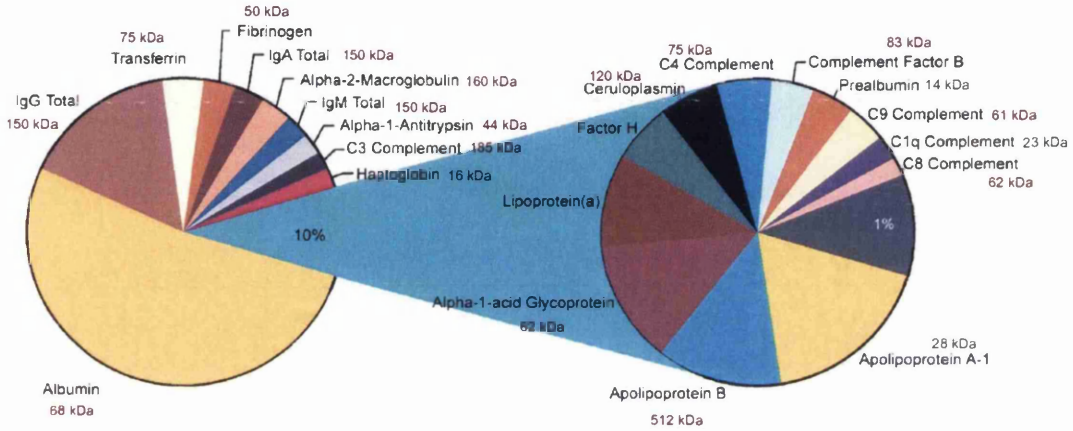


Figure 1.6: Pie chart representing the relative contribution of proteins within serum. Twenty-two proteins constitute 99% of the protein content of serum [58].

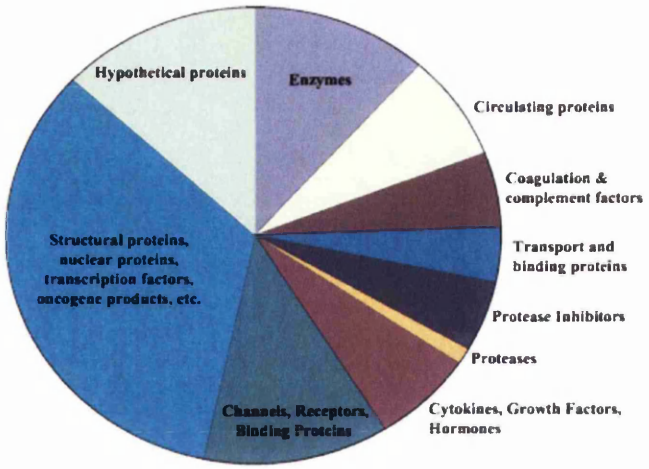


Figure 1.7: Pie chart representing the relative numbers of proteins identified within the LMW serum proteome [58].

1.2.3 Depletion of High Abundance Proteins

Solid phase extraction columns are widely used for depletion of high abundance proteins from serum and plasma. A number of different types of column are available, including, but not limited to, those based on ion-exchange, affinity ligands, dye-ligands or antibodies. Many of these are commercially available in form of columns of cartridges, microcolumns, 96-well plates and spin columns. Biological affinity separation based on antibodies or proteins is more specific, for example bacterial protein A and G have been shown to be successful at depleting IgGs from serum [68]. Using multi-affinity columns, with polyclonal antibodies for the 6 or 12 most common proteins in serum, has been used to deplete 70-95% of total serum proteins from serum [73-75]. Removal of albumin and protein G alone depletes only 70% of the total serum proteins, however only recently multi-affinity columns have become available. Even the use of the MARS columns from Agilent Technology has shown that after depleting the six most common serum proteins, the next most abundant ones become a problem and therefore highly selective columns depleting even more proteins are desirable. The Seppro™ spin columns are packed with 12 IgY antibodies coupled to microbeads [76]. High reproducibility and maintenance of the separation capacity has been observes. Over time these forms of depletion should improve protein identification and quantitation in serum.

1.2.4 The Low Molecular Weight (LMW) Proteome

As an alternative to affinity depletion, molecular weight cut-off (MWCO) centrifugal filters have also been used to remove high molecular weight proteins, such as albumin and Igs. The MWCO filters are used to separate the high and low molecular weight proteins by centrifugation. The use of denaturing acetonitrile (ACN) enables the release of proteins/peptides bound to larger proteins. Centrifugal ultrafiltration has been performed using a number of different MWCO filter sizes (10 - 50 kDa) and centrifugation speeds for 1000 – 12000 xg (Table 1.4). The technique has been demonstrated in a number of studies [30, 58, 77-80]. From the LMW fraction, Tirumalai *et al.* [58] identified 341 human serum proteins including a very low abundance protein, vasoconstrictor peptide endothelin-1, and remarkably no peptides originating from human serum albumin were identified. It is worth noting that ultra-filtration does not exclude all high molecular weight species from later analysis, peptides from proteins larger than the cut-off mass may also be identified. This may be for three reasons: one the filter has been ruptured and the filtration was unsuccessful, secondly due to its elliptical shape, a protein managed to slip through the filter or lastly the peptide is from a proteolytic fragment naturally occurring in serum [79].

Fractionation of the LMW fraction with ion-exchange-LC before subsequent RPLC-MS/MS has been used to successfully reduce the complexity of the resulting fraction and increase the number of protein identifications [58]. Harper *et al.* [81] used RPLC-MS/MS alone on the LMW fraction and identified 262 proteins from serum, including cytokines and other circulatory proteins. Proteins with higher masses than the cut-off filter were detected; however 75% of the identified proteins fell within the range of below 50 kDa.

In a similar way to albumin depletion, LMW ultrafiltration can also lead to the removal of bound proteins and peptides, since many LMW species are covalently bound to carrier proteins such as albumin and therefore may be found within the high molecular mass fraction of the serum proteome. In a paper, Zhou *et al.* [82] identified 63 proteins associated with albumin alone, and altogether 210 proteins, mapped from 378 peptides, were bound to the six most abundant proteins in serum. Only 6% of these identified proteins have previously been studied as biomarkers, the remainder may still have potential. Depletion of albumin and IgGs could result in loss of these

proteins. To get around this problem, a number of different denaturing conditions, to disrupt protein-protein interactions, have been used prior to centrifugal ultra-filtration (Table 1.4). Tirumalai *et al.* [58] showed that under the right denaturing conditions, protein-protein interactions between carrier proteins, such as albumin, and their cargo could be disrupted thus enabling greater enrichment. Denaturing of the protein-ligand bonds with 25mM NH_4HCO_3 and acetonitrile (ACN) buffer enriches the sample with proteins and peptides formerly bound to carrier proteins such as albumin. Acetonitrile is crucial in the LMW enrichment [58]. The importance of ACN in protein separation was further shown by a proof of principle study precipitating HMW proteins with ACN to gain access to the LMW proteins before centrifugal ultra-filtration [83].

Table 1.4: Centrifugal ultrafiltration in the literature for high molecular weight protein depletion. Denaturing conditions and use of MWCO filters varies in different studies. The dilution factor, type of denaturing buffer, centrifugation speed and time, filter type and pre- and post-procession are shown.

Reference	Dilition (v/v)	Denaturant	Before filtration	Ultra-filtration (xg)	Time	Filter type	After filtration
Morris 2004	3:2	20% ACN	heat for 15min @ 40C; centrifuge; condition filters with 0.1M NaOH	2000	90 min	Microcon	
Tirumalai 2003	1:5	25mM NH4HCO3, 20%ACN, pH 8.2		3000	until 90% passed	Centricon 30kDa	2DLC-MS/MS
Johnson 2004	1:4	25mM NH4HCO3, 20%ACN, pH 7.6	vortex	12,000	20min	Microcon 10kDa	
Zhou 2004	1:3	H2O	Albumin depletion, elute HSA 0.2% v/v FA + ACN 1:1, boil for 10min, dilute 1:2	1000		Centricon 30kDa	
Yeo 2004	1:1	saline solution				Centricon 50kDa	
Kaiser 2004	1:50	4M urea, o.1M NaCl, 0.0125% ammonia		3000	until 4/5 passed	Centricon 30kDa	C2 cartridge, 50%ACN, 0.5%FA
Harper 2004	3:2	20% ACN	heat for 15min @ 40C; centrifuge; condition filters with 0.1M NaOH	2000	90 min	Microcon 50kDa	
Merrell 2004	1:2	100% ACN	vortex 5 sec,stand at RT for 30 min	12,000	10min	20µ of 88% Formic acid, 2µL of 5pmol/µL mellitin, and 2µL of 5pmol/µL Glu-fibto lyophilized samples. Brought to 40µL with H20.	
Mehta 2003		50% ACN	30ul serum sepadex G25 or G50 mol sieve column for 3min at 3000xg	1000		Microcon 30kDa	

Enrichment of LMW species by releasing them from their molecular carriers is furthermore important for protein recovery from a physiologic perspective, as free phase LMW molecules should be rapidly cleared through the kidney, and this may significantly reduce the concentration of free-phase low molecular mass biomarkers to a level below detection. The abundant high molecular mass carrier proteins (such as albumin) exist above the cut-off for kidney clearance, and hence possess a half-life that is many orders of magnitude greater than small molecules [30]. Analysis of the LMW proteome may provide extra information that is not available from crude serum.

1.3 Introduction to Chromatography

To isolate proteins for analysis, samples from any source have to be fractionated. Optimally the separation should be so efficient that each fraction contains only one protein, which would make identification by mass spectrometry very simple, however in practice, complex mixtures are difficult to separate and fractions can contain many hundreds of proteins. Polyacrylamide gel electrophoresis was the first method for protein separation, several thousands of proteins could potentially be separated [84]. However this method has limitations, and more recently HPLC has been optimized for more high-throughput separation of proteins. The use of different chromatography columns allows a greater range of protein coverage, although not necessarily in one step. Both techniques are discussed in more detail below.

1.3.1 2D Gel Electrophoresis

In 1930, Tiselius [85] introduced the moving boundary method for electrophoresis of proteins. Since his pioneering work, various forms of electrophoresis have been used for the separation of protein mixtures. The steady increase in resolution can be accredited to the introduction of acrylamide gels, stacking systems, isoelectric focusing (IEF) and a variety of 2D gel electrophoretic separations [86]. The uses of 1D sodium dodecyl sulphate (SDS) and subsequent 2D-PAGE have made it possible to compare the proteome of two complex mixtures in a single separation step (Figure 1.8). Intact proteins are separated first along a pH gradient according to their isoelectric point by IEF. Then the IEF strip is laid on top of a SDS gel where the proteins are separated by molecular mass through electrophoresis. Next the gel is stained to visualise the proteins for quantitation, before they are excised for identification. The excised gel spots are digested with trypsin and the resulting peptides analysed by MS for identification from the database.

However there are limitations to the utility of gel electrophoresis, it has been shown that 2D-PAGE does not separate all proteins: acidic, basic, very large or small proteins are not separated well by 2D-PAGE. Furthermore, although 2D-PAGE is able

to resolve thousands of proteins, Gygi *et al.* [87] showed that only the most abundant proteins are visualised by staining and can be identified by MS. Hence 2D-PAGE is a useful tool for identification of less complex samples and tissue [88], but its use is problematic in serum analysis. Not only do all proteins not separate and stain in the gel, but serum is so complex that many proteins migrate to the same spot and are hard to identify. Pre-fractionation of the sample before loading it onto the gel can be employed to improve resolution, for example Chernokalskaya *et al.* [89] used centrifugal ultrafiltration for a more thorough proteome analysis of different molecular weight fractions prior to 2D-PAGE separation. This way the group identified more than 340 serum proteins.

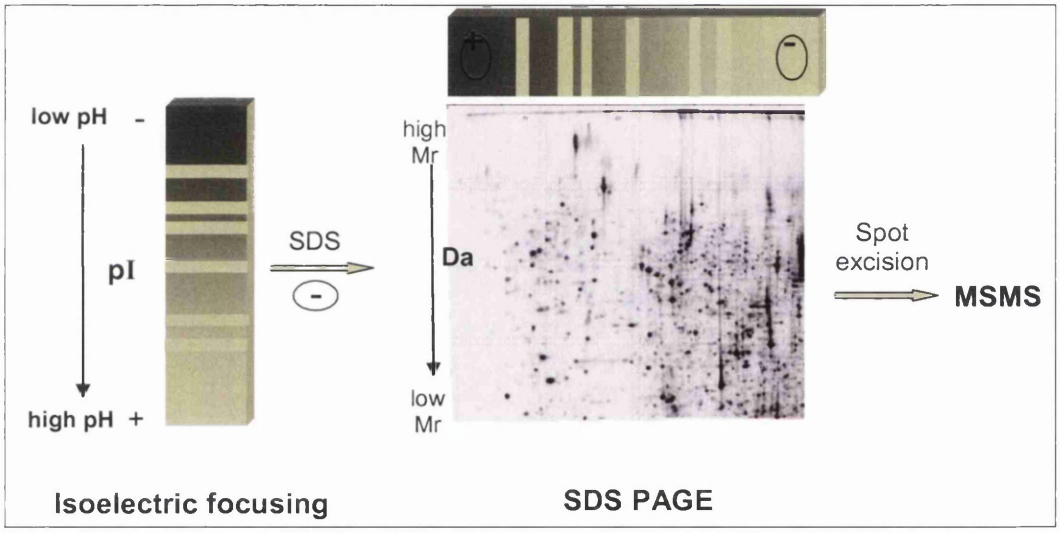


Figure 1.8: 2D Polyacrylamide Gel Electrophoresis (PAGE). Proteins are separated according to their pH by IEF and then in the second dimension depending on their molecular weight by gel electrophoresis. Protein spots are then visualized for comparison with Coomassie Blue or silver staining.

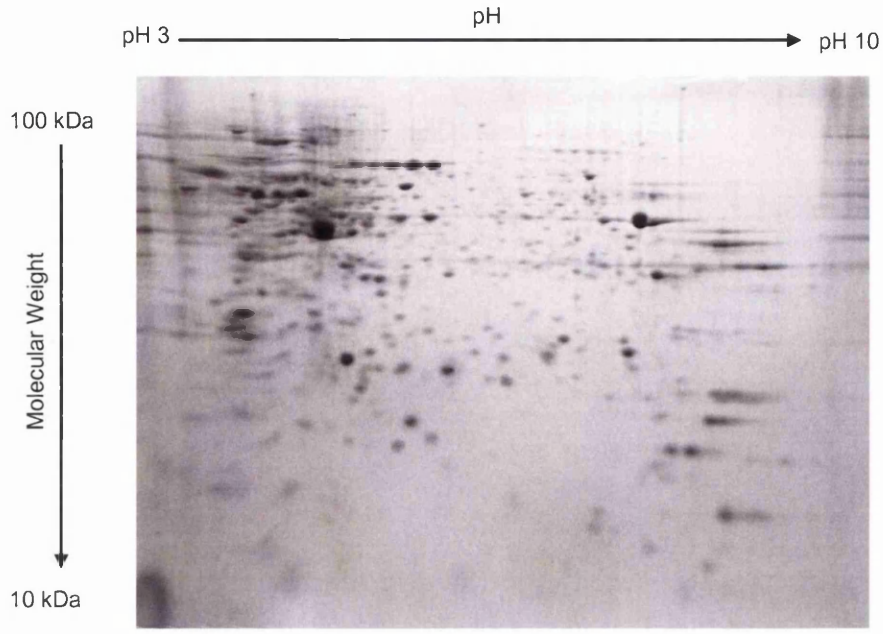


Figure 1.9: 2D-PAGE: Individual proteins separate isoelectrically along a pH gradient across the top of the gel and then dependent on their molecular weight vertically through the gel. Two gels can be compared for expression levels and spots identified by mass spectrometry.

1.3.2 High Performance Liquid Chromatography (HPLC)

Multidimensional high performance liquid chromatography (HPLC) separation can provide higher resolution and greater separation power than 2D-PAGE. HPLC fractionation can be performed on proteins and peptides (as well as other molecules), to achieve separation due to their chemical properties such as charge, polarity, size or certain affinities. The analytical column is packed with a stationary phase that binds the protein/peptide after injection. When the sample is loaded, the equilibrium is towards binding to the stationary phase. The mobile phase, a solvent or buffer that increases in concentration (salt, organic solvent or pH buffer), slowly shifts the equilibrium away from binding to the stationary phase and the molecules start to move into the mobile phase than bind to the column. Over time, different molecules are pulled off the column and collected in separate fractions, therefore separating them out. During 2D-LC separation, the eluted fractions are either collected and individually (offline) loaded on a second column or directly injected (online) onto another column as they elute from the first. To identify the proteins or peptides in the

fractions, they can then be analysed by MS. A commonly used form of multi-dimensional HPLC-MS is to use an ion-exchange column, a strong cation exchange (SCX) for peptides or weak anion exchange (WAX) for proteins, as a first step and then to separate the peptides/proteins further on a reverse phase (RP) column. Intact proteins may also be enzymatically digested before loading the second column.

In the analysis of peptides, trypsin digestion is a critical step. Prior to trypsin digestion, solubilisation and alkylation steps have been described as crucial to ensure good recovery of peptides [90-92]. Additionally, Qian *et al.* [93] described that alkylation of proteins may provide better coverage of cysteine-containing peptides and hence an increase in protein identifications.

For peptide mass fingerprinting, trypsin is the most widely used digestion enzyme, however other enzymes (e.g. chymotrypsin or Endo Arg N) or chemicals (e.g. cyanogens bromide) have been used [94]. Protein identifications can be maximised through offline pre-fractionation of the serum sample by size, polarity or charge prior to digestion to reduce the complexity. It is important to digest and analyse the fractions separately. However 2D-LC fractionation of peptides online has also been used as well as multidimensional peptide identification technology (MudPit) separation [95, 96]. Washburn *et al.* (2001) detected and identified a total of 1,484 proteins in the yeast proteome using a MudPit column, where SCX and RP LC separation are performed on the same column (Figure 1.10). Nevertheless offline separation of proteins followed by digestion of the fractions before additional RP separation appears to be more useful, as the complexity of the sample is reduced before enzymatic digestion.

Chromatographic columns are commercially available but to reduce costs and for more customized applications can be packed by hand using a high pressure pump. The use of pulled-tip columns as a part of the electrospray source allows direct spray into the mass spectrometer (Figure 1.10).

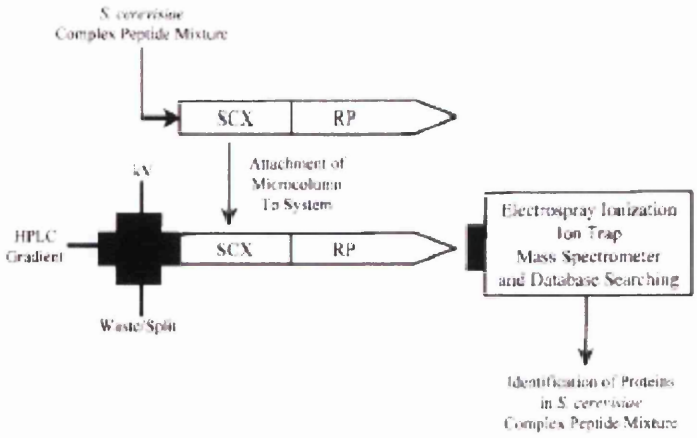


Figure 1.10: Multi-dimensional protein identification technology (MudPIT). Ion-exchange and reverse-phase chromatography is performed on the same column. Peptides are separated in two dimensions before MS analysis. Picture taken from [96].

1.4 Introduction to Mass Spectrometry (MS)

After sufficient separation, the molecules can be identified, quantitated or their structure and chemical composition elucidated by mass spectrometry. This powerful analytical technique can detect compounds at minute quantities (as little as 10^{-12} g or 10^{-15} moles for a compound of 1000 Dalton (Da) mass). This means that compounds can be identified at very low concentrations (one part in 10^{12}) in chemically complex mixtures, which is important for identification of low abundance proteins.

The technique of mass spectrometry had its beginnings in J.J. Thomson's vacuum tube where, in 1897, he demonstrated the existence of electrons and "positive rays" [84]. In his 1906 Nobel Prize-winning experiment, he discovered the electron and determined its mass-to-charge (*m/z*) ratio. However, the primary application of mass spectrometry remained in the realm of physics for nearly thirty years. Today, for molecule analysis after separation, a HPLC instrument can be coupled to the MS, so that ions are directly sprayed into the source.. This was first established in 1974 by P.J. Arpino [97]. Alternatively the fractionated proteins/peptides can be spotted on a solid plate or chip for analysis by MALDI-ToF MS or SELDI-ToF MS.

The mass spectrometer is an instrument that measures the masses of individual molecules that have been converted into ions, i.e., molecules that have been ionized. The molecular weight of an ion is not measured directly, but rather the mass-to-charge ratio (*m/z*) of the ions formed. The charge on an ion is denoted by z, and *m/z* therefore represents Da per unit of charge. In many cases, the ions encountered in mass spectrometry have just one charge (z = 1) so the *m/z* value is numerically equal to the molecular (ionic) mass in Da, as generally for MALDI-ToF MS.

For peptide identification the ion can be further fragmented in a second mass analyser using collision induced dissociation or a second ToF tube (Figure 1.11). The *m/z* ratio of the fragment ions represents fragments that have lost one or several residues. For MALDI-ToF MS, a ToF/ToF consists of two successive ToF accelerations. The first acceleration selects, isolates, and fragments (usually by collision with a neutral gas) a precursor ion of interest. The second acceleration reaccelerates the precursor ion and fragments, then measures the masses and intensities of the fragment ions.

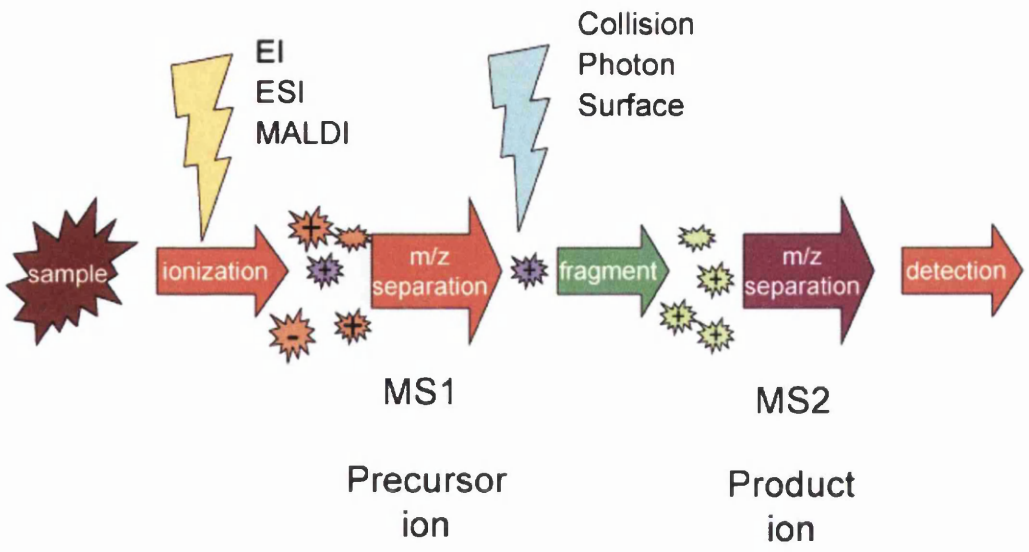


Figure 1.11: Tandem mass spectrometry. The first MS selects the precursor ions for fragmentation in the collision cell and the second MS then produces the mass spectrum of the fragment ions; diagram taken from ASMS website [98].

During tandem MS the peptide is fragmented into ions, that have lost one or several amino acids. The types of fragment ions observed in an MS/MS spectrum depend on many factors including primary sequence, the amount of internal energy, how the energy was introduced, charge state, and the type of MS instrument used. An ESI-MS using collision-induced dissociation for example produces mainly *b* and *y* ions, as it is uses low energy for fragmentation [99].

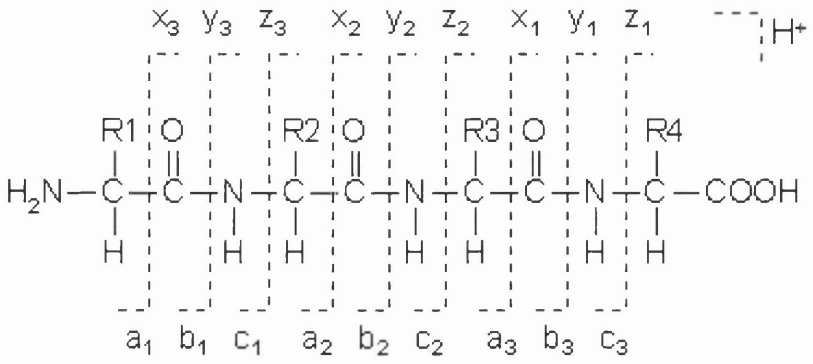


Figure 1.12: Main fragmentation path of peptides in CID/MS/MS. (Diagram was taken from Matrix Science website http://www.matrixscience.com/help/fragmentation_help.html)

Fragments will only be detected if they carry at least one charge. If this charge is retained on the N terminal fragment, the ion is classed as either *a*, *b* or *c*. If the charge is retained on the C terminal, the ion type is either *x*, *y* or *z*. The subscript indicates the number of amino acids in the fragment (Figure 1.12).

In addition to the proton(s) carrying the charge, *c* ions and *y* ions abstract an additional proton from the precursor peptide. The structures of the six singly charged fragment ions produced through cleavage in the peptide chain are shown in Figure 1.13. The mass difference between the consecutive ions with in a MS/MS spectrum is used to determine the peptide sequence of the precursor ion.

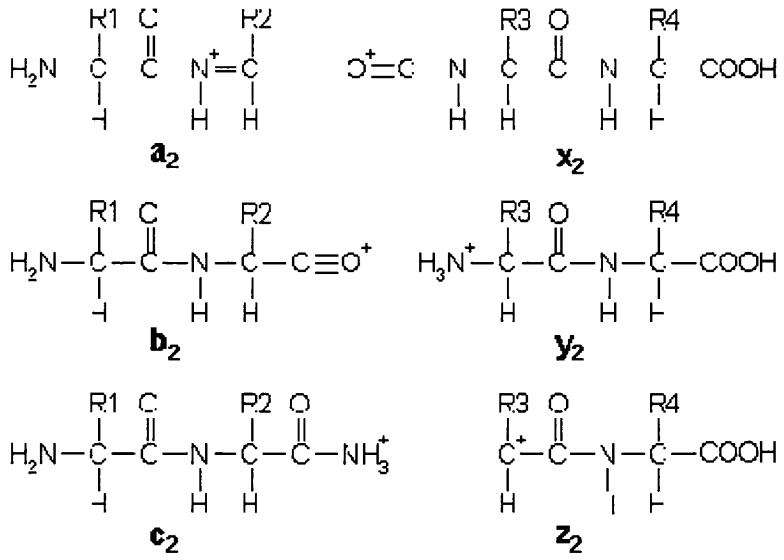


Figure 1.13: Sequence ions produced by fragmentation using a mass analyser. The structures include a single charge carrying a proton. (Diagram was taken from MatrixScience website http://www.matrixscience.com/help/fragmentation_help.html.)

For protein identification, the fragment ion pattern obtained from MS/MS can be compared to the sequence predicted for all the proteins contained in a database. In May 2007 the human UniProt database contained 468,000 protein sequences; the databases are constantly updated by the European Bioinformatics Institute (EBI) for redundancy and protein annotations [100]. Several algorithms have been developed for interpretation of MS/MS spectra from peptides [84]. For the work in this thesis, the Sequest algorithm was exclusively used to compare the observed masses with those expected from a known sequence in the database [101]. Sequest uses a unique approach to correlate the fragmentation spectrum to the peptide sequence contained in the database. The database is searched for peptides with the same mass as the precursor ion in the MS/MS spectrum. Then the observed fragmentation spectrum is compared to the predicted spectrum in the database and a similarity score is calculated. The peptides identified from the same protein are combined to predict a protein and provide a protein probability score. The more peptides from one protein are identified, the more confidence is in the protein match. Some studies claim to confidently match a protein from only one peptide; others use at least two or more peptides to identify a protein. Often the number of proteins identified in a sample depends on the search and filter stringency criteria used.

For this thesis four different types of MS were used. Two time-of-flight instruments (MALDI-ToF and SELDI-ToF MS) and two electrospray (3D-linear ion trap and Q-ToF MS), these are described in more detail below.

1.4.1 Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-ToF) MS

In 1987, Hillenkamp and co-workers [102] discovered that molecular ion species can be produced from large proteins by laser desorption, without much fragmentation, if these molecules are mixed with small organic compounds that serve as matrices. The requirements of a matrix are that it has a strong absorbance at the laser wavelength and is of low enough mass to be sublimated [103]. This process is now called matrix-assisted laser desorption/ionization mass spectrometry. The typical preparation protocol for MALDI-ToF MS is to prepare the sample in liquid form, and then to prepare a matrix solution that contains small organic compounds such as α-cyano-4-hydroxycinnamic acid or sinapinic acid. The analyte and the matrix are mixed and a small amount of solution is placed on a metal plate to dry. After the matrix crystallises, the sample plate is analysed inside the MALDI-ToF mass analyser. During the analysis process, the matrix material strongly absorbs the laser energy and quickly becomes vaporised (Figure 1.14). The proteins/peptides are embedded in the matrix and carried along in the fast vaporisation process. The molecules pick up a positive charge and travel down the time-of-flight tube, where they are analysed on basis of their *m/z* ratio (Figure 1.15).

MALDI-ToF MS can potentially obtain masses for numerous biopolymers including oligosaccharides, gangliosides, peptides and proteins that range from ~500 to 100,000 Daltons [84]. Under optimum conditions, the limit of sensitivity of tryptic peptides (below 5,000 Da) is approximately 10-50 fmol. Because of interference from the matrix, the lower limit of the mass range is about 500 Da, below this, matrix adducts dominate the spectrum.

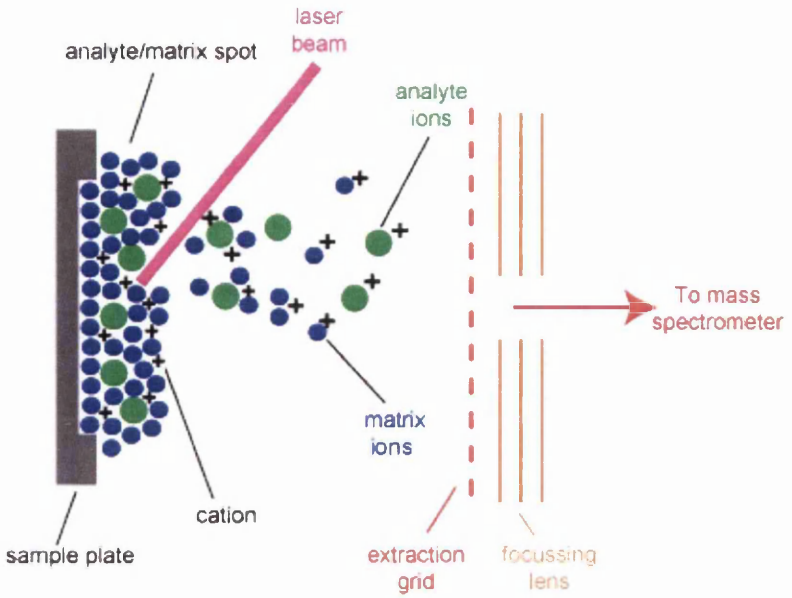


Figure 1.14: Schematic of MALDI-ToF plate and matrix vaporisation. A nitrogen laser (red line) is fired at the matrix/analyte deposited in the plate in a vacuum and the excited ions move down the time-of-flight (ToF) tube to the detector. (Diagram reproduced from University of Bristol website [104])

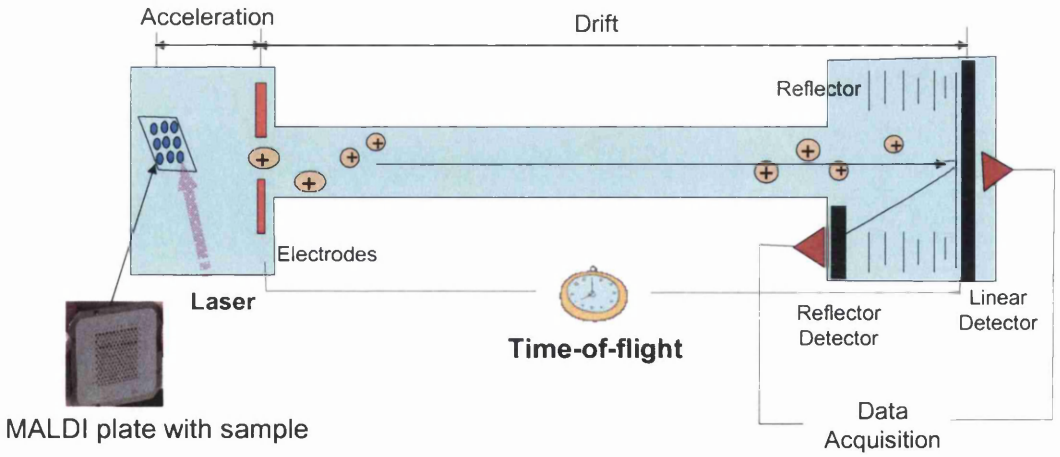


Figure 1.15: Schematic of a MALDI-ToF mass analyser. Ions travel down the ToF tube where they hit the detector (linear) or are reflected to another detector (reflectron) for more accurate mass measurements.

To enable the charged ions to move down the flight tube in the mass spectrometer, after laser excitation, they are accelerated in an electric field (i.e., 25 kV). MALDI-ToF MS can be done either in positive or negative ion mode but positive ion mode is more commonly used in proteomics experiments. MALDI-ToF MS analysis of peptides is suitable for the purpose of identification; however it is not (absolutely) quantitative, as there can be large differences in how well different proteins/peptides ionize.

For protein analysis, MALDI-ToF MS analysis is mostly performed in linear mode. The ions travel down a linear flight path and their mass/charge ratio is determined by the time it takes for them to reach the detector (Figure 1.15). Hence, this instrument is called a time-of-flight instrument. Because all the ions are exposed to the same electrical field, all similarly charged ions will have similar energies; therefore ions that have a larger mass have lower velocities and hence will require a longer time to reach the detector. MALDI-ToF is probably the most used MS technique for identification of proteins in 2D-PAGE spots.

1.4.2 Surface-Enhanced Laser Desorption/Ionization- (SELDI) ToF MS

SELDI-ToF MS is a form of soft ionization, patented by Ciphergen Biosystems Ltd. for rapid analysis of peptides, proteins and other molecules. The technique, originally described in 1993 [105], again relies on time-of-flight MS for the accurate measurement of the *m/z* ratio of peptides and proteins (Figure 1.16) but SELDI-ToF MS involves the binding of proteins and peptides present in complex biological samples, such as serum, cell lysates, tissue homogenate or culture supernatants, to ProteinChip© arrays. These arrays have certain chromatographic surfaces, similar to the interior of a HPLC column, and proteins bind to the surface. Many types of samples can be applied directly to the ProteinChip© arrays, without the need for prior removal of salts or detergents which typically interfere with other types of MS methods. An advantage of SELDI-ToF MS is that it only requires very small sample volumes [106]. Proteins and peptides in the sample bind non-covalently to the surface of the arrays depending on their biochemical properties, (e.g., acidic proteins can bind to cationic surface arrays). Unbound peptides and proteins, as well as salts, detergents and other contaminants, are then washed away from the surface of the arrays.

ProteinChip® arrays are available with a variety of chromatographic surfaces, such as reversed phase or ion-exchange, or with pre-activated surfaces (e.g. antibody capture or protein-protein interactions). After preparation of the ProteinChip® the arrays are read in the mass analyser. The MS is essentially a MALDI-ToF, a matrix is used for laser desorption/ionization. The software provided with commercially available SELDI-ToF system, allows peak alignment, normalisation and semi-quantitative comparison of two sample cohorts. The protein peaks can be visualised as peaks or gel bands and individual peaks can be highlighted for statistical analysis with in the software (Figure 1.17) or the data can be exported and further analysed using Excel or another statistical analysis program.

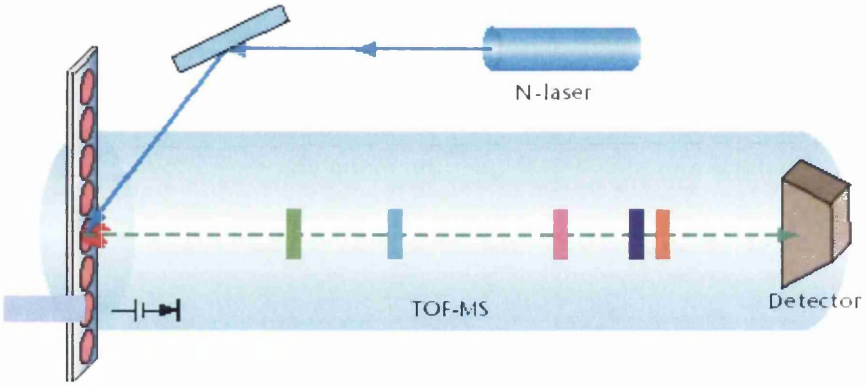


Figure 1.16: SELDI ProteinChip® array and time-of-flight mass spectrometry [106].

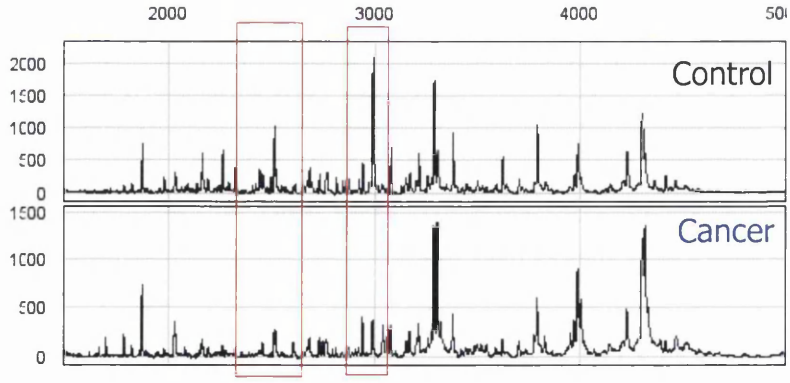


Figure 1.17: Mass spectrometry can show protein expression differences between two samples. Protein levels are quantitated as maximum peak height.

1.4.3 Electrospray Ionization Mass Spectrometry (ESI-MS)

Probably the most commonly used form of peptide identification and mapping of protein is electrspray ionization (ESI)-MS/MS. Electrospray for ESI-MS was developed in 1988 by J. Fenn where he showed the detection of a spectrum from proteins even above 20,000 Da [84]. Electrospray coupled to a HPLC system or directly injected produces ionized peptides in the liquid, by applying a voltage, as they exit the column. The ions are repulsed by the high voltage (as they carry the same polarity) and are attracted by the lower voltage of the MS. This way the droplets become smaller until only a fine mist enters the MS (Figure 1.18). As the droplets pass through a heated metal capillary, the solvent evaporates, and the droplets become over-charged and fission into ever-smaller droplets until only single, usually multiply-charged peptide ions (e.g., 2+, 3+ for tryptic peptides) are left (as opposed to MALDI-ToF MS, which only produces single charged ions from peptides).

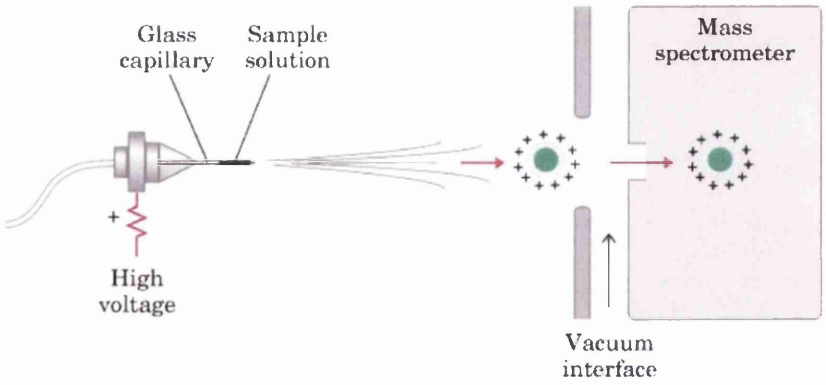


Figure 1.18: Electrospray ionization and mass spectrometry.

Even though intact proteins can be analysed, accurate mass measurements are not possible. It is further important to recognise that for large molecular weight molecules, such as proteins, the mass measured is the average mass and that the peak envelope extends over isotopic masses.

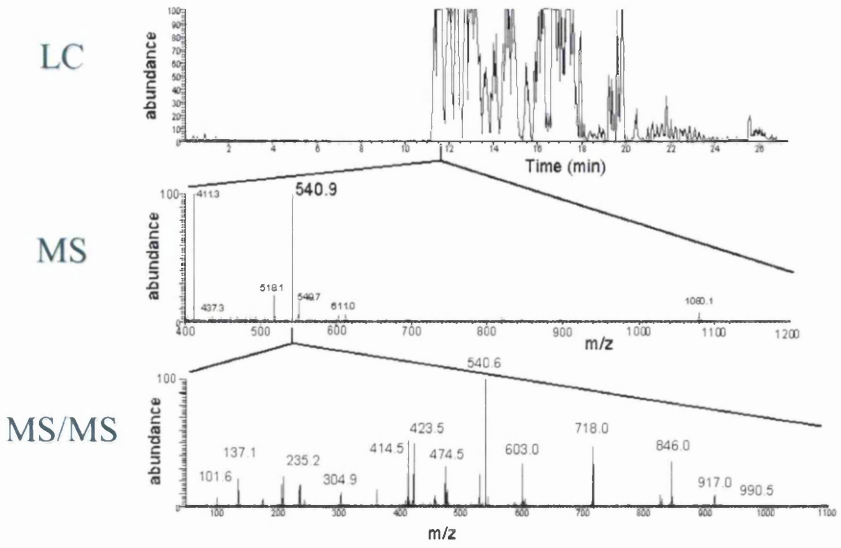


Figure 1.19: LC-MS/MS analysis. The diagram shows the transition between the basepeak chromatogram of the LC separation and the first stage of MS where parent ions are selected through to the second MS producing fragment ions in the MS/MS experiment.

In complex mixtures, the recovery of MS/MS scans and hence peptide identifications from ESI-MS/MS is low, due to the large number of peptides that have to be analysed and fragmented at one time. Hence optimising the separation by extending the elution gradient becomes important to spread molecular ions out. ESI-MS spectra of peptides are almost always recorded in positive ion mode. ESI- tandem MS analysis can be performed in "data-dependent mode", where the most intense *m/z* peaks are selected for subsequent MS/MS but then ignored for a certain amount of time (Figure 1.19). During "data-dependent mode" no peaks are analysed for MS while MS/MS is in process, this is called the "exclusion time". Alternatively, peaks are scanned for MS/MS in "selective ion fragmentation", where certain, chosen *m/z* precursors are selected by the software for MS/MS analysis. This becomes useful to identify a peptide of known *m/z*. ESI-ion trap mass spectrometers also record the mass-to-charge ratio, typically producing a spectrum in the mass range of 500-2000 Da for peptides. For peptide sequencing the *m/z* peak is further fragmented and amino acid loss is used to identify the amino acid composition; which are then used to map proteins from a database commonly using Sequest as described above [101].

1.4.4 Serum Protein Profiling using LC-MS/MS in the Literature

Attempts at serum protein identification first started in the 1970s when Anderson and Anderson [107] discovered 40 proteins in un-depleted blood serum by 2D-PAGE. Thirty years later still only 490 proteins from immunoglobulin and albumin depleted serum were identified in 1992 [108] and again in 2002 [68] by 2D-PAGE (Figure 1.20). However serum is estimated to contain many thousands of proteins, based on genomic data [68]. With the use of multidimensional HPLC separation techniques, Shen *et al.* [109] showed that, from un-depleted serum, they could conservatively identify 800 serum proteins (including all serum proteins). Using multidimensional HPLC-separations, trace components with a dynamic range of >8 orders of magnitude were detected. In a similar approach, using more conservative identification criteria, Qian *et al.* [93], confidently identified 804 different plasma proteins (not including immunoglobulins) covering a dynamic range of 6-7 orders of magnitudes. Nevertheless this is still far less than the total number of estimated plasma proteins. These studies, from different laboratories, used increasingly more conservative filtering criteria during their database searches, to minimise the number of false positive identifications. A Sequest database search can be performed at different stringency levels which determine the number of proteins identified. For example, to eliminate false positives Tirumalai *et al.* [58] searched all spectra against the human FASTA database without removing viral proteins, and each protein that also identified a virus was then removed from the protein identifications. Furthermore, proteins that were identified by only one peptide were analysed manually for their signal-to-noise ratio; and the presence of at least 3 consecutive fragment ions. Other search methods include increasing the search criteria for different charge states and reverse sequence database searching [110]. The idea is if two different peptides can be found from the same raw file and the exact same scan number then it is possible for a false positive. The data is first searched with one sequence database and then the reverse sequence database to see if there are any MS/MS scans that are mapped back to two different peptide sequences.

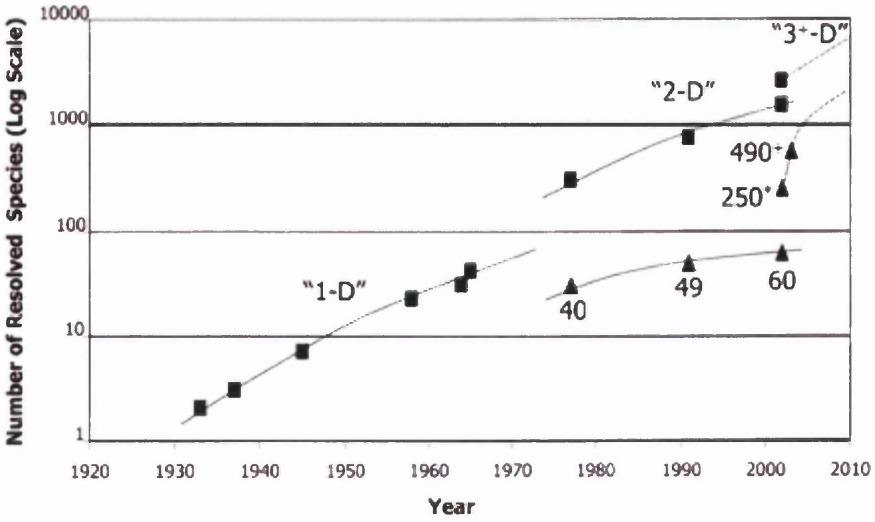


Figure 1.20: Protein Markers over 70-year period. The increase in protein species resolved and identified in plasma. Squares show the number of peaks, bands, or spots resolved based on literature. Diamonds show the number of proteins identified [52].

The use of more stringent criteria has slowed the increase in number of identifications greatly. So when Shen *et al.* [109] used the less stringent search criteria than Adkins *et al.* [68] had in their study, they identified 1600 instead of 800 (stringently searched) different proteins. This emphasises the need for powerful bioinformatics tools and stringent search criteria to confidently map proteins. It further suggests that these criteria need to be conserved and standardised across laboratories to compare results. Nonetheless, Shen *et al.* [109] identified low abundance proteins in ng/ml quantities such as human growth factor activator and alpha-fetoprotein confidently in the presence of serum albumin in the more stringent search. Although, to identify cytokines and other inflammatory mediators such as interleukin-1 and -6 (IL-1, IL-6), present in pg/ml levels, in serum or plasma, further developments are necessary. At the moment rare proteins are often only identified by a single peptide and few amino acid peaks, and only offers low confidence identification [96].

The use of different proteomic approaches results in identification of different proteins. A paper by Anderson *et al.* [67] compared protein identifications from different methods across laboratories. One of the main conclusions drawn is that with a dynamic range and complexity greater than that of any other biological material and with the technology available today, none of the methods were able to discover

anywhere near the protein distribution of the whole proteome. More concerning they could only find a small overlap of protein identifications between the different studies (Figure 1.21). As the data was collected from independent laboratories, the differences in identification may have resulted from either the presence of different proteins in the sample, or different sample preparation and separation methods, or from different MS analysis and searching tools and criteria [67]. Most likely, different technologies are capable of detecting proteins with different properties; hence a multi-technology approach has the greatest potential at discovering the maximum number of proteins.

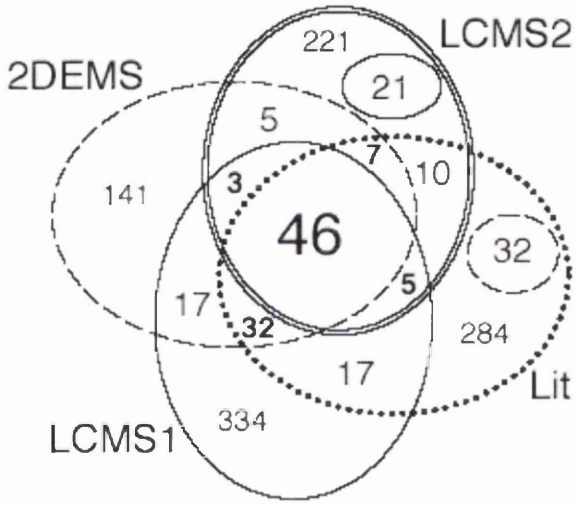


Figure 1.21: Diagram of Proteins Found in Multiple Datasets. All overlaps are shown (2-way, 3-way and 4-way) for all four input data sets: Literature search (dotted line); 2D-PAGE MS (dashed line); LC-MS (solid line) and LC-MS/MS (double solid line). Numbers represent the number of shared accessions in the respective overlapping areas. Diagram taken from [67].

1.5 Serum Protein Quantitation and Biomarker Discovery

Mass Spectrometry has its strength in the qualitative analysis and identification of molecules and not for relative quantitation of molecules in complex mixtures such as serum. Relative quantitation of comparisons of protein expression between two sample cohorts is therefore challenging and with the exception for SELDI-ToF MS data little published. In fact at the start of this project MALDI-ToF MS and LC-MS were untested for protein or peptide quantitation from clinical serum samples.

In a traditional proteomics study, proteins are first resolved by 2D-PAGE, and their expression levels are monitored by the intensities of stained spots on gels. However 2D-PAGE is not suitable for serum quantitative analysis, because low molecular weight molecules exist below the lower limit of effective resolution achieved by conventional 2D-PAGE [86]. Furthermore the large amount of albumin can mask the detection of low abundant proteins. Consequently, investigators have turned to HPLC coupled with MS. Separation methods such as multidimensional LC can provide separation power superior to that of 2D-PAGE. Mass spectrometry actually has its optimal performance in the low mass range [111]. Especially SELDI-ToF MS, analysing intact proteins, performs best on proteins below 30 kDa [112].

Mass spectrometry dominates the field of proteomics as an analytical tool. In the past few years, a number of methods based on isotopic labelling for proteins and peptides have been reported for comparing the relative abundance of proteins in two different biological groups [67, 80, 113-116].

The basis of labelling approaches for proteomic studies is the generation of two proteome pools, one unlabelled, the other isotopically tagged (now up to 8 samples can be labelled using the new iTRAQ reagents (Applied Biosystems, UK)) that behave indistinguishably during separation. In principle, isotope tags can be incorporated into proteins during cell growth [117] or after cell lyses as well as in biofluids, or even in intact animals. In a study monitoring individual protein turnover rates, Doherty *et al.* [118] fed chicken a deuterium-labelled valine diet. The use of isotopically labelled amino acids in animals however is complicated by the fact that some amino acids are already present in the organism and so only partially labelling is possible. Alternatively, proteins can be isotopically labelled at the C-terminal of the tryptic peptides using $^{16}O/^{18}O$ labelled water [119-121]. However with most of these methods only two samples can be compared, therefore often the samples and

replicates from two cohorts are often pooled for analysis. For comparison using MS, the labelled and unlabelled samples are combined during sample preparation, which allows minimal variation during separation and analysis and, thus, more accurate quantitation of the expression levels of proteins in the counterpart proteomes [122]. Most commonly the labelled peptides are separated using RP-LC-MS/MS couples to ESI-MS and MS/MS. Quantitation follows peptide identification using specialized software such as Bioworks or other commercially available programs. MALDI-ToF MS has also been used for quantitation using isotopically labelled peptides [123]. This has the advantage of singly charged peaks and being able to perform MS/MS fragmentation on peptides that showed abundance changes.

1.5.1 Peak Intensity Quantitation and Internal Standards

Although isotope labelling is a relatively sophisticated technique, it is associated with a number of problems, such as labour- and time-intensive labelling steps and high cost of the reagents, just to name a few. There are a number of approaches of label-free quantitation, in which no addition of a isotopic or chemical tag is necessary. For quantitation using an internal standard, a protein of known concentration is added to control and disease samples as an internal standard. Optimally a protein is used that does not naturally exist in the mixture. Then the mixture including the internal standard is enzymatically digested according to the standard protocol and analysed by RPLC-MS/MS. Chromatographic peptide peaks from one protein can then be combined to calculate the overall peak area or height. This way, Yeo *et al.* [124] used bovine insulin to compare plasma peptides from control and chronic asthma mice. Bondarenko *et al.* [125] claimed that the peak area of the peptide is directly proportional to the abundance of a protein in a mixture. Conversely it was noticed by [126] that different peptides ionize at different rates although from the same peptides.

Quantitation of small molecules by integration of the LC-MS extracted ion chromatogram (XIC) peaks has been used commonly in analytical chemistry. The same method can be applied to proteolytic protein digests for biomarker discovery [127, 128]. A large amount of data is produced in these experiments and the complexity of the samples analysed requires automated data analysis tools. In contrast

to pattern-based, difference based [129], or identification-based [130, 131] approaches, an analytical tool for integration of all peaks in each spectrum was designed by Higgs *et al.* [132]. However none of these papers report a comparison of real samples for marker discovery but study the methodology and evaluation of respective statistical tools.

Wiener *et al.* [129] for example used an approach to devise an algorithm that compared peptide peak intensities at each time and *m/z* ratio. The computational analysis takes into account that peak intensity measurements are variable. This has allowed the method to take into account small but significant intensity changes in low abundance peptides but ignore large but statistically insignificant changes in peptides at much greater concentrations. The method uses a *t*-test to compare the peak intensity at each time and *m/z* ratio, to compare two conditions. The experiment was carried out in tryptic digests of 19 proteins. Visual inspection of the base peak chromatogram did not show any differences between the samples. The algorithm however, was able to distinguish differences between samples that could not be detected by visual detection alone. After significantly different *m/z* peaks have been detected in a full MS analysis, "markers" can be further analysed in a second experiment where they are specifically selected by the MS for fragmentation to avoid loss in the exclusion time.
Similarly, Yeo *et al.* [124] used 3D Excel© (Windows, Microsoft) plots to visualise elution time, *m/z* and peak area. The bubble graphs identify peaks that are more intense in one sample than in the other (Figure 1.22). This may be a technique worthwhile exploring.

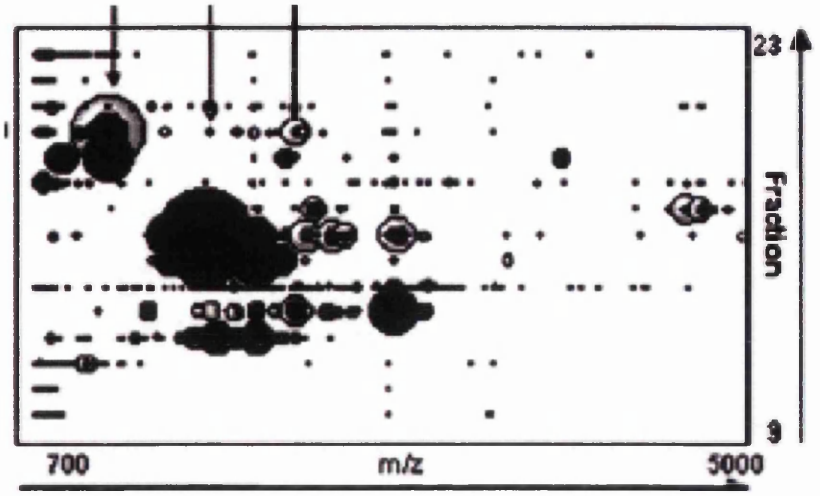


Figure 1.22: Excel 3D bubble plot depicting peptide intensities. On the x-axis the *m/z* ratio and on the y-axis elution fraction. The peak intensity is visualised in form of the size of the bubble [124].

1.5.2 Intact Protein Profiling

It could be argued that comparing peptide changes between different samples is problematic due to the fact that different peptides from the same protein may ionize differently and therefore the extrapolation back to the protein is challenging [126]. The use of intact proteins provides a more realistic measure of the actual change in protein intensity. SELDI-ToF MS has been specifically designed for high throughput detection of biomarkers. The use of the ProteinChip technology has enabled detection of a number of potential markers from serum in different diseases [62, 112, 133-135]. Using SELDI-ToF technology, biomarkers can be detected as well as more complicated protein patterns can be identified for further validation. The great advantage to using SELDI-ToF MS is the "easy" and fast application and subsequent analysis and interpretation of the data. As part of the software "normalization", peak alignment and data analysis tools are available and relative easy to use. Although SELDI-ToF MS suffers from low accuracy MS, the practicality has facilitated many studies with the necessary speed of analysis to analyse relatively large sample sets and publish a vast amount of data. MALDI-ToF MS on the other hand, although the mass accuracy is much higher and the MS technology relatively similar has not been used to the same extend for global quantitative studies from serum. For MALDI-ToF MS a

serum sample has to be fractionated for improved peak recovery by removing high abundance proteins as well as removing the high amounts of salt from serum [136, 137]. The presence of hotspots in the matrix where some proteins ionize better and produce higher peaks causes MALDI-ToF to be less reproducible than SELDI-ToF MS. However this can be overcome by accumulating a large number of spectra from each spot as well as analysing each sample in replicates, and the use of better matrix/analyte application [138]. But finally and possibly the most likely reason for the limited use of MALDI-ToF for quantitative protein profiling is the lack of readily available peak alignment tools and algorithms supporting high-throughput data analysis and interpretation of the data to elucidate potential markers.

1.6 Objectives of the Study

On the basis of the information gathered, the use of the following protocols to look for serum biomarkers of breast cancer is proposed. The use of LMW ultrafiltration for serum separation seems a good method as the first dimension of separation. It involves a single step and produces only one fraction of a sample with much reduced protein complexity. Fortunately many of the high abundance proteins are also of relative high molecular weight. The UF protocol will be investigated for reproducibility and efficiency to establish an optimised method for comparing breast cancer serum and no-cancer controls. In more detail, the application of intact protein profiling will be evaluated using SELDI-ToF MS. Furthermore, I aim to develop a robust protocol for quantitation of protein peaks using MALDI-ToF, including the programming of a peak alignment software for comparison of large sample sets. Identification of differential markers will be attempted using a MALDI-ToF/ToF as part of a collaboration formed with Applied Biosystems in Germany. The use of LC-MS for peptide quantitation will be demonstrated in a small study and the options for identification of discriminating peaks explored.

In order to maximise the chances of identifying a biomarker pattern indicative of breast cancer, the initial study group will be composed of patients with advanced metastatic breast cancer, who have not received chemotherapy. Patients with metastases to the liver were excluded. By carefully matching the control and case cohorts for factors such as age and menopausal status, we can focus on a single variable to find robust breast cancer-specific markers. Blood samples will be obtained from selected patients at Singleton hospital, and subject to standard protocols for consent and ethical approval. Issues related to sample collection, handling and storage, standardisation of protocols, availability of normal controls, access to bio-banks, clinical information, as well as ethical considerations are critical, and have been considered and dealt with from the beginning. Since sample preparation is such a key variable and concern in these types of studies, we have a standard sample collection and storage protocol, which will be adhered to strictly. In addition as a second control measure any deviation from the protocol is recorded to allow us to do post-hoc outlier identification at a later stage.

1.7 References

[1] CancerResearch_UK, (2005), *Cancer facts & figures*, accessed: 10.04.05 from: http://www.cancerresearchuk.org/aboutcancer/statistics/

[2] BreastCancerCare, (2005), *Breast Cancer*, accessed: 14.03.05 from: http://www.breastcancercare.org.uk/Home

[3] Imaginis, C., (2005), *Imaginis®*, accessed: 23.02.05 from: http://imaginis.com/breasthealth/staging.asp#survival

[4] CancerResearchUK, (2005), *Breast Cancer Factsheet – February 2004*, accessed: 10.01.06 from: http://www.cancerresearchuk.org

[5] National.Cancer.Institute, (2006), *Stage Information*, accessed: 04/11/2006 from: http://www.cancer.gov/cancertopics/

[6] WHO, in: 2nd (Ed.), WHO, Geneva, Switzerland 1981.

[7] Celis, J. E., Moreira, J. M., Gromova, I., Cabezon, T., Ralfkiaer, U., Guldberg, P., Straten, P. T., Mouridsen, H., Friis, E., Holm, D., Rank, F. and Gromov, P. (2005) Towards discovery-driven translational research in breast cancer. *Febs J* **272**, 2-15.

[8] The_American_Heritage®, (2000), *Dictionary of the English Language*, accessed: 20.03.05 from:

[9] Cheung, K. L., Graves, C. R. and Robertson, J. F. (2000) Tumour marker measurements in the diagnosis and monitoring of breast cancer. *Cancer Treat Rev* **26**, 91-102.

[10] Coveney, E. C., Geraghty, J. G., Sherry, F., McDermott, E. W., Fennelly, J. J., O'Higgins, N. J. and Duffy, M. J. (1995) The clinical value of CEA and CA 15-3 in breast cancer management. *Int J Biol Markers* **10**, 35-41.

[11] Devine, P. L., Duroux, M. A., Quin, R. J., McGuckin, M. A., Joy, G. J., Ward, B. G. and Pollard, C. W. (1995) CA15-3, CASA, MSA, and TPS as diagnostic serum markers in breast cancer. *Breast Cancer Res Treat* **34**, 245-251.

[12] Hayes, D. F., Bast, R. C., Desch, C. E., Fritsche, H., Jr., Kemeny, N. E., Jessup, J. M., Locker, G. Y., Macdonald, J. S., Mennel, R. G., Norton, L., Ravdin, P., Taube, S. and Winn, R. J. (1996) Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst* **88**, 1456-1466.

[13] Jager, W., Kramer, S., Palapelas, V. and Norbert, L. (1995) Breast cancer and clinical utility of CA 15-3 and CEA. *Scand J Clin Lab Invest Suppl* **221**, 87-92.

[14] Safi, F., Kohler, I., Rottinger, E. and Beger, H. (1991) The value of the tumor marker CA 15-3 in diagnosing and monitoring breast cancer. A comparative study with carcinoembryonic antigen. *Cancer* **68**, 574-582.

[15] Molina, R., Jo, J., Filella, X., Zanon, G., Pahisa, J., Munoz, M., Farrus, B., Latre, M. L., Escriche, C., Estape, J. and Ballesta, A. M. (1998) c-erbB-2 oncoprotein, CEA, and CA15.3 in patients with breast cancer: prognostic value. *Breast Cancer Res. Treat.* **51**, 109-119.

[16] Bast, R. C., Jr., Ravdin, P., Hayes, D. F., Bates, S., Fritsche, H., Jr., Jessup, J. M., Kemeny, N., Locker, G. Y., Mennel, R. G. and Somerfield, M. R. (2001) 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. *J Clin Oncol* **19**, 1865-1878.

[17] Molina, R., Barak, V., van Dalen, A., Duffy, M. J., Einarsson, R., Gion, M., Goike, H., Lamerz, R., Nap, M., Soletormos, G. and Stieber, P. (2005) Tumor markers in

breast cancer- European Group on Tumor Markers recommendations. *Tumour Biol* **26**, 281-293.

[18] Menard, S., Tagliabue, E., Campiglio, M. and Pupa, S. M. (2000) Role of HER2 gene overexpression in breast carcinoma. *J. Cell. Physiol.* **182**, 150-162.

[19] Krainer, M., Brodowicz, T., Zeillinger, R., Wiltschke, C., Scholten, C., Seifert, M., Kubista, E. and Zielinski, C. C. (1997) Tissue expression and serum levels of HER-2/neu in patients with breast cancer. *Oncology* **54**, 475-481.

[20] Ross, J. S., Fletcher, J. A., Linette, G. P., Stec, J., Clark, E., Ayers, M., Symmans, W. F., Pusztai, L. and Bloom, K. J. (2003) The Her-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy. *Oncologist* **8**, 307-325.

[21] Isola, J. J., Holli, K., Oksa, H., Teramoto, Y. and Kallioniemi, O. P. (1994) Elevated erbB-2 oncoprotein levels in preoperative and follow-up serum samples define an aggressive disease course in patients with breast cancer. *Cancer* **73**, 652-658.

[22] Disis, M. L., Pupa, S. M., Gralow, J. R., Dittadi, R., Menard, S. and Cheever, M. A. (1997) High-titer HER-2/neu protein-specific antibody can be detected in patients with early-stage breast cancer. *J Clin Oncol* **15**, 3363-3367.

[23] Lipton, A., Ali, S. M., Leitzel, K., Demers, L., Chinchilli, V., Engle, L., Harvey, H. A., Brady, C., Nalin, C. M., Dugan, M., Carney, W. and Allard, J. (2002) Elevated serum Her-2/neu level predicts decreased response to hormone therapy in metastatic breast cancer. *J Clin Oncol* **20**, 1467-1472.

[24] Mehta, R. R., McDermott, J. H., Hieken, T. J., Marler, K. C., Patel, M. K., Wild, L. D. and Das Gupta, T. K. (1998) Plasma c-erbB-2 levels in breast cancer patients: prognostic significance in predicting response to chemotherapy. *J Clin Oncol* **16**, 2409-2416.

[25] Le Naour, F., Misek, D. E., Krause, M. C., Deneux, L., Giordano, T. J., Scholl, S. and Hanash, S. M. (2001) Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clin Cancer Res* **7**, 3328-3335.

[26] Peyrat, J. P., Bonneterre, J., Lubin, R., Vanlemmens, L., Fournier, J. and Soussi, T. (1995) Prognostic significance of circulating P53 antibodies in patients undergoing surgery for locoregional breast cancer. *Lancet* **345**, 621-622.

[27] Lenner, P., Wiklund, F., Emdin, S. O., Arnerlov, C., Eklund, C., Hallmans, G., Zentgraf, H. and Dillner, J. (1999) Serum antibodies against p53 in relation to cancer risk and prognosis in breast cancer: a population-based epidemiological study. *Br. J. Cancer* **79**, 927-932.

[28] Hondermarck, H. (2003) Breast cancer: when proteomics challenges biological complexity. *Mol Cell Proteomics* **2**, 281-291.

[29] Cotran, R., Kumar, V. and Robbins, S., *Robbins' pathologic basis of disease*, W.B. Saunders Company, West Philadelphia 1989.

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

[31] Wright, G. L., Jr. (1974) Two-dimensional acrylamide gel electrophoresis of cancer-patient serum proteins. *Ann Clin Lab Sci* **4**, 281-293.

[32] Bhattacharya, B., Prasad, G. L., Valverius, E. M., Salomon, D. S. and Cooper, H. L. (1990) Tropomyosins of human mammary epithelial cells: consistent defects of expression in mammary carcinoma cell lines. *Cancer Res* **50**, 2105-2112.

[33] Franzen, B., Linder, S., Okuzawa, K., Kato, H. and Auer, G. (1993) Nonenzymatic extraction of cells from clinical tumor material for analysis of gene expression by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **14**, 1045-1053.

[34] Wulfkuhle, J. D., Sgroi, D. C., Krutzsch, H., McLean, K., McGarvey, K., Knowlton, M., Chen, S., Shu, H., Sahin, A., Kurek, R., Wallwiener, D., Merino, M. J., Petricoin,

E. F., 3rd, Zhao, Y. and Steeg, P. S. (2002) Proteomics of human breast ductal carcinoma in situ. *Cancer Res* **62**, 6740-6749.

[35] Luo, Y., Zhang, J., Liu, Y., Shaw, A. C., Wang, X., Wu, S., Zeng, X., Chen, J., Gao, Y. and Zheng, D. (2005) Comparative proteome analysis of breast cancer and normal breast. *Mol Biotechnol* **29**, 233-244.

[36] Adriaenssens, E., Lemoine, J., El Yazidi-Belkoura, I. and Hondermarck, H. (2002) Growth signaling in breast cancer cells: outcomes and promises of proteomics. *Biochem. Pharmacol.* **64**, 797-803.

[37] Hondermarck, H., Dolle, L., El Yazidi-Belkoura, I., Vercoutter-Edouart, A. S., Adriaenssens, E. and Lemoine, J. (2002) Functional proteomics of breast cancer for signal pathway profiling and target discovery. *J. Mammary Gland Biol. Neoplasia* **7**, 395-405.

[38] Hondermarck, H., Vercoutter-Edouart, A. S., Revillion, F., Lemoine, J., El-Yazidi-Belkoura, I., Nurcombe, V. and Peyrat, J. P. (2001) Proteomics of breast cancer for marker discovery and signal pathway profiling. *Proteomics* **1**, 1216-1232.

[39] Vercoutter-Edouart, A. S., Lemoine, J., Le Bourhis, X., Louis, H., Boilly, B., Nurcombe, V., Revillion, F., Peyrat, J. P. and Hondermarck, H. (2001) Proteomic analysis reveals that 14-3-3sigma is down-regulated in human breast cancer cells. *Cancer Res* **61**, 76-80.

[40] Umbricht, C. B., Evron, E., Gabrielson, E., Ferguson, A., Marks, J. and Sukumar, S. (2001) Hypermethylation of 14-3-3 sigma (stratifin) is an early event in breast cancer. *Oncogene* **20**, 3348-3353.

[41] Benzinger, A., Muster, N., Koch, H. B., Yates, J. R., 3rd and Hermeking, H. (2005) Targeted proteomic analysis of 14-3-3 sigma, a p53 effector commonly silenced in cancer. *Mol Cell Proteomics* **4**, 785-795.

[42] Moreira, J. M., Ohlsson, G., Rank, F. E. and Celis, J. E. (2005) Down-regulation of the Tumor Suppressor Protein 14-3-3{sigma} Is a Sporadic Event in Cancer of the Breast. *Mol Cell Proteomics* **4**, 555-569.

[43] Craven, R. A. and Banks, R. E. (2001) Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics* **1**, 1200-1204.

[44] Zhang, D. H., Tai, L. K., Wong, L. L., Sethi, S. K. and Koay, E. S. (2005) Proteomics of breast cancer: enhanced expression of cytokeratin19 in human epidermal growth factor receptor type 2 positive breast tumors. *Proteomics* **5**, 1797-1805.

[45] Wulfkuhle, J. D., McLean, K. C., Paweletz, C. P., Sgroi, D. C., Trock, B. J., Steeg, P. S. and Petricoin, E. F., 3rd (2001) New approaches to proteomic analysis of breast cancer. *Proteomics* **1**, 1205-1215.

[46] Zhang, D., Tai, L. K., Wong, L. L., Chiu, L. L., Sethi, S. K. and Koay, E. S. (2005) Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in HER-2/neu-positive breast cancer. *Mol Cell Proteomics* **4**, 1686-1696.

[47] Westley, B. and Rochefort, H. (1980) A secreted glycoprotein induced by estrogen in human breast cancer cell lines. *Cell* **20**, 353-362.

[48] Giometti, C. S., Tollaksen, S. L., Chubb, C., Williams, C. and Huberman, E. (1995) Analysis of proteins from human breast epithelial cells using two-dimensional gel electrophoresis. *Electrophoresis* **16**, 1215-1224.

[49] Trask, D. K., Band, V., Zajchowski, D. A., Yaswen, P., Suh, T. and Sager, R. (1990) Keratins as markers that distinguish normal and tumor-derived mammary epithelial cells. *Proc Natl Acad Sci U S A* **87**, 2319-2323.

[50] Worland, P. J., Bronzert, D., Dickson, R. B., Lippman, M. E., Hampton, L., Thorgeirsson, S. S. and Wirth, P. J. (1989) Secreted and cellular polypeptide patterns

of MCF-7 human breast cancer cells following either estrogen stimulation or v-H-ras transfection. *Cancer Res* **49**, 51-57.

[51] Schrohl, A. S., Christensen, I. J., Pedersen, A. N., Jensen, V., Mouridsen, H., Murphy, G., Foekens, J. A., Brunner, N. and Holten-Andersen, M. N. (2003) Tumor tissue concentrations of the proteinase inhibitors tissue inhibitor of metalloproteinases-1 (TIMP-1) and plasminogen activator inhibitor type 1 (PAI-1) are complementary in determining prognosis in primary breast cancer. *Mol Cell Proteomics* **2**, 164-172.

[52] Anderson, N. L. and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1**, 845-867.

[53] Adam, B. L., Vlahou, A., Semmes, O. J. and Wright, G. L., Jr. (2001) Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics* **1**, 1264-1270.

[54] Banez, L. L., Prasanna, P., Sun, L., Ali, A., Zou, Z., Adam, B. L., McLeod, D. G., Moul, J. W. and Srivastava, S. (2003) Diagnostic potential of serum proteomic patterns in prostate cancer. *J Urol* **170**, 442-446.

[55] Yasui, Y., Pepe, M., Thompson, M. L., Adam, B. L., Wright, G. L., Jr., Qu, Y., Potter, J. D., Winget, M., Thornquist, M. and Feng, Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4**, 449-463.

[56] Carter, D., Douglass, J. F., Cornellison, C. D., Retter, M. W., Johnson, J. C., Bennington, A. A., Fleming, T. P., Reed, S. G., Houghton, R. L., Diamond, D. L. and Vedvick, T. S. (2002) Purification and characterization of the mammaglobin/lipophilin B complex, a promising diagnostic marker for breast cancer. *Biochemistry* **41**, 6714-6722.

[57] Petricoin, E. F. and Liotta, L. A. (2004) Proteomic approaches in cancer risk and response assessment. *Trends Mol Med* **10**, 59-64.

[58] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P. and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* **2**, 1096-1103.

[59] Becker, S., Cazares, L. H., Watson, P., Lynch, H., Semmes, O. J., Drake, R. R. and Laronga, C. (2004) Surfaced-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. *Ann Surg Oncol* **11**, 907-914.

[60] Laronga, C., Becker, S., Watson, P., Gregory, B., Cazares, L., Lynch, H., Perry, R. R., Wright, G. L., Jr., Drake, R. R. and Semmes, O. J. (2003) SELDI-TOF serum profiling for prognostic and diagnostic classification of breast cancers. *Dis Markers* **19**, 229-238.

[61] Vlahou, A., Laronga, C., Wilson, L., Gregory, B., Fournier, K., McGaughey, D., Perry, R. R., Wright, G. L., Jr. and Semmes, O. J. (2003) A novel approach toward development of a rapid blood test for breast cancer. *Clin Breast Cancer* **4**, 203-209.

[62] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. and Chan, D. W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* **48**, 1296-1304.

[63] Pawlik, T. M., Fritsche, H., Coombes, K. R., Xiao, L., Krishnamurthy, S., Hunt, K. K., Pusztai, L., Chen, J. N., Clarke, C. H., Arun, B., Hung, M. C. and Kuerer, H. M. (2005) Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry. *Breast Cancer Res Treat* **89**, 149-157.

[64] Mathelin, C., Cromer, A., Wendling, C., Tomasetto, C. and Rio, M. C. (2005) Serum biomarkers for detection of breast cancers: a prospective study. *Breast Cancer Res Treat*, 1-8.

[65] Adam, P. J., Boyd, R., Tyson, K. L., Fletcher, G. C., Stamps, A., Hudson, L., Poyser, H. R., Redpath, N., Griffiths, M., Steers, G., Harris, A. L., Patel, S., Berry, J., Loader, J. A., Townsend, R. R., Daviet, L., Legrain, P., Parekh, R. and Terrett, J. A. (2003) Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *J Biol Chem* **278**, 6482-6489.

[66] Thadikkaran, L., Siegenthaler, M. A., Crettaz, D., Queloz, P. A., Schneider, P. and Tissot, J. D. (2005) Recent advances in blood-related proteomics. *Proteomics* **5**, 3019-3034.

[67] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R. and Lobley, A. (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* **3**, 311-326.

[68] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L. and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* **1**, 947-955.

[69] Hu, S., Loo, J. A. and Wong, D. T. (2006) Human body fluid proteome analysis. *Proteomics* **6**, 6326-6353.

[70] Plasma.Proteome.Institute, (2004), *Proteins in Plasma and Serum*, Institute, C. P. P. accessed: 22/02/2005 from: www. plasmaproteome.org

[71] Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W. and Guild, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics* **4**, 1175-1186.

[72] Petricoin, E. F., Ornstein, D. K. and Liotta, L. A. (2004b) Clinical proteomics: Applications for prostate cancer biomarker discovery and detection. *Urol Oncol* **22**, 322-328.

[73] Bjorhall, K., Miliotis, T. and Davidsson, P. (2005) Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **5**, 307-317.

[74] Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L. and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422-432.

[75] Misek, D. E., Kuick, R., Wang, H., Galchev, V., Deng, B., Zhao, R., Tra, J., Pisano, M. R., Amunugama, R., Allen, D., Walker, A. K., Strahler, J. R., Andrews, P., Omenn, G. S. and Hanash, S. M. (2005) A wide range of protein isoforms in serum and plasma uncovered by a quantitative intact protein analysis system. *Proteomics* **5**, 3343-3352.

[76] Huang, L., Harvie, G., Feitelson, J. S., Gramatikoff, K., Herold, D. A., Allen, D. L., Amunngama, R., Hagler, R. A., Pisano, M. R., Zhang, W. W. and Fang, X. (2005) Immunoaffinity separation of plasma proteins by IgY microbeads: meeting the needs of proteomic sample preparation and analysis. *Proteomics* **5**, 3314-3328.

[77] Georgiou, H. M., Rice, G. E. and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1**, 1503-1506.

[78] Wagner, K., Miliotis, T., Marko-Varga, G., Bischoff, R. and Unger, K. K. (2002) An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. *Anal Chem* **74**, 809-820.

[79] Morris, D. L., Jr., Sutton, J. N., Harper, R. G. and Timperman, A. T. (2004) Reversed-phase HPLC separation of human serum employing a novel saw-tooth gradient: toward multidimensional proteome analysis. *J Proteome Res* **3**, 1149-1154.

[80] Johnson, K. L., Mason, C. J., Muddiman, D. C. and Eckel, J. E. (2004a) Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometry: precision of retention time, mass, and ion abundance. *Anal Chem* **76**, 5097-5103.

[81] Harper, R. G., Workman, S. R., Schuetzner, S., Timperman, A. T. and Sutton, J. N. (2004) Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. *Electrophoresis* **25**, 1299-1306.

[82] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J., Petricoin, E. F., 3rd, Liotta, L. A., Veenstra, T. D. and Conrads, T. P. (2004) An investigation into the human serum "interactome". *Electrophoresis* **25**, 1289-1298.

[83] Merrell, K., Southwick, K., Graves, S. W., Esplin, M. S., Lewis, N. E. and Thulin, C. D. (2004) Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J Biomol Tech* **15**, 238-248.

[84] Hoffmann, E. d. and Stroobant, V., *Mass Spectrometry: Principles and Applications*, John Wiley & Sons, Ltd 2003.

[85] Tiselius, A. W. K., (1930) *The Moving Boundary Method of Studying the Electrophoresis of Proteins*, Uppsala, Sweden, PhD thesis.

[86] Issaq, H. J., Conrads, T. P., Janini, G. M. and Veenstra, T. D. (2002) Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis* **23**, 3048-3061.

[87] Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97**, 9390-9395.

[88] Fels, L. M., Buschmann, T., Meuer, J., Reymond, M. A., Lamer, S., Rocken, C. and Ebert, M. P. (2003) Proteome analysis for the identification of tumor-associated biomarkers in gastrointestinal cancer. *Dig Dis* **21**, 292-298.

[89] Chernokalskaya, E., Gutierrez, S., Pitt, A. M. and Leonard, J. T. (2004) Ultrafiltration for proteomic sample preparation. *Electrophoresis* **25**, 2461-2468.

[90] Galvani, M., Hamdan, M., Herbert, B. and Righetti, P. G. (2001) Alkylation kinetics of proteins in preparation for two-dimensional maps: a matrix assisted laser desorption/ionization-mass spectrometry investigation. *Electrophoresis* **22**, 2058-2065.

[91] Galvani, M., Rovatti, L., Hamdan, M., Herbert, B. and Righetti, P. G. (2001) Protein alkylation in the presence/absence of thiourea in proteome analysis: a matrix assisted laser desorption/ionization-time of flight-mass spectrometry investigation. *Electrophoresis* **22**, 2066-2074.

[92] Herbert, B., Galvani, M., Hamdan, M., Olivieri, E., MacCarthy, J., Pedersen, S. and Righetti, P. G. (2001) Reduction and alkylation of proteins in preparation of two-dimensional map analysis: why, when, and how? *Electrophoresis* **22**, 2046-2057.

[93] Qian, W. J., Jacobs, J. M., Camp, D. G., 2nd, Monroe, M. E., Moore, R. J., Gritsenko, M. A., Calvano, S. E., Lowry, S. F., Xiao, W., Moldawer, L. L., Davis, R. W., Tompkins, R. G. and Smith, R. D. (2005) Comparative proteome analyses of human plasma following in vivo lipopolysaccharide administration using multidimensional separations coupled with tandem mass spectrometry. *Proteomics* **5**, 572-584.

[94] Bandeira, N., Tang, H., Bafna, V. and Pevzner, P. (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Anal Chem* **76**, 7221-7233.

[95] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. and Yates, J. R., 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-682.

[96] Washburn, M. P., Wolters, D. and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-247.

[97] Arpino, P., Baldwin, M. A. and McLafferty, F. W. (1974) Liquid chromatography-mass spectrometry. II. Continuous monitoring. *Biomed Mass Spectrom* **1**, 80-82.

[98] ASMS, 2001.

[99] MatrixScience, (2006), *Instrument Specific MS/MS Ion Series Matching*, accessed: 01/05/07 from: http://www.matrixscience.com

[100] European.Bioinformatics.Institute, (2006-2007), accessed: 01/05/2007 from: http://www.ebi.ac.uk

[101] Eng, J. K., McCormack, A. L. and Yates III, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976-989.

[102] Karas, M., Bachmann, D., Bahr, U. and Hillenkamp, F. (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes* **78**, 53–56.

[103] Wu, K. J., Steding, A. and Becker, C. H. (1993) Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. *Rapid Commun Mass Spectrom* **7**, 142-146.

[104] Gates, P., (2004), *A schematic diagram of the mechanism of MALDI.*, The University of Bristol, S. o. C. accessed: 09/05/07 from: www.chm.bris.ac.uk/ms/images/maldi-mechanism.gif

[105] Hutchens, T. W. and Yip, T. T. (1993) New Desorption Strategies For The Mass-Spectrometric Analysis Of Macromolecules. *Rapid Commun. Mass Spectrom.* **7**, 576-580.

[106] Ciphergen Biosystems, I., *ProteinChip® Applications Guide Volume 1: Introductory Guide*, 2004.

[107] Anderson, L. and Anderson, N. G. (1977) High resolution two-dimensional electrophoresis of human plasma proteins. *Proc Natl Acad Sci U S A* **74**, 5421-5425.

[108] Hughes, G. J., Frutiger, S., Paquet, N., Ravier, F., Pasquali, C., Sanchez, J. C., James, R., Tissot, J. D., Bjellqvist, B. and Hochstrasser, D. F. (1992) Plasma protein map: an update by microsequencing. *Electrophoresis* **13**, 707-714.

[109] Shen, Y., Jacobs, J. M., Camp, D. G., 2nd, Fang, R., Moore, R. J., Smith, R. D., Xiao, W., Davis, R. W. and Tompkins, R. G. (2004) Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal Chem* **76**, 1134-1144.

[110] Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D. and Comb, M. J. (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol* **23**, 94-101.

[111] Yates, J. R., 3rd (2004) Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct* **33**, 297-316.

[112] Mian, S., Ugurel, S., Parkinson, E., Schlenzka, I., Dryden, I., Lancashire, L., Ball, G., Creaser, C., Rees, R. and Schadendorf, D. (2005) Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. *J Clin Oncol* **23**, 5088-5093.

[113] Conrads, T. P., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A., Steinberg, S. M., Kohn, E. C., Fishman, D. A., Whitely, G., Barrett, J. C., Liotta, L. A., Petricoin, E. F., 3rd and Veenstra, T. D. (2004a) High-resolution serum proteomic features for ovarian cancer detection. *Endocr Relat Cancer* **11**, 163-178.

[114] Conrads, T. P. and Veenstra, T. D. (2004c) The utility of proteomic patterns for the diagnosis of cancer. *Curr Drug Targets Immune Endocr Metabol Disord* **4**, 41-50.

[115] Ornstein, D. K., Rayford, W., Fusaro, V. A., Conrads, T. P., Ross, S. J., Hitt, B. A., Wiggins, W. W., Veenstra, T. D., Liotta, L. A. and Petricoin, E. F., 3rd (2004) Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml. *J Urol* **172**, 1302-1305.

[116] Rui, Z., Jian-Guo, J., Yuan-Peng, T., Hai, P. and Bing-Gen, R. (2003) Use of serological proteomic methods to find biomarkers associated with breast cancer. *Proteomics* **3**, 433-439.

[117] Chen, X., Smith, L. M. and Bradbury, E. M. (2000) Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal Chem* **72**, 1134-1143.

[118] Doherty, M. K., Whitehead, C., McCormack, H., Gaskell, S. J. and Beynon, R. J. (2005) Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates. *Proteomics* **5**, 522-533.

[119] Takao, T., Hori, H., Okamoto, K., Harada, A., Kamachi, M. and Shimonishi, Y. (1991) Facile assignment of sequence ions of a peptide labelled with 18O at the carboxyl terminus. *Rapid Commun Mass Spectrom* **5**, 312-315.

[120] Shevchenko, A., Chernushevich, I., Ens, W., Standing, K. G., Thomson, B., Wilm, M. and Mann, M. (1997) Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **11**, 1015-1024.

[121] Chen, X., Cushman, S. W., Pannell, L. K. and Hess, S. (2005) Quantitative proteomic analysis of the secretory proteins from rat adipose cells using a 2D liquid chromatography-MS/MS approach. *J Proteome Res* **4**, 570-577.

[122] Yao, X., Freas, A., Ramirez, J., Demirev, P. A. and Fenselau, C. (2001) Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842.

[123] Ji, C. and Li, L. (2005) Quantitative proteome analysis using differential stable isotopic labeling and microbore LC-MALDI MS and MS/MS. *J Proteome Res* **4**, 734-742.

[124] Yeo, S., Roh, G. S., Kim, D. H., Lee, J. M., Seo, S. W., Cho, J. W., Kim, C. W. and Kwack, K. (2004) Quantitative profiling of plasma peptides in asthmatic mice using liquid chromatography and mass spectrometry. *Proteomics* **4**, 3308-3317.

[125] Bondarenko, P. V., Chelius, D. and Shaler, T. A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74**, 4741-4749.

[126] Smith, R. D., Shen, Y. and Tang, K. (2004) Ultrasensitive and quantitative analyses from combined separations-mass spectrometry for the characterization of proteomes. *Acc Chem Res* **37**, 269-278.

[127] Wang, G., Wu, W. W., Zeng, W., Chou, C. L. and Shen, R. F. (2006) Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J Proteome Res* **5**, 1214-1223.

[128] Chelius, D. and Bondarenko, P. V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* **1**, 317-323.

[129] Wiener, M. C., Sachs, J. R., Deyanova, E. G. and Yates, N. A. (2004) Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal Chem* **76**, 6085-6096.

[130] Gao, J., Opiteck, G. J., Friedrichs, M. S., Dongre, A. R. and Hefta, S. A. (2003) Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* **2**, 643-649.

[131] Colinge, J., Chiappe, D., Lagache, S., Moniatte, M. and Bougueleret, L. (2005) Differential proteomics via probabilistic peptide identification scores. *Anal Chem* **77**, 596-606.

[132] Higgs, R. E., Knierman, M. D., Gelfanova, V., Butler, J. P. and Hale, J. E. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res* **4**, 1442-1450.

[133] Mathelin, C., Cromer, A., Wendling, C., Tomasetto, C. and Rio, M. C. (2006) Serum biomarkers for detection of breast cancers: A prospective study. *Breast Cancer Res Treat* **96**, 83-90.

[134] Li, J., Orlandi, R., White, C. N., Rosenzweig, J., Zhao, J., Seregni, E., Morelli, D., Yu, Y., Meng, X. Y., Zhang, Z., Davidson, N. E., Fung, E. T. and Chan, D. W. (2005) Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin Chem* **51**, 2229-2235.

[135] Ruetschi, U., Rosen, A., Karlsson, G., Zetterberg, H., Rymo, L., Hagberg, H. and Jacobsson, B. (2005) Proteomic analysis using protein chips to detect biomarkers in cervical and amniotic fluid in women with intra-amniotic inflammation. *J Proteome Res* **4**, 2236-2242.

[136] Baumann, S., Ceglarek, U., Fiedler, G. M., Lembcke, J., Leichtle, A. and Thiery, J. (2005) Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin Chem* **51**, 973-980.

[137] Landry, F., Lombardo, C. R. and Smith, J. W. (2000) A method for application of samples to matrix-assisted laser desorption ionization time-of-flight targets that enhances peptide detection. *Anal Biochem* **279**, 1-8.

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

CHAPTER 2

General Materials and Methods

2.1 Materials and Chemicals

All centrifugal filters and Zip-Tips were purchased from Millipore UK Ltd (Watford, UK). Tricine, glycine, tris, sinapinic acid (SA), and HPLC grade acetonitrile (ACN) were purchased from Fisher Scientific UK Ltd (Loughborough, UK), and ammonium bicarbonate (NH_4HCO_3), ß-mercaptoethanol, trifluoroacetic acid (TFA), Formic acid (FA), cytochrome C, bovine serum albumin (BSA), lysozyme, vitamin B12 and ubiquitin from Sigma-Aldrich Company Ltd. (Gillingham, UK). The BCA assay kit was purchased from Pierce (Perbio Science UK Ltd, Cramtington, UK) and ProteinChip® Arrays from Ciphergen Biosystems Ltd., (Guildford, UK). Sequence-grade trypsin was purchased from Promega (Southampton, UK) and acrylamide/ bisacrylamide and all protein standards came from Bio-Rad® (Hemel Hempstead, UK).

2.2 Serum Preparation and Handling

Human blood samples were obtained from healthy female volunteers and from female patients with metastatic breast cancer. Patients with breast cancer metastatic to the liver were excluded from the study. The project was approved by the Local Research Ethics Committee and written informed consent was obtained from all participants. The Ethics approval is shown in the Appendix (F). Human blood was collected by

venepuncture into Vacuette, gold-top serum separator tubes. Blood was allowed to clot at room temperature for 30 min and was centrifuged at 3000 x g for 5 min after which serum was collected. Serum was aliquoted into 500 µl and stored at -80°C until analyzed. Each breast cancer sample was matched with a control sample of similar age and the pair was processed simultaneously.

2.2.1 Determination of Protein Concentration

The protein concentration of all samples and fractions was determined with the Pierce BCA™ assay; here a dilution series of BSA (0.025 – 2 mg/ml) to generate a standard curve was prepared. For the mirco-assay, 10 µl of each sample and standard was mixed with 200 µl of the working reagent (50:1) in a 96-well plate. The plate was incubated floating on a water bath for 30 min at 37°C. Absorbance was measured at 550 nm in a Multiskan Ascent plate reader (Thermo Labsystems, UK). The protein concentration of each sample was read against the standard curve of BSA concentrations.

2.3 Serum Protein Pre-Fractionation

2.3.1 SDS Polyacrylamide Gel Electrophoresis (PAGE)

All protein gel electrophoresis experiments were performed using Mini-Protean II cells (Bio-Rad®, Hemel Hempstead, UK). All buffers were prepared as stock solutions prior to casting the gel. The first step was to prepare the acrylamide gel, for LMW proteins the gels are composed of 3 layers modified slightly from Schägger and von Jagow [1]. The bottom layer of acrylamide (the separating or resolving gel) comprised of about 50% of the gel height. The acrylamide concentration of the separating gel was 17% and pH 8.9. The middle layer was the Spacer gel which is 1-2 cm thick and contains 10% of acrylamide, and was buffered at pH 8.9. The top-most layer is referred to as the stacking gel, and comprised about 10% of the gel height. The stacking layer contains 4% of acrylamide and is buffered at pH 6.8. The difference in pH and acrylamide concentration at the stacking and spacer gel interface functions to compress the sample at the interface and provides better resolution and sharper bands in the separating gel. The acrylamide gel solutions for each layer were prepared with the compositions given in Table 2.1 and ammonium persulphate and TEMED were added just before pouring. Each layer was left to polymerize and, to provide a smooth surface and interface at the top of the separating gel, H_2O was placed above the gel during polymerization.

Table 2.1: Composition of separating, spacer and stacking gels.

			Separating Gels	
	Stacking Gel (4%)	**Spacer Gel (10%)**	**12%**	**17%**
19:1 acrylamide:bis-acrylamide	-	0.33 ml	-	0.84 ml
37.5:1 acrylamide:bis-acrylamide	0.2 ml	0.35 ml	1.2 ml	1.9 ml
Gel Buffer				
1 M Tris HCl pH 8.4, 0.1 % SDS	-	-	3.3 ml	3.3 ml
3 M Tris HCl pH 8.9, 0.3 % SDS	-	1.6 ml	-	-
1 M Tris HCl pH 6.8	1.25 ml	-	-	-
ddiH2O	3.5 ml	2.3 ml	1.4 ml	4.3 ml
10 % SDS	0.025 ml	-	-	-
50 % Glycerol	-	-	2.8 ml	2.2 ml
25 % ammonium persulphate	0.01 ml	0.068 ml	0.036 ml	0.013 ml
TEMED	0.005 ml	0.0017 ml	0.01 ml	0.0033 ml

Each sample was incubated for 5 min at 100°C in 5µl of sample loading buffer (0.3% SDS, 12% glycerol, 50mM Tris HCl pH 6.8, 2% ß-mercaptoethanol and 0.05% bromophenol blue) and then loaded onto the gel. The samples were run with a two buffer system, comprising of a 10x cathode buffer (1M Tris HCl pH 8.3, 1M tricine and 0.1% SDS) and 10x anode buffer (2M Tris HCl pH 8.9). The 10x stock solutions were diluted with ddiH_2O before use and the electrophoresis run at a constant current of 50 mA until the dyefront reached the bottom of the gel. A broad range and a low molecular weight protein standard marker were included.

The protein bands were fixed in a solution containing 40% methanol and 10% acetic acid for 30 min. After two 5 min ddiH_2O washes, the gels were stained with blue-silver colloidal Coomassie G-250 staining solution (added in order 20% ddiH_2O, 10% phosphoric acid, 10% ammonium sulphate, 0.12% Coomassie G-250, 60% H_2O and 20% methanol) for 12-24 hours. Gels were de-stained for 24 hours in ddiH_2O and images taken using a UVP BioDoc-It Imaging system (Cambridge, UK).

2.3.2 Centrifugal Ultrafiltration

All sample preparation and centrifugation was performed on ice or in a refrigerated Legend^{TM} T/RT swinging bucket centrifuge (Sorvall, Langenselbold, Germany) at 4°C. The centrifugal filter membranes were washed in 0.1 N NaOH and then rinsed with ddiH_2O by centrifugation for 3 min each. A serum sample was diluted with denaturing buffer (25 mM NH_4HCO_3, 20% ACN (v/v)) and incubated on ice for 60 min, with frequent shaking. The filters were then centrifuged until a minimum of 70% volume had passed through the filter. (Later into the project the HMW retentate was re-suspended to the original volume and centrifuged again until 70% of the volume had passed.) The filtrate was recovered and aliquoted into 1 ml tubes before it was lyophilized and stored at -80°C.

2.3.3 Protein Precipitation

Serum samples were precipitated using ACN, ethanol and TCA/acetone protocol for comparison. For each of the methods 20 µl of crude serum were used and the protocol performed in triplicate.

The ACN precipitation was the same as previously described [2], briefly two volumes of 100% ACN were added to 1 volume of serum and vortexed for 5 sec. The mixture was then incubated for 30 min at room-temperature and then centrifuged for 10 min at 12,000 xg.

The ethanol precipitation was performed according to the protocol described by Villanueva J. [3], here equal volumes of 100% EtOH were mixed with the serum by vortex for 1 min and then centrifuged for 10 min at 15,000 xg. The pellets (albumin) and supernatants were retained and brought to dryness in a speed vacuum centrifuge. For further analysis the supernatant was re-suspended in 20 µl of 25 mM NH_4HCO_3 and the pellet in 70 µl of 7M urea, 2M thiourea and 4% CHAPS.

For TCA/acetone precipitation a protocol published by Chen *et al.* [4] was used and 20 µl of serum were rapidly mixed with 80 µl of ice-cold acetone and 10% trichloroacetic acid (TCA) (v/v), and mixed gently by vortexing immediately. The mixture was then incubated for 90 min at -20°C and centrifuged at 15,000 xg for 20 min in a refrigerated centrifuge (4°C). The supernatant was removed and collected; the precipitated pellet (proteins) was washed with 1 ml of ice-cold acetone and incubated on ice for 15 min. Then the precipitate was centrifuged again at 15,000 xg for 20 min at 4°C and the supernatant removed and collected and the pellet lyophilized. The pooled supernatants (albumin) were precipitated again by adding 1 ml of ice-cold acetone/ TCA to the supernatant. The new pellet contains albumin. Both pellets were re-suspended in 100 µl of 7M urea, 2M thiourea and 4% CHAPS. All fractions from above were analysed by SDS-PAGE as described above.

2.3.4 Affinity Chromatography for Albumin and Protein G removal

A Vision™ BioCAD Family Perfusion Chromatography Workstation (Applied Biosystems, Warrington, UK) was used for human serum albumin (0.2 ml volume, 4 mm I.D. x 15 mm) and protein G (0.2 ml volume, 4 mm I.D. x 15 mm) depletion from serum. Before first use the Poros® anti-HSA and anti-Protein G cartridges affinity

depletion cartridges (Applied Biosystems, Warrington, UK) were equilibrated with 10 cartridge volumes (CV) of PBS (pH 7.2) at a flow-rate of 1 ml/min. For depletion the two cartridges were used in tandem, 50 µl of 10% diluted serum (in $ddiH_2O$) were injected on the anti-protein G cartridge and eluted onto the anti-HSA cartridge at a flow-rate of 0.5 ml/min. The cartridges were then washed with 30 CV (3-6 ml) of PBS (pH 7.2), before bound HSA and protein G were eluted with 10 CV of 12 mM HCl at a flow-rate of 1 ml/min. Fractions were collected automatically every minute. Finally the anti-HSA cartridge was cleaned with 10 CV of 1 M NaCl and the anti-protein G cartridge with 10 cartridge volumes CV of 2 M urea in 1 M NaCl at a flow-rate of 1 ml/min

2.3.5 Weak Anion Exchange (WAX) for Intact Protein Separation

Neat serum proteins were separated by ion exchange chromatography using a 100 mm x 4.6 mm I.D., 1000 Å PolyLC WAX column (PolyLC, Maryland, USA). The separation was performed using the Vision™ BioCAD Family Perfusion Chromatography Workstation (Applied Biosystems, Warrington, UK). After the column was equilibrated with two repeat 90 min gradients, 100% A (5% ACN in water) to 100% B (5% ACN in 0.6 M NH_4 acetate), 2x 40 µl of serum were injected and eluted at a flow-rate of 1 ml/min. 1.5 ml fractions were collected and for salt removal and drying, speed vacuum centrifuged, using a low vacuum concentrator Heto Vacuum Centrifuge (Jouan, Alterød, Denmark) for approximately 24 hours. Some fractions were pooled for SDS-PAGE analysis.

2.4 Trypsin Digestion

For subsequent LC-MS/MS peptide identification, serum proteins were digested with trypsin in solution. As a first step, protein disulphate bridges were reduced by adding 10 mM dithiothreitol (DTT) and boiling it for 10 min. The proteins were then digested, overnight at 37°C, with a protein:enzyme ratio of 50:1 trypsin (prepared at a 1 µg/µl concentration with 25 mM NH_4HCO_3). The enzyme reaction was stopped by lowering the pH to 0.1% TFA and the samples were dried and frozen in -80°C for storage.

For protein identification from SDS-PAGE gels, gel pieces were first completely de-stained with 1 ml of 50% ACN in 25mM NH_4HCO_3, pH 8.4 by vortexing it three times, each time the solution was re-applied fresh and removed. Next the gel slices were dehydrated with 50 µl of 100% ACN for 5 min and dried completely in a speed-vacuum centrifuge for 10-20 min. For proteolysis the gel pieces were covered with trypsin solution (15-20 ng/µl trypsin in 25 mM NH_4HCO_3 pH 8.4) and allowed to swell for 45 min on ice. The trypsin solution was removed and the gel slices covered with 25 mM NH_4HCO_3, pH 8.4 and incubated at 37 °C overnight with rocking. In the morning the supernatant was recovered and later combined with tryptic peptides extracted 3x with 50 µl of 70% ACN, 5% FA with 5 min sonication. Peptides were stored at -80°C or processed to Zip-Tip clean-up.

2.5 Protein and Peptide Clean-up and Concentration

Empore™ cartridges and Zip-Tips were used with the same protocol. First the C18-material was conditioned with 60% MeOH and washed with ddiH_2O. The sample was loaded by pipetting up and down through the Zip-Tip and washed with ddiH_2O, before the proteins/ peptides were elute with 80% ACN, 0.1% TFA. For storage at -20°C and further analysis the peptides were concentrated to dryness in the vacuum concentrator 5300 (Eppendorf AG, Hamburg, Germany).

2.6 SELDI-ToF MS

2.6.1 Pre-Fractionation of Intact Proteins: Using WAX Separation

To increase the number of protein peaks visualized, an anion exchange fractionation procedure was performed in which the LMW serum was separated into six different fractions (flow through, pH 9, pH 7, pH 5, pH 4, pH 3, and organic wash). For this, 40µl of LMW serum (5 mg/ml concentration) was denatured with 30 µl of U9 buffer (9 M Urea, 2% CHAPS, and 50 mM Tris-HCl, (pH 9)) by vortexing for 20 min. Two additional samples, one of all cancer and the other of all control sera were pooled, and added to investigate the effect of pooling on peak as well as biomarker discovery. The QHyper DF resin (Ciphergen, Guildford, UK) was prepared by washing three times with 5 bed volumes of 50 mM Tris-HCl (pH 9). A 50/50 slurry of resin (180 µl) in 50 mM Tris-HCl (pH 9) was then aliquoted on a 96-well filter plate and equilibrated with 3 washes of 200 µl U1 buffer (1 M Urea, 0.22% CHAPS, and 50 mM Tris-HCl (pH 9)) on a vacuum manifold (Beckman Coulter Inc., High Wycombe, UK). The whole volume of the serum/U9 mix was then added to the resin in each well of the filter plate. For use of columns the sample was added to a column each and prepared the same way as for the 96-well plate. Plates were then vortexed for 30 min to bind the serum to the anion exchange resin. Consecutively, 100 µl of wash buffer was added to each well, vortexed for 10 min, at room temperature and the eluted fraction collected via a vacuum manifold. For the pH 9.0 fractions, only one 100 µl wash was performed. For each subsequent fraction, two 100 µl washes were performed. For fraction 1, the 100 µl flow through and 100 µl pH 9.0 wash are combined into one 200

µl fraction. The wash buffers for the different fractions were 50 mM Tris-HCl, 0.1% octyl glucopyranoside (OGP), (pH 9; F1), 50 mM HEPES, 0.1% OGP (pH 7; F2), 100 mM Na-Acetate, 0.1% OGP (pH 5; F3), 100 mM Na-Acetate, 0.1% OGP (pH 4; F4), 50 mM Na-Citrate, 0.1% OGP (pH 3; F5), and 33.3% isopropanol/ 16.7% ACN/ 0.1% TFA (F6). All of the pipetting steps a Biomek 2000 laboratory workstation (Beckman Coulter Inc. High Wycombe,UK) was used. Collected fractions were stored at -20°C until final analysis.

2.6.2 Binding of LMW Proteins to ProteinChip® Arrays

Each WAX fraction was then applied to two biochemically distinct ProteinChip® array surfaces (Ciphergen, Guildford, UK). The immobilized metal affinity capture coupled with copper (IMAC-$Cu2^{+}$) and weak cation exchange (WCX, CM10) arrays were chosen to increase the proportion of the serum proteome represented on the arrays for mass spectrometric analysis. Additionally the WCX array was also analysed under stringent as well as low stringency conditions. Each cancer sample was processed together with its age-matched control sample on the same chip.

LMW serum samples were analysed on four different ProteinChip arrays to define the best to use. The IMAC-$Cu2^{+}$ chips were preloaded with 50 µl of 100 mM $CuSO_4$ per spot on a bioprocessor module, which allows simultaneous processing of 12 ProteinChip® arrays, vortexed for 5 min and rinsed with H_2O. All arrays were then equilibrated twice with 200 µl of the appropriate binding buffer (Table 2.2). Ten µl of each fraction or sample and 90 µl of the respective binding buffer were then added on each spot and vortexed for 30 min. After discarding the remaining sample, the arrays were washed three times with 200 µl of binding buffer for 5 min and two brief water rinses. All samples were analysed twice; that is, complete SELDI-ToF MS profiles of the LMW filtrates were obtained from duplicate LMW serum samples, to minimize the effects of intra-assay variation. This was performed for Q10 and CM10 arrays.

Table 2.2: Binding and washing buffers for different chromatographic chip surfaces.

Array Types:	Binding and Washing buffers:
H50	10% acetonitrile, 0.1% TFA
Q10	low: 50mM Tris-HCl, pH 9
	high: 100mM Sodium acetate, pH 6
CM10	low: 100mM Sodium acetate, pH 4
	high: 50mM Tris-HCl, pH 7
IMAC 30	0.1 M sodium phosphate, 0.5 M NaCl pH 7

2.6.3 ProteinChip Analysis, Peak Detection and Data Analysis

Sinapinic acid solution as energy absorbing matrix was prepared according to the manufacturer's instructions (Ciphergen Biosystems Inc.); 12 mg/ml in 50% ACN/ 5% TFA, and 0.6 µl of the saturated solution applied twice to each spot on the chip. ProteinChip arrays were air dried and stored at room temperature in the dark until further use. Arrays prepared in Guildford were read on the ProteinChip Reader PCS4000 model and the data analyzed with Ciphergen Express software (Ciphergen Biosystems). Later in Cardiff, these arrays and all others were read on the Protein Biological System II ProteinChip reader and analysed using the ProteinChip software Ver. 3.1 (Ciphergen Biosystems Inc., Guildford, UK). All arrays were analysed in linear mode using the following settings: 150 shots/spectrum collected in positive ion mode, laser intensity 215, detector sensitivity 10 and focus mass 7500 Da. The mass spectrometer was externally calibrated using the "All-in One" peptide mass standard (Ciphergen Biosystems Ltd., Guildford, UK) which contains vasopressin (1084.2 Da), somatostatin (1637.9 Da), bovine insulin β-chain (3495.9 Da), human insulin recombinant (5807.6 Da), and hirudin (7033.6 Da). Data analysis was performed using the Ciphergen ProteinChip® software 3.1 after baseline subtraction, peak clustering and standardisation.

Peak detection was performed using the ProteinChip Biomarker software version 3.1 (Ciphergen Biosystems Inc.). All of the spectra were compiled, baseline subtracted and normalized to the total ion current of all peaks. The part of the spectrum with *m/z* values <1,000 was not used for analysis, as the energy absorbing matrix signal generally interfered with peak detection in this area. Peaks between 1,000 and 100,000 *m/z* ratios were auto-detected with a S/N ratio of >3 and the peaks clustered using

second-pass peak selection with S/N ratio >2 and a 0.3% mass window. Mass-to-charge values that were within the 0.3% mass accuracy window were considered to be identical between replicates. The resulting peak intensity values were analysed for differences using a Mann-Whitney *U* test within the software and also exported to Excel for Student's *t*-test analysis.

2.7 MALDI-ToF MS

2.7.1 Sample Preparation and Peak Detection

Protein samples were analysed using MALDI-ToF MS after mixing 1µl of protein in 0.1% TFA with 1 µl of matrix (either sinapinic acid 10 mg/ml in 70% acetonitrile, 0.1% TFA in H_2O (w/v); or α-cyano-4-hydroxycinnamic acid 10 mg/ml in 50% acetonitrile, 0.1% TFA in H_2O (w/v)). Spectra were acquired in linear mode for positive ions using a Voyager DE-STR (Applied Biosystems, Warrington, UK) after external calibration using either calibration mixture 2 (angiotensin I: *m/z* 1297.51 ACTH (clip 1–17): *m/z* 2094.46 ACTH (clip 18–39): *m/z* 2466.72; ACTH (clip 7–38): *m/z* 3660.19 and insulin (bovine): *m/z* 5734.59) or calibration mixture 3 (insulin (bovine): 5734.59; thioredoxin (E. coli): *m/z* 11674.48 and apomyoglobin (horse): *m/z* 16952.56) from Applied Biosystems (Warrington, UK) depending on the molecular weight range acquired. Settings for spectra acquisition were as per Table 2.3. Typically 400-800 shots were accumulated per spectrum and all spectra were basepeak normalized, 2 standard deviations of noise removed and a noise filter with a correlation factor of 0.7 applied. The spectra were analysed using Data Explorer 4.0 (Applied Biosystems, Warrington, UK) and peak lists were exported for data analysis.

Table 2.3: MALDI-ToF instrument settings for spectra acquisition.

Mode of operation	Linear
Extraction mode	Delayed
Polarity	Positive
Accelerating voltage	25000 V
Grid voltage	90%
Extraction delay time	400 nsec
Number of laser shots	60/spectrum
Laser Rep Rate	20.0 Hz
Input bandwidth	25 MHz

2.7.2 Peak Alignment and Data Analysis

The code for a peak alignment tool was written in Visual Basics for Excel, the software produces a reference peak list from across all spectra in the experiment. Next the program "aligns" peak intensity values in one spreadsheet combined from all spectra in the experiment. The aligned intensity values for replicate spectra from the same sample for each mass peak were they combined by calculating an average. The data from the breast cancer and control samples was compared by means of a *t*-test in Excel (as described below: section 2.9.2). Furthermore, variations within and between the two cohorts was visualized by principal-component-analysis (PCA) described below (section 2.9.3).

2.8 Protein Profiling using LC-MS/MS

2.8.1 Column Packing

For all peptide identification work, in-house packed pulled-tip C18 reverse-phase columns (10 cm x 75 μm I.D., 3 μm, 300 Å) were used. For packing an approximately 20 cm long piece of fused silica was heated approximately 5 cm from one end using a Bunsen burner. By gently pulling until the silica came apart, a tapered tip was formed at one end. The tip was then opened by cutting the very fine end off. The fused silica was attached to a reservoir containing C18 silica-bonded stationary phase (PepMap, Dionex, Camberley, UK) suspended into a slurry with isopropanol, which in turn was connected to a high pressure pump. In order, first the pressure was raised to 2000 psi and then the solvent (methanol) turned on. Then the pressure was increased to 5000 psi, packing of the material against the narrow tip opening could be observed visually against a dark surface. After at least 12 cm were packed the pressure and solvent were turned off and the column depressurized by loosening the pump-to-reservoir connection. The packed column was then "re-packed" and equilibrated for at least 30 minutes by HPLC with 2% ACN and 0.1% TFA in ddiH2O at a flow-rate of 200 nl/min.

Using the column holder at the front of the source allows fine tuning of the positioning and distance of the column. In order to attach a 10 cm or longer column and to allow direct spray into the source the manifold where the column is attached had to be changed. A devise was build (by myself in the metal workshop) made of aluminium, consisting of a track that was attached to the front of the nano-source and a sliding manifold, where the column holder is mounted on. The construction can be seen in Figure 2.1.

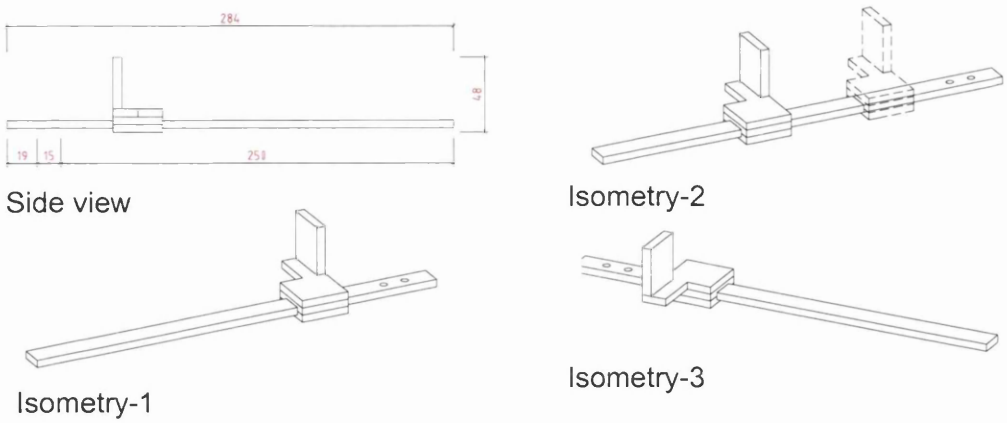


Figure 2.1 Sliding column holder for use of longer pulled-tip columns. Four isometries are shown to display the structure and use of the extension. The extension enables use of columns in variable lengths with the LCQ-Deca.

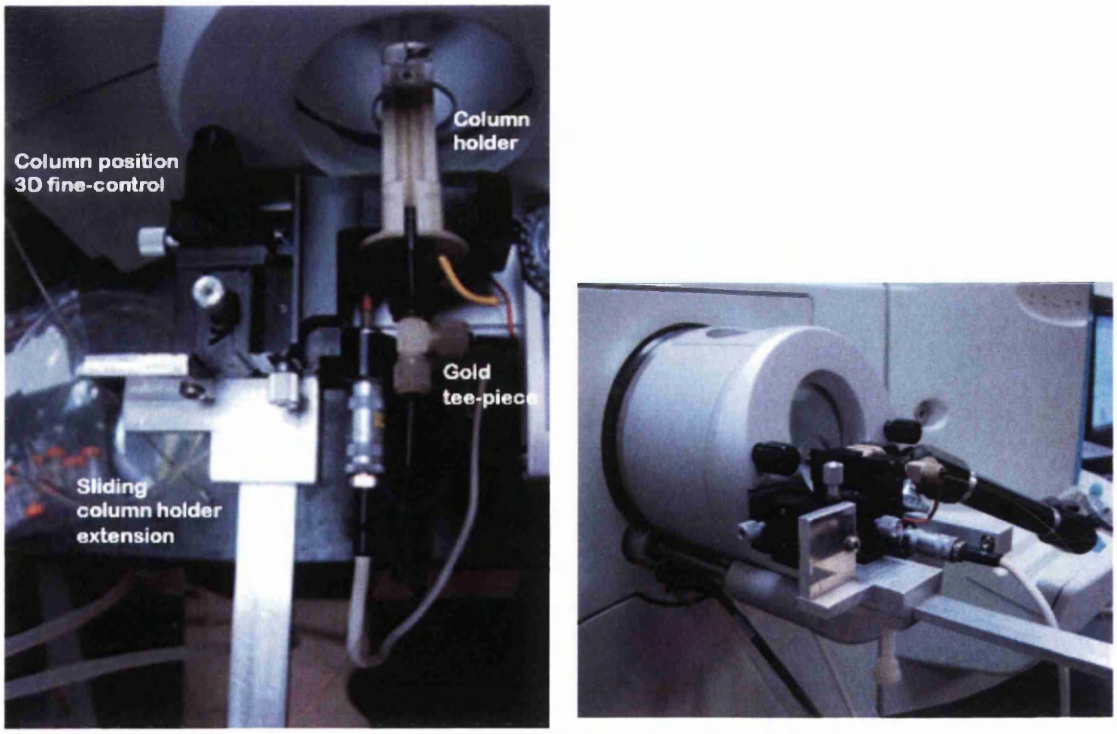


Figure 2.2: Picture of the nano-source, showing the sliding column holder attachment and the way it was attached to the nano-source.

2.8.2 LC-MS/MS Peptide Identification

For peptide identification, all serum samples analysed were first digested with trypsin described in section 2.4 and de-salted using Zip-Tips (section 2.5) or a C18-trap column. The peptide samples were injected made up in 0.1% FA in ddiH2O. For nano-LC-MS/MS experiments, after Zip-Tip desalting, an Ultimate HPLC Pump (LC-Packing, Dionex, Netherlands) was coupled to the LCQ Deca XP Plus quadruple ion trap MS (ThermoElectron, Hemel Hempstead, UK) equipped with an electrospray source. The column was connected to the Famos switch valve with a 50 μm I.D fused silica tubing via a gold tee-piece (Nanospray accessory kit, Thermo Finnigan, UK), the voltage was applied directly to the column through a gold-covered connection (Figure 2.2 and Figure 2.3). The Famos autosampler (LC-Packing, Dionex, Camberley, UK) was used to inject 8 μl of sample through a 10 μl injection loop directly onto the separation column. The autosampler was refrigerated to 10°C during the analysis. The flow-rate was 200 nl/min starting with a 15 min wash of 98% buffer A (2% ACN, 0.1% FA in H_2O) and then peptides were eluted using a stepwise gradient of 2% solvent B (0.1% FA in ACN) to 98% solvent B in 100 min. The column was then re-equilibrated for 10 min with 100% A. The electrospray voltage was held at 1.6 kV with a capillary temperature of 160°C. The mass spectrometer was operated in a data-dependent mode where each full MS scan was followed by three MS/MS scans, in which the three most abundant peptide molecular ions were dynamically selected for collision-induced dissociation (CID) using a normalized collision energy of 38%. The LC-MS/MS analysis was performed over a parent ion *m/z* range of 475-2000 Da. For online desalting using a C18-trap column (5 cm x 300 μm I.D. Dionex, Netherlands) the set-up and conditions described above were the same, except for here the sample was loaded onto the trap at a flow-rate of 30 μl/min for 3 min, using the Switchos pump and then eluted after column switching straight onto the C18-analytical column at a flow-rate of 200 nl/min.

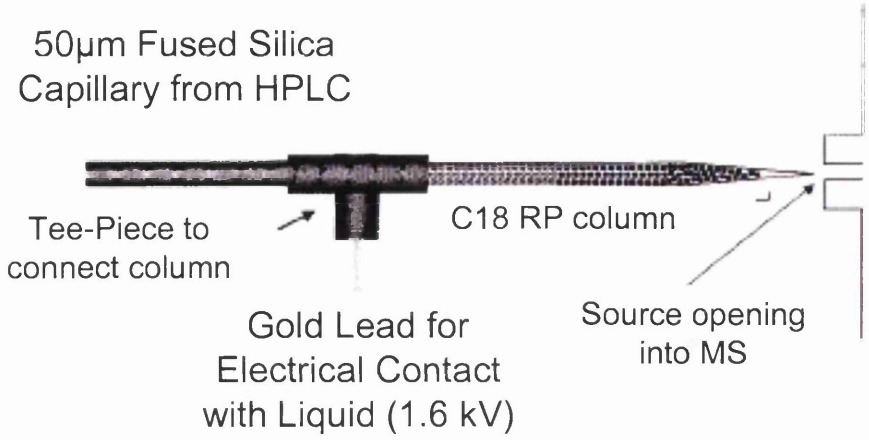


Figure 2.3: Schematic diagram of a pulled-tip C18 RP-column attached to the HPLC for online LC-MS/MS of peptides.

2.8.3 Sequest Searches for Peptide Identification and Protein Mapping

Tandem MS .raw files acquired by Xcalibur™ 1.4 SR1, were searched against the human FASTA database with TurboSequest within the Bioworks Browser Ver. 3.2 (Thermo Finnigan, Hemel Hempstead, UK). The settings for Sequest searching were set as seen in Table 2.4. The amino acid sequences and peptide identifications were then filtered and the peptide results exported into Excel.

Table 2.4: Bioworks Browser settings for TurboSequest searching.

Database	**human.fasta (date)**
Instrument	LCQ Deca XP
Enzyme	Fully Tryptic (KR)
Precursor mass	Monoisotopic using AMU
Fragment mass	Monoisotopic using AMU
Intensity threshold	50000 AMU
Missed cleavage sites	2
Precursor tolerance	1.4 AMU
Peptide tolerance	2 AMU
Fragment tolerance	1 AMU
Mass range	500.00 - 3500.00

2.8.4 Q-ToF Tandem MS for Peptide Identification

For tandem MS fragmentation of peptide peaks for amino acid sequencing, a LMW serum protein digest was separated by capillary HPLC separation using a in-house packed C18 reverse-phase column (15 cm, 300 μm I.D., 3μm, 300 Å) PepMap (Dionex, Camberley, UK). The peptides were eluted over 60 min from 98% A (2% ACN, 0.1% FA in water) to 100% B (0.1% FA in ACN) at a flow-rate of 4 μl/min and analysed online in a Q-ToF Ultima (Waters, UK). For tandem MS analyses the parameters were as follows: capillary voltage was set to 3.5 kV, the cone voltage to 35 V, with the cone gas flow to 40 litres/ hr. The source and desolvation temperature were set to 100°C and 120°C, respectively. MS/MS was performed by collision with argon gas (pressure 20 psi) with a collision energy of 32 eV at a scan time for 1 s and a scan range between 50 – 1600 Da. All spectra were analysed using MassLynx Version 4 (Micromass Ltd, Walters, UK).

2.9 General Data Analysis

2.9.1 Standardisation

Before statistical analysis the data was standardised using the same method as used by the "normalization factor method" as calculated as part of the Ciphergen ProteinChip® software 3.1. Standardisation compensates for varying levels of total protein in the samples or spectra intensity variations. It was assumed that on average, the number of proteins being expressed is the same across all samples being standardized. Also, the number of proteins whose expression levels change is few relative to the total number of protein peaks in the spectra. The method uses the "Total Ion Current" standardization which uses the area under the spectrum for standardization.

First the average of all peak intensities (TIC) in the standardization range was calculated by dividing the total intensity of all peaks by the number of peaks in the spectrum. Secondly a "normalization coefficient" (NC) was calculated which takes an average across the TIC of all selected spectra. Finally the "normalization factor" (NF) for each spectrum was calculated by dividing the NC by the TIC for each spectrum. The NF for each spectrum was used to multiply each peak intensity value within the standardization range. For the SELDI-ToF MS data this was performed as part of the

ProteinChip® software. For MALDI-ToF MS data this was manually calculated in Excel.

2.9.2 Unpaired Student's *t*-test

The Student's *t*-test calculates the difference between the means of the breast cancer sample group and the control samples for every peak in the spectrum. To determine whether the breast cancer and control samples have an equal variance an F-test was performed in Excel. When the variance was found to be significantly different (*p*-value <0.05) a type III *t*-test was performed, otherwise a type II *t*-test used. The difference between the two tests is the degrees of freedom used to determine whether the *t*-test result is significant or not. In excel the data from one group was compared to the data of the other group, selecting an unpaired *t*-test of either type II or type III, dependent on the result of the F-test calculated before.

2.9.3 Principal Component Analysis (PCA)

Principal component analysis is a multivariate projection method designed to extract and display the systematic variation in a data set. Even data characterised by thousands of variables can be reduced to just a few information rich plots. For this study, the multivariate analysis package SIMCA-P 10 (Umetrics, Umeå, Sweden) was used. The peak intensities and *m/z* value for every sample (including each replicate) were transposed and imported into SIMCA-P, here the primary (*m/z* value) and secondary (sample name) observer were defined. An un-supervised PCA analysis was performed and the scatter plot for PC1 and PC2 was exported in from of a 3D-plot.

2.10 References

[1] Schägger, H. and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal Biochem* **166**, 368-379.

[2] Merrell, K., Southwick, K., Graves, S. W., Esplin, M. S., Lewis, N. E. and Thulin, C. D. (2004) Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J Biomol Tech* **15**, 238-248.

[3] Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C. and Tempst, P. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal Chem* **76**, 1560-1570.

[4] Chen, Y. Y., Lin, S. Y., Yeh, Y. Y., Hsiao, H. H., Wu, C. Y., Chen, S. T. and Wang, A. H. (2005) A modified protein precipitation procedure for efficient removal of albumin from serum. *Electrophoresis* **26**, 2117-2127.

CHAPTER 3

Serum Sample Preparation and Pre-Fractionation

It has been established that of all the proteomes, particularly in relation to cancer development, serum or plasma may provide the greatest source of biological or medical information [1-4]. Chan *et al* [5] reported that the majority of lower abundance proteins identified in serum represent species secreted or shed by cells as a result of signalling, necrosis, apoptosis, and haemolysis. This great variety of proteins present makes serum the most informative proteome for clinical applications. Most cells interact with serum or shed their content into the serum after they have been damaged or following cell death and so the protein content of serum is thought to reflect the overall profile of an individual. The serum protein concentration ranges from 60-80 mg/ml, however 99% of the content is made up of only 22 proteins [6]. Despite these highly abundant proteins the serum proteome is one of the most complex, as the serum proteome exhibits a dynamic range of protein concentrations of up to eight orders of magnitude. The vast majority of such proteins are at such low abundance that they make up a small percentage of the entire protein content of serum [4]. The complexity of the serum proteome however may not only provide a vast amount of information about the disease state but also complicates the analytical process.

The highly abundant proteins such as human serum albumin (>50% total protein concentration), immunoglobulin G, antitrypsin, immunoglobulin A, transferrin, and haptoglobin [7] mask the detection of the remaining proteome during 2D gel electrophoresis and mass spectrometric analysis. To study the serum proteome, proteins such as immunoglobulin G (IgG) and albumin are often eliminated as the first step in the analytical protocol [1, 7-12]. Removal of some serum proteins can be achieved either by affinity chromatography, precipitation, magnetic beads or centrifugal ultrafiltration (UF). Many of these techniques have been utilised for

albumin removal. However, this step has been hypothesised to concomitantly remove potentially informative proteins/peptides associated with these highly abundant proteins targeted for depletion [12, 13]. Albumin and α2-macroglobulin especially act as molecular sponges, binding the low molecular weight (LMW) species and transporting them through serum [13, 14]. Nonetheless, under denaturing conditions, through addition of acetonitrile or isopropanol, protein-protein bonds are broken and LMW species released [4, 15, 16]. An important initial step in this project was to optimise a practical and reproducible pre-fractionation protocol after evaluating different techniques (albumin depletion, precipitation, weak anion exchange (WAX) separation and centrifugal ultrafiltration). This is described in this first result chapter.

3.1 SDS Gel Electrophoresis for Protein Visualisation

SDS-PAGE was used for visualisation purposes and quality control of the pre-fractionation results; here this method was optimised for LMW proteins. Glycine–SDS-PAGE, also known as Laemmli–SDS-PAGE [17], and tricine–SDS-PAGE [18] are based on glycine-Tris and tricine-Tris buffer systems, respectively, and are commonly used for separating proteins. The acrylamide gels used, are characterised by the total percentage concentration of both monomers (acrylamide and the crosslinker bisacrylamide) and the ratio of the concentration of the crosslinker relative to the acrylamide concentration (37:1 or 19:1). Together, Laemmli–SDS-PAGE and tricine–SDS-PAGE cover the protein mass range 1–500 kDa. However, the more commonly used Laemmli-SDS-PAGE is optimal for proteins >30 kDa whereas tricine-SDS-PAGE enables separation of proteins <30 kDa [18]. A direct comparison of the resolution capacity of a 17% acrylamide tricine–SDS-PAGE and Laemmli–SDS-PAGE in the low molecular mass range is shown in Figure 3.1. The tricine buffer enables better resolution of the LMW proteins forming sharper bands. The different separation characteristics of the two techniques are directly related to the strongly differing pK values of the functional groups of glycine and tricine in the electrophoresis buffer, that define the electrophoretic mobilities of the trailing ions (glycine and tricine) relative to the electrophoretic mobility of proteins [18].

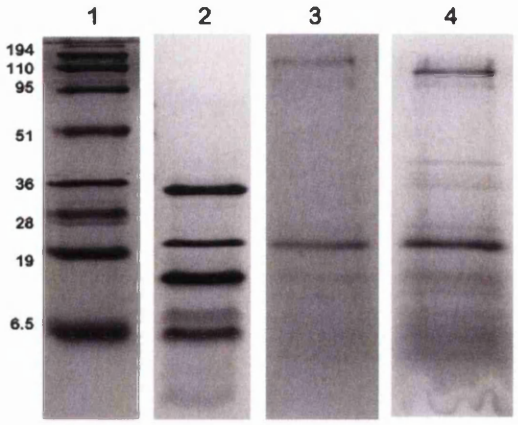


Figure 3.1: Comparison of SDS-PAGE with 17% acrylamide gels using a glycine and tricine running buffer. Shown is the resolution of the same concentration of LMW proteins by separation using glycine (lane 3) and tricine (lane 4) running buffer. Lanes 1 and 2 show molecular weight markers separated with the tricine running buffer.

A uniform acrylamide tricine–SDS gel covers a narrow mass range, for example a 16% gel covers the range 1–70 kDa and offers high resolution, especially for the small protein and peptide range. These uniform acrylamide tricine–SDS gels are almost exclusively used to separate very small proteins and peptides. Uniform high-acrylamide Laemmli gels cannot be used to access the small protein range, because the stacking limit in the Laemmli system is too high, and small proteins usually appear as smearing bands near the gel front. In a less convenient way, however, the small protein and peptide range can be accessed by making use of gradient gels that continuously de-stack proteins according to decreasing mass during electrophoresis. Laemmli-type gradient gels, for example, 8–16% acrylamide gels for the range 6–250 kDa, cover a wide range of masses [18]. To simplify gel casting, in the absence of a gradient gel pourer, a "layer" gel was designed, as described in the materials and methods (section 2.3.1). The separating gel, approximately 2/3 of the gel, consists of a 17% acrylamide/bisacrylamide at a 19:1 ratio and on top a "spacer" layer of 10% acrylamide/bisacrylamide which separates larger proteins between 70-200 kDa. And the up most layer a regular 4% stacking gel. This optimised SDS-PAGE gel used in combination with a tricine-Tris buffer could separate LMW proteins very well and was useful for visualisation and control purposes throughout the thesis.

3.2 Serum Pre-Fractionation Methods for MS Analysis

3.2.1 Affinity Chromatography

One of the most commonly cited methods of depleting highly abundant proteins from serum samples is affinity removal [8, 9, 19]. At the time when this project was started, very little information was available on depletion kits and their effectiveness. Although with hindsight, the Poros® cartridges were not the best kit to use, as only two proteins are removed, we tested them for serum albumin and protein G removal. The cartridges were used as described by the manufacturer, except for the addition of 20% ACN to break protein-protein interactions, and detailed in the Materials and Methods (section 2.3.4). A crude serum sample was injected onto the cartridges and washed with PBS. The antibodies immobilised on the cartridges bind HSA and protein G, which were eluted after 15 minutes with 12 mM HCl. Fractions from the flow-through (FT) and the eluted fractions were collected and analysed using SDS-PAGE and MALDI-ToF MS. The Poros® depletion cartridges showed little specificity for albumin and protein G, all lanes showed gel bands for albumin, including lane 3 (Figure 3.2 b) which should have contained proteins of the FT only. It was first suspected that the large amount of albumin in the FT caused the lane to be smeared. However, MALDI-ToF MS analysis showed that relatively little HSA was present in the FT (Figure 3.2 c). Hence the smearing may have been due to the high salt content from the PBS, which despite Zip-Tip concentration may have interfered with SDS-PAGE. More importantly, very few protein peaks were shown in the MALDI-ToF mass spectrum of the FT, in fact most of the peaks appear to be related (Figure 3.2 c). The major protein peak at *m/z* 14132.40 appears to be singly charged, forming a dimer at *m/z* 28280.07 and a doubly changed peak at *m/z* 7058.01 of the same protein. Besides these three peaks the spectrum contains very few peaks of significant intensity.

The eluted fraction in lane 7 should contain mainly albumin and protein G but also clearly shows many other protein bands in the gel of a range of molecular weights (Figure 3.2b). The same was observed by MALDI-ToF MS (Figure 3.2c).

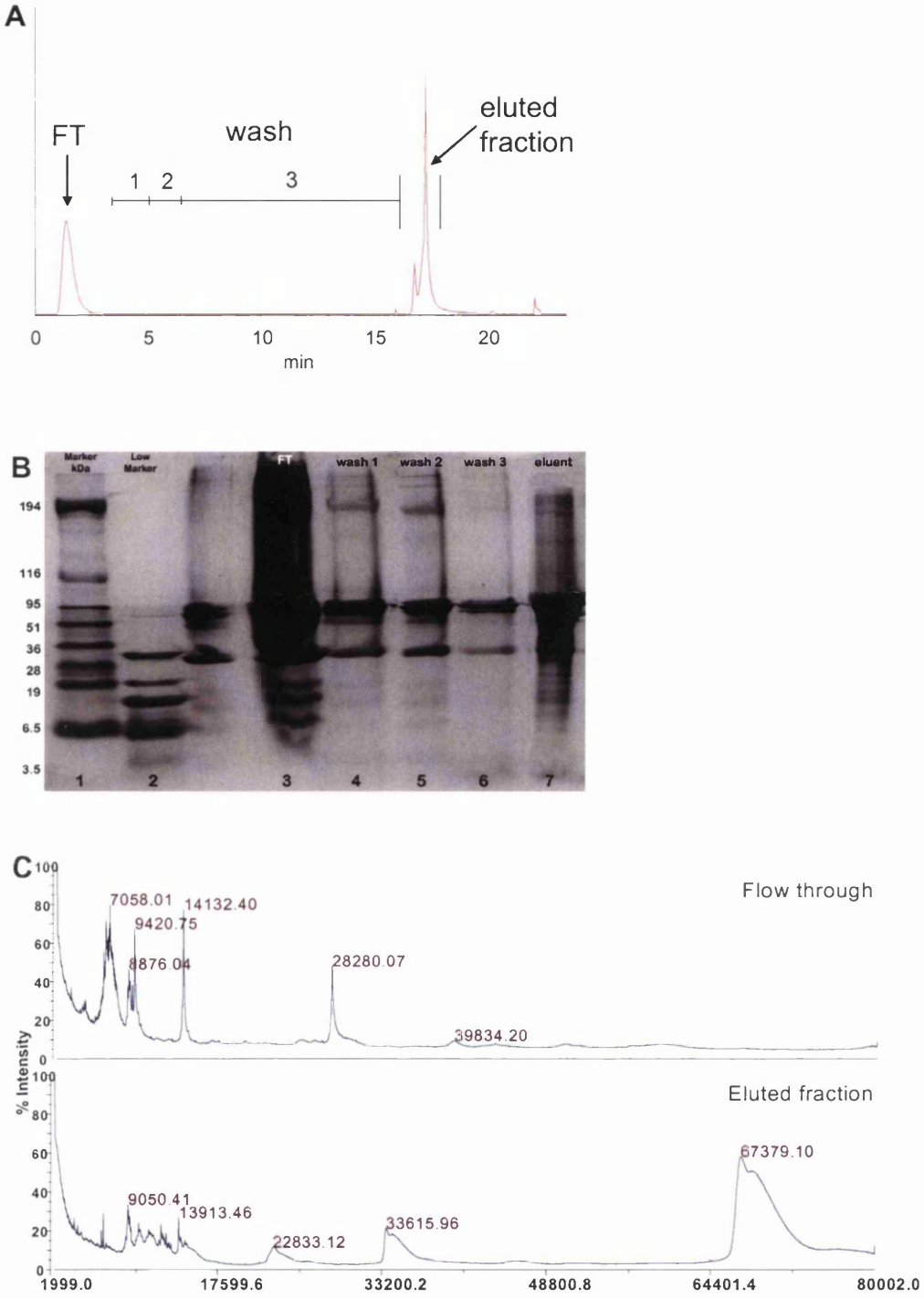


Figure 3.2: Immunoaffinity depletion of HSA and protein G from serum. (**A**) HSA and protein G bind to the anti-HSA and protein G cartridges, while un-bound proteins were washed onto the flow-through (FT) shown at 3 min. The bound proteins eluted after 17 min with HCl. (**B**) Protein fractions were collected and separated on a 17% tricine-SDS-PAGE. In lanes 1 and 2 molecular weight markers are shown, in lane 3 and in lanes 4-6 proteins from the FT and proteins that wash of the cartridges later are separated, respectively. In lane 7 proteins that were eluted from the cartridges are shown. (**C**) The MALDI-ToF MS spectrum shows proteins from the FT fraction and proteins that bound to the cartridges in the eluted fractions.

These findings suggest that, despite diluting the serum to account for the capacity of the affinity cartridges, albumin was still present to a large degree in the "depleted" FT fraction and that many other proteins were removed, as they non-specifically bind or were removed alongside HSA. This is in accordance with other reports using immunoaffinity, where it was clearly reported that many proteins bind to these carrier proteins and are concomitantly removed [13, 20]. Trying to improve specificity in removing albumin, different buffers, PBS with and without 20% ACN and NH_4HCO_3 with and without 20% ACN, were tested. The results showed that PBS with 20% ACN to break protein-protein interactions returned the most proteins in the depleted fractions. None of the other buffers tested improved the recovery of LMW and low abundant proteins (data not shown). The high salt concentrations of PBS and HCl caused down-stream problems with MS and SDS-PAGE which make the method impractical and not very high-throughput. Additionally, at the time, there was little data available on longevity and reproducibility of these affinity depletion systems. By now more accounts of immunoaffinity depletion using the Multiple Affinity removal System from Agilent can be found [8, 11, 21]. The multi-affinity depletion cartridge appears to successfully remove the 6 most abundant proteins from serum, however the FT also requires desalting and treatment prior to further analysis [19, 22].

3.2.2 Protein Precipitation

Protein precipitation with trichloroacetic acid (TCA)/acetone, acetonitrile or ethanol has also been useful for removal of albumin and other highly abundant serum proteins without the loss of the remaining peptides and proteins. The goal was that albumin remains soluble while the other proteins precipitate in the case of TCA/acetone or albumin precipitates while the other proteins and peptides remain soluble for ethanol and acetonitrile precipitation. Organic solvents are miscible and reduce the water activity around the protein and the dehydrated hydrophobic areas make it less soluble. As the water is displaced, oppositely-charged areas of different proteins become attracted and aggregate together. Albumin precipitation in serum relies on the "Cohn process" [23], where albumin is extracted and recovered from blood plasma. The process is based on the differential solubility of albumin and other plasma proteins

based on pH, organic solvent concentration, temperature, ionic strength, and protein concentration. Albumin has the highest solubility and lowest electric point of all the major plasma proteins. This makes it the final product to be precipitated in 40% ethanol, or separated from its solution as a paste [23]. This may be due to its size and isoelectric point; large proteins aggregate earlier as they have more surface area to interact with and their surface can be oppositely charged [24]. Using our method of albumin precipitation, one commonly used for removal of albumin, albumin actually specifically precipitates using 100% ethanol. Although that is the reverse to what was described by the Cohn process, the process still applies as albumin appears to have specific characteristics and behaves differently to the other serum proteins. For albumin removal TCA appears to bind to albumin and render it soluble in organic solvents [25], so that all other proteins precipitate and albumin remains in the solution. It is not clear why albumin in particular remains soluble in organic solvents while other proteins do not. It may be due to the degree of TCA binding to the protein, as albumin may have a large number of TCA binding sites. Chen *et al.* [25] tested the effect of different organic solvents and acids to optimise the protocol and showed convincing results. Three different methods of precipitation were tested as they were described in the literature and in the Materials and Methods (section 2.3.3). A 20 µl sample of human serum was precipitated as previously described using acetonitrile [20], ethanol [26] and TCA/acetone [25]. Although differences between the precipitation methods were observed, none of them showed sufficient removal of high abundant proteins (Figure 3.3). Albumin was still present in all the depleted fractions and further many other bands are visible in the fraction that should contain removable proteins only. In addition, very few proteins remained in the depleted fraction. TCA/acetone precipitation was most specific towards removing albumin from serum (Figure 3.3, lanes 6-7), similar to the results found by Chen Y. *et al.* [25]. However strong denaturing agents were necessary to re-solubilise the precipitated proteins, which results in down-the-line complications such as protein quantification and MS compatibility.

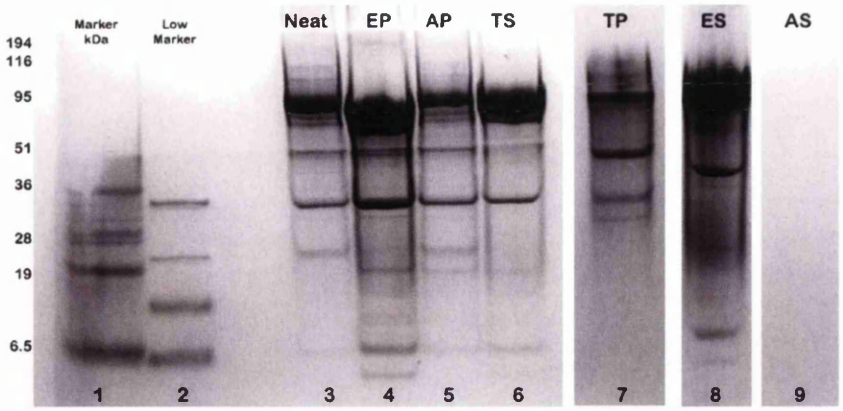


Figure 3.3: Serum preparation to reduce the protein complexity studied using 17% tricine SDS PAGE gels, stained with colloidal Coomassie blue. Proteins in the precipitate and the supernatant after solid-phase extraction using acetonitrile (albumin pellet: AP and protein supernatant: AS) are shown in lanes 5 and 9, ethanol (albumin pellet: EP and protein supernatant: ES) in lanes 4 and 8, and TCA/acetone (protein pellet: TP and albumin supernatant: TS) in lanes 6 and 7. Neat serum was run in lane 3 for comparison.

3.2.3 Weak Anion Exchange (WAX) Chromatography

Weak anion exchange chromatography is a widely used method for separating proteins prior to enzymatic digestion and MS analysis [24, 27, 28]. Ionic exchange separates proteins by their different net charge at a specific pH. The stationary phase inside the column binds proteins by electrostatic interactions. Anion exchange columns have a positively charged matrix, where a single positively charged nitrogen atom is immobilised on a chromatography support. Proteins are ionized in solution, depending on the pH of the buffer and the pI of the protein. When the pH of the buffer is greater than the pI of the protein, this will become negatively charged and bind to the anion exchange column. Generally proteins with a pI <8 will bind to an anion exchange column; for example albumin has a pI of 4.6. As described in more detail in the Materials and Method (section 2.3.5), serum was loaded onto the column and as the salt concentration (ammonium acetate) increased with the gradient, more proteins eluted from the column. Fractions were collected and separated by SDS-PAGE. The majority of proteins bound to the column with great affinity and required a higher concentration of ammonium acetate (Figure 3.4, lanes 11, 13 and 14). Some highly abundant proteins (such as albumin) were eluted in single fractions (Figure 3.4, lanes 6, 7 and 10). SDS-PAGE analysis of the fractions collected showed that, despite

desalting, the fractions that were pooled had a much higher salt concentration and caused smearing in the gel (Figure 3.4, lane 12). Although WAX separation was able to fractionate the serum proteins successfully, it was decided that this form of pre-fractionation is not optimal for preparing a large amount of samples. There are at least 10 fractions that would have to be collected and desalted for trypsin digestion and LC-MS/MS analysis. Desalting of the fraction was a critical step as the high salt concentrations used in the elution gradient interfered with trypsin digestion and MS analysis. It was later discovered that desalting is possible using a high vacuum centrifugal concentrator (Jouan, Alterød, Denmark), however without this instrument desalting is a limiting factor for subsequent analysis. As the endpoint of this experiment, removal of albumin, was shown using SDS-PAGE, no further investigations testing precipitation, affinity depletion or WAX were carried out.

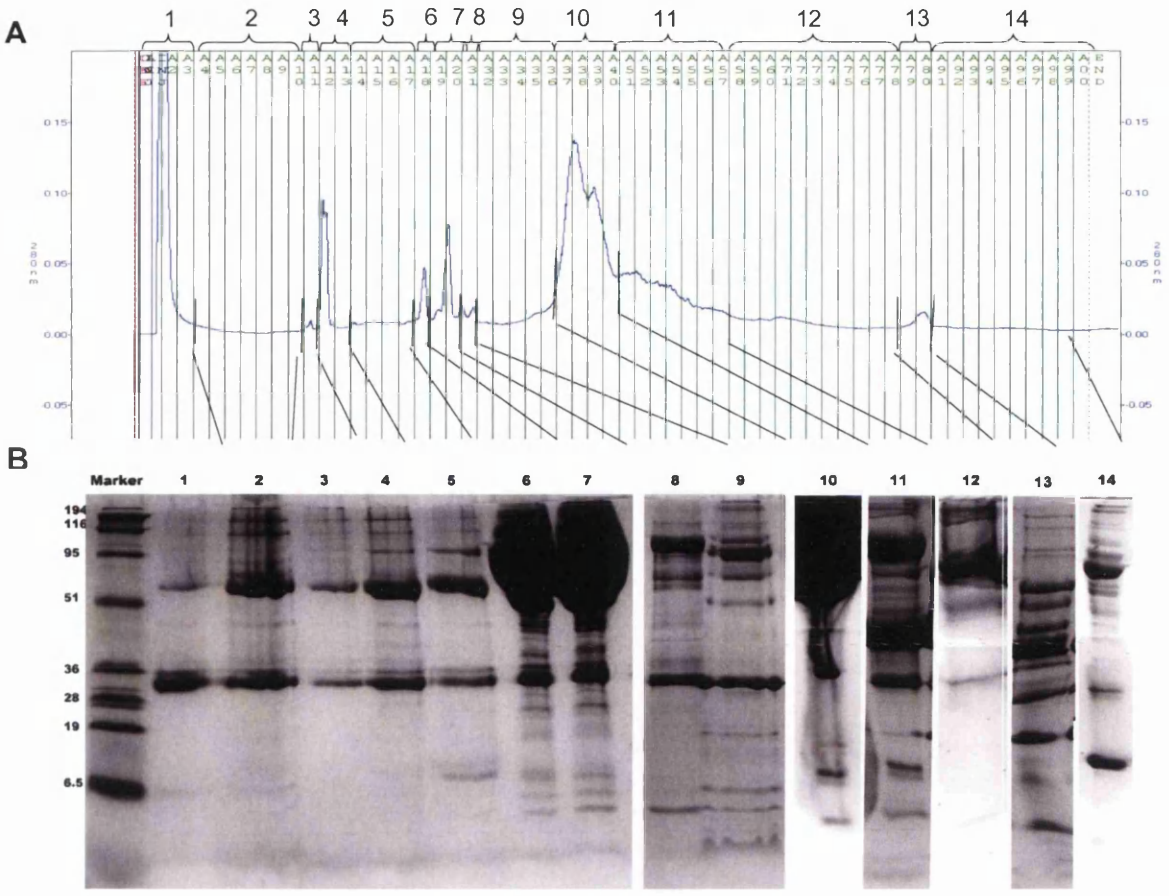


Figure 3.4: WAX separation (4.6 mm I.D x 100 mm, 1000 Å) 90 min gradient 5% ACN, 0.6 M NH_4 Acetate) of neat serum proteins (**A**). Fractions were collected and pooled for visualisation by tricine-SDS-PAGE. Some of the fractions appear smeared due to an increased protein and salt concentration (**B**).

3.2.4 Centrifugal Ultrafiltration (UF)

Ultrafiltration is a sieving mechanism creating two fractions. A retentate and a filtrate are formed, separating larger from smaller proteins, respectively. UF has no absolute cut-off point as some larger proteins pass through the membrane and some small ones stay behind. As proteins are not truly spherical, contain bound water when in serum or are engulfed by a sheath of fluid, their actual size can be changed when passing through the membrane. The shape especially affects the diffusion coefficient of individual proteins [29]. As soon as the trans-membrane pressure is exposed by centrifugation, the solvent flux starts and the solvent is pushed towards the membrane surface. The local solute concentration increases causing a concentration polarisation

effect. Concentration can be taken to the extent that larger proteins aggregate and form a thixotropic gel (also called gel layer or filter cake). UF is a dynamic process and formation of the gel layer reduces the amount of proteins passing through the membrane. The surface charge increases due to higher concentrations of protein and electrostatic interactions cause particle-particle bonds, which reduce the passage of smaller proteins. However protein aggregation is reversible and by re-diluting the proteins in the retentate, interactions can be broken. This can be explained by a reduction of Gibbs free energy as the solute is dispersed. Prior to ultrafiltration acetonitrile was added to the sample in order to release proteins and peptides bound to albumin [4, 15, 20]. As the majority of highly abundant proteins are of higher molecular weight, they could be removed using UF (Figure 1.6 in the Introduction and Table 3.2). The process is not very labour intensive as only one fraction is produced with a relatively small subset of serum proteins. After lyophilisation and desalting, this fraction can be used directly for MS analysis. SDS-PAGE can be performed on the fraction without any further preparation. Figure 3.5 shows a good separation of high and low molecular weight proteins. For comparison crude serum was run in lane 3 and to show reproducibility the filtrate and retentate of two different UFs are shown. The UF fractions showed no loss of LMW proteins to the HMW fraction and no HMW proteins passing through the membrane (Figure 3.5). UF is shown to be more efficient for the removal of HMW proteins such as albumin and immunoglobulins than Poros® affinity chromatography or precipitation. And furthermore this pre-fractionation method is higher-throughput as up to 24 filters can be prepared at one time in a large bench top centrifuge. It was therefore decided to use UF for further preparation of serum samples in MS biomarker discovery. The rest of this chapter will deal with the evaluation and optimisation of UF to ensure the use of the most effective and reproducible conditions for centrifugal UF for removal of highly abundant proteins.

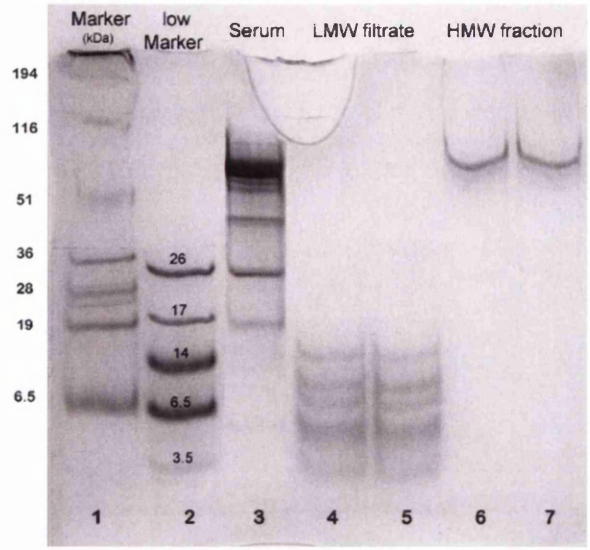


Figure 3.5: Serum samples separated by centrifugal ultrafiltration. Lanes 1 and 2: molecular weight markers, lane 3: unfiltered serum (25 µg), lanes 4 and 5: LMW serum filtrate (50 µg) and lanes 6 and 7: HMW retentate (25 µg).

3.3 Optimisation of UF for Biomarker Discovery

3.3.1 Evaluation of Different Centrifugal Filters

Following discussion with Millipore, a number of MWCO membranes were tested and compared. All filters were used according to the manufacturer's specification. The aim of the filtration was to either completely remove or sufficiently reduce the levels of albumin and IgG in the serum sample, and to amplify the LMW protein concentration. Additionally, as blocking of some filters had been observed, minimal blocking and easy use of the filters was studied.

For serum fractionation Millipore recommended Amicon® Ultra-15 and Centriprep® centrifugal units. Centricon® Plus-20 and Centriplus® centrifugal filters were not recommended, as they were apparently designed for samples with low protein concentration (Figure 3.6). Nevertheless, the two latter filter types are the most commonly used in the literature for serum UF for proteomics studies [4, 14, 30-33].

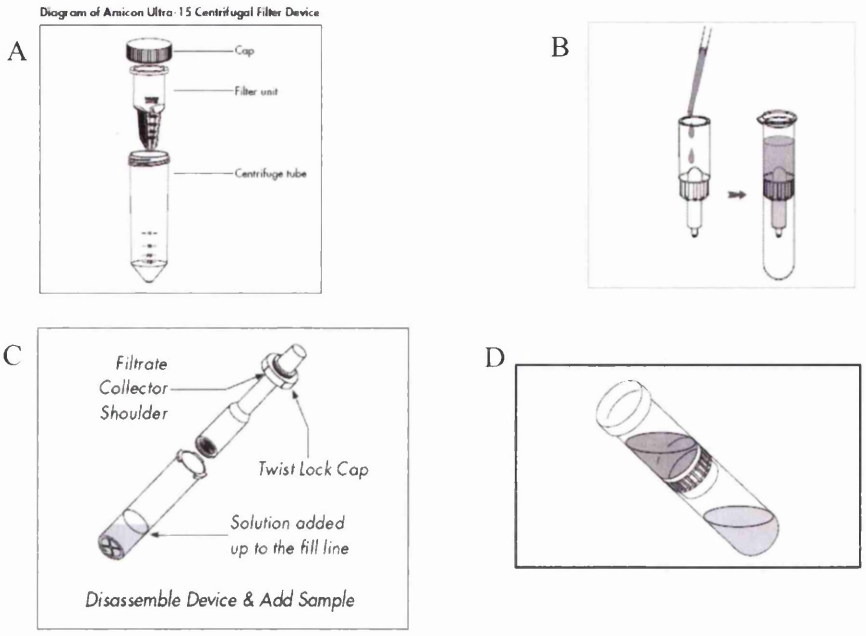


Figure 3.6: Millipore centrifugal ultrafiltration devices. The filtration mechanism of each filter is shown. (**A**) The Amicon Ultra filters are cross-flow membranes so that residue collecting at the bottom of the tube during centrifugation cannot block the membrane and the solvent passed tangentially. (**B**) In Centricon Plus filters the membrane is located in the middle of the device and the serum collects around the filter. Again this is a cross-flow membrane. (**C**) Centriprep filters work by a reverse centrifugal mechanism. The serum sample is in the lower tube; another tube with a dead-end filter at the end is placed close to the bottom and during centrifugation the LMW fraction is pushed directly through the filter up into the inner tube. (**D**) Centriplus devices have a dead-end membrane in the bottom of the upper tube; by centrifugal force small proteins are pushed directly through the membrane in to the lower cup. (Diagrams were taken from Millipore manuals.)

For comparison of the filters, the centrifugation speed and duration recommended by the manufacturer was used as shown in Table 3.1; assuming that these conditions would be optimal. The membranes were prepare as recommended by the manufacturer and described in the Materials and Methods (section 2.3.2). Each filter was processed in duplicate. An aliquot of the same serum sample was used for all filters.

Additionally to the different filters a new centrifugation protocol was also tested in an attempt to achieve better protein recovery using Centriplus® filters only. The serum

sample was diluted and filtered twice to increase the recovery of LMW species especially those bound to albumin and other HMW proteins. The concentration of the serum was reduced to avoid blocking of the filters and the centrifugation speed lowered to 750 xg. The protein concentration of each LMW filtrate, from each filter type, was determined using a BCA assay and the filtrate was further analysed using tricine-SDS-PAGE and MALDI-ToF MS. Except for the Amicon Ultra (30 kDa) filters, all protein concentration recoveries were reproducible (Table 3.1).

Table 3.1: Running conditions used for each of the different UF membranes, also shown is the expected recovery of cytochrome C (cyto C) and protein concentration recovered in the LMW filtrate with standard errors across 2 replicates.

Filter	MWCO	dilution	speed	time	expected cyto C retention	LMW filtrate recovery (%)	standard error
Amicon A10	10 kDa	1:1 x2	2000 xg	40 min	95	3.4	0.95
Amicon Ultra	30 kDa	1:5 x2	3000 xg	38 min	35	3.0	1.40
Amicon Ultra	50 kDa	1:5	3000 xg	38 min	35	1.6	0.10
Centricon Plus	30 kDa	1:20 x2	2000 xg	36 min	---	2.9	0.35
Centriprep	30 kDa	1:5	1500 xg	38 min	15	1.5	0.00
Centriplus	30 kDa	1:5	2000 xg	16 hours	75	2.4	0.19
Centriplus	50 kDa	1:5	2000 xg	16 hours	10	2.3	0.00
Centriplus	30 kDa	1:20 x2	750 xg	16 hours	75	2.6	0.41
Centriplus	50 kDa	1:20 x2	750 xg	16 hours	10	2.8	0.95

Tricine-SDS-PAGE and MALDI-ToF MS analysis showed that most of the filters allowed some albumin to pass through the membrane (Figure 3.7 and Figure 3.9). However, the results also showed that the filters recommended by Millipore (Amicon Ultra A30 and A50, Centriprep® and Amicon® Ultra A10) were permeable to albumin and other proteins larger than the molecular cut-off value. Centriprep® columns showed no proteins at all in the LMW fraction. The Centriplus® filters on the other hand managed to retain the major portion of albumin and, when analysed by tricine gel electrophoresis, LMW protein band intensities were increased compared to neat serum and the filtrates from the other UF membranes (Figure 3.7 and Figure 3.9). Although not recommend for samples as concentrated as serum, the Centriplus® filters performed significantly better than the other filters tested. Most convincingly, MALDI-ToF MS showed the superiority of the Centriplus® filters in recovering

LMW proteins *m/z* <6000 (Figure 3.9). Many more LMW peaks are shown in the spectrum and no peaks >10 kDa. In comparison, the spectra from the other filters show less LMW peaks and the Amicon filtrates contain may HMW proteins including albumin. As the LMW region was the main region of interest, Centriplus® filters with a MWCO level of 50 kDa were chosen to be used for further analysis. Furthermore using the new protocol, higher concentrations of LMW proteins were obtained of a greater number of proteins.

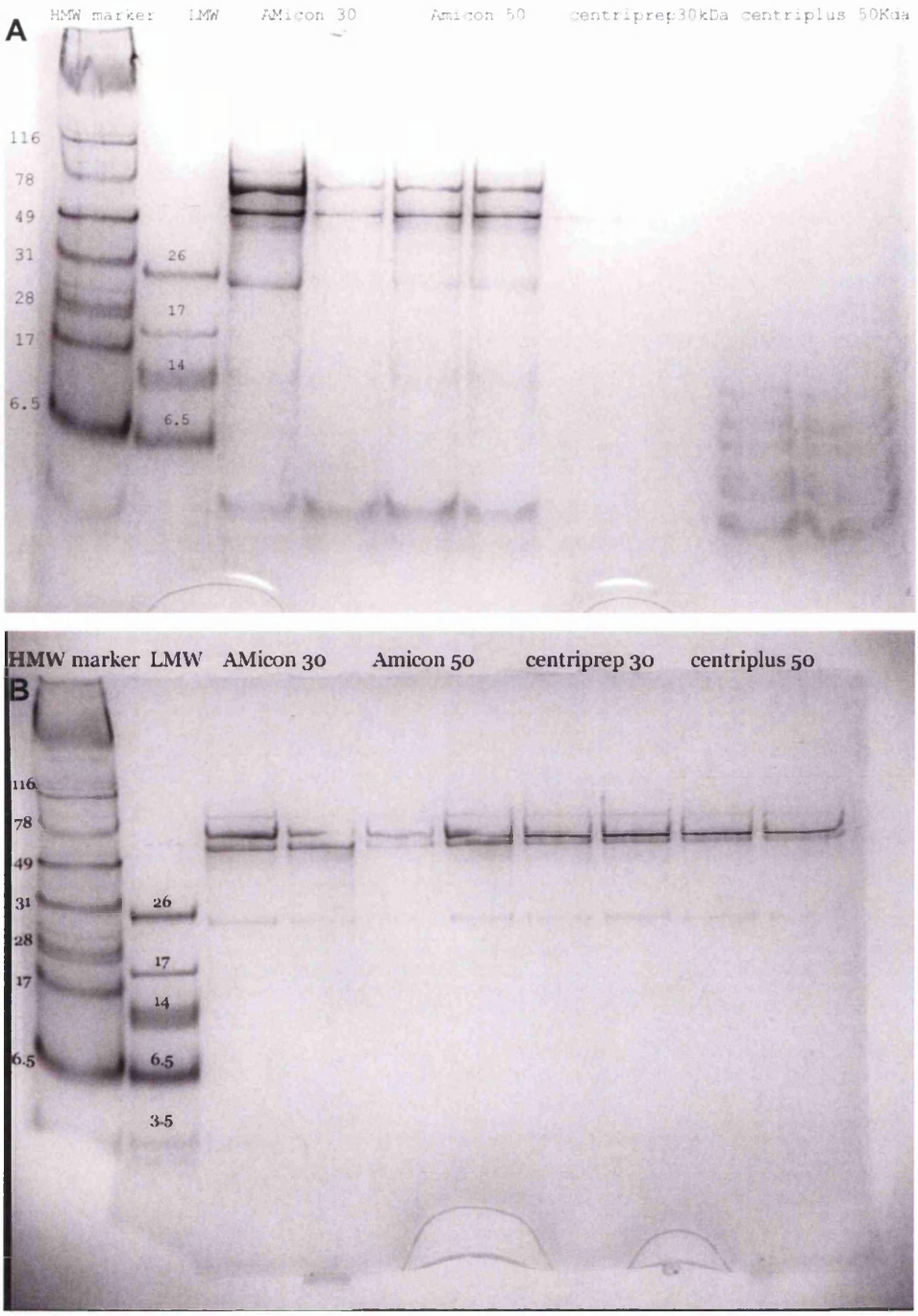


Figure 3.7: SDS-PAGE separation of the filtrate and retentate for comparison of the performance of different centrifugal ultrafiltration membranes. **(A)** LMW filtrates, lane 1 and 2: molecular weight markers; lanes 3-4: Amicon Ultra (30 kDa), lanes 5-6: Amicon Ultra (50 kDa), lane 7-8: Centriprep (30 kDa) and lane 9-10: Centriplus (50 kDa). **(B)** HMW retentate of the same membranes as in A.

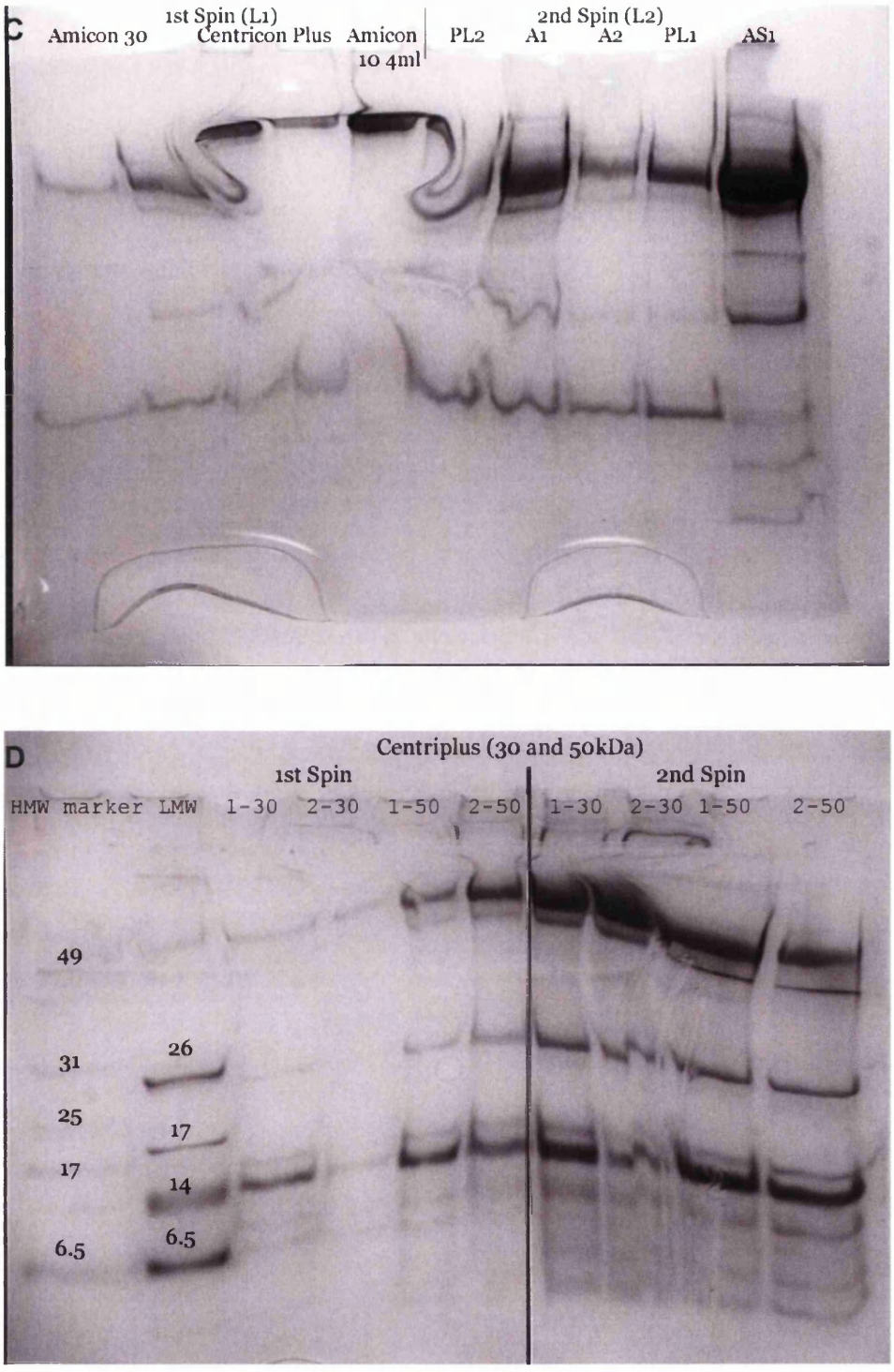


Figure 3.8: SDS-PAGE separation of the filtrate and retentate for comparison of the performance of different centrifugal ultrafiltration membranes continued. **(C)** LMW filtrate from other filters: Lane 1-2: Amicon Ultra (30 kDa) spin 1; lanes 3-4: Centricon Plus (30 kDa) spin 1, lane 5: Amicon Ultra (10 kDa) spin 1; lanes 7-8: Amicon Ultra (30 kDa) spin 2; lanes 6 and 9: Centricon Plus (30 kDa) spin 2, and lane 10: Amicon Ultra (10 kDa) spin 2. **(D)** Comparison of Centriplus filters with different MWCO 30 and 50 kDa. Lanes 1 and 2: molecular weight markers; lanes 3-4: spin 1 of 30 kDa, lanes 5-6: spin 1 of 50 kDa filters, lanes 7-8: spin 2 of 30 kDa filters; lanes 9-10: spin 1 of 50 kDa filters.

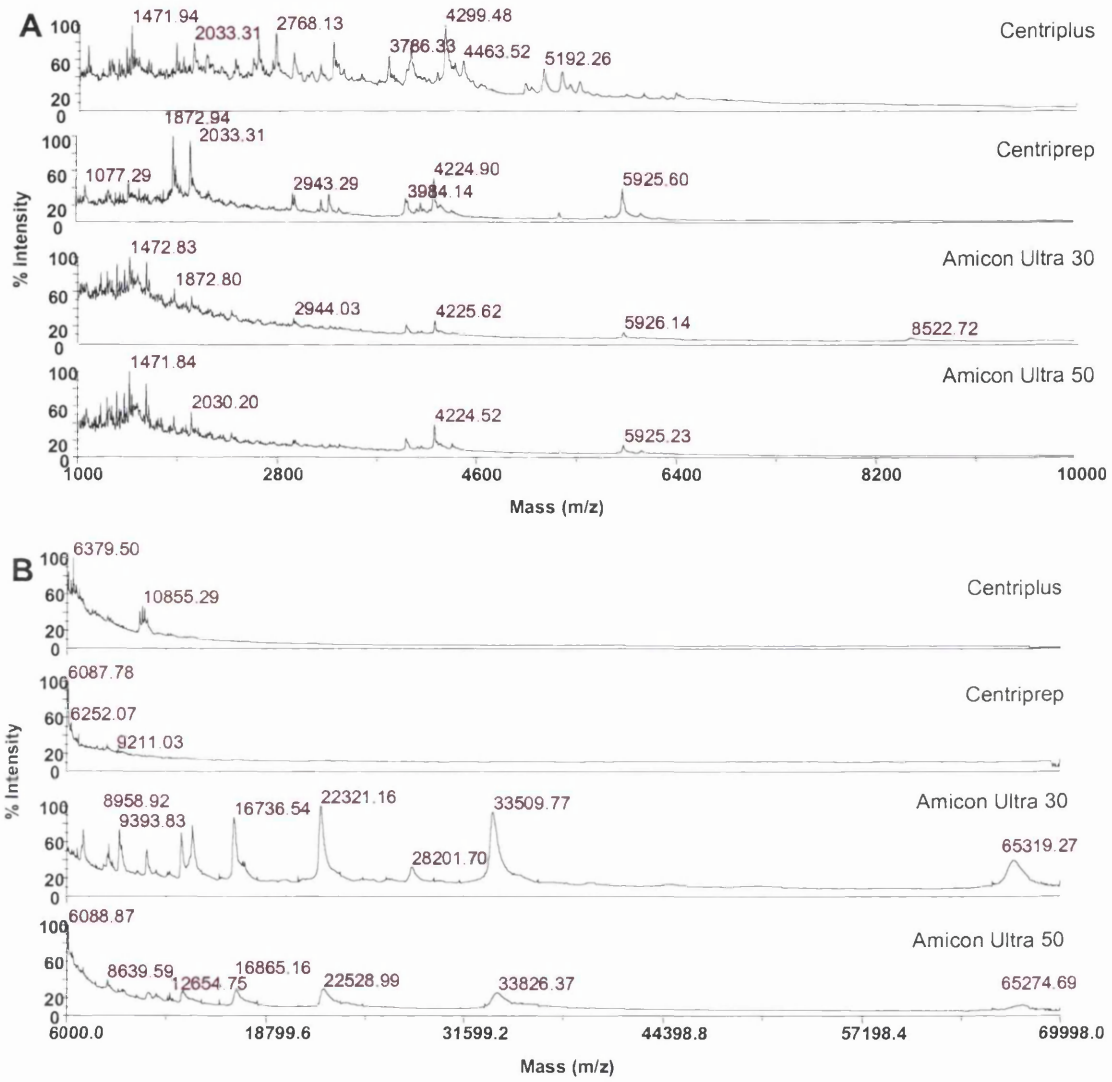


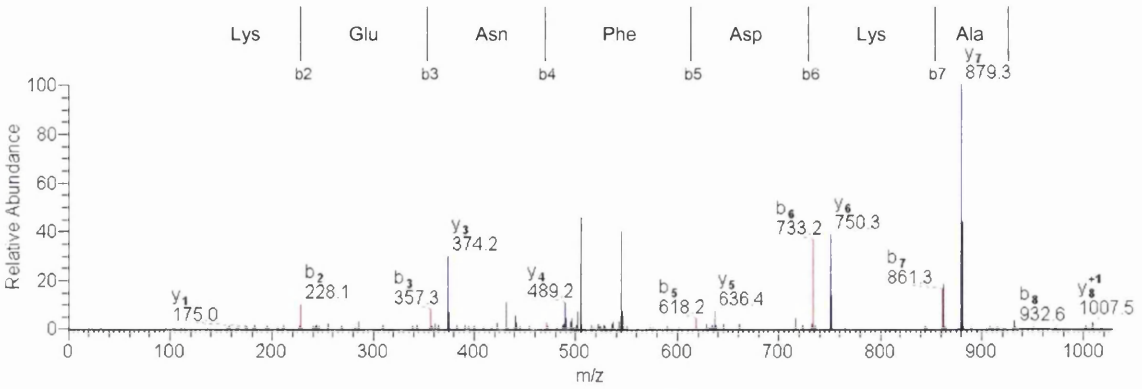
Figure 3.9: MALDI-ToF analysis of LMW filtrates form different Millipore membranes tested. Two mass ranges (**A**) 1000-10000 Da and (**B**) 6000-70000 Da are shown.

3.3.2 Centrifugation Speed and Protein Concentration

For maximum recovery of LMW proteins 50 kDa MWCO filters were used throughout the remaining study, as according to the manufacturer these allow the majority of LMW proteins to pass through the membrane whilst efficiently retaining high molecular weight proteins. Earlier experiments, using serum samples at a concentration of 12 mg/ml, resulted in the formation of a gel layer and protein loss.

This gel layer has previously been described to contain albumin and fibrinogen [34, 35]. However, to resolve this, the serum sample was diluted to a concentration of 3 mg/ml and the retentate re-diluted for a second round of filtration. The centrifugation speed during UF influences the recovery and retention of proteins. The recommended centrifugation speed is 3000 xg, however it has been hypothesised that a lower speed may allow more time for small proteins/peptides to be released from larger molecules. In a report by Georgiou [30] the filters were used at very high centrifugation speeds, which resulted in insufficient retention of high abundance proteins, 64% of proteins passed through the filter into the filtrate [30]. Another report, using UF for LMW proteome purification used the recommended centrifugation speed of 3000 xg; here 0.2% LMW proteins of the total serum proteome were isolated [4], however it was also estimated that the LMW proteome should make up at least 1% of all serum proteins. In this study, lowering the centrifugation speed from 3000 xg to 750 xg increased the recovery of LMW proteins by 0.7% (determined by BCA protein assay). Using a more dilute starting material, lower spin speed and multiple filtrations, we were able to recover 2.8% of the total protein content. The SDS-PAGE analysis confirmed that proteins isolated were of LMW as no bands were visible in the upper region of the gel (Figure 3.6).

To investigate whether the lower centrifugation speed would benefit protein identification, each filtrate was analysed by LC-MS/MS. For this, un-filtered crude serum and the LMW filtrates from UF at 3000 xg and 750 xg were analysed by 1D-LC-MS/MS after enzymatic digestion with trypsin. The MS/MS spectra were searched and analysed against the UniProt human FASTA database using TurboSequest through Bioworks 3.2 [36] as described in the Materials and Methods (section 2.8). The results shown are all from fully tryptic peptides and filtered within the Bioworks Browser for peptide probability (P = 0.001), for high stringency cross correlation (X_{corr}) 1+, 2+, 3+ = 1.9, 2.5, 3.2 and delta correlation (deltaCn) = 0.08. For increased confidence in the proteins mapped from identified peptides, the result filters were set relatively stringent, so only fragmentation spectra with good *b* and *y* ion sequences were identified from the database (Figure 3.10). As a result some proteins may have been lost from the identification list. To increase protein identifications each protein digest was analysed by LC-MS/MS in duplicate and the results combined.



	AA	B	Y	
1	V	100.08	-	9
2	K	228.17	1007.53	8
3	E	357.21	879.43	7
4	N	471.26	750.39	6
5	F	618.32	636.35	5
6	D	733.35	489.28	4
7	K	861.45	374.25	3
8	A	932.48	246.16	2
9	R	-	175.12	1

Figure 3.10: Tandem MS analysis of the precursor *m/z* 554.94, all *b* and y ions were accounted for. This peptide was matched to plasma retinol binding protein (RETBP_HUMAN), which occurs at relatively low serum concentrations (3.17 µg/ml).

The proteins matched by TurboSequest are listed in Table 3.2. While 27 proteins were detected in crude serum, 13 were albumin or immunoglobulin. In addition, half (13) of the proteins were matched by only 1 peptide. As shown, 17 of the 24 proteins with information on serum concentration levels were from high abundance proteins [37]. Only 4 low abundance proteins were detected and the dynamic range of protein concentrations covered 4 orders of magnitude with pregnancy zone protein precursor the lowest reported protein concentration of 8.4 µg/ml. The most abundant protein detected, as expected, was serum albumin identified with 45 unique peptides.

In comparison, after UF the lowest abundant protein detected was prothrombin precursor which has a reported serum concentration of 1.2 ng/ml. The dynamic range of protein concentrations in the LMW filtrate (750 xg) was shown to be over a range of 7 orders of magnitude. Peptide mass fingerprinting further emphasised the need for albumin depletion and protein pre-fractionation prior to further analysis.

Identification of proteins in the LMW filtrate after 750 xg compared to 3000 xg showed an increase of 22% more proteins in the 750 xg LMW filtrate. Further, comparison of the LC-MS/MS results from the LMW filtrates with those of un-filtered crude serum showed many proteins were removed from the LMW filtrate, including 12 immunoglobulins (Figure 3.11). In conclusion, a spin speed of 750 xg, lower sample concentration and multiple filtrations were found to be the optimal UF conditions for an efficient serum sample preparation and were used for future analysis.

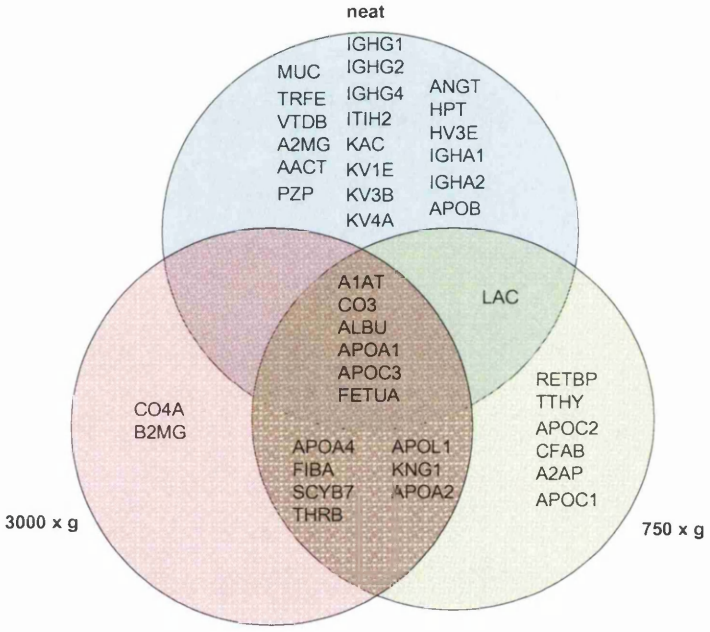


Figure 3.11: The Venn diagrams illustrate the overlap of protein identifications after LC-MS/MS analysis from different processing procedures. Comparing neat serum and two different centrifugation speeds (3000 xg and 750 xg). An increase in protein identifications was observed in UF filtrates using the lower spin speed protocol. Additional proteins were identified that are usually masked during proteomic analysis of neat serum.

Table 3.2: Proteins detected by RP-LC-MS/MS in crude serum and serum UF at 3000 x g and 750 x g in comparison. Highly abundant proteins in serum are highlighted in **bold** (the information on serum abundance was taken from Polanski and Anderson [37] and high and low levels from Tanaka *et al.* [38]).

Accession no	Protein	SwissProt ID	Protein probability	Molecular mass	Matched peptides	Abundance (pg/ml)
crude serum						
P02768	**Serum albumin precursor**	**ALBU_HUMAN**	**1.47E-12**	**69322**	**45**	**4.10E+10**
P02787	**Serotransferrin precursor**	**TRFE_HUMAN**	**8.44E-14**	**77000**	**15**	**4.00E+09**
P01009	**Alpha-1-antitrypsin precursor**	**A1AT_HUMAN**	**3.20E-07**	**46707**	**6**	**1.40E+09**
P01023	**Alpha-2-macroglobulin precursor**	**A2MG_HUMAN**	**4.75E-09**	**163174**	**6**	**1.80E+09**
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**1.30E-10**	**30759**	**6**	**1.40E+09**
P01857	**Ig gamma-1 chain C region**	**IGHG1_HUMAN**	**1.82E-11**	**36083**	**6**	**high**
P00738	**Haptoglobin precursor**	**HPT_HUMAN**	**2.37E-07**	**45177**	**4**	**1.25E+09**
P01876	**Ig alpha-1 chain C region**	**IGHA1_HUMAN**	**4.08E-08**	**37631**	**4**	**high**
P01842	**Ig lambda chain C regions**	**LAC_HUMAN**	**2.26E-10**	**11230**	**4**	**high**
P02774	Vitamin D-binding protein precursor	VTDB_HUMAN	1.41E-08	52929	3	?
P01877	**Ig alpha-2 chain C region**	**IGHA2_HUMAN**	**4.08E-08**	**36485**	**2**	**high**
P01859	**Ig gamma-2 chain C region**	**IGHG2_HUMAN**	**1.37E-06**	**35862**	**2**	**high**
P01861	**Ig gamma-4 chain C region**	**IGHG4_HUMAN**	**1.82E-11**	**35918**	**2**	**high**
P01834	**Ig kappa chain C region**	**KAC_HUMAN**	**3.33E-08**	**11602**	**2**	**high**
P01011	Alpha-1-antichymotrypsin precursor	AACT_HUMAN	1.04E-04	47621	1	?
P02765	Alpha-2-HS-glycoprotein precursor	FETUA_HUMAN	5.71E-06	39300	1	6.10E+08
P01019	Angiotensinogen precursor	ANGT_HUMAN	1.45E-07	53121	1	
P04114	Apolipoprotein B-100 precursor	APOB_HUMAN	1.27E-07	515242	1	
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	4.22E-05	10846	1	1.20E+08
P01024	**Complement C3 precursor**	**CO3_HUMAN**	**1.49E-04**	**187045**	**1**	**high**
P01766	**Ig heavy chain V-III**	**HV3E_HUMAN**	**1.17E-08**	**13218**	**1**	**high**
P01597	**Ig kappa chain V-I**	**KV1E_HUMAN**	**7.17E-07**	**11654**	**1**	**?**
P01620	**Ig kappa chain V-III**	**KV3B_HUMAN**	**1.26E-10**	**11768**	**1**	
P01625	**Ig kappa chain V-IV**	**KV4A_HUMAN**	**1.61E-05**	**12632**	**1**	**?**
P01871	Ig mu chain C region	MUC_HUMAN	1.95E-04	49526	1	-
P19823	Inter-alpha-trypsin inhibitor	ITIH2_HUMAN	7.16E-04	106370	1	low
P20742	Pregnancy zone protein precursor	PZP_HUMAN	1.27E-08	163732	1	8.36E+06
UF at 3000 xg						
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	5.65E-13	45372	11	
P02671	**Fibrinogen alpha chain precursor**	**FIBA_HUMAN**	**1.97E-09**	**94914**	**8**	**high**
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**3.07E-08**	**30759**	**6**	**1.40E+09**
P01024	**Complement C3 precursor**	**CO3_HUMAN**	**3.13E-11**	**187045**	**3**	**high**
P02768	**Serum albumin precursor**	**ALBU_HUMAN**	**1.85E-08**	**69322**	**3**	**4.10E+10**
P02765	Alpha-2-HS-glycoprotein precursor	FETUA_HUMAN	8.22E-14	39300	2	6.10E+08
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	2.58E-09	10846	2	1.20E+08
P00734	Prothrombin precursor	THRB_HUMAN	3.28E-07	69992	2	1.20E+03
P02652	Apolipoprotein A-II precursor	APOA2_HUMAN	1.91E-04	11168	2	2.44E+08
P02775	**Platelet basic protein precursor**	**SCYB7_HUMAN**	**4.40E-09**	**13885**	**1**	**5.94E+09**
O14791	Apolipoprotein-L1 precursor	APOL1_HUMAN	1.37E-07	43900	1	
P01009	**Alpha-1-antitrypsin precursor**	**A1AT_HUMAN**	**2.33E-09**	**46707**	**1**	**1.40E+09**
P0C0L5	Complement C4-A precursor	CO4A_HUMAN	7.84E-06	192672	1	
P01042	Kininogen-1 precursor	KNG1_HUMAN	8.62E-05	71900	1	
P61769	Beta-2-microglobulin precursor	B2MG_HUMAN	1.49E-07	13706	1	2.05E+06
UF at 750 xg						
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	5.69E-12	45372	13	
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**1.21E-10**	**30759**	**11**	**1.40E+09**
P02671	**Fibrinogen alpha chain precursor**	**FIBA_HUMAN**	**1.01E-08**	**94914**	**7**	**high**
P01024	**Complement C3 precursor**	**CO3_HUMAN**	**1.30E-09**	**187045**	**5**	**high**
P02765	Alpha-2-HS-glycoprotein precursor	FETUA_HUMAN	1.88E-12	39300	3	6.10E+08
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	1.20E-12	10846	3	1.20E+08
P01009	**Alpha-1-antitrypsin precursor**	**A1AT_HUMAN**	**8.54E-10**	**46707**	**3**	**1.40E+09**
P02652	Apolipoprotein A-II precursor	APOA2_HUMAN	3.43E-06	11168	3	2.44E+08
P02775	Platelet basic protein precursor	SCYB7_HUMAN	9.48E-10	13885	2	5.94E+09
P00734	Prothrombin precursor	THRB_HUMAN	2.05E-10	69992	2	1.20E+03
P02766	Transthyretin precursor	TTHY_HUMAN	4.60E-09	15877	2	3.00E+08
P02768	**Serum albumin precursor**	**ALBU_HUMAN**	**1.22E-04**	**69322**	**1**	**4.10E+10**
P00751	Complement factor B precursor	CFAB_HUMAN	1.39E-07	85479	1	
P01842	**Ig lambda chain C regions**	**LAC_HUMAN**	**9.88E-09**	**11230**	**1**	**high**
P02655	Apolipoprotein C-II precursor	APOC2_HUMAN	1.65E-05	11277	1	
O14791	Apolipoprotein-L1 precursor	APOL1_HUMAN	1.02E-06	43900	1	
P02753	Plasma retinol-binding protein precursor	RETBP_HUMAN	7.18E-05	22995	1	3.17E+07
P08697	Alpha-2-antiplasmin precursor	A2AP_HUMAN	2.17E-05	54531	1	
P01042	Kininogen-1 precursor	KNG1_HUMAN	8.09E-05	71900	1	
P02654	Apolipoprotein C-I precursor	APOC1_HUMAN	9.52E-04	9326.1	1	6.10E+07

The number of detected proteins in the LMW filtrate was relatively low and since the magnitude of protein concentrations is large it was assumed that the filtrate is still too complex for 1D-LC-MS/MS. Therefore the LMW serum sample was first separated by tricine-SDS-PAGE and all bands were excised and trypsin digested (Figure 3.12). The extracted peptides were then analysed by LC-MS/MS and identified against the human FASTA database using Bioworks v. 3.2 as described above. This may give a more comprehensive account of what is actually present in the LMW filtrate.

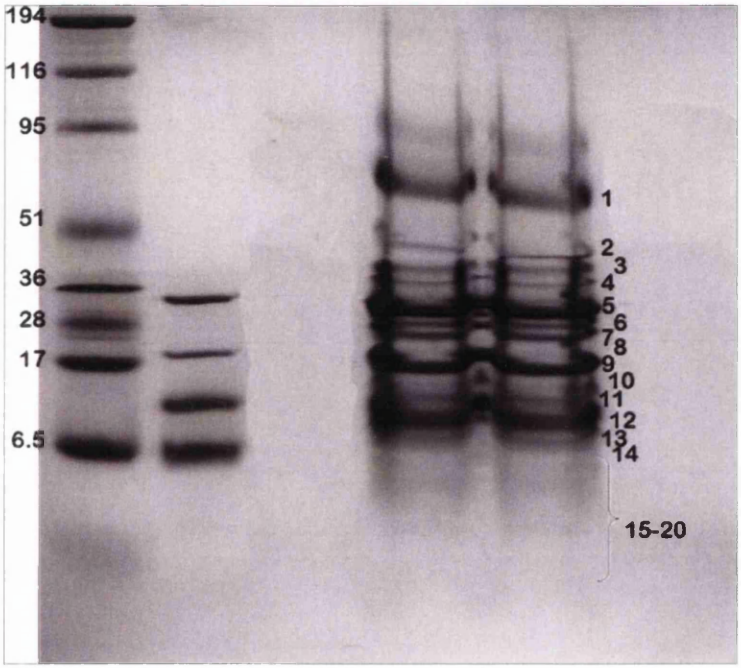


Figure 3.12: SDS-PAGE of LMW serum proteins run in two lanes. The individual MW bands were excised and individually digested for peptide mass fingerprinting by LC-MS/MS. Bands 15-20 were excised every 2mm across both lanes for increased recovery of proteins.

In total, 47 proteins were detected from all the bands excised. In each band an obvious protein was detected with a high number of peptides and the "correct" molecular weight, highlighted in **bold** (Table 3.3). Some low abundance proteins not detected in crude serum were identified including prothrombin precursor, as before. Above, to increase the number of identifications from the LMW filtrates, the peptide results from two independent 1D-LC-MS/MS runs were combined. This was not possible for the gel bands; otherwise even more proteins with higher confidence may have been detected. Proteins with a very high molecular weight were detected in some of the

lower gel bands. It is possible that these proteins originate from split of subunits or fragments during the UF in the denaturing buffer. Moreover, the presence of a much reduced proportion of some highly abundant HMW proteins in the filtrate was to be expected as UF is a dynamic process and the shape of a molecule can cause it to slip through the pores of the membrane. However, albumin and other HMW proteins were removed satisfactorily, allowing the detection of other less abundant proteins during MS or electrophoresis (Table 3.2 and Figure 3.12).

Table 3.3: Proteins detected from gel bands of LMW serum by LC-MS/MS. 47 unique proteins were detected, proteins that match the molecular weight of the gel band are marked in bold, abundance levels in serum were taken from www.plasmaproteome.org [37] and **high**/low indicators from [38]. High abundance proteins (> 1.0 x10^9 pg/µl) are marked in **bold**.

Accession no	Protein	SwissProt ID	Protein probability	Molecular mass	Matched peptides	abundance (pg/ml)
Protein gel band 1						
P01042	**Kininogen-1 precursor**	**KNG1_HUMAN**	**9.12E-04**	**71900.1**	**1**	
P02768	**Serum albumin precursor**	**ALBU_HUMAN**	**1.03E-09**	**69321.6**	**17**	**4.10E+10**
P02790	**Hemopexin precursor**	**HEMO_HUMAN**	**2.45E-06**	**51643.3**	**2**	
P01009	Alpha-1-antitrypsin precursor	A1AT_HUMAN	7.45E-11	46707.1	3	**1.40E+09**
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	1.33E-08	45371.5	1	**high**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	4.16E-09	30758.9	2	**1.40E+09**
Protein gel band 2						
Q5TAX3	Zinc finger CCHC domain-containing protein 11	ZCH11_HUMAN	7.36E-04	185015.3	1	
P02768	Serum albumin precursor	ALBU_HUMAN	7.38E-08	69321.6	1	**4.10E+10**
P14136	**Glial fibrillary acidic protein, astrocyte**	**GFAP_HUMAN**	**7.49E-08**	**49849.7**	**1**	
P06727	**Apolipoprotein A-IV precursor**	**APOA4_HUMAN**	**8.37E-12**	**45371.5**	**18**	**high**
P25311	Zinc-alpha-2-glycoprotein precursor	ZA2G_HUMAN	5.99E-08	33850.9	3	
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	1.62E-08	30758.9	2	**1.40E+09**
Protein gel band 3						
P01024	Complement C3 precursor	CO3_HUMAN	1.53E-05	187045.3	1	**high**
Q14624	Inter-alpha-trypsin inhibitor heavy chain H4 precursor	ITIH4_HUMAN	3.38E-12	103294.1	6	low
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	4.11E-07	94914.3	1	**2.72E+09**
P00751	Complement factor B precursor	CFAB_HUMAN	3.33E-07	85478.6	2	**high**
P01042	Kininogen-1 precursor (Alpha-2-thiol proteinase	KNG1_HUMAN	8.25E-04	71900.1	1	
P01009	**Alpha-1-antitrypsin precursor**	**A1AT_HUMAN**	**1.42E-10**	**46707.1**	**1**	**1.40E+09**
P06727	**Apolipoprotein A-IV precursor**	**APOA4_HUMAN**	**4.69E-07**	**45371.5**	**5**	**high**
P02760	**AMBP protein precursor [Contains: Alpha-1-microglobulin**	**AMBP_HUMAN**	**7.79E-07**	**38974.0**	**1**	
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**2.58E-12**	**30758.9**	**13**	**1.40E+09**
P00915	Carbonic anhydrase 1	CAH1_HUMAN	1.80E-06	28721.3	1	low
P24592	Insulin-like growth factor-binding protein 6 precursor	IBP6_HUMAN	4.76E-11	25306.2	1	
P02753	Plasma retinol-binding protein precursor	RETBP_HUMAN	5.55E-07	22995.3	1	3.17E+07
P41222	Prostaglandin-H2 D-isomerase precursor	PTGDS_HUMAN	1.30E-07	21015.4	1	
P02766	Transthyretin precursor	TTHY_HUMAN	4.03E-08	15877.1	4	3.00E+08
Protein gel band 4						
Q14624	Inter-alpha-trypsin inhibitor heavy chain H4 precursor	ITIH4_HUMAN	3.67E-06	103294.1	1	low
P06727	**Apolipoprotein A-IV precursor**	**APOA4_HUMAN**	**1.48E-08**	**45371.5**	**11**	**high**
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**9.83E-09**	**30758.9**	**7**	**1.40E+09**
P00915	**Carbonic anhydrase 1**	**CAH1_HUMAN**	**1.07E-06**	**28721.3**	**4**	low
P02753	Plasma retinol-binding protein precursor	RETBP_HUMAN	3.90E-05	22995.3	1	3.17E+07
P41222	Prostaglandin-H2 D-isomerase precursor	PTGDS_HUMAN	8.96E-08	21015.4	2	3.00E+08
P02766	Transthyretin precursor	TTHY_HUMAN	5.88E-08	15877.1	3	3.00E+08
P01842	Ig lambda chain C regions	LAC_HUMAN	1.32E-08	11229.5	1	**high**
Protein gel band 5						
P98160	Basement membrane-specific heparan sulfate proteoglycan core protein precursor	PGBM_HUMAN	2.87E-04	468528.2	1	
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	8.72E-06	94914.3	2	**2.72E+09**
P01009	Alpha-1-antitrypsin precursor	A1AT_HUMAN	1.52E-09	46707.1	1	**1.40E+09**
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	7.90E-12	45371.5	2	**high**
P02649	Apolipoprotein E precursor	APOE_HUMAN	4.73E-09	36131.8	1	**?**
P02647	**Apolipoprotein A-I precursor**	**APOA1_HUMAN**	**3.46E-11**	**30758.9**	**18**	**1.40E+09**
P00746	**Complement factor D precursor**	**CFAD_HUMAN**	**2.49E-05**	**26986.8**	**1**	**high**
P22352	**Glutathione peroxidase 3 precursor**	**GPX3_HUMAN**	**1.82E-08**	**25489.0**	**4**	
P02753	**Plasma retinol-binding protein precursor**	**RETBP_HUMAN**	**5.68E-07**	**22995.3**	**3**	3.17E+07
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**2.99E-12**	**15877.1**	**4**	3.00E+08
P02775	Platelet basic protein precursor	SCYB7_HUMAN	6.06E-06	13885.4	1	low
P01623	Ig kappa chain V-III region WOL	KV3E_HUMAN	9.34E-08	11738.9	5	**high**
P01842	Ig lambda chain C regions	LAC_HUMAN	7.55E-09	11229.5	2	**high**

Accession no	Protein	SwissProt ID	Protein probability	Molecular mass	Matched peptides	abundance (pg/ml)
Protein gel band 6						
P98160	Basement membrane-specific heparan sulfate proteoglycan core protein precursor	PGBM_HUMAN	4.77E-07	468528.2	1	
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	1.54E-10	94914.3	2	**2.72E+09**
P01042	Kininogen-1 precursor	KNG1_HUMAN	1.34E-05	71900.1	1	
P02768	Serum albumin precursor	ALBU_HUMAN	9.70E-06	69321.6	1	**4.10E+10**
P01009	Alpha-1-antitrypsin precursor	A1AT_HUMAN	1.59E-10	46707.1	2	**1.40E+09**
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	1.88E-10	45371.5	18	**high**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	7.64E-10	30758.9	8	**1.40E+09**
P22352	**Glutathione peroxidase 3 precursor**	**GPX3**	**1.57E-05**	**25489.0**	**3**	
P16035	**Metalloproteinase inhibitor 2 precursor**	**TIMP2_HUMAN**	**7.66E-12**	**24383.1**	**1**	3.40E+04
P02753	**Plasma retinol-binding protein precursor**	**RETBP_HUMAN**	**3.06E-08**	**22995.3**	**6**	3.17E+07
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**3.23E-11**	**15877.1**	**5**	3.00E+08
P02775	Platelet basic protein precursor	SCYB7_HUMAN	1.29E-08	13885.4	1	low
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	1.16E-08	10845.5	1	1.20E+08
Protein gel band 7						
Q9H8L6	Multimerin-2 precursor	MMRN2_HUMAN	1.53E-08	104352.1	1	
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	3.14E-06	94914.3	2	**2.72E+09**
P01042	Kininogen-1 precursor	KNG1_HUMAN	2.27E-06	71900.1	2	
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	1.65E-08	45371.5	7	**high**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	6.46E-08	30758.9	2	**1.40E+09**
P02753	**Plasma retinol-binding protein precursor**	**RETBP_HUMAN**	**4.93E-08**	**22995.3**	**6**	3.17E+07
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**1.02E-08**	**15877.1**	**6**	3.00E+08
Protein gel band 8						
P02787	Serotransferrin precursor	TRFE_HUMAN	7.39E-07	76999.7	1	**4.00E+09**
P02768	Serum albumin precursor	ALBU_HUMAN	7.17E-08	69321.6	4	**4.10E+10**
P01857	Ig gamma-1 chain C region	IGHG1_HUMAN	1.57E-10	36083.2	1	**high**
P02647	Apolipoprotein A-I precursor	_HUMAN	5.37E-07	30758.9	3	**1.40E+09**
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**6.29E-09**	**15877.1**	**4**	3.00E+08
Protein gel band 9						
P02768	Serum albumin precursor	ALBU_HUMAN	4.93E-06	69321.6	3	**4.10E+10**
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	1.17E-04	45371.5	1	**high**
P01857	Ig gamma-1 chain C region	IGHG1_HUMAN	1.64E-11	36083.2	1	**high**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	6.75E-04	30758.9	1	**1.40E+09**
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**6.53E-08**	**15877.1**	**6**	3.00E+08
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	3.17E-09	10845.5	1	1.20E+08
Protein gel band 10						
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	5.07E-07	94914.3	1	**2.72E+09**
P01009	Alpha-1-antitrypsin precursor	A1AT_HUMAN	6.72E-09	46707.1	1	**1.40E+09**
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	1.49E-07	45371.5	3	**high**
P02766	**Transthyretin precursor**	**TTHY_HUMAN**	**3.84E-09**	**15877.1**	**5**	3.00E+08
P01034	**Cystatin C precursor**	**CYTC_HUMAN**	**5.05E-11**	**15789.1**	**3**	3.20E+05
P07737	**Profilin-1**	**PROF1_HUMAN**	**2.46E-06**	**14913.5**	**1**	
P02775	Platelet basic protein precursor	SCYB7_HUMAN	5.47E-11	13885.4	1	low
P02655	Apolipoprotein C-II precursor	APOC2_HUMAN	3.49E-07	11276.8	1	
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	5.50E-07	10845.5	1	1.20E+08
Protein gel band 11						
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	3.86E-08	45371.5	1	**high**
P02766	Transthyretin precursor	TTHY_HUMAN	1.46E-06	15877.1	1	3.00E+08
P02775	**Platelet basic protein precursor**	**SCYB7_HUMAN**	**1.16E-08**	**13885.4**	**2**	low
P61769	**Beta-2-microglobulin precursor**	**B2MG_HUMAN**	**3.98E-05**	**13705.9**	**1**	2.05E+06
P02655	Apolipoprotein C-II precursor	APOC2_HUMAN	1.86E-04	11276.8	1	
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	4.45E-10	10845.5	1	1.20E+08
Protein gel band 12						
P01042	Kininogen-1 precursor	KNG1_HUMAN	4.64E-05	71900.1	1	
P02775	**Platelet basic protein precursor**	**SCYB7_HUMAN**	**5.94E-09**	**13885.4**	**2**	low
P02656	Apolipoprotein C-III precursor	APOC3_HUMAN	2.76E-09	10845.5	1	1.20E+08

Accession no	Protein	SwissProt ID	Protein probability	Molecular mass	Matched peptides	abundance (pg/ml)
Protein gel band 13						
P02656	**Apolipoprotein C-III precursor**	**APOC3_HUMAN**	**3.80E-09**	**10845.5**	**1**	1.20E+08
Protein gel band 14						
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	3.82E-06	94914.3	4	**2.72E+09**
P01042	Kininogen-1 precursor	KNG1_HUMAN	5.00E-04	71900.1	1	
P48677	Glial fibrillary acidic protein homolog	GFAP_HUMAN	2.92E-06	41824.5	1	
P02649	Apolipoprotein E precursor	APOE_HUMAN	1.26E-07	36131.8	1	**?**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	1.51E-09	30758.9	4	**1.40E+09**
P02775	Platelet basic protein precursor	SCYB7_HUMAN	4.02E-08	13885.4	2	low
P02655	**Apolipoprotein C-II precursor**	**APOC2_HUMAN**	**6.08E-06**	**11276.8**	**4**	
P02652	**Apolipoprotein A-II precursor**	**APOA2_HUMAN**	**1.35E-04**	**11167.9**	**1**	2.44E+08
P02656	**Apolipoprotein C-III precursor**	**APOC3_HUMAN**	**6.92E-10**	**10845.5**	**3**	1.20E+08
P02776	**Platelet factor 4 precursor**	**PLF4_HUMAN**	**4.25E-05**	**10837.9**	**1**	9.70E+03
Protein gel band 15						
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	5.49E-09	30758.9	1	**1.40E+09**
P02652	**Apolipoprotein A-II precursor**	**APOA2_HUMAN**	**8.74E-04**	**11167.9**	**1**	2.44E+08
P02656	**Apolipoprotein C-III precursor**	**APOC3_HUMAN**	**4.49E-10**	**10845.5**	**1**	1.20E+08
P02654	**Apolipoprotein C-I precursor**	**APOC1_HUMAN**	**7.77E-05**	**9326.1**	**1**	6.10E+07
Protein gel band 16						
P02671	Fibrinogen alpha chain precursor	FIBA_HUMAN	3.45E-05	94914.3	1	**2.72E+09**
P00734	Prothrombin precursor	THRB_HUMAN	6.05E-05	69992.2	3	1.20E+03
P06727	Apolipoprotein A-IV precursor	APOA4_HUMAN	3.86E-04	45371.5	1	**high**
P02647	Apolipoprotein A-I precursor	APOA1_HUMAN	1.03E-08	30758.9	2	**1.40E+09**
P10124	Secretory granule proteoglycan core protein precursor	PGSG_HUMAN	7.99E-06	17612.6	1	
P02766	Transthyretin precursor	TTHY_HUMAN	1.40E-08	15877.1	2	3.00E+08
P02775	Platelet basic protein precursor	SCYB7_HUMAN	1.28E-04	13885.4	1	low
P02652	**Apolipoprotein A-II precursor**	**APOA2_HUMAN**	**1.54E-07**	**11167.9**	**2**	2.44E+08
P02656	**Apolipoprotein C-III precursor**	**APOC3_HUMAN**	**6.04E-10**	**10845.5**	**1**	1.20E+08
P02654	**Apolipoprotein C-I precursor**	**APOC1_HUMAN**	**6.29E-06**	**9326.1**	**2**	6.10E+07
Protein gel band 17						
P01042	Kininogen-1 precursor	KNG1_HUMAN	5.23E-05	71900.1	1	
P00734	Prothrombin precursor	THRB_HUMAN	1.92E-05	69992.2	1	1.20E+03
P02765	Alpha-2-HS-glycoprotein precursor	FETUA_HUMAN	1.32E-07	39299.7	2	6.10E+08
P02655	**Apolipoprotein C-II precursor**	**APOC2_HUMAN**	**1.34E-04**	**11276.8**	**1**	
Protein gel band 19						
P01024	Complement C3 precursor	CO3_HUMAN	6.61E-07	187045.3	1	**high**
P01042	Kininogen-1 precursor	KNG1_HUMAN	3.30E-04	71900.1	1	
P02765	Alpha-2-HS-glycoprotein precursor	FETUA_HUMAN	4.79E-06	39299.7	2	6.10E+08
Protein gel band 20						
P20742	Pregnancy zone protein precursor	PZP_HUMAN	1.08E-04	163732.1	2	8.38E+06
P01023	Alpha-2-macroglobulin precursor	A2MG_HUMAN	1.88E-07	163174.3	7	**1.80E+09**
P01876	Ig alpha-1 chain C region	IGHA1_HUMAN	1.31E-06	37630.7	2	**high**
P01857	Ig gamma-1 chain C region	IGHG1_HUMAN	9.49E-10	36083.2	3	**high**
P01781	**Ig heavy chain V-III region**	**HV3T_HUMAN**	**3.52E-06**	**12722.2**	**2**	
P01597	**Ig kappa chain V-I region**	**KV1E_HUMAN**	**7.72E-10**	**11653.8**	**8**	

3.3.3 Investigation of Possible Contamination from the Filter Material

It has been hypothesised that the centrifugal filters could contaminate the serum sample with foreign molecules (e.g. glycerine) that would be detectable during MALDI-ToF MS or in a SDS-PAGE gel. To test this, and to ensure the UF process does not interfere with subsequent analysis, the denaturing buffer was filtered 3x and collected each time. This was then lyophilised, loaded on a gel and analysed by MALDI-ToF MS highly concentrated. Neither tricine-SDS PAGE nor MALDI-ToF MS analysis detected any contaminants in the buffer. This was important when identifying MS peaks to make sure all peaks originate from the serum sample.

3.4 Discussion and Conclusions

In this chapter some fundamental decisions on pre-fractionation of serum samples were made. This was important so that sample preparation could be efficient during biomarker experiments. It was established that un-filtered crude serum is too complex for analysis of all proteins in one step. Pre-fractionation is necessary to remove high abundance proteins such as albumin and IgGs. For visualisation of the proteins, SDS-PAGE for LMW proteins was optimised and it was found that a combination of high percentage separating gels with a spacer layer and tricine-Tris electrophoresis buffer can significantly improve resolution of LMW bands compared to Laemmli gels run with glycine-Tris buffer. For removal of high abundance proteins UF was compared to albumin depletion, protein precipitation and WAX separation. Centrifugal UF was found to be superior compared to alternative depletion methods. Depletion was possible in a single step, producing one fraction without the need for high salt buffers. However all methods suffered from non-specific removal of other proteins and incomplete removal of albumin. Optimising UF membranes were tested and the centriplus filters were found to be most effective. Here an optimised UF protocol using 50 kDa MWCO filters, at a reduced centrifugation speed (750 xg), collecting two successive filtrates was proposed. The use of UF enabled detection of proteins with a range of serum concentrations over 7 orders of magnitude. Low abundance proteins of ng/ml were detected in LMW filtrates, 3 orders of magnitude lower than detected in crude serum. Hence we established that UF is a good method for removal of albumin and that the LMW sub-proteome has a small enough complexity for proteomic analysis. As a next step the reproducibility of UF has to be validated to test their usefulness for biomarker discovery.

3.5 References

[1] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L. and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. Mol Cell Proteomics **1**, 947-955.

[2] Zheng, X., Baker, H. and Hancock, W. S. (2006) Analysis of the low molecular weight serum peptidome using ultrafiltration and a hybrid ion trap-Fourier transform mass spectrometer. J Chromatogr A **1120**, 173-184.

[3] Schrader, M. and Schulz-Knappe, P. (2001) Peptidomics technologies for human body fluids. Trends Biotechnol **19**, S55-60.

[4] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P. and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. Mol Cell Proteomics **2**, 1096-1103.

[5] Chan, K. C., Lucas, D. A., Hise, D., Schaefer, C. F., Xiao, Z., Conrads, T. P., Janini, G. M., Beutow, K. H., Issaq, H. J. and Veenstra, T. D. (2004) Analysis of the Human Serum Proteome. Clinical Proteomics **1**, 101-226.

[6] Plasma.Proteome.Institute, (2004), Proteins in Plasma and Serum, Institute, C. P. P. accessed: 22/02/2005 from: www. plasmaproteome.org

[7] Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W. and Guild, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. Proteomics **4**, 1175-1186.

[8] Bjorhall, K., Miliotis, T. and Davidsson, P. (2005) Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. Proteomics **5**, 307-317.

[9] Hinerfeld, D., Innamorati, D., Pirro, J. and Tam, S. W. (2004) Serum/Plasma depletion with chicken immunoglobulin Y antibodies for proteomic analysis from multiple Mammalian species. J Biomol Tech **15**, 184-190.

[10] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., Schatz, C. R., Miller, S. S., Su, Q., McGrath, A. M., Estock, M. A., Parmar, P. P., Zhao, M., Huang, S. T., Zhou, J., Wang, F., Esquer-Blasco, R., Anderson, N. L., Taylor, J. and Steiner, S. (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. Proteomics **3**, 1345-1364.

[11] Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L. and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. Proteomics **3**, 422-432.

[12] Quero, C., Colome, N., Prieto, M. R., Carrascal, M., Posada, M., Gelpi, E. and Abian, J. (2004) Determination of protein markers in human serum: Analysis of protein expression in toxic oil syndrome studies. Proteomics **4**, 303-315.

[13] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J., Petricoin, E. F., 3rd, Liotta, L. A., Veenstra, T. D. and Conrads, T. P. (2004) An investigation into the human serum "interactome". Electrophoresis **25**, 1289-1298.

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

[15] Chertov, O., Biragyn, A., Kwak, L. W., Simpson, J. T., Boronina, T., Hoang, V. M., Prieto, D. A., Conrads, T. P., Veenstra, T. D. and Fisher, R. J. (2004) Organic solvent extraction of proteins and peptides from serum as an effective sample preparation for detection and identification of biomarkers by mass spectrometry. Proteomics **4**, 1195-1203.

[16] Orvisky, E., Drake, S. K., Martin, B. M., Abdel-Hamid, M., Ressom, H. W., Varghese, R. S., An, Y., Saha, D., Hortin, G. L., Loffredo, C. A. and Goldman, R. (2006) Enrichment of low molecular weight fraction of serum for MS analysis of peptides associated with hepatocellular carcinoma. Proteomics **6**, 2895-2902.

[17] Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680-685.

[18] Schägger, H. and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. Anal Biochem **166**, 368-379.

[19] Huang, H. L., Stasyk, T., Morandell, S., Mogg, M., Schreiber, M., Feuerstein, I., Huck, C. W., Stecher, G., Bonn, G. K. and Huber, L. A. (2005) Enrichment of low-abundant serum proteins by albumin/immunoglobulin G immunoaffinity depletion under partly denaturing conditions. Electrophoresis **26**, 2843-2849.

[20] Merrell, K., Southwick, K., Graves, S. W., Esplin, M. S., Lewis, N. E. and Thulin, C. D. (2004) Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. J Biomol Tech 15, 238-248.

[21] Martosella, J., Zolotarjova, N., Liu, H., Nicol, G. and Boyes, B. E. (2005) Reversed-phase high-performance liquid chromatographic prefractionation of immunodepleted human serum proteins to enhance mass spectrometry identification of lower-abundant proteins. J Proteome Res **4**, 1522-1537.

[22] Li, X., Gong, Y., Wang, Y., Wu, S., Cai, Y., He, P., Lu, Z., Ying, W., Zhang, Y., Jiao, L., He, H., Zhang, Z., He, F., Zhao, X. and Qian, X. (2005) Comparison of alternative analytical techniques for the characterisation of the human serum proteome in HUPO Plasma Proteome Project. Proteomics **5**, 3423-3441.

[23] Lin, J. J., Meyer, J. D., Carpenter, J. F. and Manning, M. C. (2000) Stability of human serum albumin during bioprocessing: denaturation and aggregation during processing of albumin paste. Pharm Res **17**, 391-396.

[24] Scopes, R. K., Protein Purification: Principles and Practice, Springer-Verlag, Boston 1994, pp. 85-87.

[25] Chen, Y. Y., Lin, S. Y., Yeh, Y. Y., Hsiao, H. H., Wu, C. Y., Chen, S. T. and Wang, A. H. (2005) A modified protein precipitation procedure for efficient removal of albumin from serum. Electrophoresis **26**, 2117-2127.

[26] Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C. and Tempst, P. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. Anal Chem **76**, 1560-1570.

[27] Rai, A. J., Stemmer, P. M., Zhang, Z., Adam, B. L., Morgan, W. T., Caffrey, R. E., Podust, V. N., Patel, M., Lim, L. Y., Shipulina, N. V., Chan, D. W., Semmes, O. J. and Leung, H. C. (2005) Analysis of Human Proteome Organization Plasma Proteome Project (HUPO PPP) reference specimens using surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry: multi-institution correlation of spectra and identification of biomarkers. Proteomics **5**, 3467-3474.

[28] Hood, B. L., Zhou, M., Chan, K. C., Lucas, D. A., Kim, G. J., Issaq, H. J., Veenstra, T. D. and Conrads, T. P. (2005) Investigation of the mouse serum proteome. J Proteome Res **4**, 1561-1568.

[29] Williams, P., (1997) Protein ultrafiltration: a colloidal interaction approach, Chemical Engineering, Swansea, **Ph.D.**

[30] Georgiou, H. M., Rice, G. E. and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. Proteomics **1**, 1503-1506.

[31] Wagner, K., Miliotis, T., Marko-Varga, G., Bischoff, R. and Unger, K. K. (2002) An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. Anal Chem **74**, 809-820.

[32] Morris, D. L., Jr., Sutton, J. N., Harper, R. G. and Timperman, A. T. (2004) Reversed-phase HPLC separation of human serum employing a novel saw-tooth gradient: toward multidimensional proteome analysis. J Proteome Res **3**, 1149-1154.

[33] Johnson, K. L., Mason, C. J., Muddiman, D. C. and Eckel, J. E. (2004) Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometry: precision of retention time, mass, and ion abundance. Anal Chem **76**, 5097-5103.

[34] Clark, W. R. and Gao, D. (2002) Low-molecular weight proteins in end-stage renal disease: potential toxicity and dialytic removal mechanisms. J Am Soc Nephrol 13 Suppl 1, S41-47.

[35] Rockel, A., Hertel, J., Fiegel, P., Abdelhamid, S., Panitz, N. and Walb, D. (1986) Permeability and secondary membrane formation of a high flux polysulfone hemofilter. Kidney Int **30**, 429-432.

[36] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. and Yates, J. R., 3rd (1999) Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol **17**, 676-682.

[37] Polanski, M. and Anderson, N. L. (2006) A list of candidate cancer biomarkers for targeted proteomics. Biomarker Insights **2**, 1-48.

[38] Tanaka, Y., Akiyama, H., Kuroda, T., Jung, G., Tanahashi, K., Sugaya, H., Utsumi, J., Kawasaki, H. and Hirano, H. (2006) A novel approach and protocol for discovering extremely low-abundance proteins in serum. Proteomics **6**, 4845-4855.

CHAPTER 4

Centrifugal Ultrafiltration: Reproducibility and Efficiency

Despite the wide use of ultrafiltration (UF) for serum proteome fractionation [1-8], this approach has not been studied for efficiency and reproducibility in detail, only one report [9] testing the reproducibility of protein recovery from centrifugal membranes has been published. At the time the work for this chapter was carried out, Tammen *et al.* [9] had not published their data and also a completely different approach had been taken. It is particularly important to use a robust and reproducible sample preparation protocol for protein profiling and biomarker discovery.

The optimised UF protocol for sample preparation significantly improved detection of low molecular weight proteins from complex mixtures compared to manufacturer's recommendations (results were shown in Chapter 3). The results further emphasised the need for multiple consecutive filtrations. In the present chapter the performance of the UF membranes and whether they pose a potential source of variation was investigated. The aim was to show that UF is suitable as a pre-fractionation method for biomarker discovery protein profiling purposes. Three experiments, as outlined in Figure 4.1, were carried out to assess in detail how markers of different molecular weight and whole serum can be separated by the optimised UF process. First a simple mixture of standard markers with varying masses was separated by ultrafiltration. This allowed estimation of protein concentrations passing through the filter without the interference of complex mixtures and the high total protein concentrations present in biological samples. Secondly, human serum samples were spiked with standards in the filtrate for recovery analysis by absorbance quantitation. This enabled monitoring of the filter performance under "genuine" conditions. Finally, the reproducibility of

recovery of serum proteins naturally occurring in the sample was studied by SDS-PAGE, LC-MS/MS and MALDI-ToF MS. All three methods were used because, as shown, the results from each method were complementary and support the argument, showing reproducibility across filtrations.

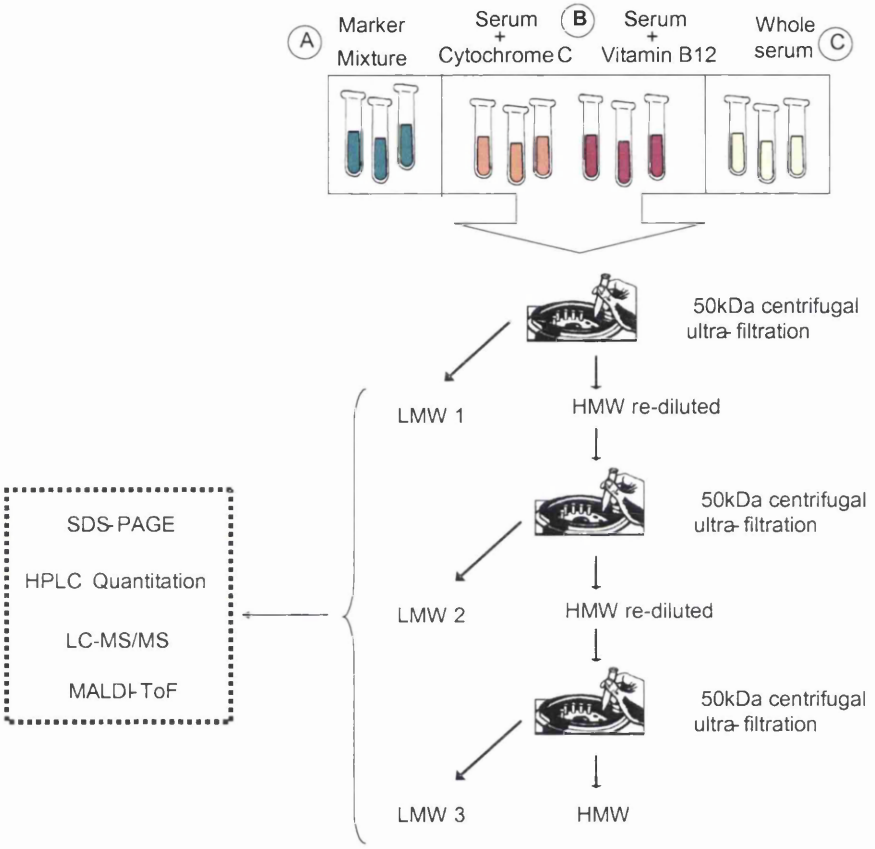


Figure 4.1: A schematic description of the experiment setup. (**A**) A simple mixture of standard markers, (**B**) human serum spiked with either cytochrome C or vitamin B12 and (**C**) neat human serum were all separated by centrifugal ultrafiltration. Each HMW retentate was made up to the original volume in denaturing buffer and filtered again, this was repeated twice. All the LMW filtrates and the final HMW retentate were collected and analysed by SDS-PAGE for successful separation, HPLC for protein recovery and LC-MS/MS for serum protein identification.

4.1. The Marker Mixture

A simple mixture of 5 markers, containing 235 nmol/ml ubiquinin, 1481 nmol/ml vitamin B12, 121 nmol/ml BSA, 161 nmol/ml cytochrome C and 142 nmol/ml lysozyme (final concentration) was prepared in denaturing buffer (25mM NH_4HCO_3 and 20% ACN (v/v)) and pre-fractionated by UF. Each marker was prepared as a 2 mg/ml concentration except for BSA, which was at a concentration of 8 mg/ml to mimic the excess of albumin in human serum. For direct comparison the markers were prepared in the same denaturing buffer as used for serum. For subsequent quantitation of each marker in the marker mixture, the mixture was separated by reverse-phase chromatography, using an in-house packed Poros 20 R2 reverse-phase separation column (100 mm x 4.6 mm I.D, 20 μm particle size) (Applied Biosystems, Warrington, UK) at a flow rate of 5 ml/min, on a Vision™ BioCAD Family Perfusion Chromatography Workstation (Applied Biosystems, Warrington, UK).

The HPLC chromatogram in Figure 4.2 b shows that baseline separation of each of the components in the marker mixture was achieved, so that the individual markers could be quantified individually without contamination from the other markers. Standard curves of UV absorbance versus a serially diluted marker mixture of known concentration were prepared for each marker, and to avoid introduction of variation during the chromatography of the standards, each dilution was separated in triplicate. The standard curves for each of the markers are shown in Figure 4.2 a.

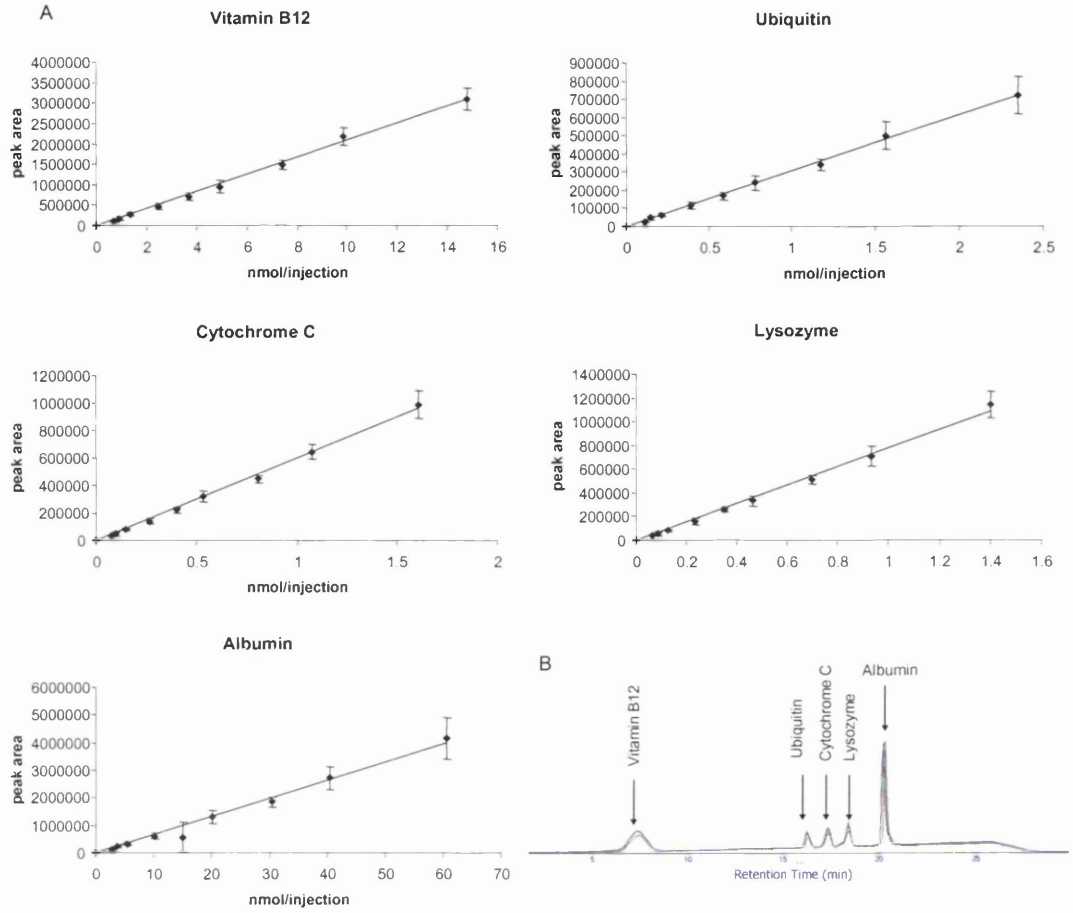


Figure 4.2: Standard curves generated from the UV detector response of the marker mixture diluted in series by HPLC separation. (**A**) Each marker was diluted to produce a standard curve for subsequent quantitation. The S.E bars illustrate the accuracy of the separation and peak area analysis. (**B**) Each of the markers elutes from the reverse phase column at a different retention time, a good separation was observed. Three replicate HPLC separations of the same sample were overlaid in the graph to demonstrate reproducibility.

The concentration of each marker was estimated by calculating it from the standard curves. Each marker achieved a high squared correlation coefficient (R^2 greater than 0.98), indicating a high degree of accuracy in forming the standard curves (Table 4.1). The C.V. of each dilution in each marker across 3 replicate RP-separations was calculated and showed good consistency (Table 4.1).

Table 4.1: The coefficient of variance (C.V.) across the three replicates of each dilution step in the calibration curve for each marker was calculated as well as the square correlation coefficient (R^2) for the best fit of the standard curve to ensure accurate estimation of the marker concentration.

	Mw (Da)	Average C.V. (%)	R2
vitamin B12	1350	14	0.997
albumin	67000	25	0.985
cytochrome C	12300	13	0.997
lysozyme	14100	12	0.995
ubiquitin	8500	13	0.998

From these standard curves the sample recovery in each fraction, after UF, was determined and it was noted that consecutive spins of UF can significantly increase the protein recovery (Figure 4.3 a). After one round of UF a maximum of only 50% (of the total amount in the combined LMW fractions) of each marker was recovered in the filtrate. For example, a total of 85% of the original amount of ubiquitin was recovered after three rounds of UF. However, only 45% of ubiquitin was pushed through the membrane during the first filtration. The second filtration step considerably improved the overall recovery of LMW proteins, yet protein recovery did not significantly benefit from a third filtration step. The requirement for more than one round of UF is explained by the fact that during UF the concentration of a compound in a mixture remains the same, e.g. as 1/3 of the mixture volume is retained, 1/3 of vitamin B12 remains above the membrane. However, as the concentration in the HMW retentate increases some proteins may form aggregates or be forced through the membrane. To re-dissolve the proteins the retentate was re-suspended and filtered again.

Combining the successive filtrates, the recovery of the proteins in the centrifuged marker mixture was shown to be highly reproducible although dependent on molecular weight (Figure 4.3 b).

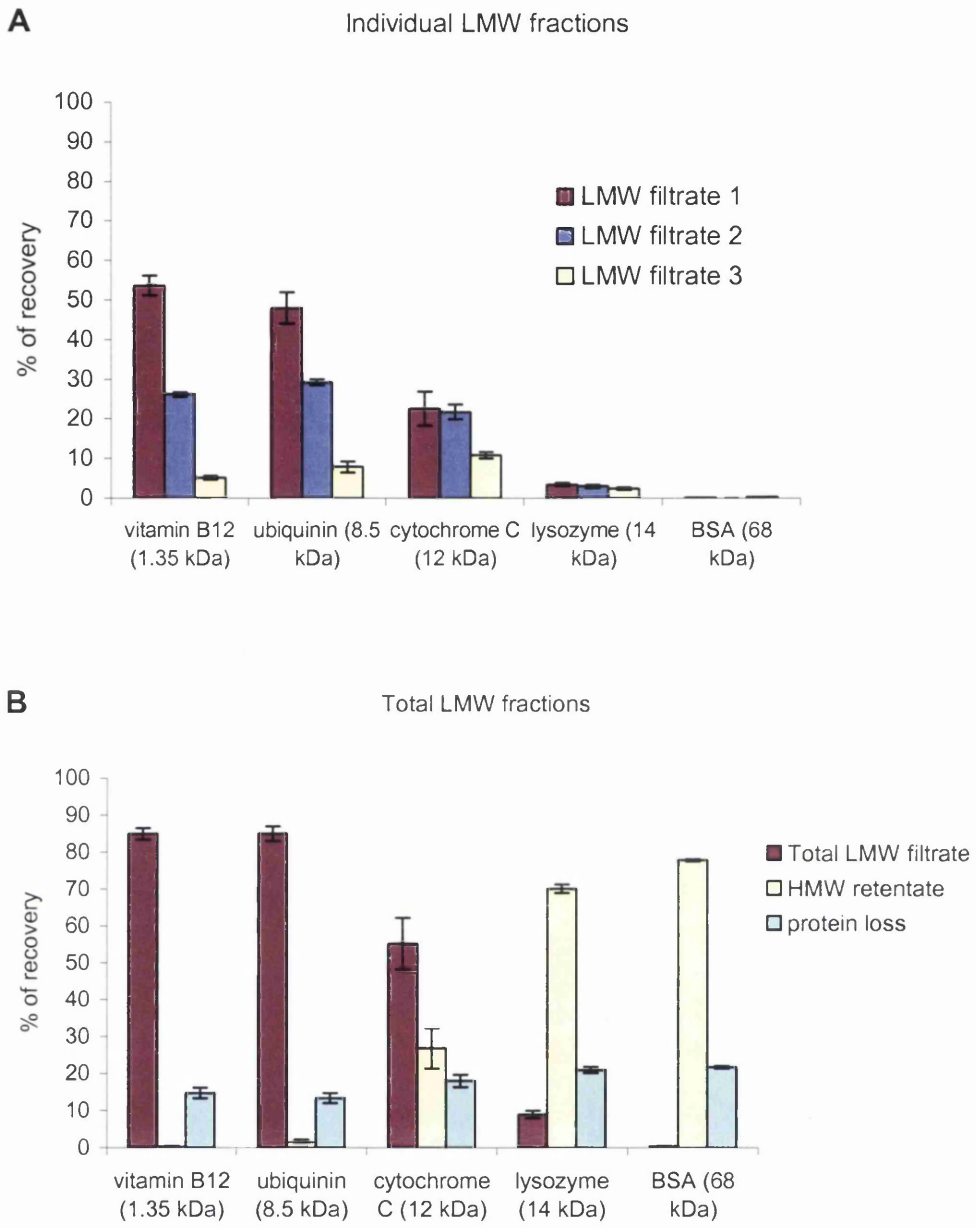


Figure 4.3: Recovery of the markers in each fraction, subjected to ultrafiltration. (**A**) The marker recovery in each individual filtrate is shown. Standard error bars are shown for each filtrate. (**B**) The individual LMW filtrates were combined and recovery in each fraction (i.e. LMW filtrate and HMW retentate) was recorded by HPLC. From this the marker loss of each marker was calculated. The variation between the replicate filtration runs is shown as ± SE.

The C.V.s of recovery across the replicate filtrates of each marker was calculated and is shown in Table 4.2. As expected, vitamin B12 and ubiquitin, passed through the membrane readily. Despite cytochrome C and lysozyme having a molecular weight

below the pore size of the filters, neither passed through the membrane as efficiently as vitamin B12 or ubiquitin (Figure 4.3 b).

However, MALDI-ToF MS analysis of the whole marker mixture showed that lysozyme forms a dimer, with a molecular weight of 28,193 Da (Figure 4.4). The literature confirmed that lysozyme exists as a dimer even under denaturing conditions [10, 11]. Hence protein recovery is dependent on a number of factors additional to molecular weight, which may include pH, tertiary shape and solubility.

Table 4.2: Marker recoveries from the mixture. The yield of markers in the filtrate was dependent on molecular weight, the CV of recovery across three replicate filtrations showed high reproducibility.

Source/ Marker	Marker Mixture	
	Filtrate yield (%)	Replicate CV (%)
Vitamin B12	85	1.8
Ubiquitin	85	2.3
Cytochrome C	55	13
Lysozyme	9	1.7
BSA	0.5	0.4

Furthermore, although cytochrome C has a molecular mass of 12.4 kDa, it did not pass through the filter completely, whereas ubiquitin, which has a mass of 8.5 kDa passed through the filter with almost 90% efficiency. Additionally, the reproducibility of the recovery for cytochrome C was not as good as the other markers (Table 4.2). However, cytochrome C is amphiphilic and due to the water binding to the molecule when in solution it may be variable in shape and size during filtration [12]. Furthermore, amphiphiliates can bind to biological membranes by burrowing their hydrophobic part into them. This could be an explanation for the variable behaviour of cytochrome C during the UF. As explained in Chapter 3 (section 3.2.4) polarisation of concentrated molecules at the membrane can block filtration of concentrated proteins. Passage of proteins depends on more than MW and so certain proteins may pass with more difficulty than others.

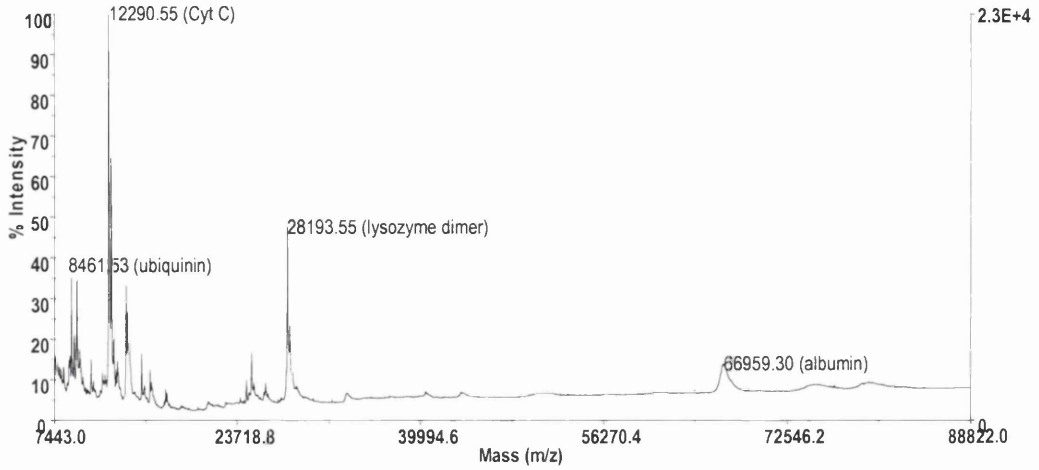


Figure 4.4: MALDI-ToF analysis of the marker mixture. Dimerisation of lysozyme at 28,193 Dalton is evident.

4.2. Markers Spiked into Serum

4.2.1. Choice of Markers for UF Evaluation

Reproducible recovery of serum proteins can not necessarily be extrapolated from the results using a simple mixture, as the complexity of serum may influence the performance of the MWCO membranes. To investigate the reproducible recovery of proteins from serum, standards (markers) can be spiked into the sample prior to UF. Insulin is a native protein in serum and at a MW of 5.6 kDa it was expected to be filtered successfully. To quantify its recovery, insulin with a fluorescent FITC tag was purchased and spiked into human serum. Since 1 mole of FITC equates 1 mole of insulin, it should have been directly correlated. The results however were very inconsistent and high amounts of the fluorescence were detected in the HMW retentate. To investigate the quality and MW of the insulin-FITC it was analysed by MALDI-ToF MS. The insulin was suspended in NH_4HCO_3 at a concentration of 175 pmol/µl and then desalted using Millipore filter disc for 30 minutes. MALDI-ToF MS analysis showed 2 peaks at *m/z* 6125 and *m/z* 6513, these have a mass difference of 390 Da and 778 Da compared to unlabelled insulin (Figure 4.5). A single peak for insulin-FITC tag at *m/z* 6125 was expected. MALDI-ToF MS in reflectron mode

revealed that the FITC tag could be removed by fragmentation (Figure 4.5 b). The major peak in the spectrum was insulin without the tag, however there were two more peaks indicating that possibly two tags are attached to FITC labelled insulin. It was therefore concluded that two FITC-tags were attached to some insulin molecules. Later it was confirmed that our particular batch of FITC labelled insulin had an average of 1.6 mol of FITC per mol of insulin (Sigma-confirmed). However, the complex with two FITC tags could be still small enough to pass through the filter. To explain the retention of FITC fluorescence above the filter, it could be assumed that the FITC tag became detached in the denaturing solution during UF and re-bound to a HMW molecule within serum. Chemical bonds are dynamic and N-H and N-C bonds (bond between FITC and insulin) break and reform readily in solution; hence the insulin-FITC was not suitable for testing of the UF reproducibility.

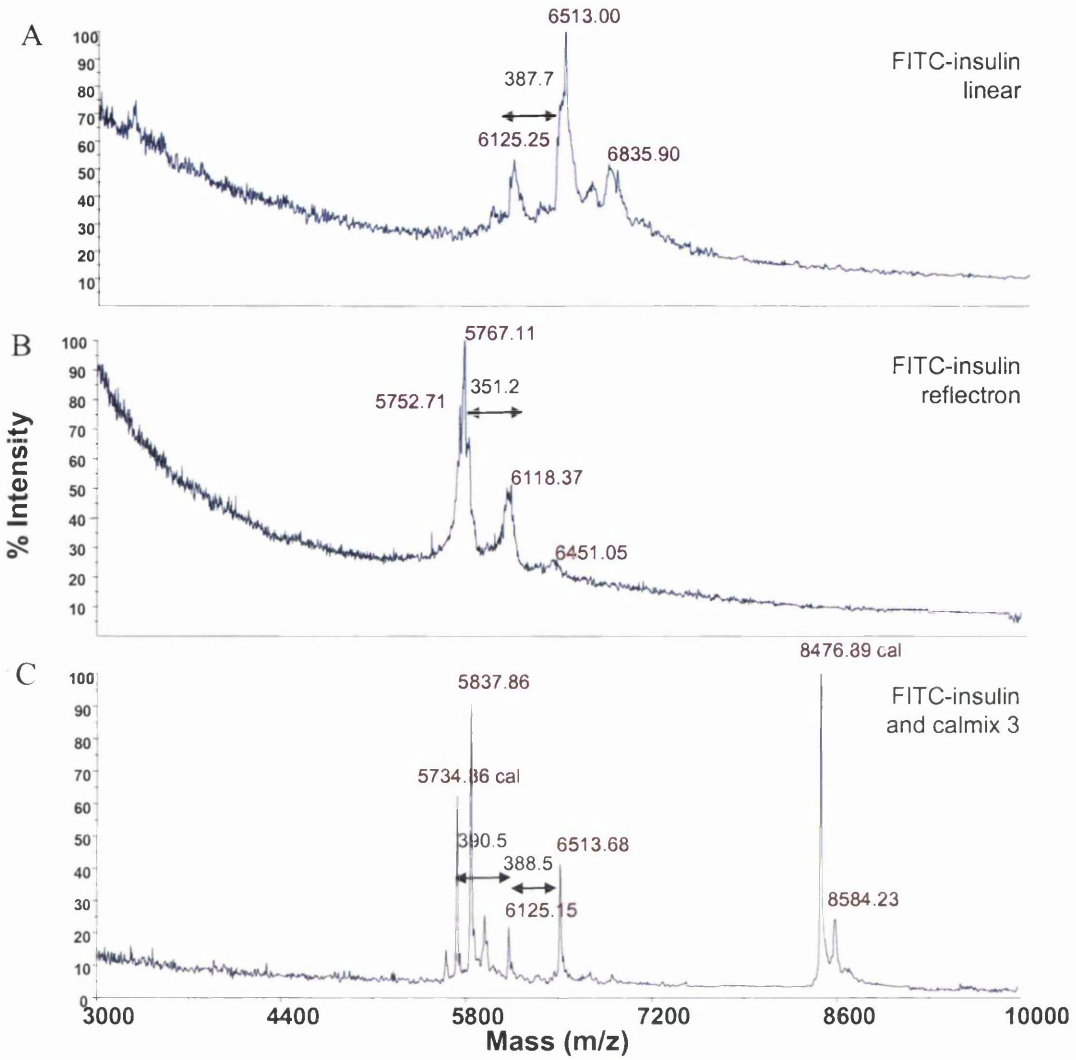


Figure 4.5: MALDI-ToF MS analysis of FITC-labelled insulin. (**A**) Shown is FITC-labelled insulin analysed in positive linear mode. (**B**) The same spot was analysed in positive reflectron mode to break the NH_2 bond and reveal the partially and unlabelled insulin molecule. (**C**) For calibration purposed the analyte was mixed with calibration mixture 3 and the insulin and apomyoglobin +2 peaks are labelled, insulin peaks tagged with one and with two FITC are also visible.

Markers with natural spectral absorbance different to that of serum were therefore chosen, as these will be stable. Hence cytochrome C and vitamin B12 were each spiked into human serum samples before UF (Figure 4.1 b). The markers were chosen in order to mimic medium size proteins suspected to pass through the membrane (cytochrome C) and smaller serum peptides (e.g. vitamin B12 has a similar mass to fibrinopeptide A). Both markers have a unique absorbance at 413 nm and 361 nm, respectively; allowing their specific quantitation in serum samples due to absorbance

spectral differences between un-spiked serum and the two marker proteins. Three filters were loaded with spiked serum and one un-spiked serum sample was processed simultaneously for use as a blank and a control. This was repeated for cytochrome C and vitamin B12. Recovery of the marker in the filtrate was quantitated relatively by photo spectrometry for their unique absorbance, which had been determined by scanning of the pure marker in the photo spectrometer for the wavelengths where the molecule absorbs the maximum light. Recovery of both markers showed excellent reproducibility (Figure 4.6 and Table 4.3). Although 50% of cytochrome C passed through the filter when in the relatively simple marker mixture, only 30% were recovered in the filtrate when spiked into serum. In addition, when cytochrome C was filtered through the membrane in denaturing buffer alone, 80% of the protein passed through the filter after only one filtration. These results suggest that the more complex the sample, the lower the recovery of cytochrome C in the LMW filtrate. This confirmed the theory of a filter cake or gel layer that causes blockage of the membrane, where proteins aggregate, especially as the sample concentration increases. Cytochrome C may be prone to that due to its amphiphilic properties. However, vitamin B12 passed through the membrane with 90% efficiency, as in the simple mixture. Both the marker mixture and the markers spiked into serum suffered from an unaccounted loss, which may also be explained by the formation of a gel layer on top of the membrane, where molecules bind to the filter or become lodged in the pores. The precise mechanism of this action however is still a matter of debate [13-15]. Nevertheless, UF of small proteins and peptides (<10 kDa) did not suffer from insufficient recovery (Figure 4.3).

Table 4.3: Marker recoveries in serum. The yield of markers in the filtrate of LMW serum was dependent on molecular weight, the C.V. of recovery across three replicate filtrations showed high reproducibility. *To asses the effect of sample complexity cytochrome C was filtered in buffer alone at a concentration of 1mg/ml, this showed high recovery.

	Serum	
	Filtrate yield (%)	Replicate CV (%)
Vitamin B12	90	0.8
Cytochrome C	30	1.8
Cytochrome C in Buffer*	80	---

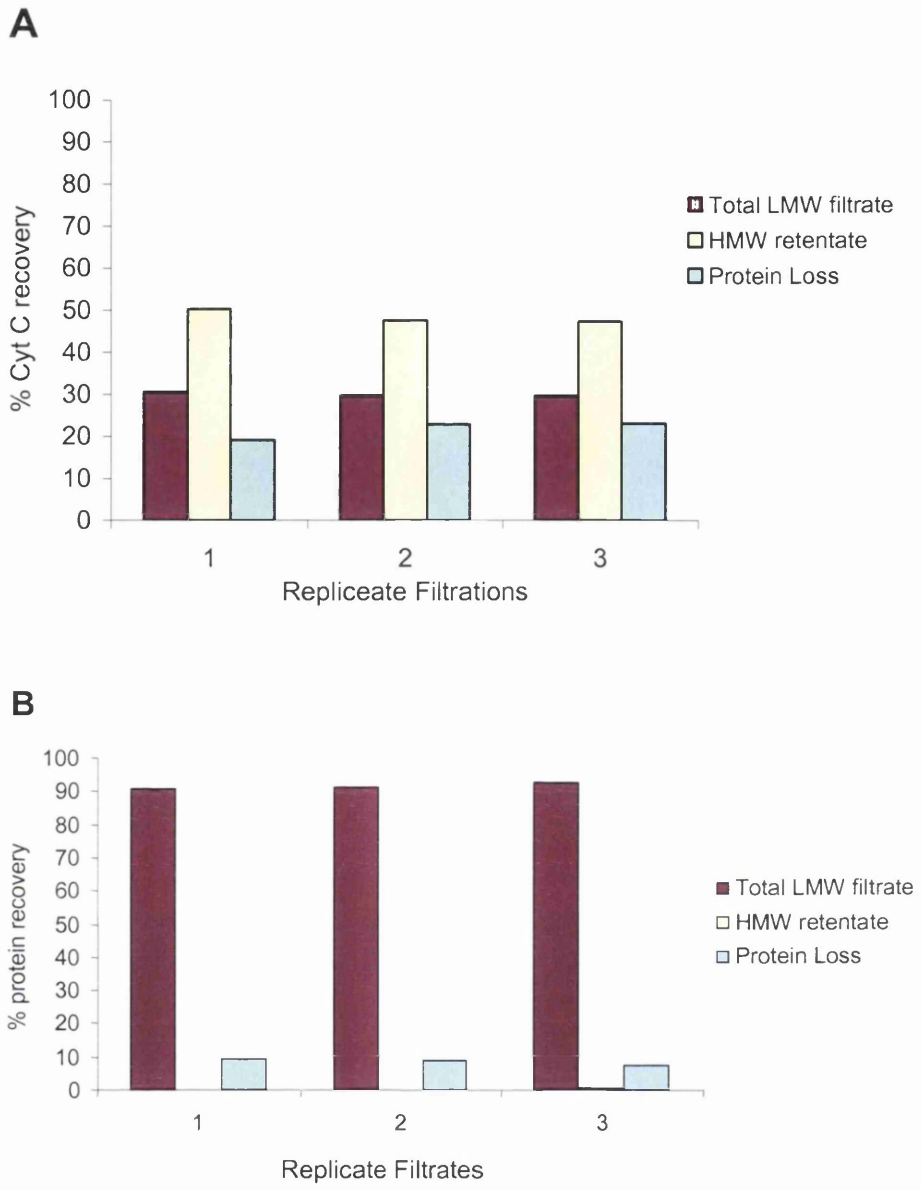


Figure 4.6: Recovery of markers spiked into human serum after centrifugal ultrafiltration. The marker recovery of the combined LMW filtrates and the HMW was determined by absorbance. (**A**) The total LMW protein recovery of cytochrome C was reproducible with a C.V. of 1.8%. (**B**) Vitamin B12 recovery in the LMW filtrate across 3 filters in reproducible (C.V. = 1%). Less than 1% of vitamin B12 remains in the HMW retentate.

4.3. Serum Protein Analysis for Reproducible Recovery from UF

Finally, the reproducibility of UF was further shown analysing un-spiked LMW serum proteins (from combined filtrates) by SDS-PAGE, MALDI-ToF MS and LC-MS/MS. It was important to confirm that native serum proteins also behave reproducibly during UF. This was complicated by the limitations MS techniques encounter. A combination of all three visualisation methods provided a more complete picture. MALDI-ToF MS is not strictly quantitative as hotspots in the matrix can influence the peak intensities across an entire spectrum. However this can be improved by acquisition of a large number of spectra and replicate analyses of the same sample. Although the chromatography across replicate LC-MS/MS separations should be identical, variations due to column pressure, solvent splitting, carry over and factors unknown can occur. Standardisation of the peak intensity in the basepeak chromatogram to the same scale, can to a certain extent remove basepeak intensity variation. SDS-PAGE showed to be the most reproducible and robust method for visualisation and comparing protein concentrations across replicate protein samples. However, only relatively abundant and large proteins are visualised well by SDS-PAGE.

4.3.1. MALDI-ToF MS and SDS-PAGE

Analysis and comparison of the serum proteins from the three replicate filters by MALDI-ToF MS and SDS-PAGE demonstrated that the ultrafiltration step was reproducible for all proteins passing through the filter (Figure 4.8 and Figure 4.7). Whilst acknowledging that MALDI-ToF MS is not truly quantitative, it is encouraging that the same protein peaks (i.e. no additional or missing peaks) were present in the spectrum of each of the replicate filtrates and that the peaks were at similar relative intensity between samples. The third spectrum appears to have an overall lower peak intensity, making it look as if the spectra were not reproducible, however the same peaks are present just at lower intensity.

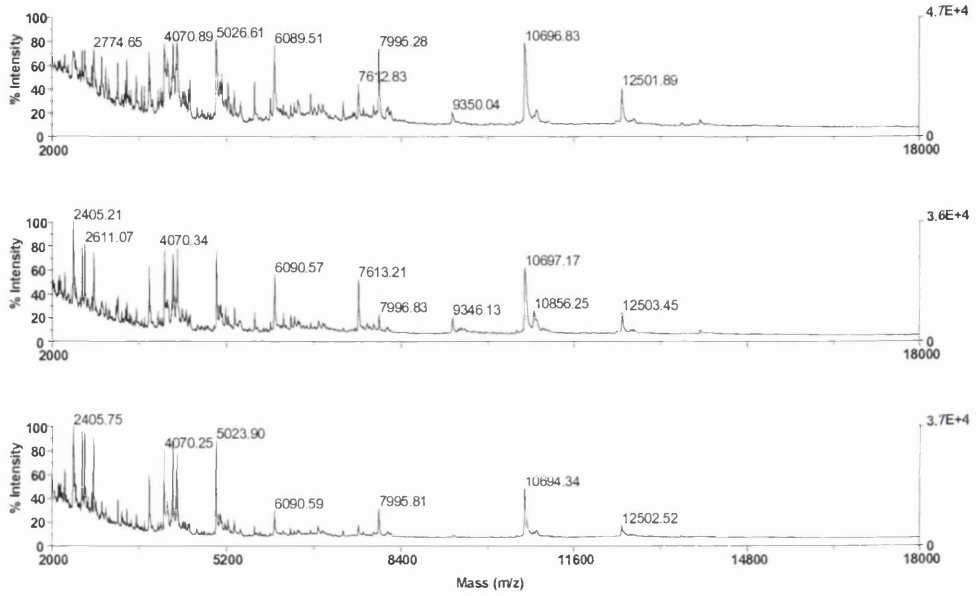


Figure 4.7: Serum protein analysis by MALDI-TOF. Three replicate filtrates of serum were analysed to show reproducibility of the centrifugal ultrafiltration procedure.

The same three LMW serum samples, separated by SDS-PAGE, showed identical protein bands at equal intensity across the three filtrates. Small proteins ionize easier during MALDI-ToF MS and so no peaks are visible for proteins larger than 12 kDa, where in the SDS-PAGE these proteins bands are clearly visible. A very strong band was visible at 15 kDa and again at 28 kDa. Alternatively these proteins may not bind well to C18 Zip-Tips, which were used for desalting of the LMW proteins. This gave emphasis to the reason why we used more that one visualisation tool to assess the UF. Furthermore this also showed that different methods will give complementary results for protein identifications and quantitation. This is in agreement with the findings by Anderson *et al.* [16] comparing different proteome analysis technologies.

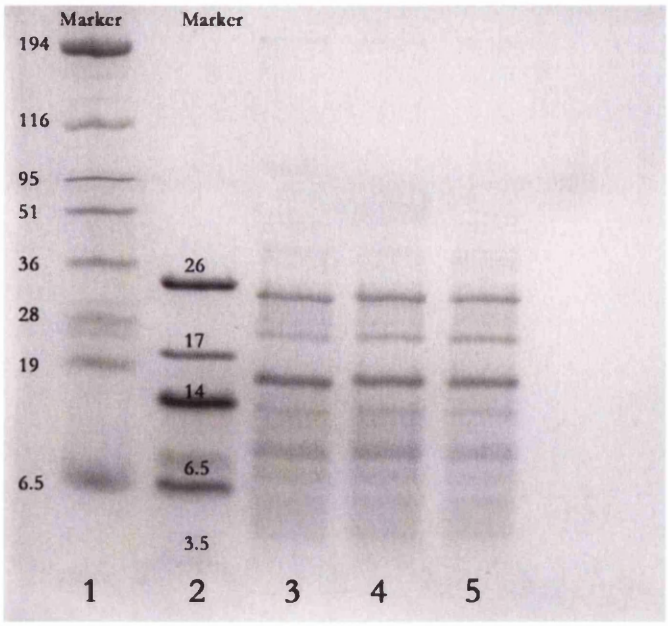


Figure 4.8: Recovery of LMW proteins after centrifugal ultrafiltration of serum. Serum protein analysis by 17% tricine SDS-PAGE, stained with colloidal Coomassie blue. The separation of the three replicate LMW filtrates shows reproducibility of the ultrafiltration procedure. Band intensities are comparable between lanes 3 to 5.

4.3.2. LC-MSMS Protein Profiling of Replicate Serum Filtrates

LC-MS/MS analysis of the three replicate LMW filtrates was performed in duplicate to account for run-to-run variation during the peptide identification. The scale of the peak intensity axis of the basepeak chromatograms were standardised to 2.0 $x10^9$ for 3000 xg and 3.0 $x10^9$ for 750 xg peak intensity, respectively (Figure 4.9). The spectra were similar except for one replicate in each sample group (3000 xg: 2b and 750 xg: 3b). Although there was some variation between the basepeak chromatograms of each of the 3 filtrates, the same, if not more variation could be seen between the two replicate LC-MS separations of the same LMW sample. Nevertheless, sufficient reproducibility was observed and confirms the usefulness of the UF process.

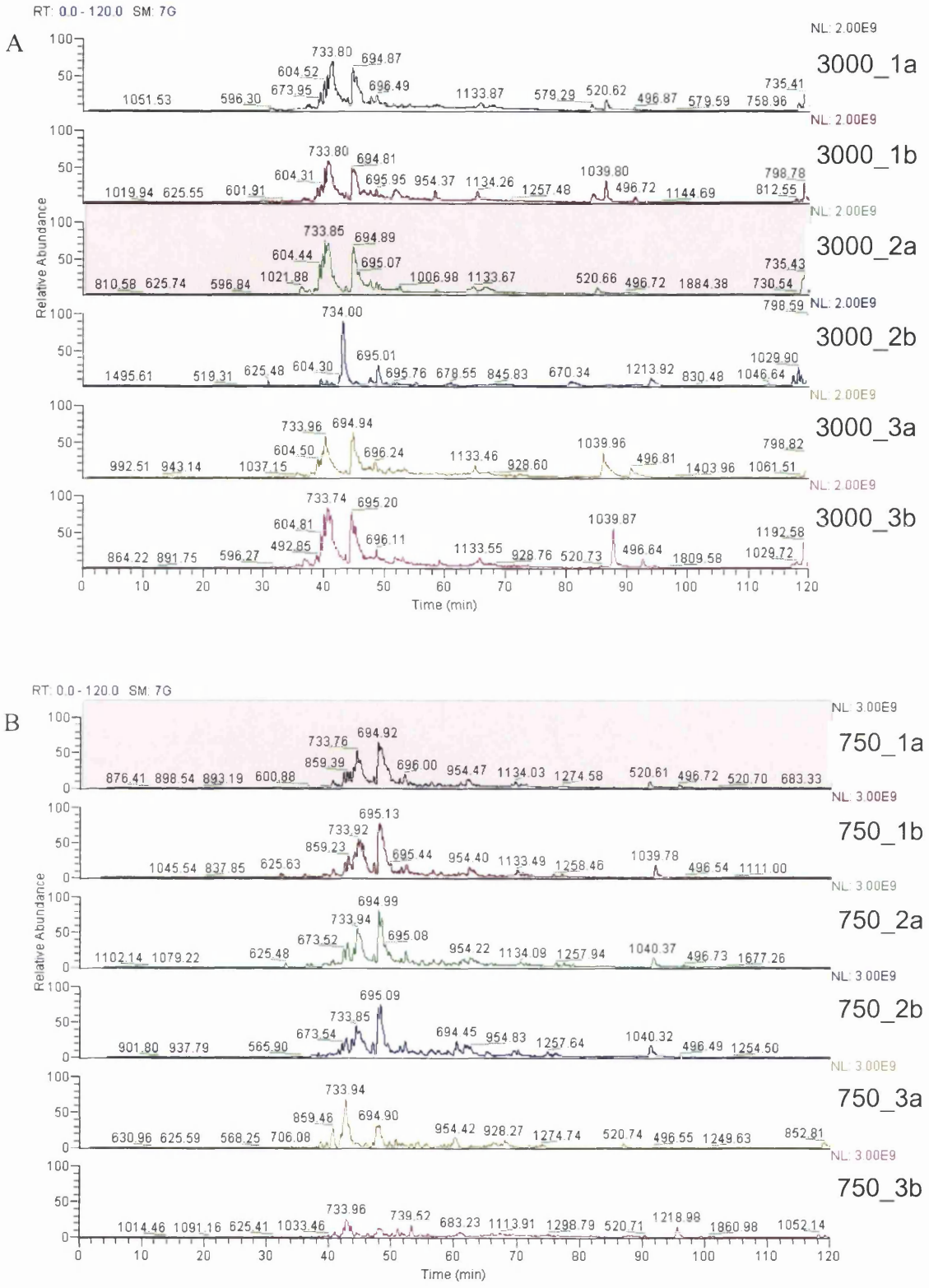
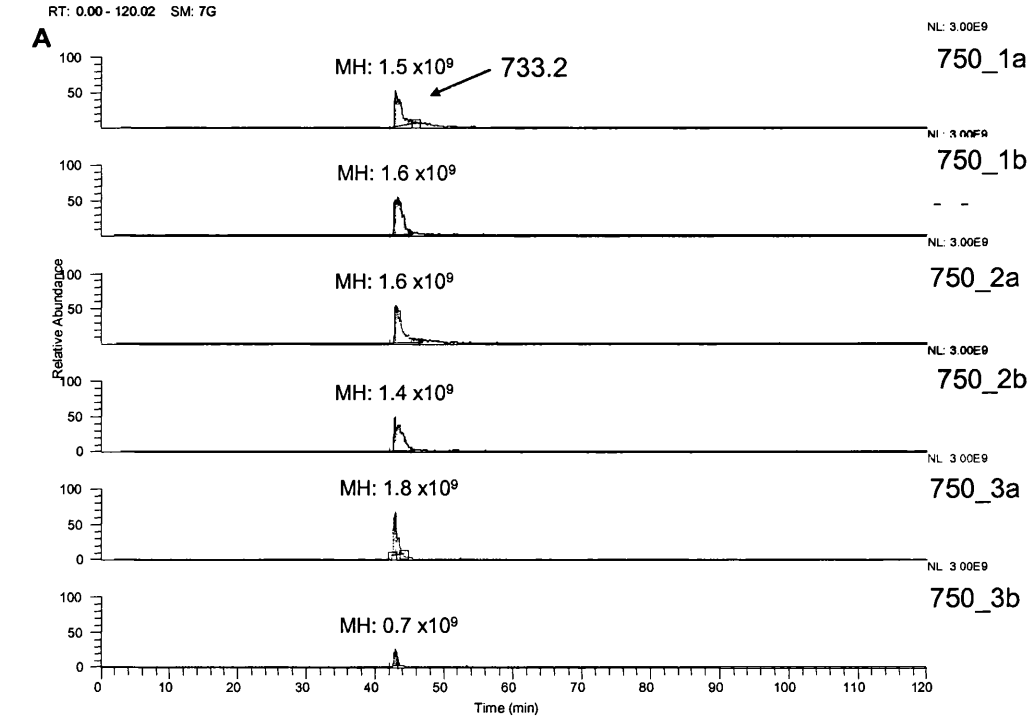


Figure 4.9: Basepeak chromatograms for the three replicate filtrates separated by RP-LC-MS/MS. Each filtrate was separated twice by LC-MS/MS for the samples centrifuged at 3000 xg (**A**) and 750 x g in (**B**).

Analysis of the peak intensity of two individual peaks confirmed the results shown by the basepeak chromatograms in Figure 4.9. The peak intensity across the replicates was reproducible except for the "outlier" 750_3b shown in Figure 4.10. Standardisation of the spectra against the average of the peak intensity could be used to adjust for low overall signal variation. Furthermore by using a large number of technical replicates, outliers could be excluded. Reproducibility across the three UF was shown despite some variation introduced during the LC-MS analysis.



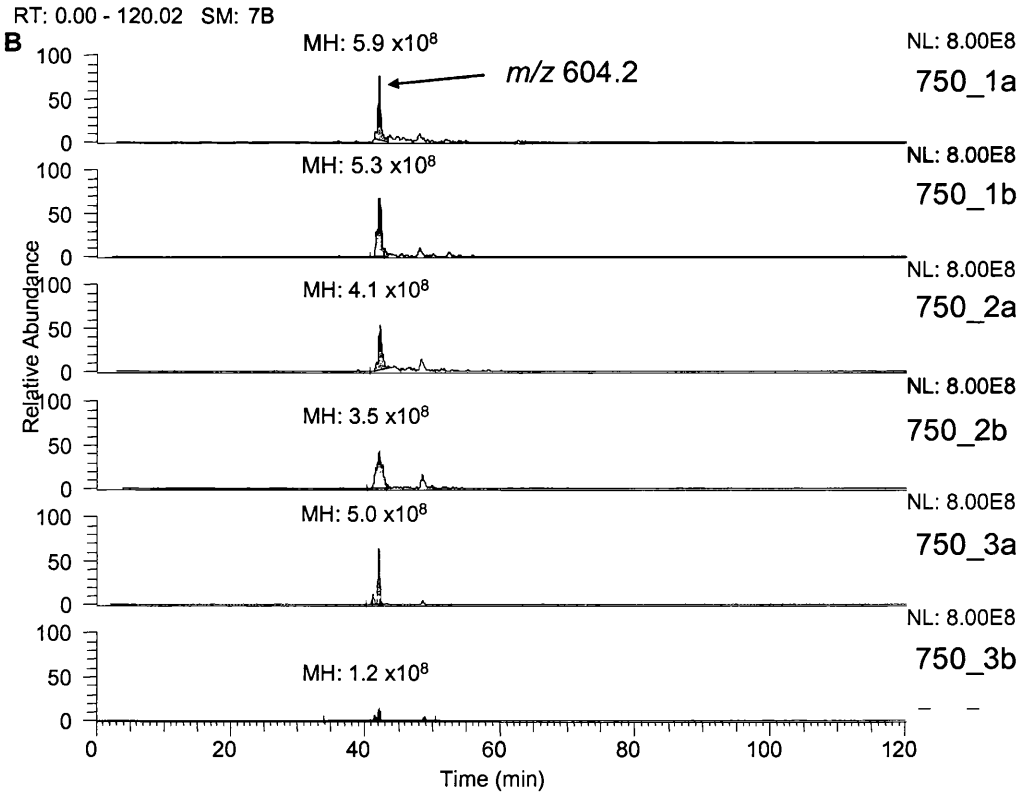


Figure 4.10: The extracted ion chromatogram for individual peaks is shown across the 3 LMW replicates and their RP-separation replicates. Two examples, *m/z* 733.2 (**A**) and *m/z* 604.2 (**B**) are shown.

The human FASTA database using Sequest was searched for matches with the tandem MS spectra from the LC-MS/MS experiment as described in Chapter 3 (section 3.3.2) and the Materials and Methods (section 2.8). The results from the two replicate RP-separations were combined and each replicate compared against the others. The overlap of protein matches from LC-MS/MS analysis was 50% across all three replicate filtrates centrifuged at 750 xg (Figure 4.11). Hence half the proteins were found in all three replicate filtrations. It is widely accepted that the overlap of protein matches across replicate LC-MS/MS analysis is low [17, 18]. Zheng [18] only found a 15% overlap between 3 replicate LC-MS/MS analysis of the same serum sample analysed using an LTQ-FT MS. In this study, the processing of the same serum peptide sample in duplicate by LC-MS/MS resulted in a of 61% overlap of protein identifications from two replicate LC-MS/MS runs, hence an overlap in protein identifications of 50% across three filtration experiments can be considered as an acceptable indicator of reproducibility.

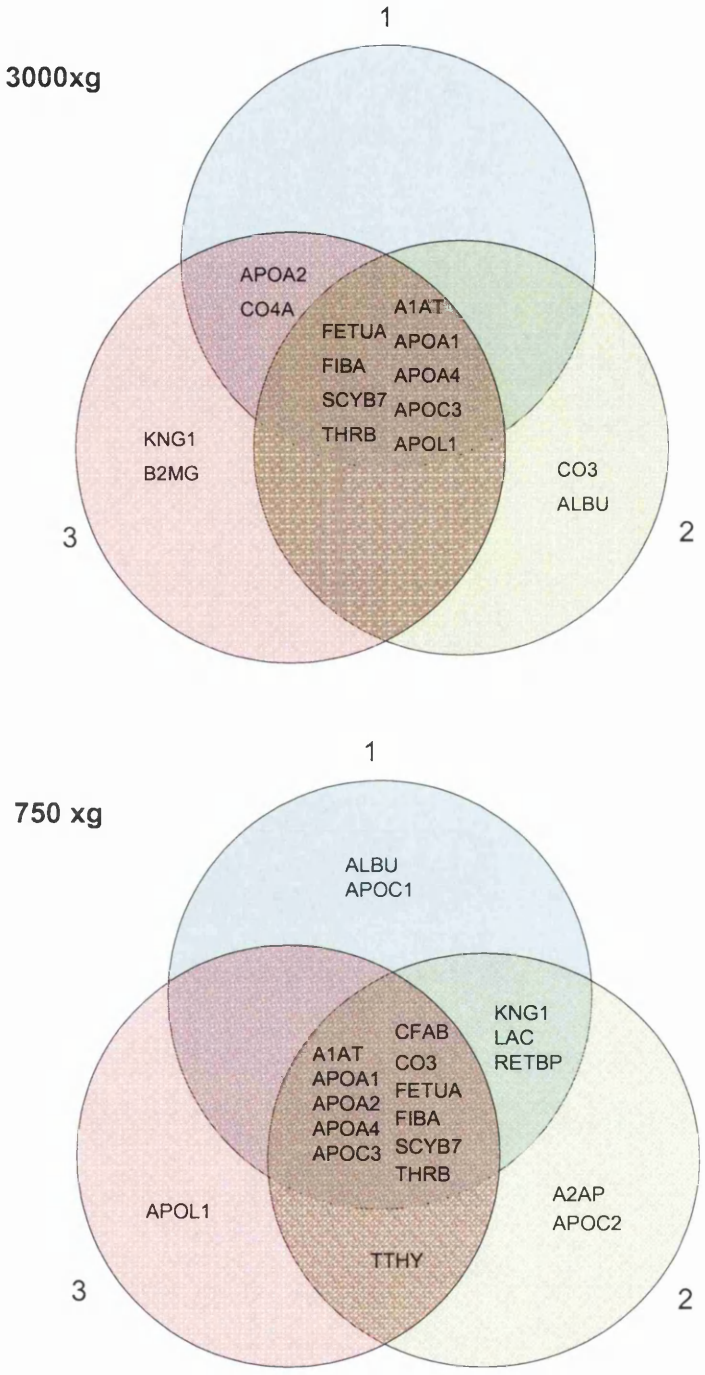


Figure 4.11: Protein identifications from LC--MS/MS analysis. Proteins identified after bottom-up protein profiling from three replicate ultrafiltrations. The Venn diagrams illustrate the overlap of protein identifications across three replicates centrifuging the filters at 3000 xg and 750 xg. An increase of protein identifications was observed in UF filtrates using the lower spin speed protocol. The Venn diagram shows a 50% overlap between all three centrifugal filters for the samples centrifuged at 750 xg.

The presence of a much reduced proportion of some highly abundant HMW proteins in the filtrate was to be expected as UF is a dynamic process and the shape of a molecule can cause it to slip through the pores of the membrane. However, albumin and other HMW proteins were removed satisfactorily, allowing the detection of other less abundant proteins during MS or electrophoresis (Table 4.5). Serum albumin was only detected in one of the three filtrates from the 3000 xg and 750 xg UF, and only with 3 and 2 peptides, respectively. In conclusion, no extra variation was introduced during the UF step, confirming the usefulness of this pre-processing technique to reduce the complexity of serum for MS analysis. Furthermore it could be concluded that more than one LC-MS/MS experiment was necessary to ensure reproducible detection of proteins in each sample, similar to the results previously reported by Scherl, A. [19].

Table 4.4: Proteins detected by RP-LC-MS/MS. Each of the three replicate filtrates at 3000 xg is shown for comparison for the reproducibility. The results from the replicate LC-MS/MS runs were merged. This data is complementary to the Venn diagram in Figure 4.11.

SwissProt ID	Protein	Accession no	Protein probability	Molecular mass	Matched peptides
UF membrane 1					
APOA4	Apolipoprotein A-IV precursor	P06727	5.65E-13	45371.5	6
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	7.88E-13	39299.7	2
SCYB7	Platelet basic protein precursor	P02775	8.32E-10	13885.4	2
APOL1	Apolipoprotein-L1 precursor	O14791	3.35E-09	43900.0	1
FIBA	Fibrinogen alpha chain precursor	P02671	5.91E-09	94914.3	2
APOC3	Apolipoprotein C-III precursor	P02656	1.04E-08	10845.5	2
APOA1	Apolipoprotein A-I precursor	P02647	1.92E-08	30758.9	5
A1AT	Alpha-1-antitrypsin precursor	P01009	6.79E-07	46707.1	1
CO4A	Complement C4-A precursor	P0C0L4	7.84E-06	192649.5	1
THRB	Prothrombin precursor	P00734	2.38E-04	69992.2	1
APOA2	Apolipoprotein A-II precursor	P02652	5.34E-04	11167.9	1
UF membrane 2					
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	4.85E-13	39299.7	2
APOA4	Apolipoprotein A-IV precursor	P06727	2.99E-11	45371.5	9
CO3	Complement C3 precursor	P01024	3.13E-11	187045.3	3
APOC3	Apolipoprotein C-III precursor	P02656	8.22E-10	10845.5	2
SCYB7	Platelet basic protein precursor	P02775	1.20E-09	13885.4	2
FIBA	Fibrinogen alpha chain precursor	P02671	2.04E-09	94914.3	3
THRB	Prothrombin precursor	P00734	5.07E-09	69992.2	1
APOA1	Apolipoprotein A-I precursor	P02647	7.74E-09	30758.9	1
APOL1	Apolipoprotein-L1 precursor	O14791	8.24E-09	43900.0	1
ALBU	Serum albumin precursor	P02768	1.85E-08	69321.6	3
A1AT	Alpha-1-antitrypsin precursor	P01009	2.58E-08	46707.1	1
UF membrane 3					
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	8.22E-14	39299.7	2
APOA4	Apolipoprotein A-IV precursor	P06727	1.09E-10	45371.5	8
FIBA	Fibrinogen alpha chain precursor	P02671	1.97E-09	94914.3	7
A1AT	Alpha-1-antitrypsin precursor	P01009	2.33E-09	46707.1	1
APOC3	Apolipoprotein C-III precursor	P02656	2.58E-09	10845.5	2
SCYB7	Platelet basic protein precursor	P02775	4.40E-09	13885.4	1
APOL1	Apolipoprotein-L1 precursor	O14791	1.37E-07	43900.0	1
B2MG	Beta-2-microglobulin precursor	P61769	1.49E-07	13705.9	1
THRB	Prothrombin precursor	P00734	3.28E-07	69992.2	2
KNG1	Kininogen-1 precursor	P01042	8.62E-05	71900.1	1
APOA2	Apolipoprotein A-II precursor	P02652	1.91E-04	11167.9	2
CO4B	Complement C4-B precursor	P0C0L5	2.23E-04	192671.6	1
CO4A	Complement C4-A precursor	P0C0L4	2.23E-04	192649.5	1

Table 4.5: Proteins detected by RP-LC-MS/MS. Each of the three replicate filtrates at 750 xg is shown for comparison for the reproducibility. The results from the replicate LC-MS/MS runs were merged. This data is complementary to the Venn diagram in Figure 4.11.

SwissProt ID	Protein	Accession no	Protein probability	Molecular mass	Matched peptides
UF membrane 1					
APOC3	Apolipoprotein C-III precursor	P02656	1.20E-12	10845.5	3
APOA4	Apolipoprotein A-IV precursor	P06727	3.22E-12	45371.5	13
SCYB7	Platelet basic protein precursor	P02775	2.49E-11	13885.4	2
APOA1	Apolipoprotein A-I precursor	P02647	1.36E-10	30758.9	9
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	5.86E-10	39299.7	3
A1AT	Alpha-1-antitrypsin precursor	P01009	8.54E-10	46707.1	2
FIBA	Fibrinogen alpha chain precursor	P02671	6.21E-09	94914.3	8
CO3	Complement C3 precursor	P01024	5.79E-08	187045.3	2
CFAB	Complement factor B precursor	P00751	1.39E-07	85478.6	1
THRB	Prothrombin precursor	P00734	1.55E-07	69992.2	2
LAC	Ig lambda chain C regions	P01842	2.30E-06	11229.5	1
APOC2	Apolipoprotein C-II precursor	P02655	4.13E-05	11276.8	1
RETBP	Plasma retinol-binding protein precursor	P02753	7.18E-05	22995.3	1
KNG1	Kininogen-1 precursor	P01042	1.01E-04	71900.1	1
ALBU	Serum albumin precursor	P02768	1.22E-04	69321.6	2
APOA2	Apolipoprotein A-II precursor	P02652	7.91E-04	11167.9	1
APOC1	Apolipoprotein C-I precursor	P02654	9.52E-04	9326.1	1
UF membrane 2					
APOA4	Apolipoprotein A-IV precursor	P06727	4.71E-13	45371.5	10
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	1.88E-12	39299.7	2
APOA1	Apolipoprotein A-I precursor	P02647	1.13E-11	30758.9	9
FIBA	Fibrinogen alpha chain precursor	P02671	2.62E-11	94914.3	4
APOC3	Apolipoprotein C-III precursor	P02656	2.32E-10	10845.5	3
SCYB7	Platelet basic protein precursor	P02775	1.38E-09	13885.4	1
A1AT	Alpha-1-antitrypsin precursor	P01009	7.06E-09	46707.1	3
LAC	Ig lambda chain C regions	P01842	9.88E-09	11229.5	1
CFAB	Complement factor B precursor	P00751	1.55E-06	85478.6	1
APOA2	Apolipoprotein A-II precursor	P02652	3.43E-06	11167.9	2
CO3	Complement C3 precursor	P01024	1.44E-05	187045.3	2
APOC2	Apolipoprotein C-II precursor	P02655	1.65E-05	11276.8	1
A2AP	Alpha-2-antiplasmin precursor	P08697	2.17E-05	54531.2	1
TTHY	Transthyretin precursor	P02766	2.68E-05	15877.1	1
KNG1	Kininogen-1 precursor	P01042	8.09E-05	71900.1	1
RETBP	Plasma retinol-binding protein precursor	P02753	1.25E-04	22995.3	1
THRB	Prothrombin precursor	P00734	2.38E-04	69992.2	1
UF membrane 3					
APOA4	Apolipoprotein A-IV precursor	P06727	5.90E-11	45371.5	12
APOA1	Apolipoprotein A-I precursor	P02647	1.21E-10	30758.9	9
THRB	Prothrombin precursor	P00734	2.05E-10	69992.2	2
SCYB7	Platelet basic protein precursor	P02775	9.48E-10	13885.4	2
CO3	Complement C3 precursor	P01024	1.30E-09	187045.3	3
A1AT	Alpha-1-antitrypsin precursor	P01009	2.15E-09	46707.1	1
TTHY	Transthyretin precursor	P02766	4.60E-09	15877.1	1
APOC3	Apolipoprotein C-III precursor	P02656	9.61E-09	10845.5	3
FIBA	Fibrinogen alpha chain precursor	P02671	1.01E-08	94914.3	4
FETUA	Alpha-2-HS-glycoprotein precursor	P02765	1.04E-08	39299.7	2
APOL1	Apolipoprotein-L1 precursor	O14791	1.02E-06	43900.0	1
CFAB	Complement factor B precursor	P00751	3.00E-06	85478.6	1
APOA2	Apolipoprotein A-II precursor	P02652	7.11E-04	11167.9	2

4.4. The Use of Multiple Filtrations Improving Protein Recovery

The advantage of using multiple consecutive filtrations for increased protein recovery had already been shown using the marker mixture. This was further investigated, analysis the consecutive filtrates of serum by MALDI-ToF MS. Figure 4.12 shows the different protein peaks present in each of the filtrates. In the second filtrate protein peaks additional to the first filtrate are clearly visible. The third round of UF did not significantly improve protein recovery (Figure 4.12). The change in protein peaks detected by MALDI-ToF MS can be partly explained by the fact that as the protein concentration in the retentate decreased by re-dilution, LMW proteins were dissociated from each other and then pass through the filter. Although only a small percentage of serum proteins filtered through the membrane, the equilibrium of proteins binding to each other may have changed and proteins trapped the first time, may have had the opportunity to pass through the filter during the second filtration.

It was not by coincidence which proteins pass through the membrane the first and which the second time as the recovery of proteins across three replicate was reproducible, even for individual filtration steps.

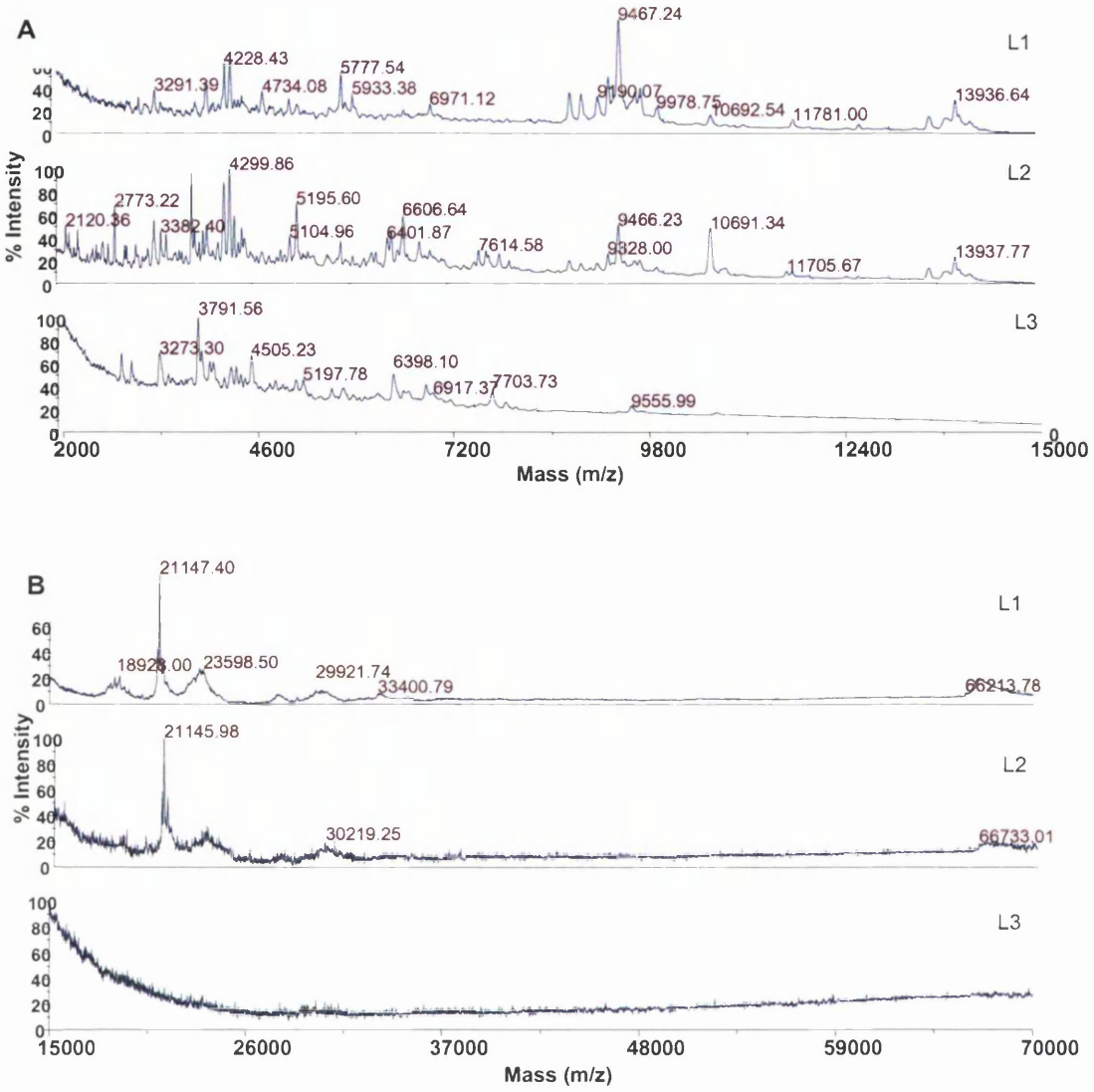


Figure 4.12: Consecutive LMW filtrations of serum. Each filtrate was analysed separately (L1, L2 and L3) for a low (**A**) and a higher mass range (**B**). Different proteins were detected in each of the filtrates.

To investigate any concerns regarding *in vitro* degradation of the serum proteins during the extended sample preparation, crude serum left at 4°C for 24 hours was compared with an aliquot of the same sample kept frozen. Each sample was analysed by SDS-PAGE and MALDI-ToF MS, if proteolytic degradation was present, we would expect to see more LMW peaks and less or broader bands in the HMW region. However the SDS-PAGE in Figure 4.13 for the serum samples incubated at 4°C for 24 and 0 hours looks almost identical, no evidence of degradation is visible. This was further confirmed by MALDI-ToF MS analysis of the same samples (Figure 4.14). This is in agreement with reports by West-Nielsen *et al.* [20] and Traum *et al.* [21],

investigating serum sample preparation and the effects of sample storage on protein degradation. Their results showed that storage of serum at 4°C for 24 hours does not cause significant degradation of serum proteins, however room temperature had an effect after only 4 hours.

Additionally, the effect of addition of ACN was demonstrated, to prove that this has no detrimental effect on the proteome (Figure 4.13 and Figure 4.15). The effect of ACN in the sample is visible before and after 24 hour incubation in the loss of a band at approximately 200 kDa, marked with an arrow (Figure 4.13). Furthermore the MALDI-ToF spectra in Figure 4.15 show a larger number of LMW peaks than the serum samples incubate without the addition of ACN. This may be due to dissociation of protein-protein interaction or the denaturing of a protein dimer which results in smaller masses, confirming the denaturing effect of ACN.

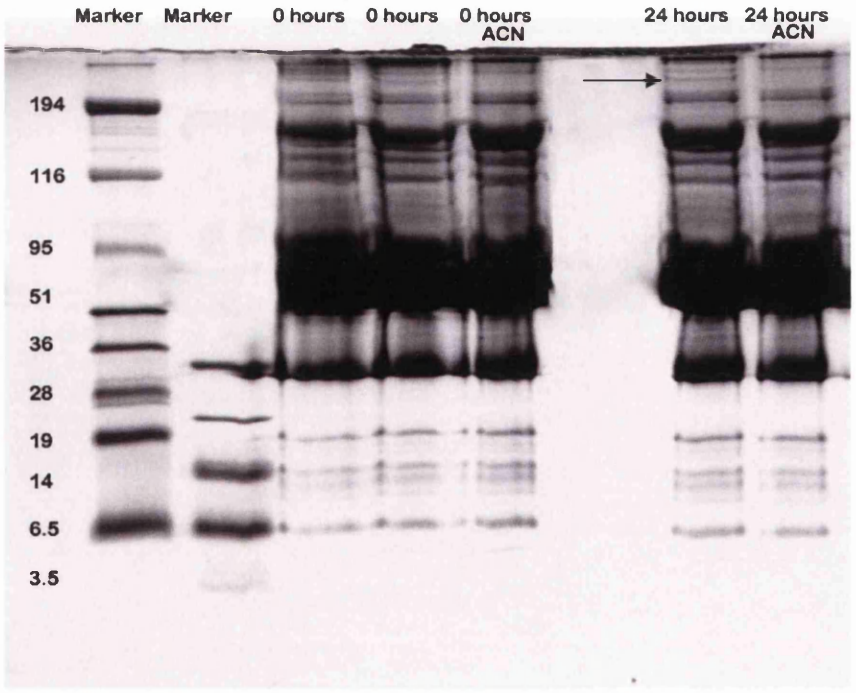


Figure 4.13: SDS-PAGE of crude serum (from the same aliquot) diluted in denaturing buffer (25 mM NH_4HCO_3, 20% ACN) and in non-denaturing 25 mM NH_4HCO_3. The serum samples were incubated for 0 and 24 hours at 4°C. 24 hour storage at 4°C has no noticeable effect on the serum proteins the presence of ACN in the sample appears to affect one band at ~200 kDa, marked with an arrow. This band is also visible in the 0 hour incubation without ACN.

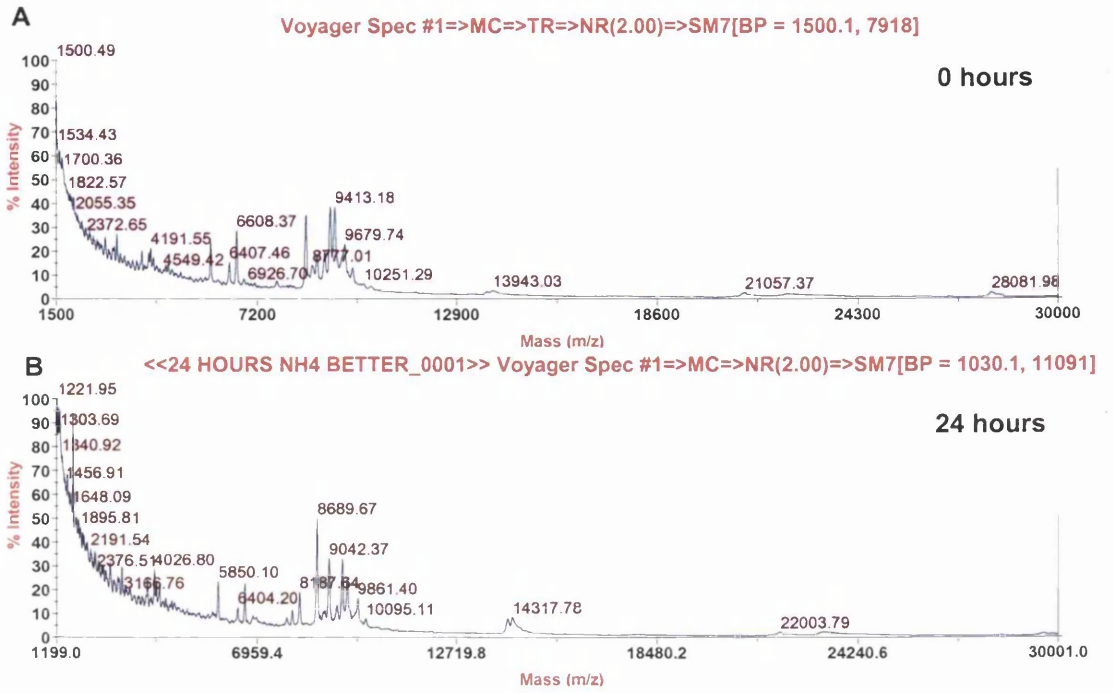


Figure 4.14: MALDI-ToF MS spectra of serum samples in NH_4HCO_3; showing that there is no noticeable difference between a fresh serum sample (**A**) and after 24 hours at 4°C (**B**).

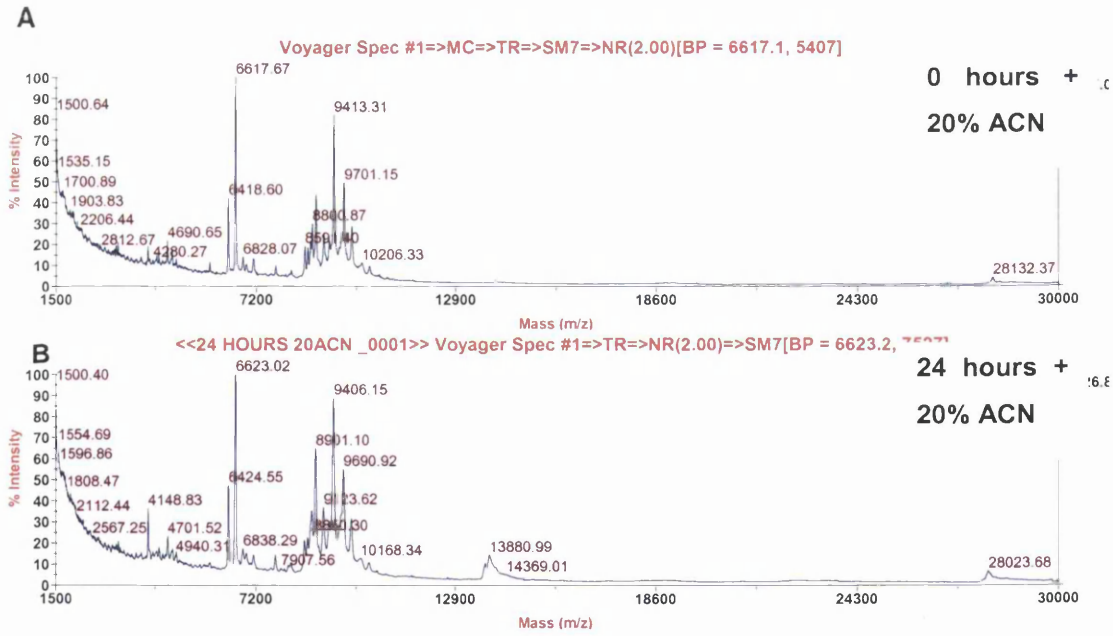


Figure 4.15: MALDI-ToF MS spectra of serum samples in denaturing buffer; showing that there is no noticeable difference, after addition of 20% ACN, between a fresh serum (**A**) and after 24 hours at 4°C (**B**).

4.5. Discussion and Conclusions

The reproducibility and robustness of the MWCO filters was tested and showed to be satisfactory. This was shown using a simple marker mixture where all markers passed the filter in a reproducible manner. This also showed that filtration is dependent on MW but also on pH, solubility and shape of the molecule. This was particularly obvious in the case of cytochrome C for shape and lysozyme for pH. Furthermore, the choice of marker was critical as solubility, stability and retention are important for quantitation. Spiking of cytochrome C and vitamin B12 into serum further showed the significance of the marker choice as vitamin B12 was not affected by the concentration of the sample but filtration of cytochrome C was reduced from 50 to 30% in complex serum. However, the C.V.s of the markers in serum were less than 2%, showing that despite some marker loss, recovery was consistent and reproducible. For LMW serum proteins, reproducibility of the recovery was assessed using MALDI-ToF MS, SDS-PAGE and LC-MS/MS. The three methods showed to be complementary each showing additional data to support the results. MALDI-ToF analysis was particularly good for visualising small proteins and peptides; however there is variation in the spectra, therefore it will be important for a large number of replicates and accumulation of many shots and spectra. Coomassie stained SDS-PAGE is known to be quantitative and showed good reproducibility of medium and larger serum proteins. The basepeak chromatograms of the LC-MS separation did not appear as robust as the other methods however looking at individual peptides in the extracted ion chromatograms showed that peptide recovery was reproducible. Hence the variation was not due to irreproducibility in the UF process. Variations in the basepeak intensity could be reduced by statistical standardisation and large numbers of technical replicates. It can be concluded that no further variation was introduced during this sample processing step and that UF of small proteins and peptides (<10 kDa) did not suffer from insufficient recovery (Figure 4.3), and so these results are encouraging and show that UF can enrich the LMW sub-proteome for MS analysis. In this chapter we furthermore confirmed that serum preparation on ice or at cold conditions does not cause significant degradation of the serum proteins. This is in accordance with findings more recently published [20-22]. In summary we have shown that UF is a robust sample pre-processing method to enrich the LMW sub-proteome for subsequent biomarker discovery in serum.

4.6. References

[1] Chernokalskaya, E., Gutierrez, S., Pitt, A. M. and Leonard, J. T. (2004) Ultrafiltration for proteomic sample preparation. *Electrophoresis* **25**, 2461-2468.

[2] Georgiou, H. M., Rice, G. E. and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1**, 1503-1506.

[3] Harper, R. G., Workman, S. R., Schuetzner, S., Timperman, A. T. and Sutton, J. N. (2004) Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. *Electrophoresis* **25**, 1299-1306.

[4] Johnson, K. L., Mason, C. J., Muddiman, D. C. and Eckel, J. E. (2004) Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometry: precision of retention time, mass, and ion abundance. *Anal Chem* **76**, 5097-5103.

[5] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P. and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* **2**, 1096-1103.

[6] Wagner, K., Miliotis, T., Marko-Varga, G., Bischoff, R. and Unger, K. K. (2002) An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. *Anal Chem* **74**, 809-820.

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

[8] Morris, D. L., Jr., Sutton, J. N., Harper, R. G. and Timperman, A. T. (2004) Reversed-phase HPLC separation of human serum employing a novel saw-tooth gradient: toward multidimensional proteome analysis. *J Proteome Res* **3**, 1149-1154.

[9] Tammen, H., Schulte, I., Hess, R., Menzel, C., Kellmann, M. and Schulz-Knappe, P. (2005) Prerequisites for peptidomic analysis of blood samples: I. Evaluation of blood specimen qualities and determination of technical performance characteristics. *Comb Chem High Throughput Screen* **8**, 725-733.

[10] Ermakova, E. (2005) Lysozyme dimerization: brownian dynamics simulation. *J Mol Model (Online)* **12**, 34-41.

[11] Scott, G. and Mowrey-McKee, M. (1996) Dimerization of tear lysozyme on hydrophilic contact lens polymers. *Curr Eye Res* **15**, 461-466.

[12] Hase, T., Harabayashi, M., Kawai, K. and Matsubara, H. (1987) A carboxyl-terminal hydrophobic region of yeast cytochrome c1 is necessary for functional assembly into complex III of the respiratory chain. *J Biochem (Tokyo)* **102**, 411-419.

[13] Rockel, A., Hertel, J., Fiegel, P., Abdelhamid, S., Panitz, N. and Walb, D. (1986) Permeability and secondary membrane formation of a high flux polysulfone hemofilter. *Kidney Int* **30**, 429-432.

[14] Clark, W. R. and Gao, D. (2002) Low-molecular weight proteins in end-stage renal disease: potential toxicity and dialytic removal mechanisms. *J Am Soc Nephrol* **13 Suppl 1**, S41-47.

[15] Williams, P., (1997) *Protein ultrafiltration: a colloidal interaction approach*, Chemical Engineering, Swansea, Ph.D.

[16] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R. and Lobley, A. (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* **3**, 311-326.

[17] Washburn, M. P., Wolters, D. and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-247.

[18] Zheng, X., Baker, H. and Hancock, W. S. (2006) Analysis of the low molecular weight serum peptidome using ultrafiltration and a hybrid ion trap-Fourier transform mass spectrometer. *J Chromatogr A* **1120**, 173-184.

[19] Scherl, A., *17th IMSC*, Prague 2006.

[20] West-Nielsen, M., Hogdall, E. V., Marchiori, E., Hogdall, C. K., Schou, C. and Heegaard, N. H. (2005) Sample handling for mass spectrometric proteomic investigations of human sera. *Anal Chem* **77**, 5114-5123.

[21] Traum, A. Z., Wells, M. P., Aivado, M., Libermann, T. A., Ramoni, M. F. and Schachter, A. D. (2006) SELDI-TOF MS of quadruplicate urine and serum samples to evaluate changes related to storage conditions. *Proteomics* **6**, 1676-1680.

[22] Hsieh, S. Y., Chen, R. K., Pan, Y. H. and Lee, H. L. (2006) Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. *Proteomics* **6**, 3189-3198.

CHAPTER 5

Biomarker Discovery using MALDI-ToF MS

Serum proteomics has generated considerable amount of excitement among oncologists and analytical chemists in recent years [1-5], with the promise of developing high-throughput blood tests for breast cancer, through mass spectrometry-based profiling and biomarker discovery. Using the sample preparation method optimised in chapter 4, the LMW sub-proteome from a cohort of breast cancer patients was used for protein profiling using MALDI-ToF MS. Most data published studying biomarkers from serum has been generated using SELDI-ToF MS, which, in contrast to MALDI-ToF MS, was specifically designed for biomarker discovery. The SELDI-ToF equipment comes with a spectra alignment program as well as statistical analysis. This however is not available for MALDI-ToF analysis. Hence spectral alignment appeared to be the most challenging problem in terms of signal processing and a program was therefore developed especially for this purpose. In this chapter, MALDI-ToF MS was optimised for protein profiling, comparing breast cancer serum proteins with those from non-cancer controls. The use of MALDI-ToF is cheaper in consumables and faster in analysis than SELDI-ToF MS and additionally the mass analyser is more sensitive with greater mass accuracy, which may provide greater confidence in the results.

As described in the Introduction (section 1.5), few published studies are available discussing biomarker discovery from intact proteins, using MALDI-ToF MS. However, during the course of this thesis, Villanueva and his co-workers have published some studies that focused on the use of magnetic beads for serum

peptidome analysis [6-9]. The use of magnetic beads was shown to enable sample clean-up and removal of albumin in one step. The beads were coated with a reverse-phase chromatographic resin that binds polar proteins and peptides. Another paper by Callesen *et al.* [10] used MALDI-ToF MS for protein profiling of breast cancer serum. The serum samples were desalted and the peptidome isolated using SPE cartridges. In their study three "markers" were detected (Figure 5.1). Although markers of the same mass (*m/z* 4.3, 8.1 and 8.9 kDa) had been convincingly discovered using SELDI-ToF MS by Li *et al* [2, 11]; and again in a independent study also using SELDI-ToF MS [12], closer inspection of the results by Callesen *et al.* [10] left some doubt as to the discriminatory capacity of these markers (Figure 5.1). The overall spectrum intensity of the control sample (Figure 5.1 b) appears lower. In our results two of the peaks were also present but remained unchanged between breast cancer and control spectra (Figure 5.2). We therefore had to develop the technique for biomarker discovery using MALDI-ToF MS on serum samples effectively from basic principles due to a lack of previously published/established techniques.

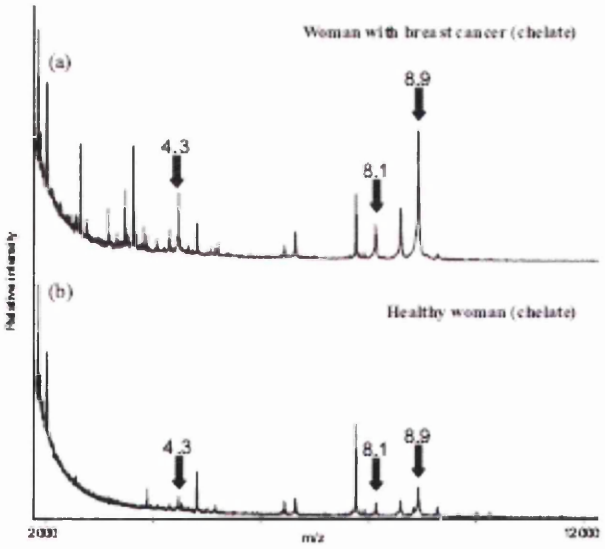


Figure 5.1: MALDI-ToF spectra were taken from Callesen et al. [10], showing a breast cancer serum sample in (**a**) and a control sample in (**b**). The peaks with *m/z* ratios of 4.3, 8.1 and 8.9 kDa were reported as biomarkers in this study; however the overall intensity across all peaks appears lower in the control spectrum. The "markers" do not appear convincing compared to results obtained in our study.

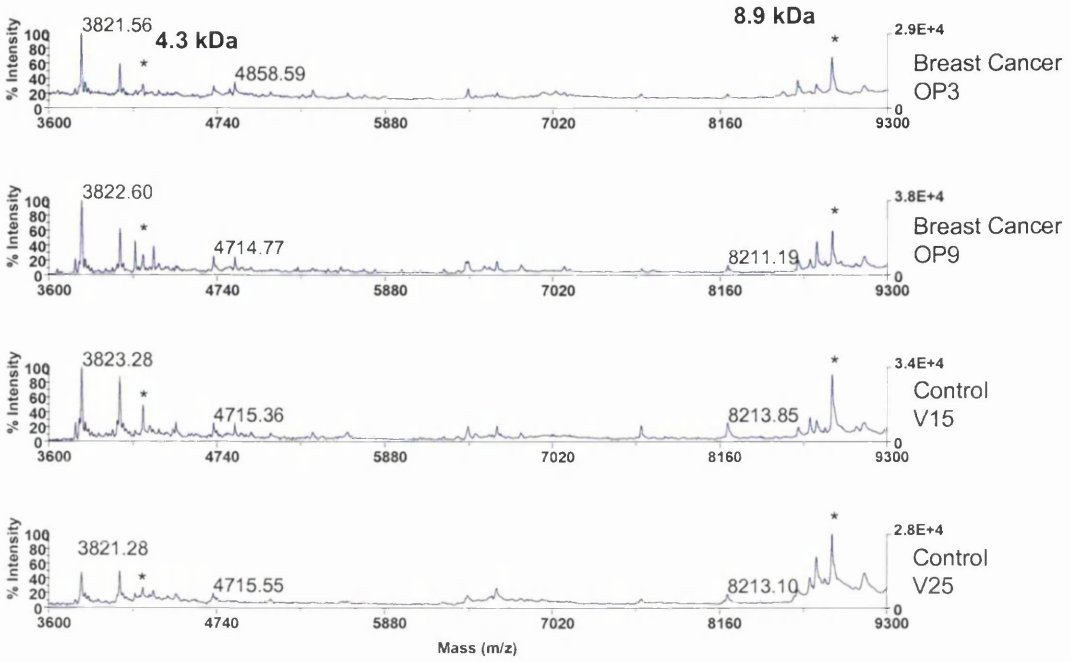


Figure 5.2: MALDI-ToF spectra from two breast cancer and two control samples. The peaks for *m/z* 4.3 and 8.9 kDa (marked *) were unchanged between the two cohorts in our results. No mass peak was observed at a mass of 8.1 kDa in any of the spectra.

In this chapter LMW proteins from breast cancer patients and control serum samples were compared for protein profiling using MALDI-ToF MS. In the first comparative experiment, only one replicate of LMW filtrates per serum sample was used; from these results a number of potential markers were retrieved and areas where the technique could be improved were highlighted. This sample set was named "S1". A second profiling study was performed on a largely overlapping set of samples, but using LWM filtrates from triplicate ultrafiltrations of the same serum sample. These results were potentially validating the S1 results however, as many factors of the sample preparation were changed, the results are more complementary than fully confirmatory. The second sample set and experiment are termed "S2" throughout the rest of the thesis. Some potential markers were discovered from both sample sets and encouragingly some overlapped. The discovery of these markers is exciting and identification was attempted using MALDI-ToF-ToF peptide fragmentation. From this, we were able to get peptide identifications with homology to three proteins.

5.1. Protein Profiling using MALDI-ToF MS: Sample Set 1 (S1)

Serum samples from breast cancer patients were collected and the LMW sub-proteome was analysed using MALDI-ToF MS. Specimens were linked to database records but were anonymised. Blood collection, serum preparation, storage and handling, UF and mass spectrometry were carried out in exactly the same way for breast cancer patient samples as for the control samples. In fact, in each experiment, the 8 metastatic breast cancer and 8 control sera were ultra-filtered and analysed simultaneously. Each breast cancer sample was matched with a control of a similar age. The age distribution of both sample cohorts for sample set 1 (S1) is shown in Figure 5.3.

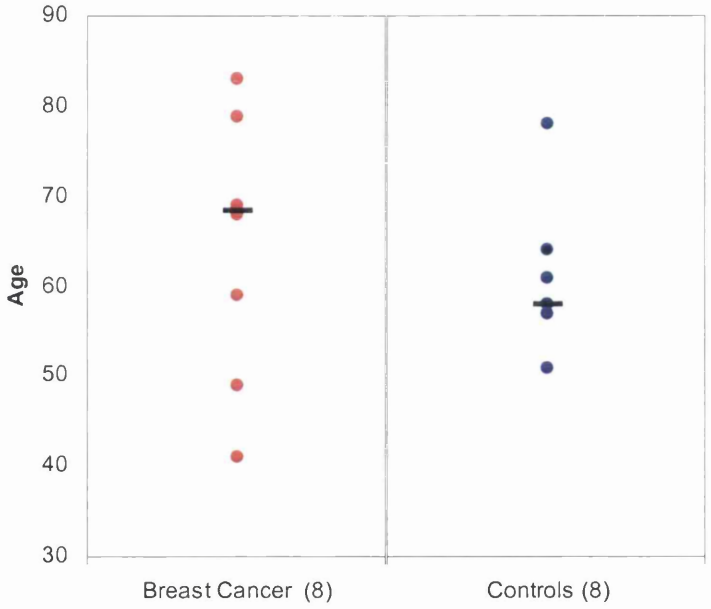


Figure 5.3: Age distribution of healthy controls and breast cancer patients in experiment S1. Equal numbers of patients and controls were used from each sample cohort which were individually age-matched. Black horizontal lines represent medians for each group.

All samples from both cohorts were prepared using the un-optimised/original UF method, as described in the Materials and Methods (section 2.3.2) and Chapter 3 (section 3.3.2), collecting one single filtrate from serum, which had been diluted 1:5 and filtered at a centrifugation speed of 3000 xg using a 30 kDa MWCO membrane. For MALDI-ToF MS, each LMW serum sample was Zip-Tipped as three aliquots

(Figure 5.4). The eluted fraction were dried and spotted with αCHCA and the spectra were analysed in linear mode using a Voyager STR DE (Applied Biosystems, Warrington, UK), as described in the materials and methods.

For each set of triplicate samples, 8 spectra were accumulated in a mass range of 1000-7000 Da. Each spectrum was recorded in automatic mode and, before every sample, the mass spectrometer was externally calibrated from an adjacent spot, using Cal Mix 2 (angiotensin I *m/z* 1297.51, ACTH clip 1-17 *m/z* 2094.46, ACTH clip 17-39 *m/z* 2456.66, ACTH clip 4-38 *m/z* 3660.19 and insulin *m/z* 5734.59) from Applied Biosystems.

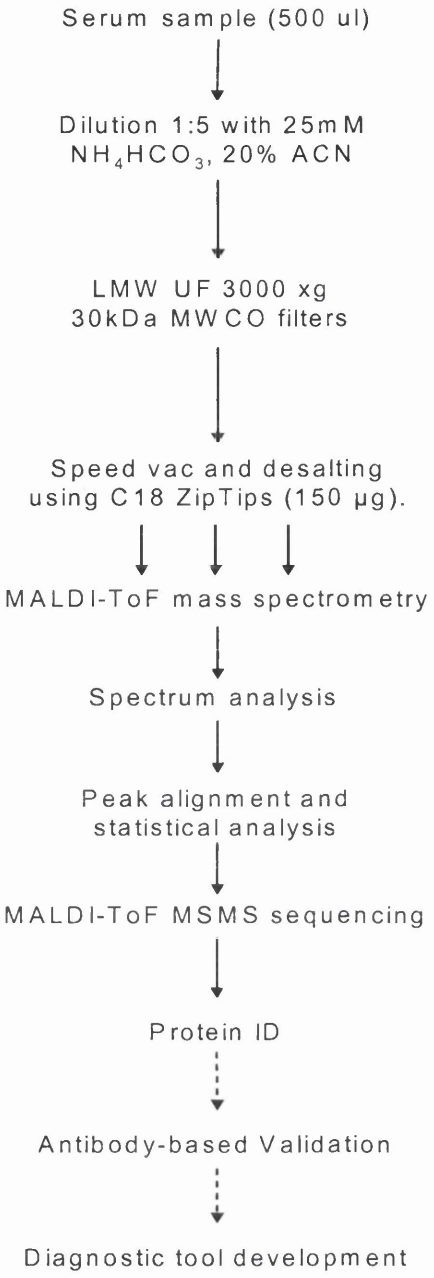


Figure 5.4: Experiment flowchart for protein profiling of LMW serum samples. Serum samples were prepared by centrifugal ultrafiltration and LMW serum proteins were profiled by MALDI-ToF MS. Mass spectra were manually calibrated, baseline-corrected, noise was removed and the spectra were smoothed before the centroid peak intensities were exported. For profiling, the peaks were standardised by total ion signal and peak intensity ratios of the averages of cancer and control were compared as well as *t*-tests performed on individual peak intensities. The target proteins were identified by MALDI-ToF MS/MS. Potentially, following identification, antibodies could be developed for validation of the biomarkers.

Each spectrum was checked visually and two spectra (V13b and V13c) had to be removed prior to further analysis, as these spectra were empty (Figure 5.5). The reproducibility of the remaining triplicate MS spectra can be seen in two examples (Figure 5.6) and the full data are shown in Appendix B.

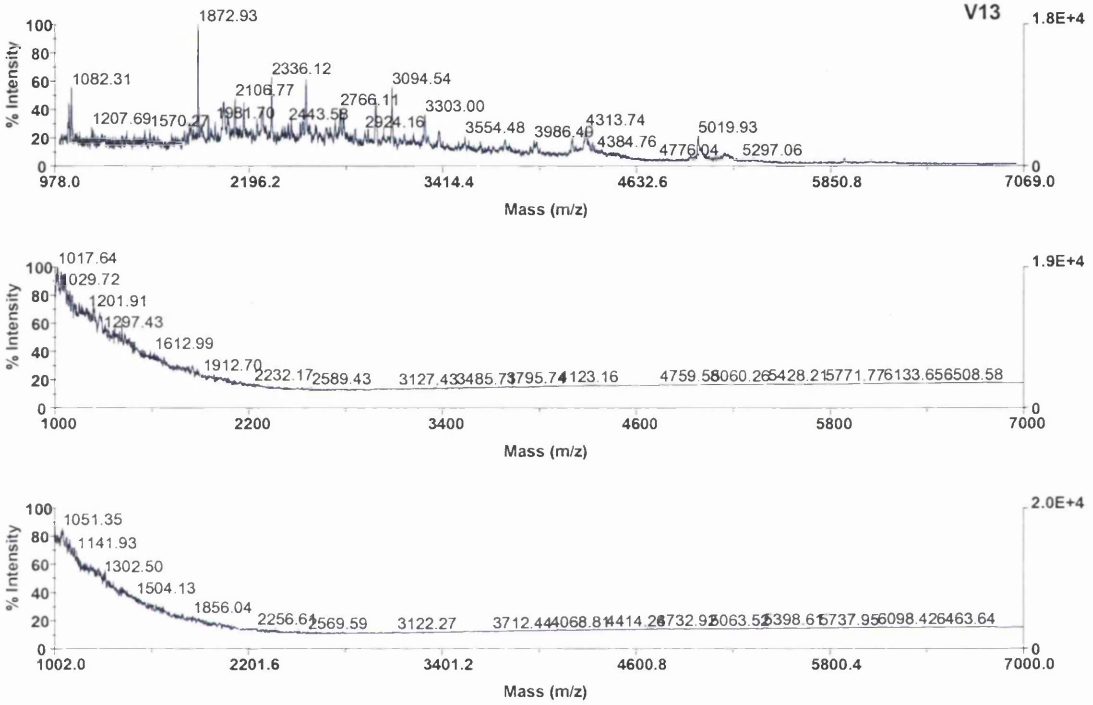


Figure 5.5: Three aliquots of sample V13 were each cleaned using Zip-Tips and analysed by MALDI-ToF MS. The data for the bottom two spectra were removed as the spectra appear to contain no peaks.

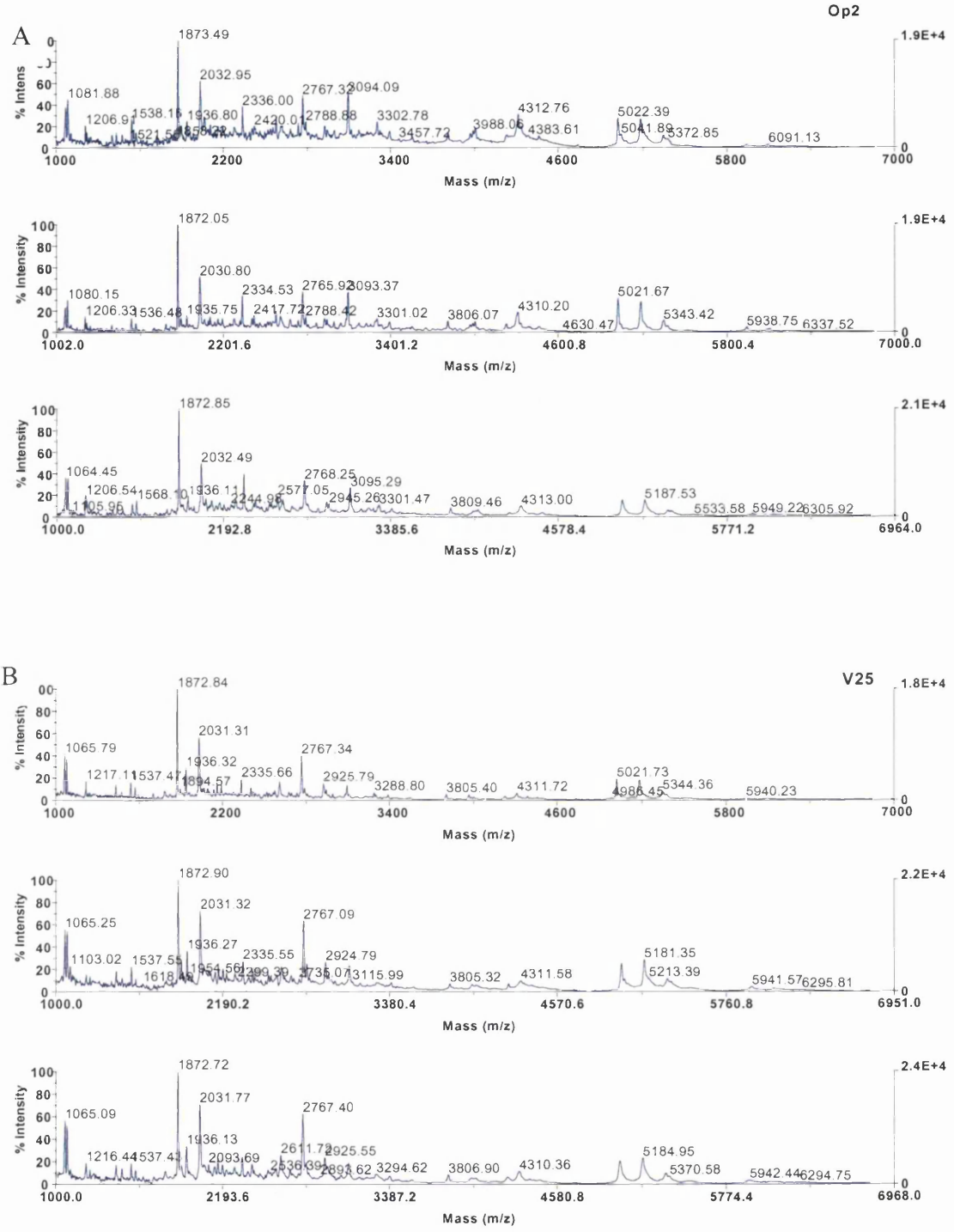


Figure 5.6: Two representative examples of MALDI-ToF MS spectra from LMW serum sample of the S1 sample set. The replicate Zip-Tip eluates analysed by MALDI-ToF MS are shown in one figure for each sample. The spectra have undergone baseline subtraction, noise removal and Gaussian smoothing.

Using the Data Explorer™ software (Applied Biosystems, Warrington, UK), each spectrum was manually calibrated (*m/z* 1873, 3094, 5022), baseline-corrected, the peaks were centriod and the peak detection performed as described in the materials and method (section 2.7.1). Peaks with a minimum intensity and area of 1 and that were detected above a signal to noise threshold of 100, were labelled and defined as a mass peak. The sequence of spectrum processing can be seen in Figure 5.7. The centroid peak intensities were exported as these contain the information of the peak intensity as well as of the peak area. Especially in larger molecular weight peaks this is important where the peak width may be increased. Manual calibration is essential as, despite external calibration, the mass accuracy of the Voyager STR DE was relatively poor. A mass accuracy of 10,000 ppm was observed, which means that a small mass peak of, for example 1873 Da, could be displayed at anything between 1850 and 1890 Da, or for a higher mass of 5080 Da, values could be recorded between 5030 and 5130 Da. Without manual calibration many peaks which are in fact different would be recorded as the same peak, or a peak that is essentially the same in two samples recorded as two different peaks, creating false peaks.

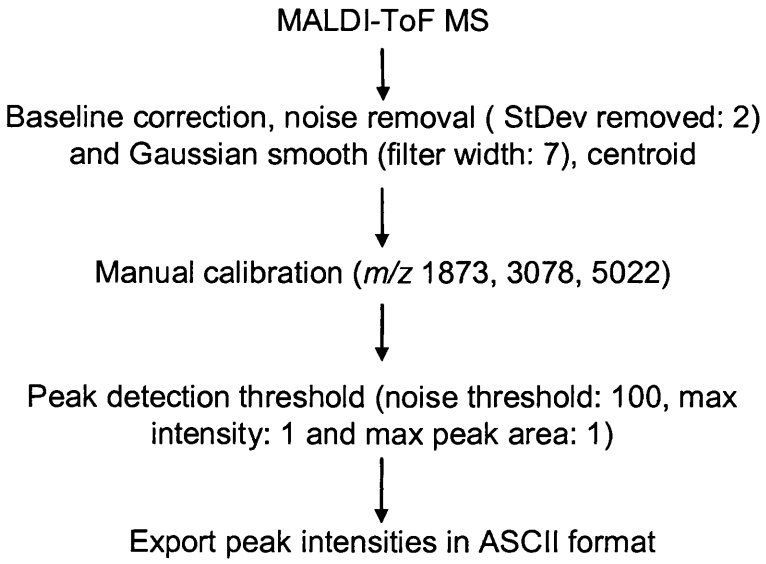


Figure 5.7: MALDI-ToF MS spectrum processing and peak detection in Data Explorer ™.

After manual calibration in Data Explorer™, the mass accuracy was significantly improved (Figure 5.6). Interestingly, visualisation of the mass accuracy in Figure 5.8 showed that the precision increased with molecular weight. Low mass peaks showed a mass accuracy of approximately 5000 ppm where as mass values greater than 4000 Da showed a mass accuracy of 2000 ppm and better. This is unexplained since the calibration file covered all the masses in the spectrum. The peak intensities were exported in the form of ASCII files for further analysis and biomarker detection.

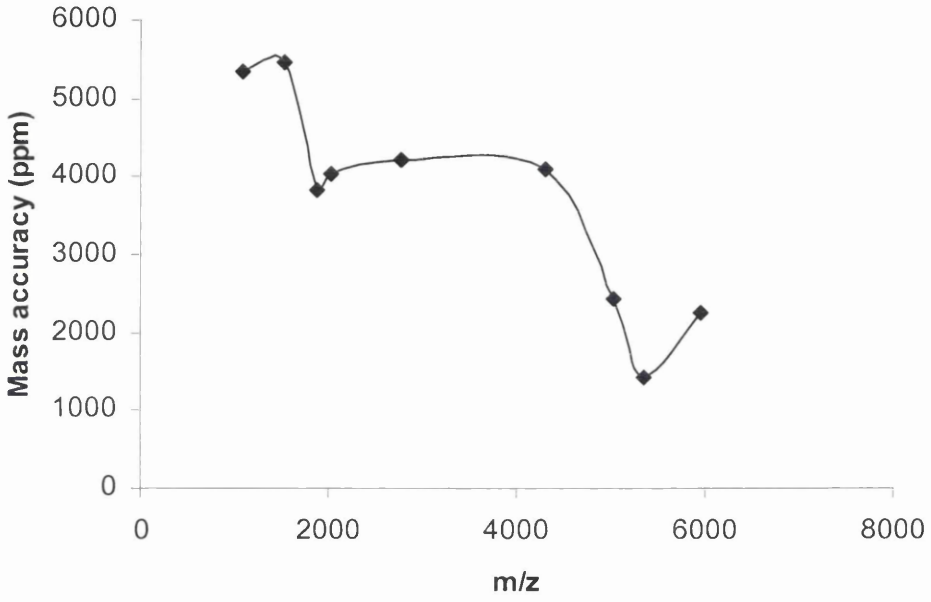


Figure 5.8: Correlation of mass accuracy against *m/z* of individual mass peaks. The mass accuracy appears to increase with increasing mass.

5.2. Data Standardisation to Remove Variation in the Spectra

Prior to any data analysis, the peak intensities for each spectrum were standardised; this is important since, as can be seen from Figure 5.9, the overall spectrum intensity can vary between spectra. For example the spectrum for OP1 has a lower intensity than that of sample OP2, rows one and two of the breast cancer spectra in Figure 5.9. This variation could have been introduced by a number of factors including, but not limited to, hotspots in the matrix that produce higher peak intensities, pipetting errors and application of different volumes or concentrations of sample [13, 14]. Finally the clean-up process using Zip-Tips may also introduce variations in the amount of protein eluted. However these variations should be minor, and not biologically significant. The standardisation described below can remove such differences. Any standardisation technique relies on making some assumptions: that the average number of proteins expressed is the same across all samples being standardised, and that the number of proteins whose expression levels change is small relative to the total number of protein peaks in the spectra. These assumptions should hold true for these samples, since the total number of peaks is very large, and the samples are biologically similar. The "Total Ion Current" (TIC) standardisation was used, which measures peak intensities of all peaks in each spectrum and calculating a "normalisation factor" (NF) that is used to bring the data closer together [15].

First the average of all peak intensities across each spectrum was calculated. Secondly a "normalisation coefficient" was calculated which takes an average across all the results from step 1 (similar to a grand average). Finally the NF for each spectrum was calculated by dividing the "normalisation coefficient" by the average peak intensity from step 1 for each spectrum. This is a standard automated procedure for other techniques (e.g. Ciphergen Biomarker Wizard software for SELDI-ToF MS analyses), but has to be performed manually in Excel for MALDI-ToF data.

The NF for each spectrum was then used to multiply each peak intensity value within the standardisation range. At this point it is worth mentioning again that the analysis (Zip-Tipping and MALDI-ToF MS) was performed in triplicate to minimise the effects of variation.

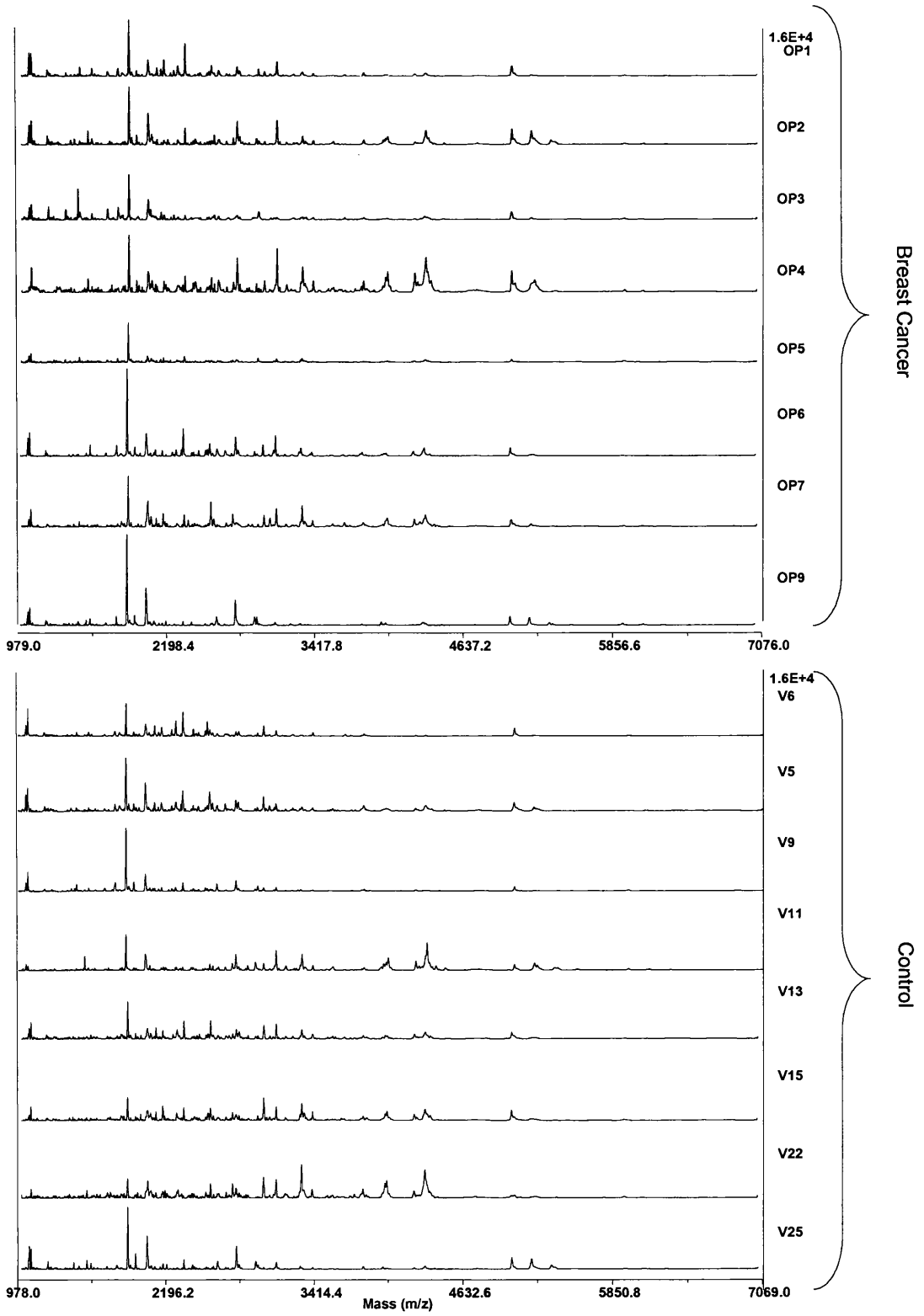


Figure 5.9: MALDI-ToF MS spectra of all LMW serum samples from S1. The triplicate spectra have been averaged using Data Explorer™.

5.3. Peak Alignment and Data Analysis: A New Software Tool

One of the most challenging aspects of protein profiling using MALDI-ToF MS is the lack of mass alignment tools to compare peak intensities of particular *m/z* values across all the spectra recorded. This is especially "tricky" since the *m/z* values for the same mass peak can differ from one spectrum to the next, as explained above with regards to mass accuracy. For statistical analysis the peak intensities for all peaks across all spectra had to be combined in one data sheet.

At the start of the project, there were no software packages for quantitative analysis of MALDI-ToF MS peaks commercially available. The complexity of the data recovered from serum, even just for the LMW fraction, is very large and requires an automated method for the alignment of peaks across all spectra to be able to compare the same peaks across all spectra analysed. I therefore created a peak alignment tool in Microsoft Visual Basic for Applications (VBA) for Excel; which was later named "*mz*Align". As a first step, a range for each *m/z* value was generated, with a value of 2500 ppm above and below the *m/z* value, in the reference mass list of the "mastersheet". In Excel the mastersheet is the spreadsheet where all masses from across all spectra are combined as the reference mass list (Figure 5.10). Using a VBA code, the mass list of each spectrum was then searched for a "match" in the *m/z* ranges of the reference mass list. If this was not present, the mass range of the "new" *m/z* value was added to the reference list. In this way, a mass list of all masses across all spectra could be compiled, but without duplicate *m/z* values. As this method is not completely foolproof and some mass ranges overlap or other mass ranges are too large, *m/z* values could be missed or could appear in more than one *m/z* range. The mass range was therefore increased in this second round for all *m/z* values in the newly compiled reference mass list to 5000 ppm. Using an "if clause" [=if (B3<=C2,C2+0.1,B3)] in Excel, the range was now changed only if an overlap occurred. As a control, a second programmed code was created using the reference mass list to align the *m/z* value from each spectrum in the mastersheet. This alignment was manually checked and, if necessary, the *m/z* ranges corrected. This step is labour-intensive and could be improved. A third code then searched through all the spectra as before and added the standardised peak intensity value for each peak next to the reference mass list in the mastersheet. The full code for each of the actions possible in *mz*Align is shown in Appendix A.

A screenshot of the *mz*Align program is shown in Figure 5.10. The second column shows the reference mass list (*m/z*) combined from across all spectra, the third and fourth column show the lower and upper mass boundaries, and the remaining columns show peak intensities that were filled in from across all spectra. Finally the buttons that can be pressed for creation of the reference mass list (Master List), alignment of the *m/z* values (*m/z* values) and alignment of either the raw or "normalised" peak intensities can also be seen. The performance of *mz*Align can be monitored, using the two progress bars for both the number of lines completed in the reference mass list and the number of sheets (sample/spectra) searched. The progress bars were included to allow the user to monitor the analysis process as this can be lengthy. Once all the peak intensity values from each spectrum have been aligned in the mastersheet next to their corresponding *m/z* value, statistical data analysis can be performed.

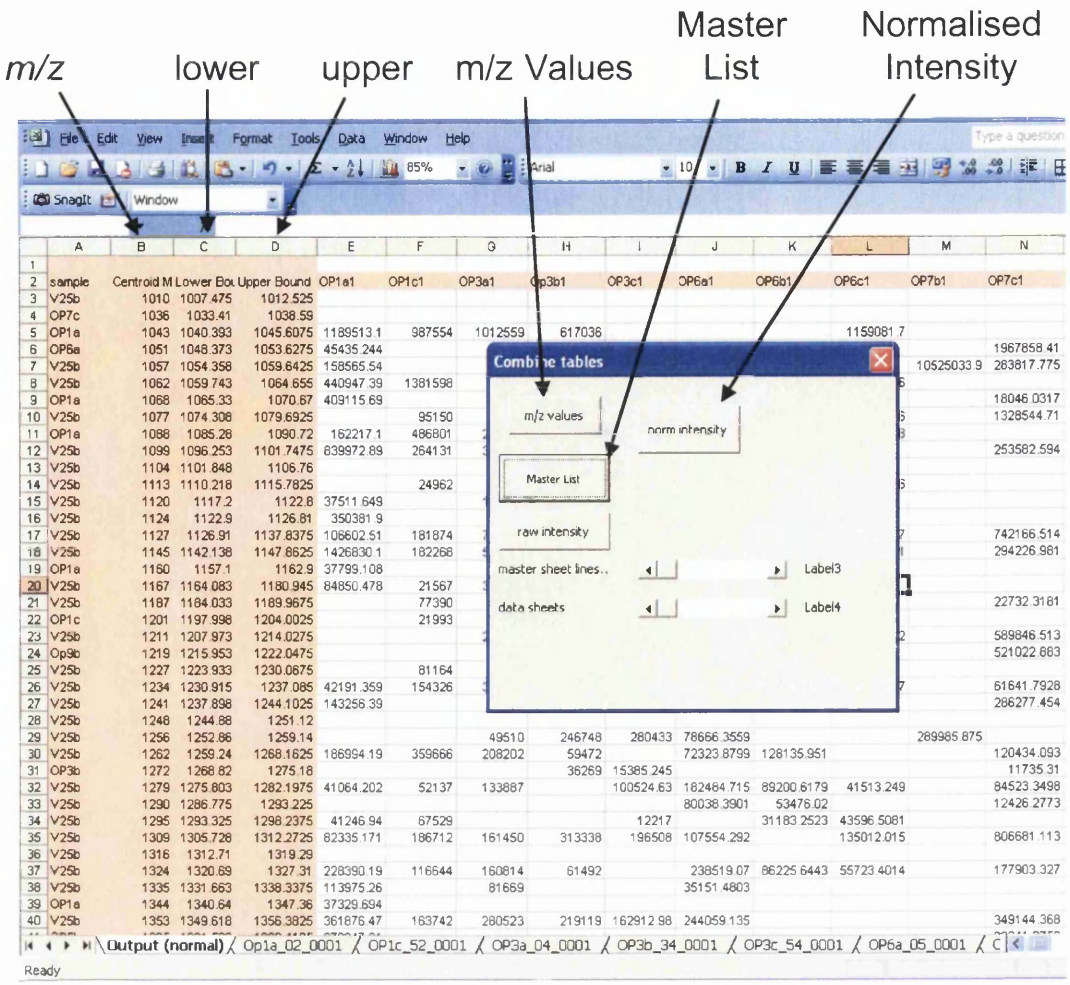


Figure 5.10: The "mastersheet" of the *mz*Align program created in VBA for peak alignment from MALDI-ToF spectra. The first part of the program creates a reference mass list of all masses across all spectra (seen in column B). The other programs search through all spectra using the reference master list and will add a value for either *m/z* value, raw intensity or the normalised intensity, where necessary.

For the analysis of MALDI-ToF MS spectra and comparison of peak intensities *mz*Align proved to be invaluable, enabling automated, high-throughput analysis of all mass spectra acquired as part of an experiment. In this way, 272 *m/z* values were combined from S1 into one table. After alignment the data could be further analysed in Excel or exported into another statistical software package such as SPSS. Later in the project, we obtained an evaluation software package from Applied Biosystems, called Markerview (MV), which performs a similar alignment to *mz*Align. In this program, the mass tolerance for creation of a reference *m/z* list can be defined during import of the spectra. However, corrections of the individual masses in the reference

list cannot be performed after peak alignment. A mass tolerance of 2000 ppm was recommended, however this resulted in the generation of too many peaks and did not correctly align masses that were the same between spectra. The mass accuracy across all the spectra was earlier calculated to have an average of 4000 ppm after manual calibration (Figure 5.8). Furthermore, use of this latter mass tolerance setting in Markerview resulted in a reasonable number of mass peaks and was most similar to the mass list generated using *mz*Align. Using Markerview and this mass tolerance setting, 274 *m/z* values were aligned into one table.

As a first step of statistical analysis, the data was tested for normality using the Kolmogorov-Smirnov test in SPSS. The vast majority of the data was found to be normal. As advised by a statistician from the library information service centre (University of Wales, Swansea) during an SPSS one-on-one course, since >90% of the data was normal, the entire dataset could be regarded as normal for further analysis. In fact, it would be expected for a large normal dataset to contain some data points that deviate from normality. A Student's *t*-test, for that reason, was then performed on the data. To simplify, the order of events involving data analysis using *mz*Align and Markerview is described in Figure 5.11.

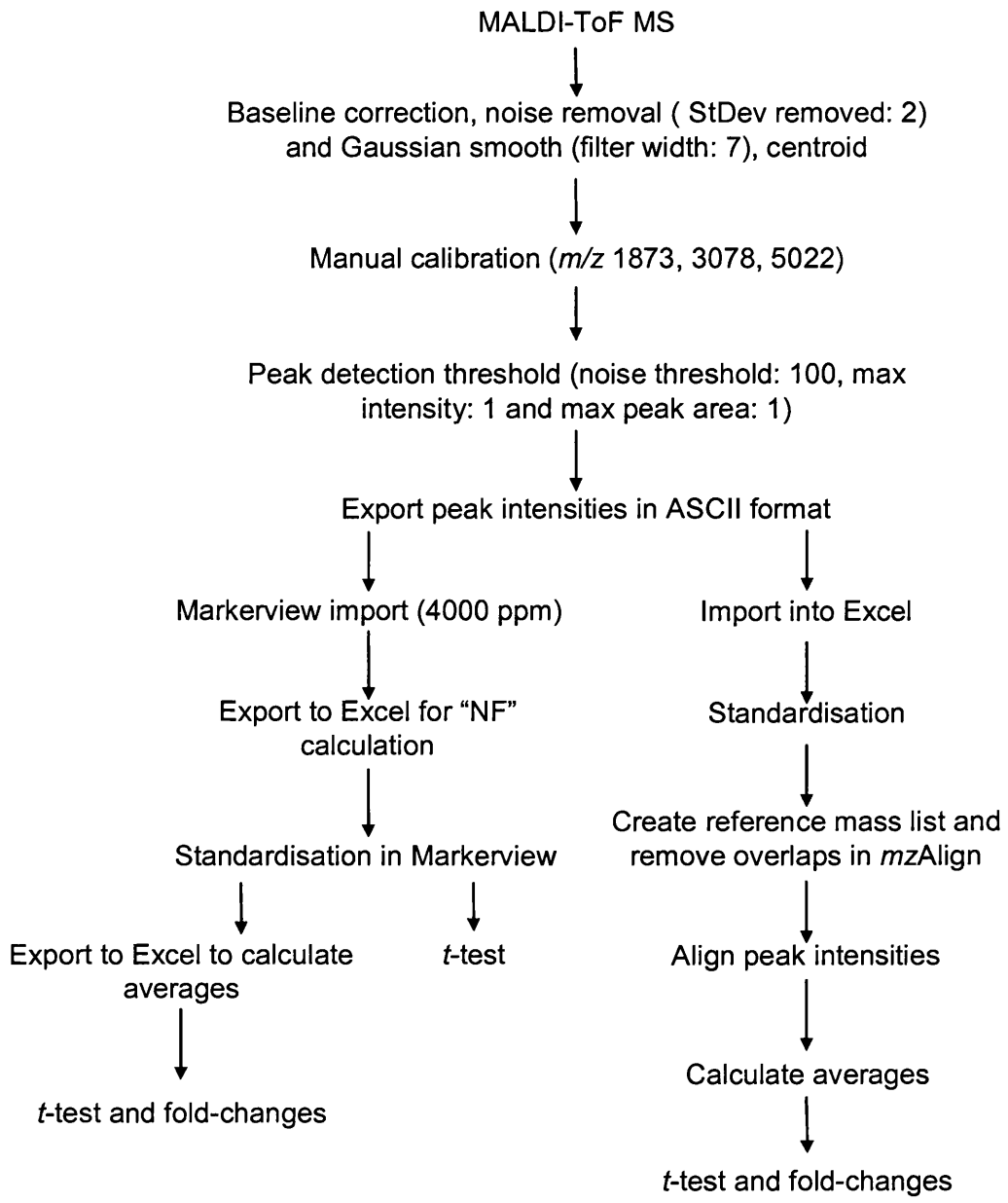


Figure 5.11: Data analysis and spectra processing. The data was analysed using our custom VBA alignment tool (*mz*Align) and the commercially available Markerview software. Both software programs use alignment to generate a reference mass list and to align all peak intensity values in rows for downstream statistical analysis. Prior to alignment, a "normalisation factor" (NF) was calculated in Excel for both tools, to correct for variations in each spectrum.

5.3.1. Biomarkers discovered in Sample Set S1

For biomarker discovery, the standardised data was analysed, comparing control and breast cancer samples. Before statistical analysis could be performed, all missing data points in the averaged dataset were replaced with a value of "1". This was justified by the assumption that if no peak is present in any of the three replicate spectra from the same sample, then it could be assumed that this protein peak is not present in the particular serum sample. Missing data points were replaced with a value of 1, and not zero, as the statistical algorithms cannot deal with the value of zero, as it causes division by zero errors. The value of "1" is significantly lower than the intensity values for "present" peaks. Values that are missing across all the technical replicates are therefore accepted to be "zero" but were replaced with "1". However the datapoints that are only missing in one or two of the three replicate spectra are instead combined as a single averaged value, reflecting the fact that they are present in at least one sample, meaning that there is in fact a compound present.

Principal components analysis (PCA) is a statistical technique that is used to show clustering or grouping of samples dependent on biological variation between two sample cohorts, without prior knowledge of the identity of each sample. Un-supervised PCA analysis of the technical replicates for S1 revealed two separate groups with the controls and breast cancer samples in separate clusters. Furthermore it showed that the replicates are actually relatively close together and that most of the variation therefore results from the biological samples, rather than from the technique itself (Figure 5.12).

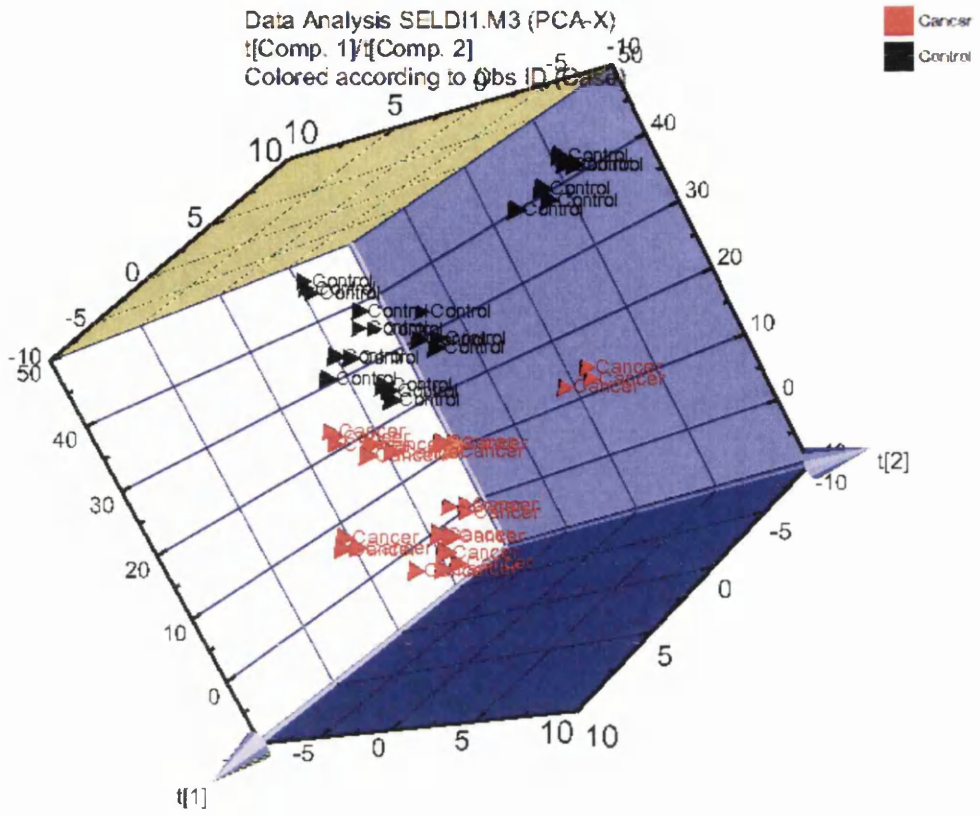


Figure 5.12: Un-supervised principal components analysis of MS-based serum protein profiling data derived from healthy controls and metastatic breast cancer patients.

The variance for each of the protein peaks combined in the reference peak list was examined. For this, all technical replicates from each LMW serum sample were averaged and an un-paired Student's *t*-test was performed as described in the Materials and Methods (section 2.9.2). The Student's *t*-test compares the peak intensities from the control and breast cancer serum samples and a *p*-value was calculated for each protein comparison. Prior to the *t*-test, the F-test statistic was calculated for each *m/z* value, to determine whether the two sample groups have equal variances. If equal variance was observed, a type II *t*-test was performed; if the variance was not equal, a type III *t*-test was used. This was computed in Excel as part of an "if" statement. Using a cut-off value of $p < 0.05$, 12 significantly different proteins were retrieved from the averaged sample sets (Table 5.1 and Table 5.2).

Table 5.1 shows the *m/z* value of each peak found to be different between breast cancer and control samples. Columns 2 and 3 show the number of spectra in which this peak was found, numbers in red indicate that this peak was found in less than 25% of spectra from each cohort. Table 5.1 shows the significant *p*-values for peaks from the un-averaged data, and, for comparison, the *m/z* peaks that were found to be discriminant using the averaged data. However, only differential peaks ($p < 0.05$) from the averaged data (shown in Table 5.2) are further discussed and the spectra analysed. Finally the results from alignment using Markerview followed by statistical analysis in Excel and in Markerview alone are shown in both tables, and the reasons for these different forms of analysis are discussed below.

Table 5.1: Protein peaks found to be differentially expressed in sample set S1. In columns two and three, the number of spectra (n Cancer and n Control) in which each peak was found is shown. If this number is shown in red then this peak was found in less than 25% of spectra from each cohort. In the last column, the results from the "*t*-test" in Markerview are shown. The software does not provide a fold-change value; therefore up- or down-regulation of the breast cancer samples (C) is documented. C down: peak intensity lower in breast cancer samples compared to controls and C up: peak intensity higher in breast cancer samples compared to controls.

			Alignment VBA, t-test excel				Alignment Markerview, t-test excel				t-test Markerview	
Centroid Mass	n Cancer (24)	n Control (22)	p-value VBA all	fold-change	p-value averages	fold-change	p-value MV all	fold-change	p-value averages	fold-change	p-value averages	fold-change
1064	10	8	0.014	1.5	0.019	1.7	0.025	1.4	0.014	**2.9**		
1185	11	9	0.032	1.4	0.023	1.4	0.026	1.3				
1273	10	10	0.027	1.8								
1313	14	9	0.037	**2.1**	0.035	**1.9**						
1341	13	11							0.028	**1.9**		
1355	11	11	0.033	1.3			0.004	1.2				
1365	9	10	0.047	**1.9**								
1378	11	10	0.016	**3.2**	0.045	**2.4**	0.018	**2.7**			0.045	**C up**
1391	1	6							0.048	**-7.0**	0.033	**C down**
1400	14	15					0.044	**1.8**				
1468	16	13	0.027	1.1			0.049	1.0				
1511	10	11	0.037	-1.5								
1591	18	14	0.049	1.7	0.047	1.6						
1608	16	8					0.022	1.6				
1776	10	10	0.008	1.0	0.002	**2.0**	0.018	-1.1	0.040	**1.9**		
1905	24	21	0.039	1.6								
1969	21	19					0.014	-1.3				
2034	14	8	0.046	**-2.7**							0.039	**C down**
2140	22	20					0.039	-1.2			0.019	**C down**
2441	19	21									0.046	**C down**
2556	18	19	0.037	-1.6	0.040	-1.7					0.009	**C down**
2565	4	3					0.027	-1.5			0.037	**C down**
2633	4	9									0.040	**C down**
2717	21	22									0.001	**C down**
2826	20	22	0.021	**-2.1**			0.004	**-2.4**	0.030	-1.5		
2925	13	12			0.038	1.5					0.029	**C down**
2963	14	21									0.000	**C down**
2995	19	21	0.002	**-1.9**	0.022	**-3.1**	0.001	**-2.0**	0.015	**-3.6**	0.023	**C down**
3338	20	22									0.005	**C down**
3432	16	21									0.024	**C down**
3594	13	16									0.010	**C down**
3850	21	21	0.017	-1.2	0.021	-1.6	0.027	-1.2			0.035	**C down**
4018	13	17									0.025	**C up**
4069	13	5									0.010	**C down**
4457	5	15									0.039	**C up**
5101	10	6	0.016	-1.1	0.022	**3.0**	0.030	**-2.7**	0.043	**2.7**	0.016	**C down**
5897	0	4			0.038	-26141.33			0.020	**-10.0**	0.026	**C down**
6123	0	5									0.032	**C up**
6278	6	1			0.036	11.1			0.036	**11.9**		
6992	2	3	0.001	1.2	0.002	1.7	0.000	**1.9**				

The Markerview software was mainly used for peak *m/z* alignment and the *t*-test was calculated separately in Excel. However the far right column shows *p*-values retrieved from a *t*-test performed as part of the Markerview analysis process. Interestingly, Markerview shows statistically significant peaks for *m/z* values that have no data points for one of the two groups (e.g. *m/z* 5897: peak absent in all cancer samples, but present in 4 controls). Despite the claim to be using a *t*-test, this indicates that Markerview must be using another form of statistical test or normalisation, since a *t*-test cannot be performed on groups where one arm has no values. In addition, when exporting the peak intensities aligned using Markerview, missing values were replaced with a "0" by the software, which would again make a *t*-test impossible.

Table 5.2: Statistical analysis of discriminating peaks derived from the averaged peak intensities from serum protein profiling of breast cancer patients and healthy controls. Aligned using *mz*Align and Markerview, number of spectra (n Cancer and n Control) in each cohort shown in red were found in less than 25% of spectra.

			Alignment VBA, t-test excel		**Alignment MView, t-test excel**	
Centroid Mass	n Cancer (24)	n Control (22)	p-value averages	fold-change	p-value averages	fold-change
1064	10	8	0.019	1.7	0.014	**2.9**
1185	11	9	0.023	1.4		
1313	14	9	0.035	**1.9**		
1341	13	11			0.028	**1.9**
1378	11	10	0.045	**2.4**		
1391	1	6			0.048	**-7.0**
1591	18	14	0.047	1.6		
1776	10	10	0.002	**2.0**	0.040	**1.9**
2556	18	19	0.040	-1.7		
2826	20	22			0.030	-1.5
2925	13	12	0.038	1.5		
2995	19	21	0.022	**-3.1**	0.015	**-3.6**
3850	21	21	0.021	-1.6		
5101	10	6	0.022	**3.0**	0.043	**2.7**
5897	0	4	0.038	**-26141**	0.020	**-10.0**
6278	6	1	0.036	**11.1**	0.036	**11.9**
6962	2	3	0.002	1.7		

The Markerview software can also provide an interesting alignment control option, where individual peaks can be visualised across all samples, as well as visualisation of the peak intensities across all spectra (Figure 5.13). Further evidence for using the averaged data for statistical analysis was provided by the peak intensity visualisation

which forms part of Markerview. This clearly showed that use of the averaged results was important to remove potential variation introduced by single outliers. For example, as seen in Table 5.1, the peak for *m/z* 5101 was calculated to be down-regulated in breast cancer samples from the non-averaged data, whereas the averaged data shows this peak to be significantly up-regulated in breast cancer. Looking at the individual spectra themselves and the Markerview visualisation, this protein peak may be up-regulated in breast cancer, but is certainly not down-regulated (Figure 5.13 and Figure 5.14).

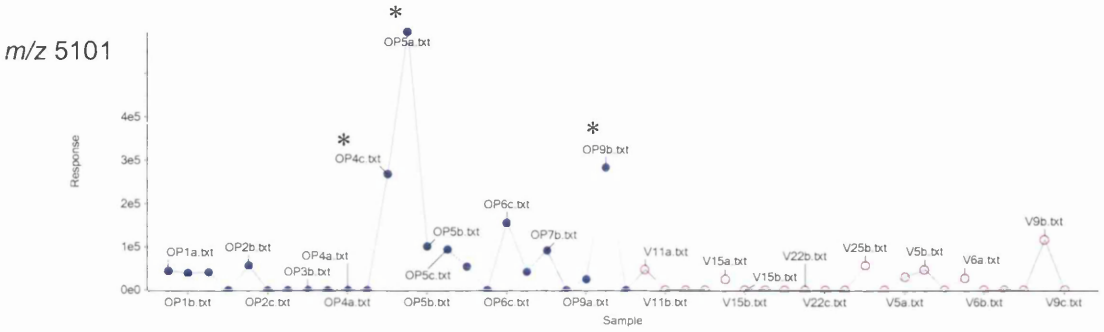


Figure 5.13: Markerview visualisation of the peak intensities across all spectra aligned for *m/z* 5101, showing skewing of the result by outliers (marked *).

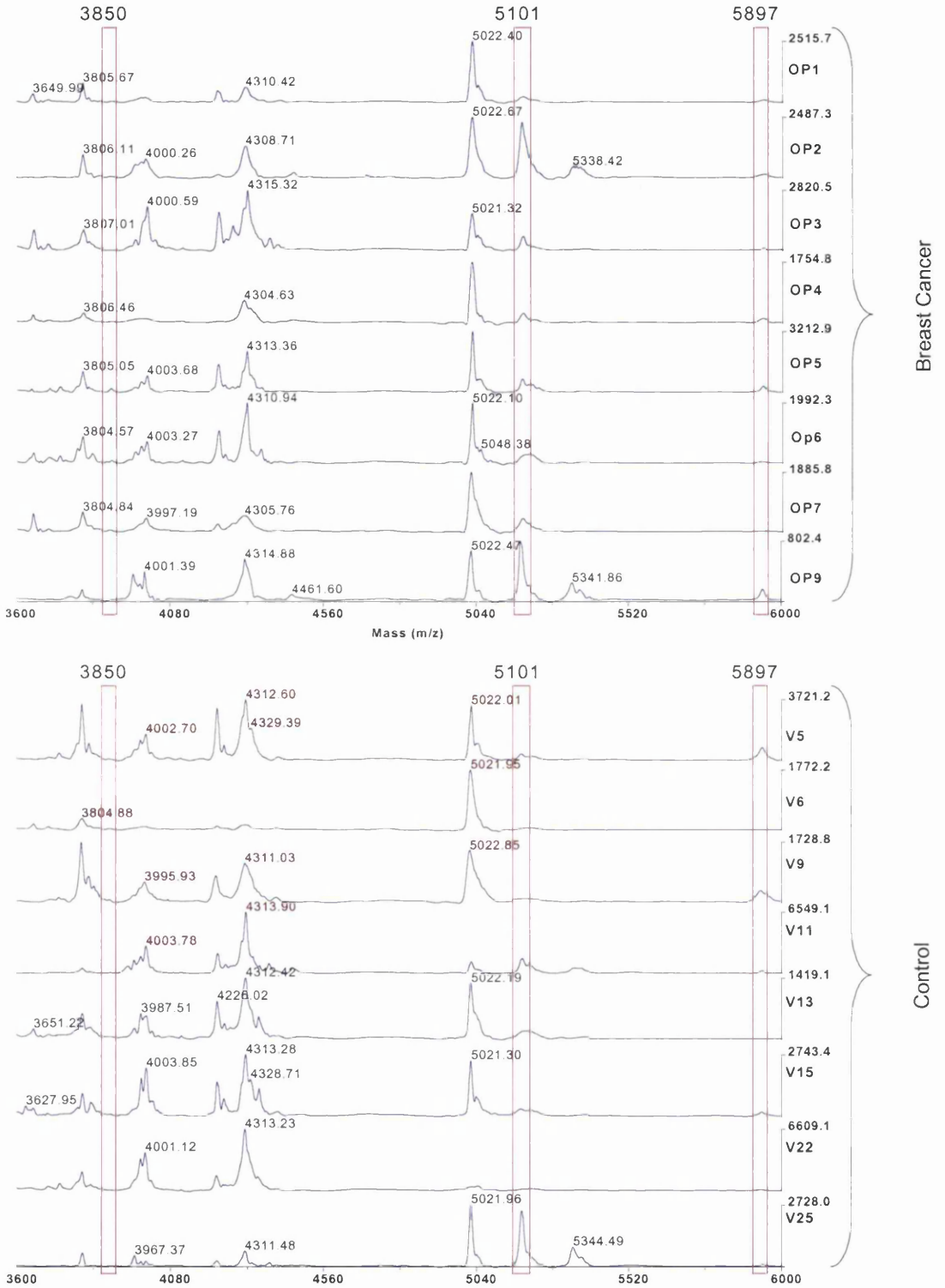


Figure 5.14: MALDI-ToF MS spectra aligned in Data Explorer from all samples across each clinical cohort. The absence of a peak for *m/z* 3850 and the discriminating peaks for *m/z* 5101 and 5897 are boxed in red.

These graphs further highlight the importance of calculating averages across replicates to reduce variation. Peak *m/z* 1608, in Figure 5.15 for example, was originally retrieved on the basis of having a significant *p*-value after calculating the variance taking into account all the data points (after alignment in Markerview), but was not significant when the averaged data was used. Figure 5.15 shows that 17 peak intensities were aligned with *m/z* 1608 in the cancer group, of which only four had elevated peak intensities. The remaining 13 peaks had roughly the same intensity as those aligned from the control spectra. Averaging removed the variation introduced by these outliers. Furthermore analysis of the spectra in Data Explorer™ in Figure 5.16 showed that *m/z* 1608 has a peak intensity that is too low to be seen in the spectra.

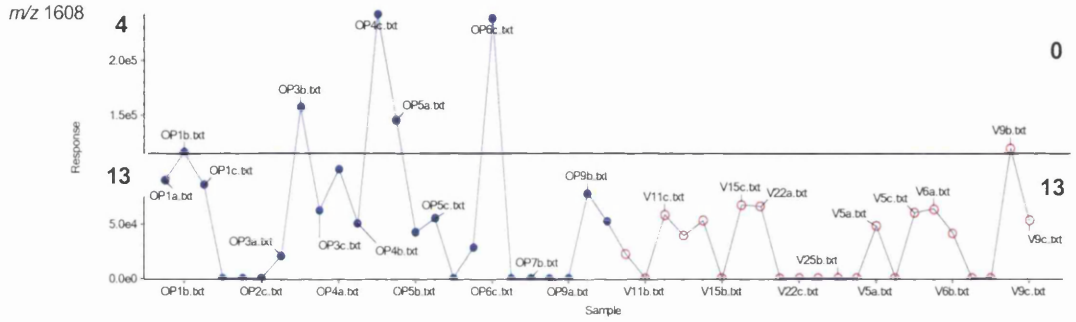


Figure 5.15: Markerview visualisation of the peak intensities across all spectra aligned for *m/z* 1608. In the breast cancer cohort, shown in blue (OP1 – OP9), four peaks have a greater intensity than any of the controls; however the remaining 13 peaks had a similar peak intensity to the controls (V5 – V25). The line shows the intensity of the highest control peak.

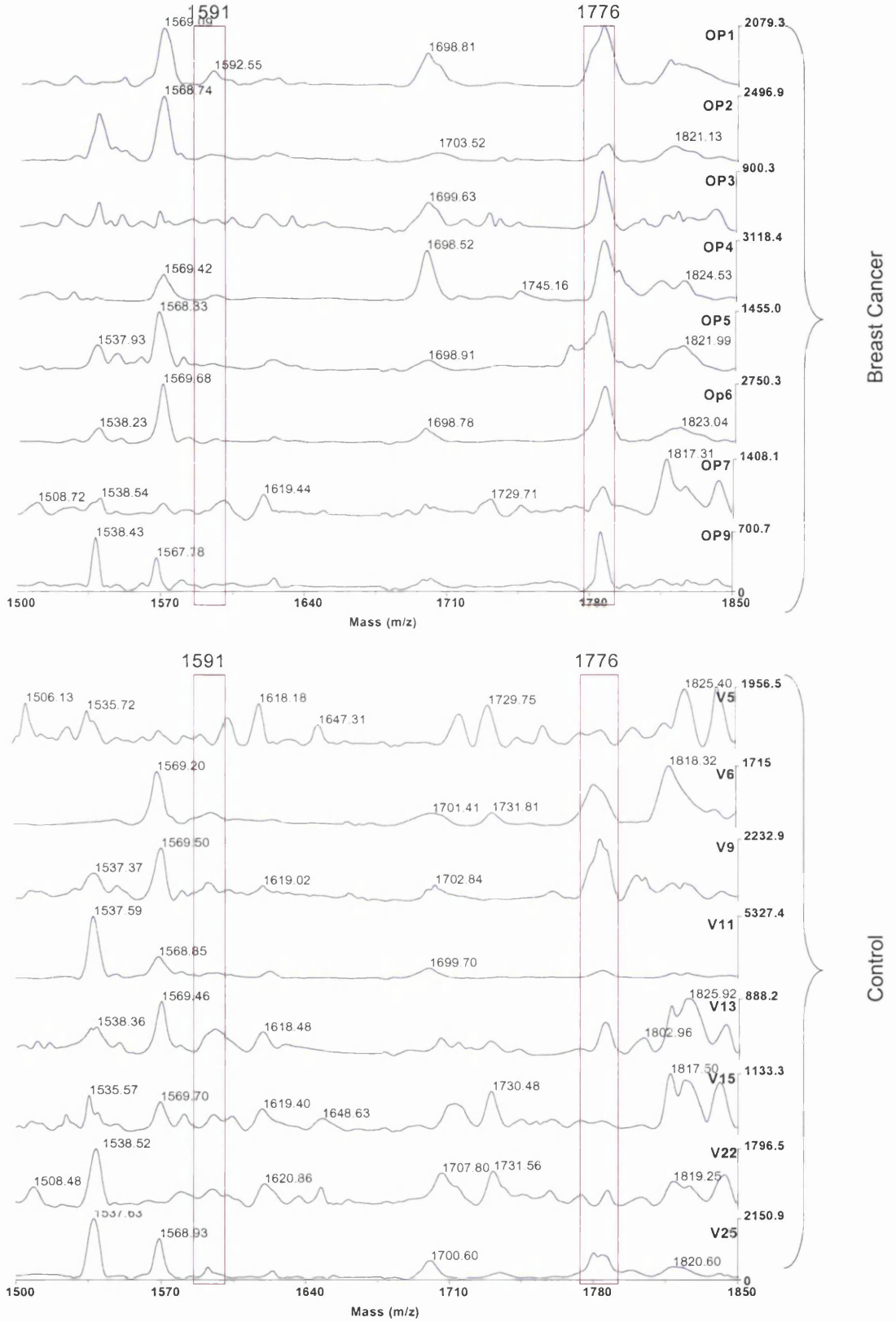


Figure 5.16: MALDI-ToF MS spectra aligned in Data Explorer from all samples across each clinical cohort. The mass range of *m/z* 1500 to 1850 shows that *m/z* 1776 was significantly up-regulated and *m/z* 1591 significantly reduced in the breast cancer samples.

In total, from the averaged data, using both alignment tools, 17 discriminating peaks were detected (13 using *mz*Align and 9 using Markerview). Of these, four were the same between *mz*Align and Markerview (Table 5.2). To verify the significance of discriminating peaks, each of the 17 protein peaks was visually inspected as described for *m/z* 5101 above (Figure 5.14 to Figure 5.19). It was shown that some peaks differ dramatically between the two groups (*m/z* 1064, 1776, 2556, and 2995) whereas others show more subtle differences (*m/z* 1391, 1591, 1925, 5101 and 5897), and yet others have a very low signal-to-noise ratio and are probably not significant (*m/z* 1185, 1313, 1341, 1368, 2826, 3850, 6278 and 6962). The peaks in this last category were too small (low intensity) to perform tandem MS analysis for identification and may therefore not be robust enough to classify as potential markers. Two examples are shown in Figure 5.15 and Figure 5.19, but the peaks are actually too low to be detected and are therefore not boxed, but the areas where the peaks would be can be seen in the figures.

The discriminant peaks will now be discussed in order of increasing molecular weight. Looking at the discriminant peaks for *m/z* 1064 and 1391, an overall greater intensity was observed in the breast cancer samples. Although the intensity of most peaks in the spectrum <1400 are variable across samples within the breast cancer and control groups, the peaks at *m/z* 1064 and 1391 are significantly and visually different, whereas the intensity of the non-discriminating neighbouring peak (*m/z* 1082) was the same between the breast cancer and control groups (Figure 5.17).

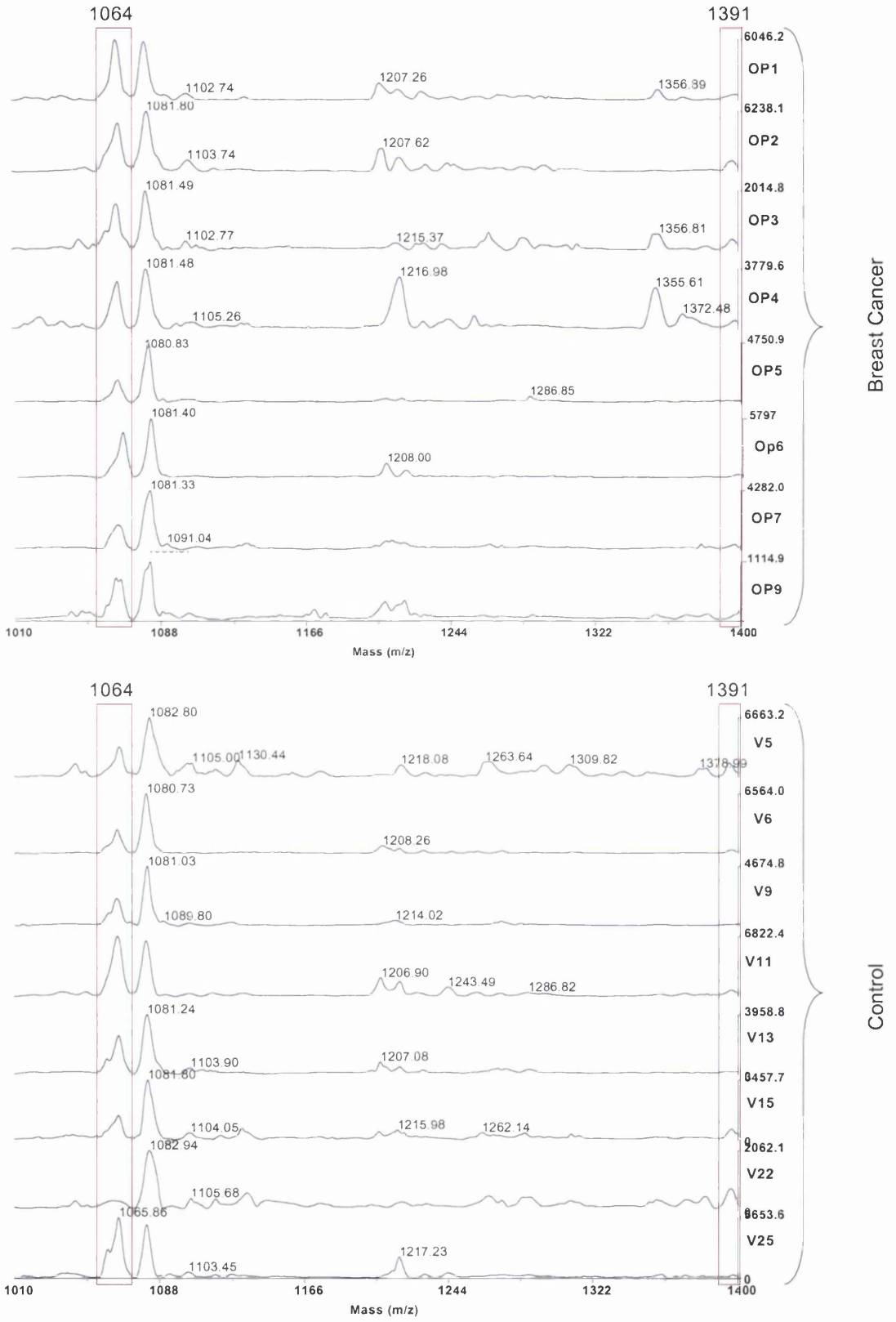


Figure 5.17: MALDI-ToF MS spectra from all samples in the mass range 1010-1400 Da.

Figure 5.18 shows the alignment of a truly discriminatory peak, *m/z* 2995. Here a significant number of mass peaks in the control sample group were of greater intensity than those from the breast cancer cohort. The Markerview graph is similar to what was observed in the spectra in Figure 5.19; although the peak is present in the breast cancer spectra, it occurs more frequently in the control samples and has a higher intensity. The same is true and can be seen for *m/z* 2556 (Figure 5.19). Also seen in Figure 5.19; although *m/z* 2826 has a significant *p*-value, the peak intensity is too low for visualisation. Furthermore, the peak at *m/z* 2925 does not look convincingly different between control and breast cancer samples, which was possibly to be expected as the groups were only 1.5-fold different in their peak intensities.

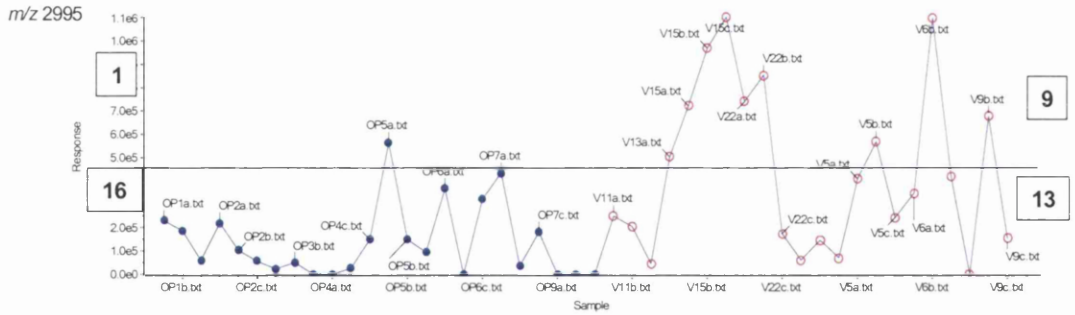


Figure 5.18: Markerview visualisation of the peak intensities across all spectra aligned for *m/z* 2995. In the control cohort (V5 – V25), shown in red, 9 samples show a higher peak intensity for *m/z* 2995 than the control samples in blue (OP1 – OP9). A line shows the intensity of the highest breast cancer peak.

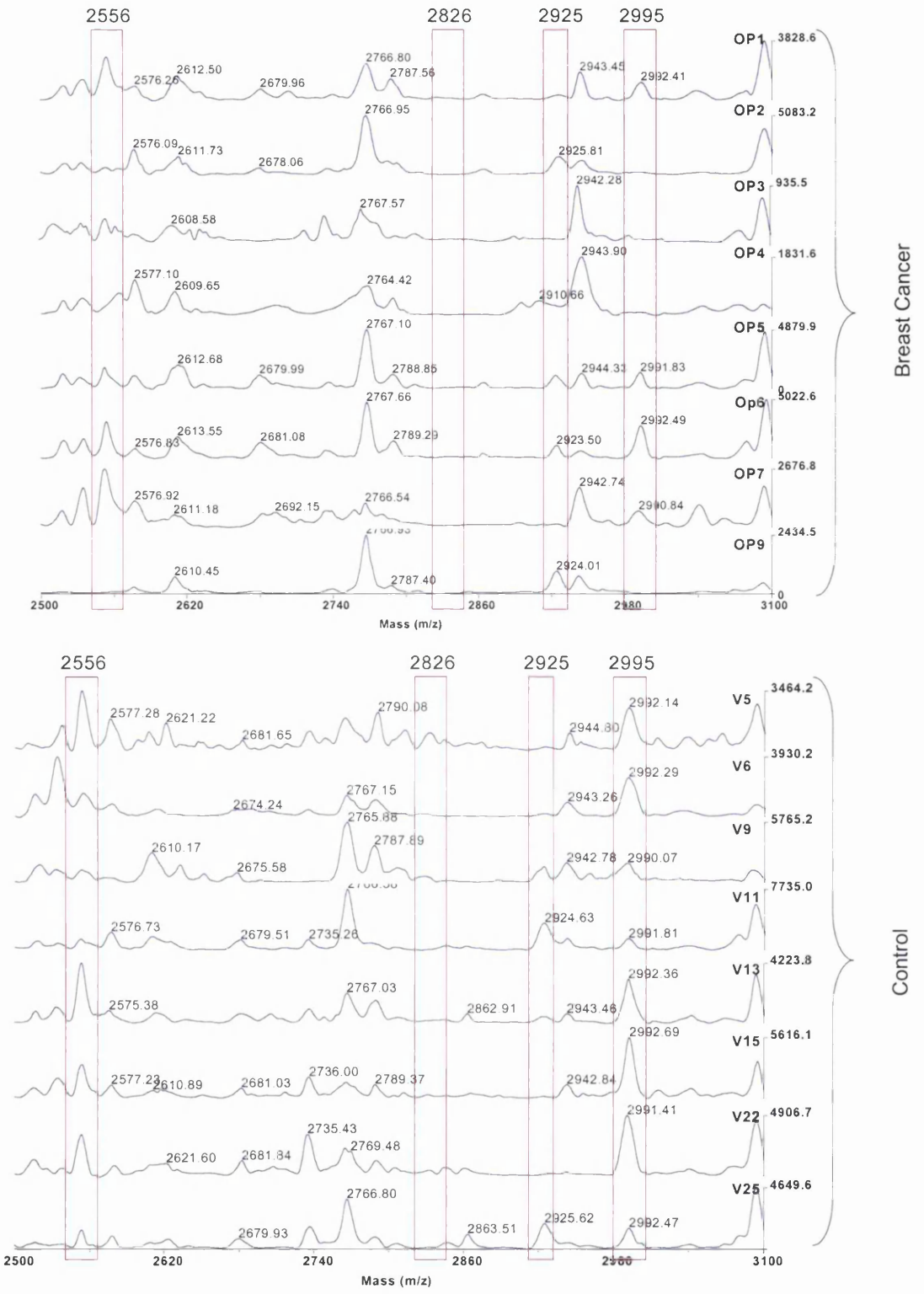


Figure 5.19: MALDI-ToF MS spectra aligned in Data Explorer from all samples across each clinical cohort. The mass range of *m/z* 2500 to 3100 is shown, the peaks for *m/z* 2995 appear to be more convincingly discriminating than the other peaks in the mass range.

Some discriminating peaks were only retrieved from the averaged data. Some of these were present/absent calls, i.e. expressed in one clinical cohort only (e.g. Figure 5.20 and Figure 5.21). The peak for *m/z* 6278 was increased in 6 breast cancer samples but in only one of the controls at a low intensity. For these peaks the missing values in the averaged data had to be replaced with "1" and a type III *t*-test was calculated allowing for unequal variance.

m/z 6278

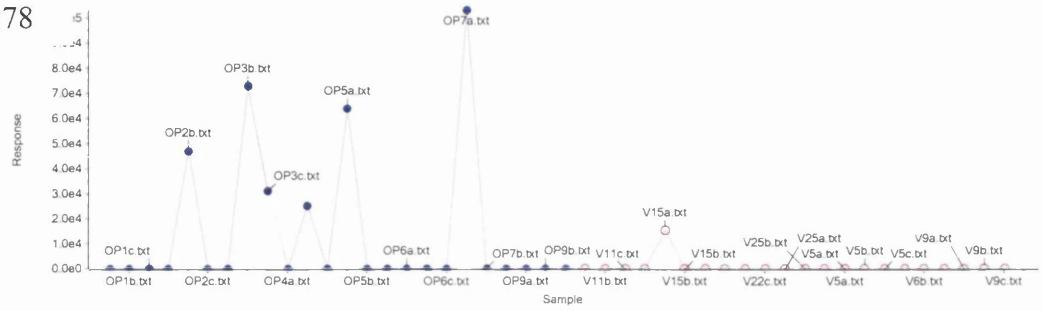


Figure 5.20: Markerview analysis of *m/z* 6278 shows protein peaks that are different between the two clinical sample cohorts. This peak is an example of peaks with a *p*-value that could not be calculated using a *t*-test in Excel, due to no peaks aligned from the control group, but was identified as significant using Markerview.

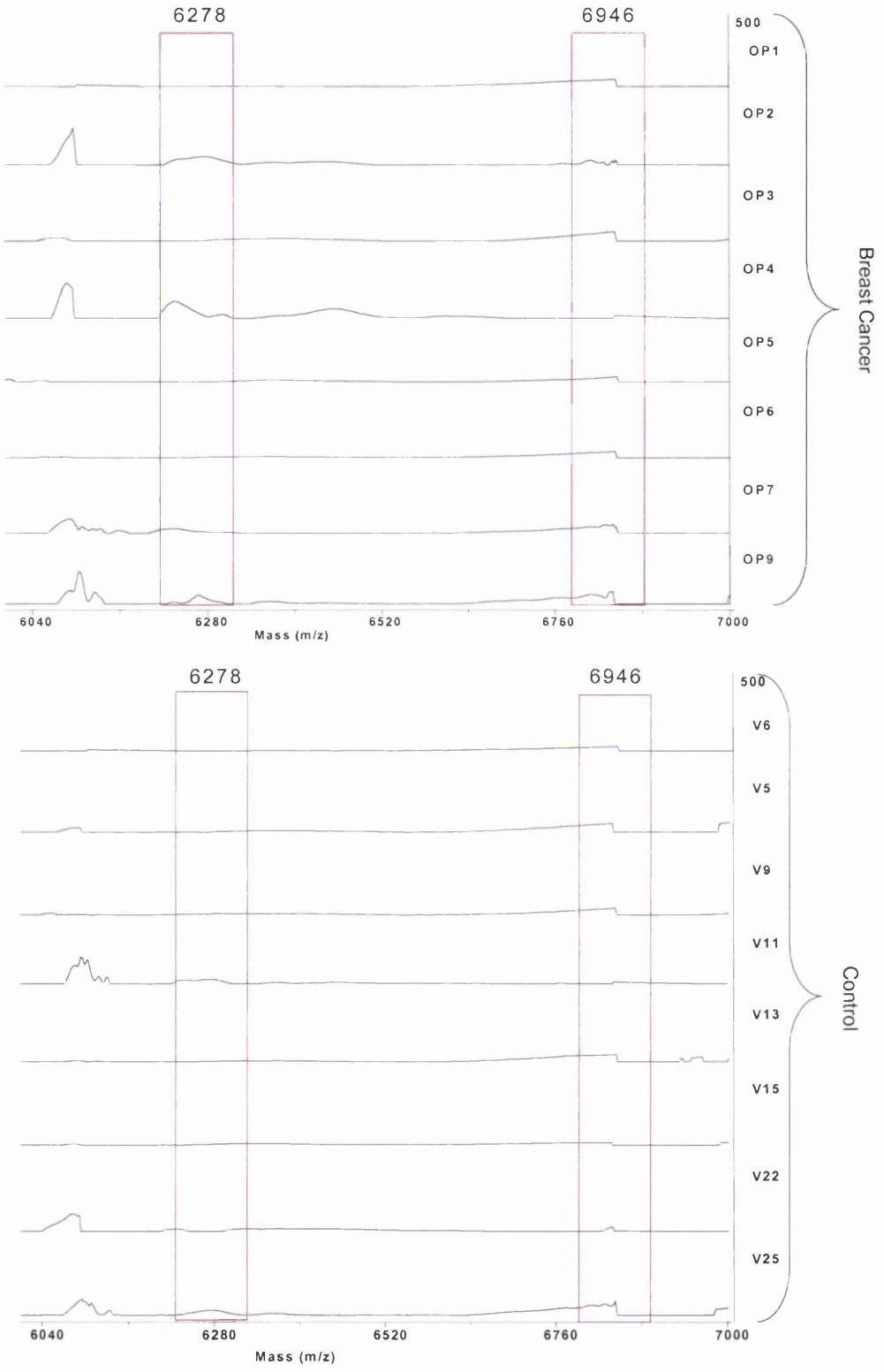


Figure 5.21: MALDI-ToF MS spectra aligned in Data Explorer from all samples across each clinical cohort. The mass range of *m/z* 6020 to 7000, showing *m/z* 6278 only to be present in some of the breast cancer samples and that *m/z* 6962 may not be a real peak.

During the course of this experiment, a number of issues were highlighted that could be further optimised. Although the results were interesting and may be useful, as the centrifugal ultrafiltration (UF) was only performed once, we could not be clear whether potential markers were found as a result of a clinical difference between the sample groups or due to the sample preparation step. The UF step therefore needs to be performed in triplicate. Furthermore, although the sample clean-up had been tested and optimised; and the LMW serum samples were desalted and concentrated using Zip-Tips from the same volume of LMW filtrate as for SELDI-ToF analysis, the actual capacity of the Zip-Tips had not been defined. This was tested in the following section.

In addition, the spectra generated in this experiment showed a lower number of peaks than expected, with surprisingly few low molecular weight peaks present. A thorough check of the specifications of the MWCO filters revealed that actually only 10% of proteins with a molecular weight >12 kDa are expected to pass through a 30 kDa MWCO filter, despite their name. Hence, as described previously (Chapter 3), the UF step was optimised and in the remainder of the experiments 50 kDa MWCO membranes were used. Finally, using the larger MWCO filters, larger proteins were recovered in the filtrate and therefore a different matrix than αCHCA may have to be used. The optimal matrix composition was determined in the next section.

5.4. Optimisation of the Sample Preparation Procedures

5.4.1. Optimisation of Sample Concentration

To analyse human serum by mass spectrometry, the sample first has to be desalted, as salt interferes with the ionization process. For this, C18 chromatography in the form of Zip-Tips is often used: the C18 chains bind proteins and peptides and allow salts to be washed off with acidified water. C18 Zip-Tips have a binding capacity of approximately 5 μg of protein. For larger sample quantities, SPE cartridges, also packed with C18 reverse-phase silica, are recommended. Here the binding capacity of C18 Zip-Tips as well as SPE cartridges was investigated to find the optimal conditions for salt removal and protein concentration.

Different amounts of LMW protein were Zip-Tipped and directly applied onto a MALDI-ToF plate using the sandwich preparation method, where matrix, sample and then matrix again are applied separately and each layer is left to dry before the next is applied, using sinapinic acid as matrix (although it was later discovered that mixing sample and matrix 1:1 before spotting provides better results). The Zip-Tip flow-through (FT) and the TFA wash of each sample were collected and desalted using Millipore membrane discs before MALDI-ToF MS analysis. The membrane discs remove salts through a different process than C18 chromatography: the sample is spotted onto the "filter paper" which floats on a water bath. The salts pass through the paper by osmosis leaving the purified sample behind. This provides an alternative method for desalting than the C18 chromatography for the proteins that passed through the Zip-Tips without binding.

To check the optimal binding capacity of the C18 Zip-Tips, a LMW filtrate with a protein concentration of 85 μg/ml was Zip-Tipped from a range of volumes (500, 250, 100 and 10 μl), as described in the Materials and Methods (section 2.5), to give varying final protein amounts of 42.5, 21.25 μg, 8.5 and 0.85 μg. The bound peptides were eluted in 3 μl of acidified water and spotted with sinapinic acid. The spectra are shown in Figure 5.22 a-c; they are very similar and no peaks appeared to be lost when smaller volumes were concentrated. The results show that the protein recovery did not improve with protein amounts greater than 20 μg (Figure 5.22). This may therefore be the maximum capacity of the 10 μl Zip-Tips. At the lower end of the volume

spectrum, after concentrating 100 µl (i.e. 8.5 µg of protein), the same protein peaks were visible in the spectrum, just at a slightly lower intensity, whereas, in the filtrate taken from 10µl of serum, i.e. 0.85 µg of protein, many peaks were lost and the spectrum showed different peaks. It was therefore determined that 8.5 µg of protein is the optimal protein amount for loading, taking into account the limited sample availability.

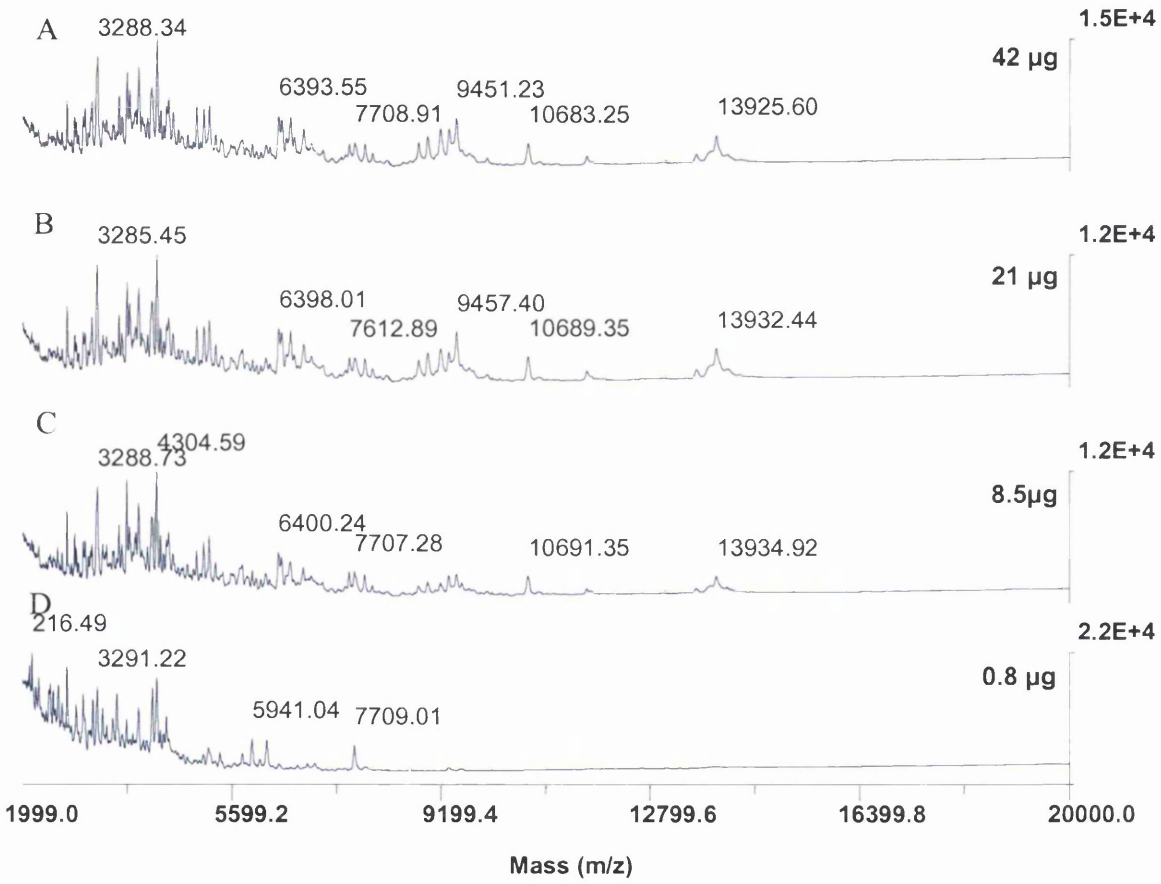


Figure 5.22: MALDI analysis of different volumes of Zip-tipped LMW serum. **A**: 500µl of 85 µg/ml LMW filtrate, **B**: 250 µl, **C**: 100 µl and **D**: 10 µl of sample.

When analysing the FT and the 0.1% TFA wash, it became apparent that many proteins, especially those of larger molecular weight, did not actually bind to the Zip-Tips. Overloading may be the reason why proteins that were also present in the eluted fraction also appeared in the FT (Figure 5.23). Larger proteins especially serum albumin do not bind well to C18 chromatography (Figure 5.24).

Due to this poor binding capacity of albumin to C18 Zip-Tips and SPE cartridges have actually been used for albumin depletion in the past [10]; (personal communication Matharoo-Ball, B. 2006) and also in form of magnetic beads covered in C8-oligo silica [6, 8]. However throughout this study we used C18 Zip-Tips purely for desalting.

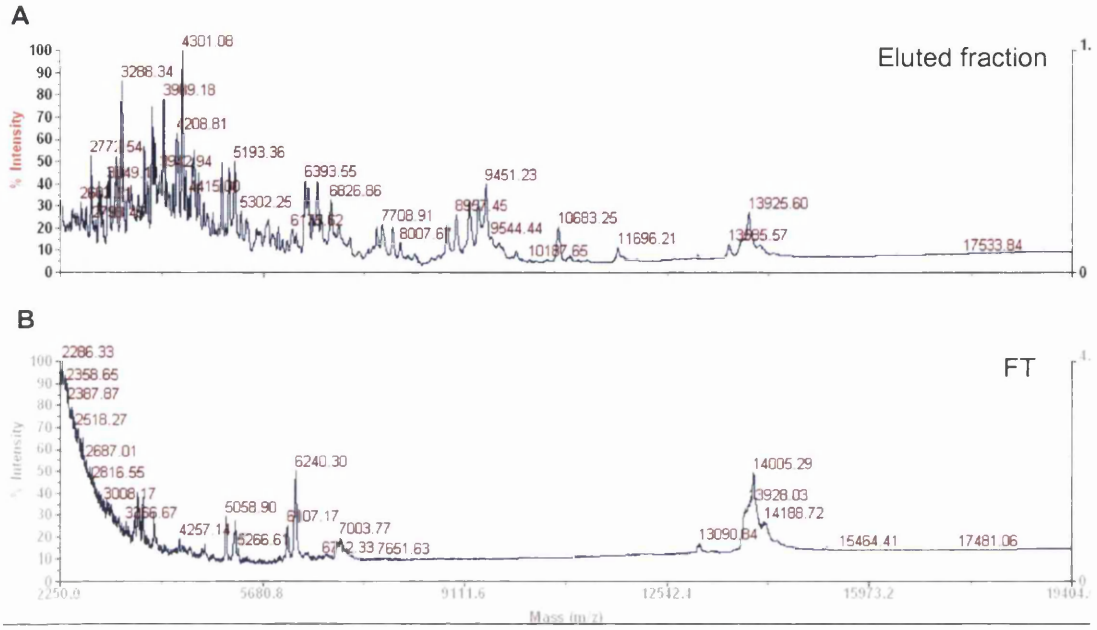


Figure 5.23: MALDI-ToF MS analysis of proteins after Zip-Tip clean-up. Panel **A** shows peaks from the eluted fraction of the Zip-Tip and panel **B** shows proteins in the FT that were not bound to the C18 Zip-Tips.

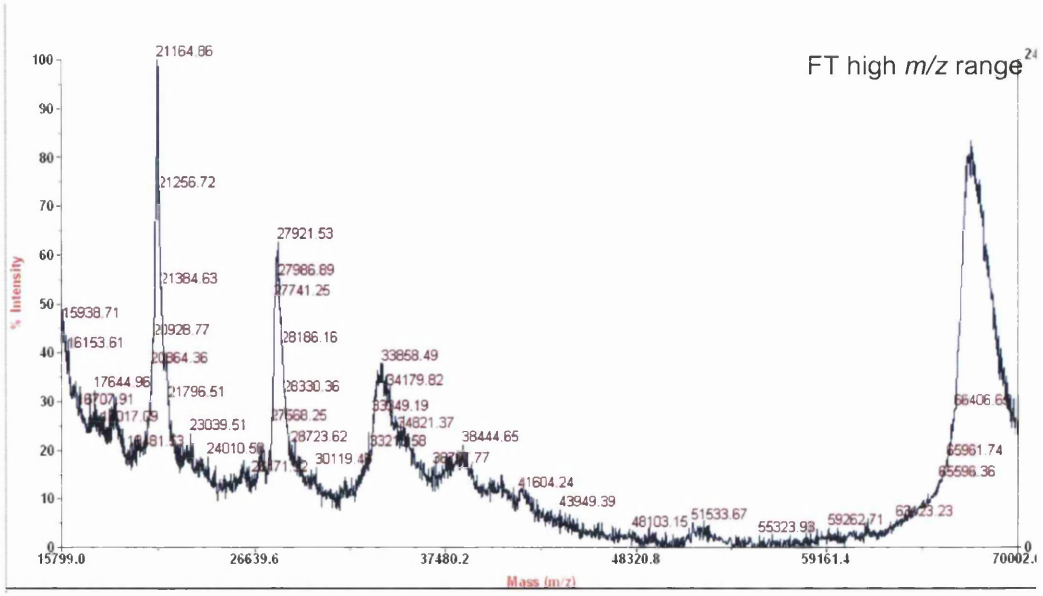


Figure 5.24: MALDI-ToF analysis (mass range 15- 70 kDa) of the FT of C18 Zip-Tips. Larger molecular weight proteins bind poorly to the C18 Zip-Tips and are present in the FT. The eluted faction of the Zip-Tips contains no proteins and produced an empty spectrum (not shown).

5.4.2. Comparison of C18 Zip-Tips with C18 SPE cartridges

Here, 1 ml of the same LMW serum sample (85 µg/ml) as used previously was desalted using C18 SPE cartridges from Empore™ (3M, Bracknel, UK) for direct comparison with C18 Zip-Tips. Analysis of the eluted fraction from these cartridges by MALDI-ToF MS did not produce the same number of peaks as when these samples were desalted with C18 Zip-Tips (Figure 5.25). In particular, no peaks were detected in the 7-70 kDa mass region (Figure 5.25b). Many proteins did not bind to the SPE cartridges but instead appeared in the FT, which leads to the conclusion that the SPE cartridges did not bind protein as well as the Zip-Tips tested. Considering the fact that C18 material is used in both devices, it is possible that the make of this particular type of cartridge was inferior. It was therefore decided that the Empore™ SPE cartridges were not suitable for desalting LMW serum samples, and C18 Zip-Tips from Millipore (Watford, UK) were chosen for further analysis.

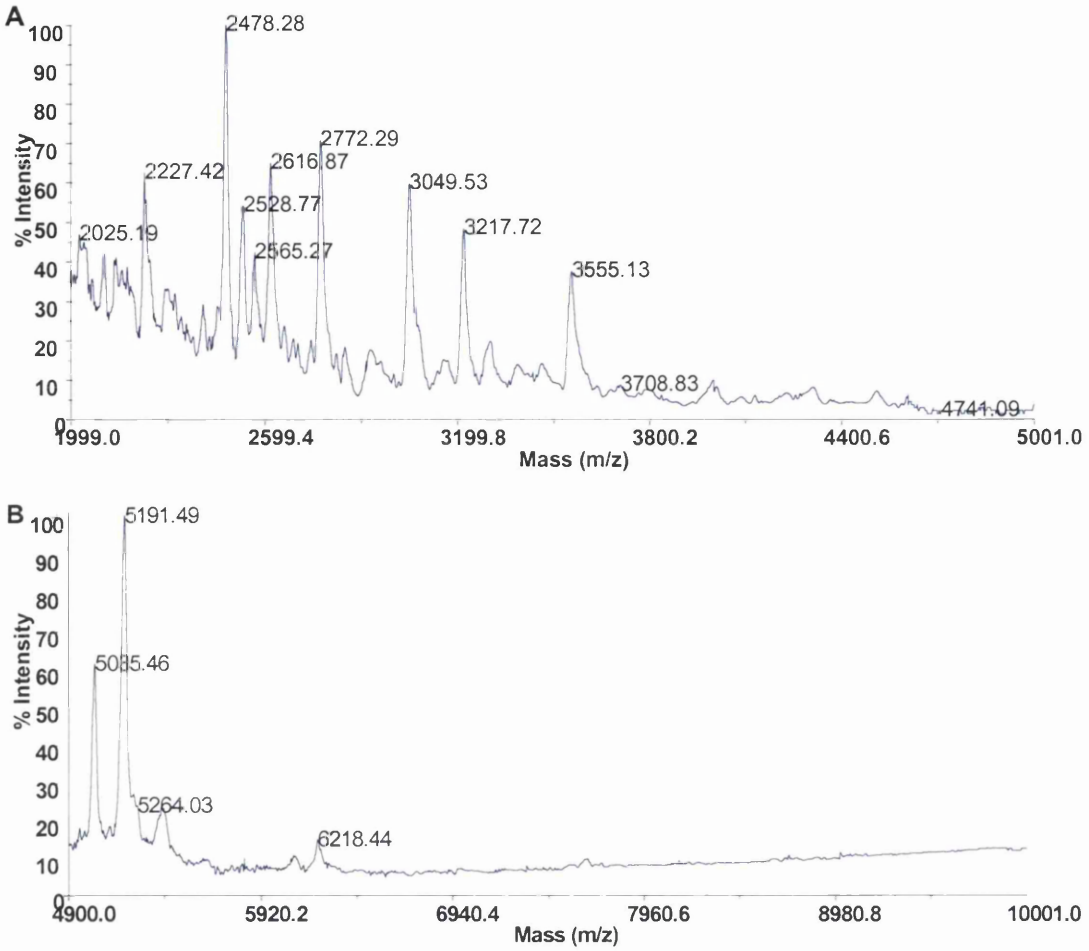


Figure 5.25: MALDI-ToF analysis of the eluted fraction from SPE C18 clean-up. The fractions show fewer peaks than after C18 Zip-Tip concentration, when analysed by MALDI-ToF MS. The mass range between 2000 and 5000 Da is shown in **A** and larger molecules in **B**.

5.4.3. MALDI-ToF Matrices

One of the most important aspects of MALDI-TOF MS is good sample preparation [16, 17]. Many variables influence the integrity of a good, homogeneous MALDI sample and this includes, but is not limited to, the concentration of the matrix and analyte, choice of matrix, analyte sample history (i.e. exposure to strong ionic detergents, formic acid), the sample/matrix application and solubility and compatibility of matrix and analyte solutions. A number of different matrices are available for MALDI-ToF MS analysis. Not all samples work well with every MALDI-ToF matrix since each peptide/protein has a unique structure. Additionally each matrix compound has its own unique physical properties and interacts with the analyte molecules in a unique manner. In addition, to select a matrix to obtain good qualitative spectra, it is important to produce a homogeneous spot, in order to get reproducible spectra, especially during automated acquisition of mass spectra. In this section, commercially available matrices and their mixtures (Table 5.3) were tested for the optimisation of a matrix for analysis of intact LMW proteins for protein profiling. Critical qualities examined were homogeneity of spread of the matrix/analyte mixture, homogeneity of intensities across the spot (i.e. no hotspots), good signal intensity at low laser energy and a high signal-to-noise ratio. The matrices tested and their applications are described in Table 5.3.

.

Table 5.3: MALDI-ToF MS matrices tested to find the optimal matrix for quantitative protein analysis from LMW serum. Sinapinic acid (SA), α-cyano-4-hydroxycinnamic acid (CHCA), 2,5-dihydroxybenzoic acid (DHB) and L-(–)-fucose (6-deoxy-L-galactose) (fucose) were prepared at different concentrations and as mixtures.

Matrix No	Matrix	Conc	Solvent
1	SA	10mg/ml	30/70 ACN, 0.1% TFA
2	SA	30mg/ml	50/50 MeOH
3	SA - fucose	8mg/ml each	50/49/1 0.1% TFA,ACN,Acetone
4	CHCA	10mg/ml	50/50 ACN, 0.1% TFA
5	CHCA - FA	10mg/ml	70/30 ACN, 5% FA
6	CHCA - fucose	8mg/ml each	50/49/1 0.1% TFA,ACN,Acetone
7	DHB	20mg/ml	30/70 ACN, 0.1% TFA
8	DHB - fucose	8mg/ml each	50/49/1 0.1% TFA,ACN,Acetone
9	fucose	10mg/ml	50/50 ACN, 0.1% TFA
14	DHB-CHCA	1:1 (v:v)	matrix 7 and 4
15	DHB - CHCA - fucose	1:1 (v:v)	matrix 7 and 6
11	SA - CHCA	1:1 (v:v)	matrix 1 and 4
12	SA - CHCA	1:2 (v:v)	matrix 1 and 4
13	SA - CHCA	2:1 (v:v)	matrix 1 and 4

Each matrix (Table 5.3) was evaluated for suitability with LMW serum proteins by analysing a low (2-5 kDa) and a high (5-16 kDa) mass region. Digital photographs were also taken of the matrix spots to show how well each matrix/analyte mixture spread across the MALDI plate when spotted (Figure 5.26); additionally each spot was examined for spot appearance, homogeneity of the peak intensity across the spot and the optimum laser intensity required to produce good mass spectra (Table 5.4). The results showed that, to produce the same peak intensity, SA required a higher laser intensity than either αCHCA or DHB, making these latter two the better choices. However DHB did not produce a homogeneous spot (Figure 5.26: matrix 7). Addition of L-(–)-fucose (6-deoxy-L-galactose) to other matrices had previously been reported to improve homogeneity of the spot [18-20]. However when fucose was added to each

of the three matrices, these results could not be confirmed, as none of the mixtures showed improved spot homogeneity (Table 5.4 and Figure 5.26: matrices 3, 6 and 15).

Table 5.4: MALDI spot properties for different matrices in respect of production of protein peaks, homogeneity of matrix and optimum laser intensity required for good peak intensity at low and high MW ranges. Each matrix was prepared as explained in Table 5.3, matrix 1 (sinapinic acid) and matrix 4 (αCHCA) were applied mixed with the analyte before application to the plate (m) and by the sandwich method (s) where matrix is applied then analyte and then matrix again.

Matrix No	Matrix	Spot appearance	Intensity homogeneity	Laser intensity (low/high *m/z*)
1m	SA		good	3031 / 3202
1s	SA		worse than mix	3067 / 3209
2	SA	no spot		
3	SA - fucose			3102 / 3245
4m	CHCA	homogeneous	good	2289 / 2353
4s	CHCA	worse than mix	worse than mix	2183 / 2318
5	CHCA - FA			2225 / 2340
6	CHCA - fucos	funny bubbles	good	2389 / 2546
7	DHB	crystals outside	worse high MW peaks	2781 / 2924
8	DHB - FUCOSE	worse than 7	worse than 7	3138 / 3138
9	fucose	no peaks		
14	DHB-CHCA	not homogeneous	good peaks	2297 / 2632
15	DHB - CHCA FUCOSE	less bubbles than 6	less peaks than 6	2746 / 3174
11	SA - CHCA	good	very good	2232 / 2389
12	SA - CHCA	worse than 11	worse than 11, sharp peaks	2318 / 2389
13	SA - CHCA	good	very good	2460 / 2746

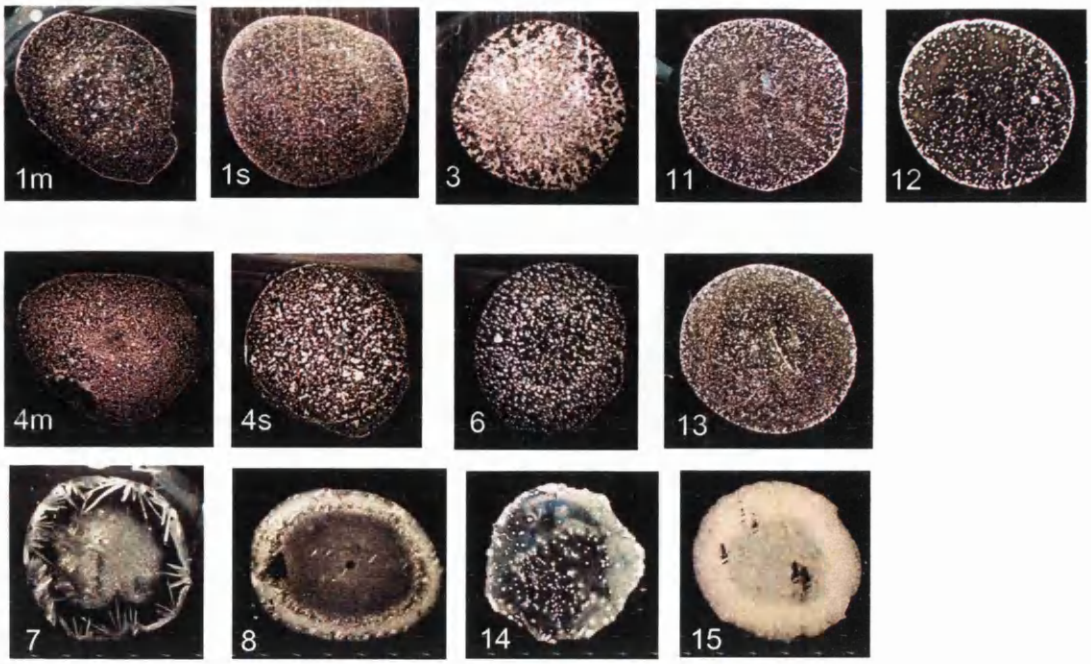


Figure 5.26: Digital photographs of MALDI matrices mixed with LMW serum proteins and peptides. Top row, SA; middle: CHCA; and bottom row: DHB matrices and mixtures.

To investigate the different matrices further, the signal-to-noise ratio (S/N) of a selection of protein peaks was calculated using Data Explorer™ (Applied Biosystems). To obtain the S/N ratio, a peak was highlighted and the noise level was calculated automatically from an area in the spectrum that has a flat baseline and no obvious peaks. The results are shown in Table 5.5 and Figure 5.27. In the low mass region, matrix 11 (SA-αCHCA) produced a good clean spectrum, similar to those seen with αCHCA alone (Figure 5.27). However, the S/N ratio was better for matrix 11 than for αCHCA alone. Additionally, it was observed that, for the mixture, the peak intensity increased the longer the laser was applied; which was not observed in any of the other matrices. This is important when accumulating a large number of shots in automated fashion, to retain a reproducible mass spectrum. This mixture also performed well with proteins of higher molecular weight, however SA alone mixed before application gave a more homogeneous spot and the S/N ratio of the peaks in the high molecular weight were higher than the other matrices (Table 5.5). According to these results, for analysis of LMW proteins recovered from 50 kDa MWCO filters, where more high molecular weight peaks are present, SA appears to be more suited. Although proteins < 3000 Da had lower S/N ratios using SA, protein peaks with a *m/z* greater than 5000 appeared to ionize better using SA. The peak intensity was kept

relatively similar for all spectra acquired for direct comparison, however, it was not possible to obtain a good peak intensity for the <5000 Da peaks using SA. As this is conflicting with the results obtained for spot homogeny and S/N ratios, a mix of αCHCA and SA may be the right choice. The mixture performed well with small and medium sized proteins. For the mixture the .bic file settings in the Voyager software were set to the default for αCHCA and SA to test the best conditions. The αCHCA setting produces sharper peaks and better S/N ratios than the SA setting, this could be expected, as the spectra look more similar to the spectra produced by αCHCA alone. In conclusion, SA-CHCA produced the best spectra for the higher mass range and should be used for future MALDI-ToF analysis.

	Signal-to-noise ratio								Signal-to-noise ratio					
	CHCA				DHB				SA					
Protein m/z	CHCA FA	5% CHCA - fucose	CHCA mix 1:1	CHCA sandwich	DHB 20mg/ml	DHB-CHCA	DHB-CHCA-FUCOSE	DHB-FUCOSE	SA 10mg/ml mix	SA - fucose	SA sandwich	SA-CHCA 1:1 chca settings	SA-CHCA 1:2 chca settings	SA-CHCA 2:1 chca settings
2322.4	53.7	55.7	67.2	75	65.6	88.6	20.4	59.6	4.1	1.7	1.8	112.7	172.8	16.5
2685.8	9.4	20.5	14.2	9.9	8.1	18.5	4.6	7	4.5	9.9	5.8	27.5	14.9	25
3079.9	70.7	142.6	97.7	89.9	54.2	128.3	4.1	17.8	4.4	5.6	5.8	143.1	173.3	164.7
3790.8	16.4	24.8	2.5	8.3	8.7	8.5	2.2	5.8	30.5	50	9.5	30.5	10.7	14.7
4460.8	17.9	25	15.7	11.6	7.2	11.2	1.3	5.7	7.2	17.9	3.8	22.8	6.8	23.9
5193.2	245.1	283.5	227.9	338.1	91.8	210.8	3.1	7.4	190.9	84	118.2	244.1	339.1	174.4
6600.8	61.8	156.3	36.3	112	22.1	134.2	4.2	10.1	165.9	53.3	60.8	174.7	106.3	144.6
8962.9	29.2	10.6	18	25.8	16.8	35.9	4.2	11.7	39.7	11.8	8.2	32.1	15.5	22
9178.4	32.3	46.1	39.9	53.9	7.3	55.2	4.6	6	55.2	28.5	27.8	67.6	48.5	77.3
10686.3	58.9	98.9	45.5	75.7	43.4	95.8	4.7	7.8	140.2	64.5	63.7	97.3	67.4	87.5
13932.9	5.2	3.9	6.5	3.8	13.3	19.6	4	4.9	33.4	18.2	17.2	7.1	15.5	6.6

Table 5.5: The signal-to-noise ratio (S/N) for a number of protein peaks for each matrix and mixture. The highest S/N for each peak is highlighted in green.

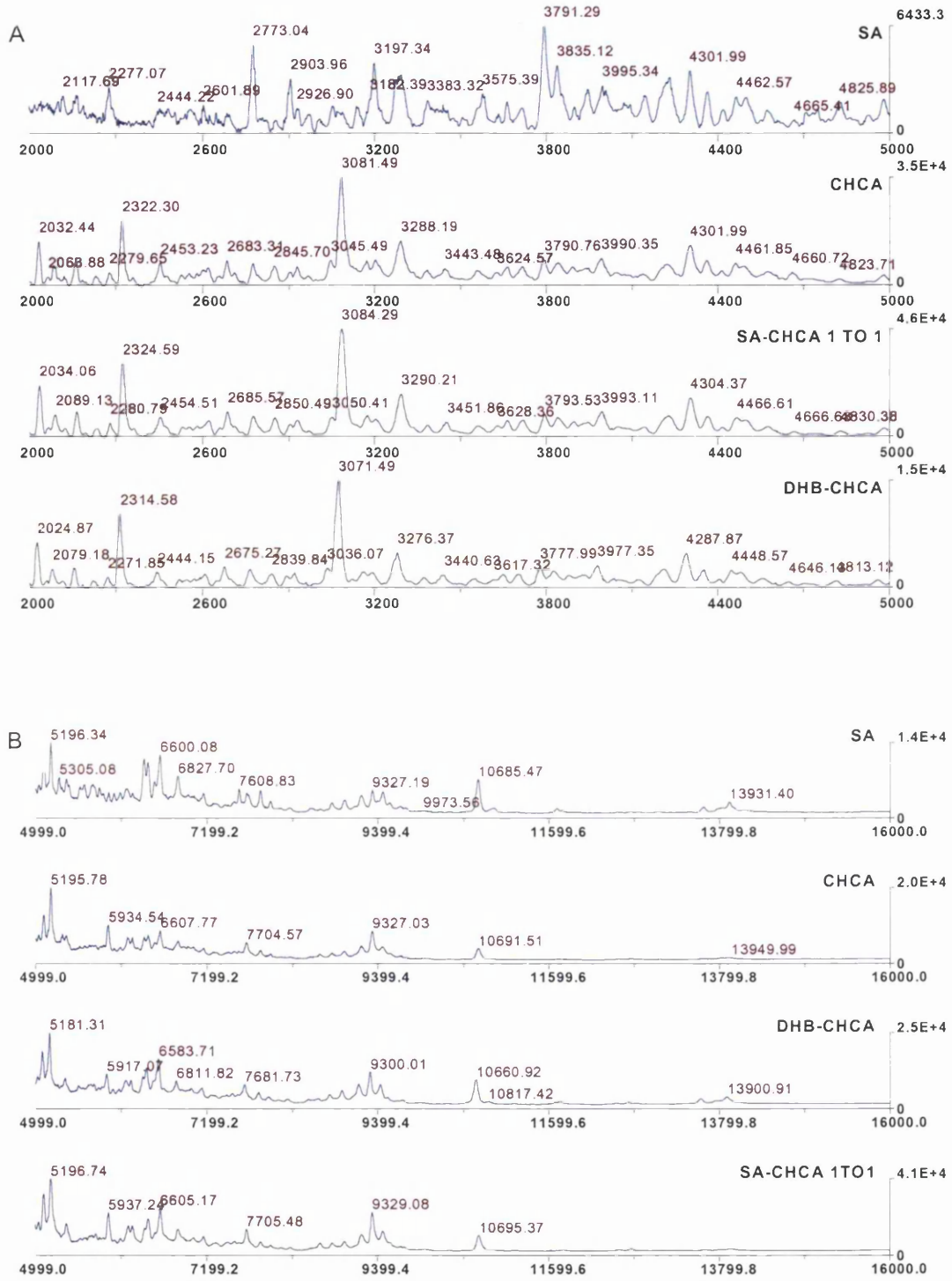


Figure 5.27: MALDI-ToF MS spectra of LMW serum with different matrices chosen according to their signal-to-noise ratio from Table 5.5. In the upper panel (**A**), SA, CHCA, SA-αCHCA and DHB-αCHCA are shown for the mass range *m/z* 2000 – 5000 Da. In the lower panel (**B**), for the mass range *m/z* 5000 – 16000 Da, SA, αCHCA, DHB-αCHCA, and SA-αCHCA are shown.

5.5. Protein Profiling on Sample Set S2

In light of the sample preparation improvements made above, we carried out a second experiment (S2), using a lower sample concentration with 20 μg of protein, and the sample was mixed in the tube with αCHCA-SA matrix mixture. Due to limited sample availability, in the second sample set, two different cancer and 5 different control serum samples were used compared to sample set S1 (Table 5.6).

Table 5.6: Samples used in each of the two experiments, S1 and S2. Not enough serum was available to repeat the analysis using exactly the same samples.

	S1	S2		S1	S2
Cancer	OP1	OP1	*Control*	V5	V5
Cancer	OP2	OP2	*Control*	V6	V6
Cancer	OP3	OP3	*Control*	V9	-
Cancer	OP4	-	*Control*	-	V10
Cancer	OP5	OP5	*Control*	V11	-
Cancer	OP6	-	*Control*	V13	-
Cancer	OP7	OP7	*Control*	V15	V15
Cancer	OP9	OP9	*Control*	-	V17
Cancer	-	OP10	*Control*	-	V19
Cancer	-	OP11	*Control*	-	V20
			Control	V22	-
			Control	V25	V25

It is also noteworthy that the samples were analysed using a slightly different UF protocol to incorporate the optimisations described above. The differences in sample preparation between sample sets S1 and S2 are outlined in Figure 5.28 and the protocol is described as the optimised UF protocol in Chapter 3 (section 3.3.2). Briefly, the serum samples were diluted further to avoid blockage of the membrane, a 50 kDa MWCO was used at a lower centrifugation speed (750 xg) to allow more proteins to be recovered and finally the HMW retentate was re-suspended and filtered again to increase the recovery of proteins normally found bound to HMW carrier proteins. Using 50 kDa MWCO filters resulted in the presence of peaks of a higher molecular weight in S2 than in S1. To gain more confidence in the results, the UF process was performed in triplicate and because Zip-Tips have a low binding capacity, 20 μg of LMW protein is sufficient for analysis and was used in this sample set. Although these changes do not allow a direct comparison of the two experiments, it

was considered important to show all the results especially since many of the differential protein peaks were retrieved from both experiments. To allow the analysis of proteins >5000 Da, minor differences in the MS analysis were made, these are shown in Table 5.7. Additionally the spectra were externally calibrated using Cal Mix 3 (insulin *m/z* 5734.59, thioredoxin *m/z* 11674.48, apomyoglobin *m/z* 16952.56) during automatic acquisition. Spectra processing and data analysis were the same as described for sample set S1.

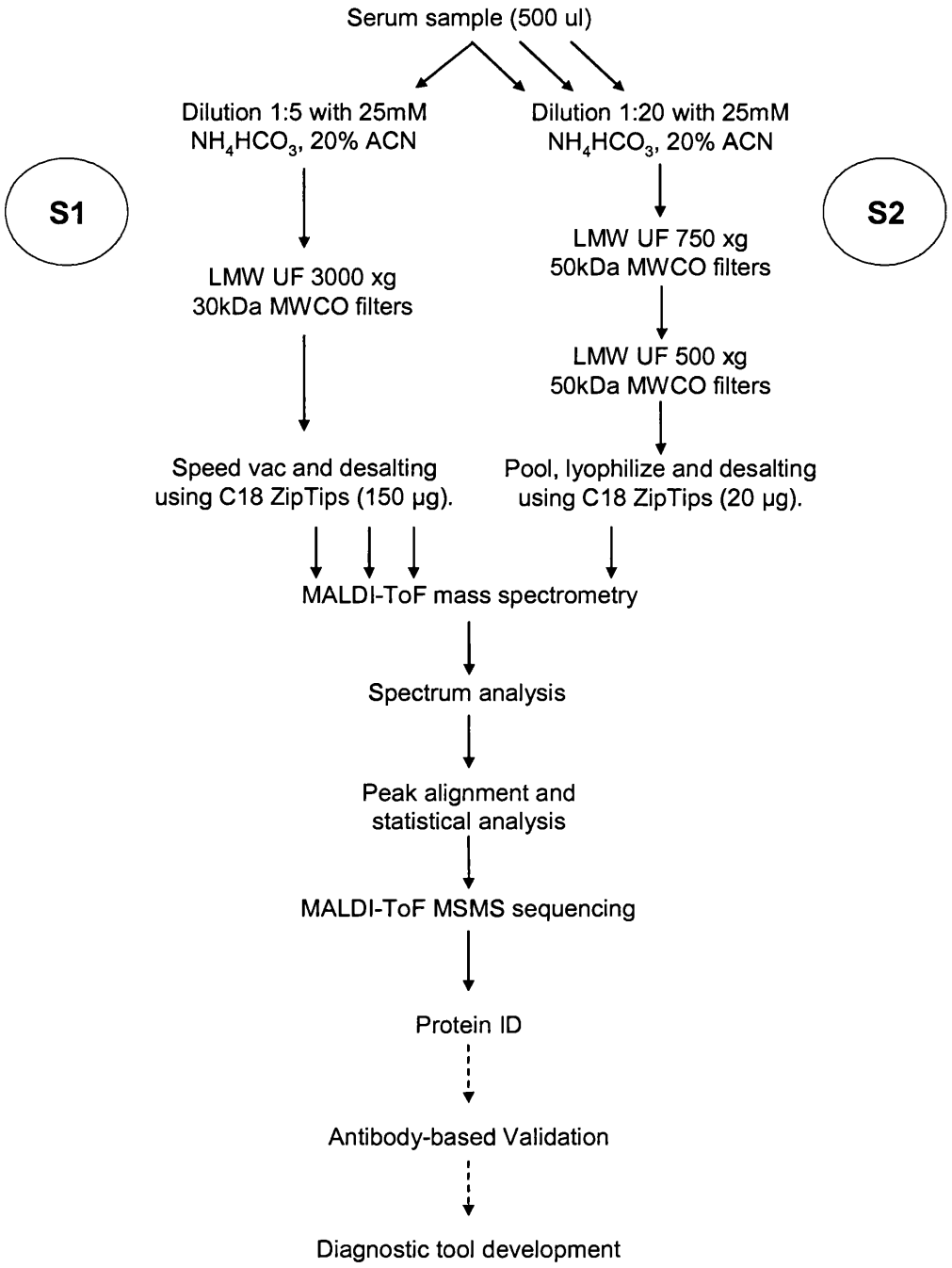


Figure 5.28: Experiment flow chart for protein profiling of LMW serum samples. Serum samples were prepared by UF and LMW serum proteins were profiled by MALDI-ToF MS. Mass spectra were manually calibrated, baseline-corrected, noise was removed and the spectra were smoothed before the centroid peak intensities were exported. For profiling, the peaks were standardised by total ion signal and peak intensity ratios of the averages of cancer and control samples were compared as well as *t*-tests carried out on individual peak intensities. Target proteins were then identified by MALDI-ToF MS/MS. Upon identification, antibodies could be developed for validation of the biomarkers.

Table 5.7: Additional MALDI-ToF instrument settings to those described in the materials and method (section 2.8.1), these settings for automated acquisition of spectra were used for S2, compared to S1. The settings were slightly different to allow MS analysis of *m/z* peaks up to 20 kDa recovered from UF using 50 kDa MWCO membranes.

	S1	S2
Acquisition mass range	1000 - 7000 Da	1000 - 20000 Da
Laser intensity	2400 - 2600	2400 - 2600
Calibration matrix	α-CHCA	SA and α-CHCA mix
Low mass gate	700 Da	1000 Da
Min intensity	10000	30000
Max intensity	55000	50000

The majority of the cancer samples were the same between S1 and S2. As before in S1, the control samples were chosen to create a match from a volunteer of a similar age. The median age of both groups was much closer in this second sample set, compared to S1, an extra reason why the two experiments could not be directly compared. The age distribution of both breast cancer patients and controls for S2 can be seen in Figure 5.29.

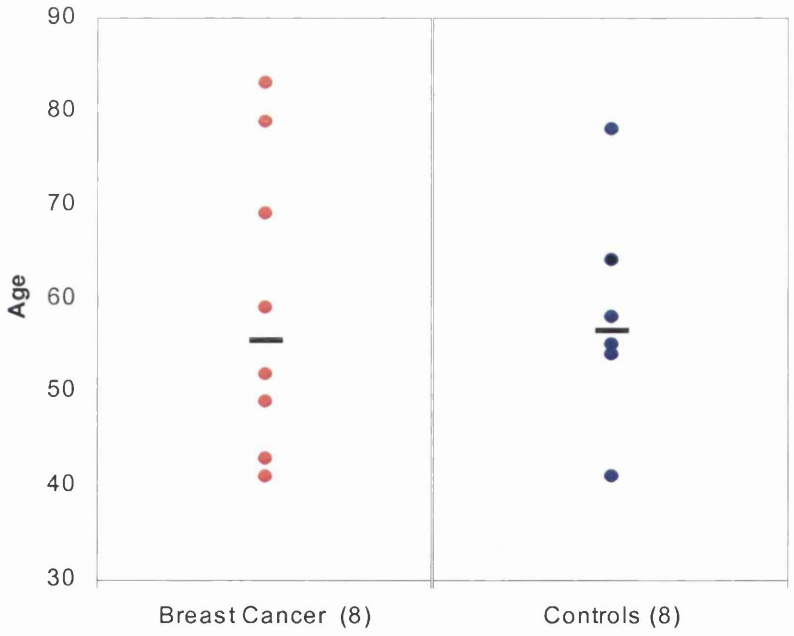


Figure 5.29: Age distribution of healthy controls and breast cancer patients in experiment S2. Equal numbers of patients and controls were used from each cohort and were individually age-matched. Black horizontal lines represent medians for each group.

The MALDI-ToF MS spectra for each sample from S2 are shown in the Appendix (D). An example of one triplicate mass spectrum is shown in Figure 5.30. Despite some variation in the peak intensities, the spectra are comparable and show good reproducibility. Spectra that were different from the majority of the spectra and especially from the other replicates of the same serum sample were identified as outliers and 5 spectra were removed from the analysis. In addition, in the case of V15a shown in Figure 5.31, the calculated NF of this spectrum was more than twice the average NF, due to an overall lower spectrum intensity (1.4×10^4 compared to 3.0×10^4) and was therefore removed from further analysis. The other spectra and replicates showed good mass accuracy after manual calibration and reproducible peak intensity between the replicates of each serum sample.

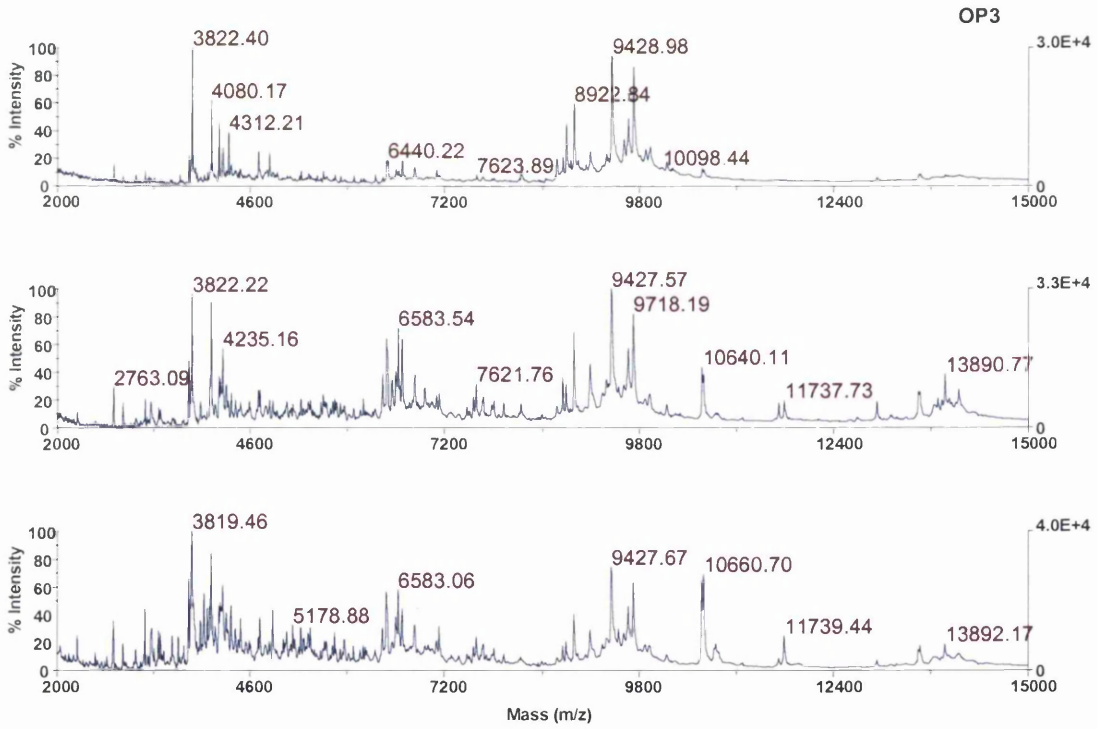


Figure 5.30: An example of the reproducibility across three filtrates from the same serum sample. Each serum sample in S2 was prepared and analysed in triplicate.

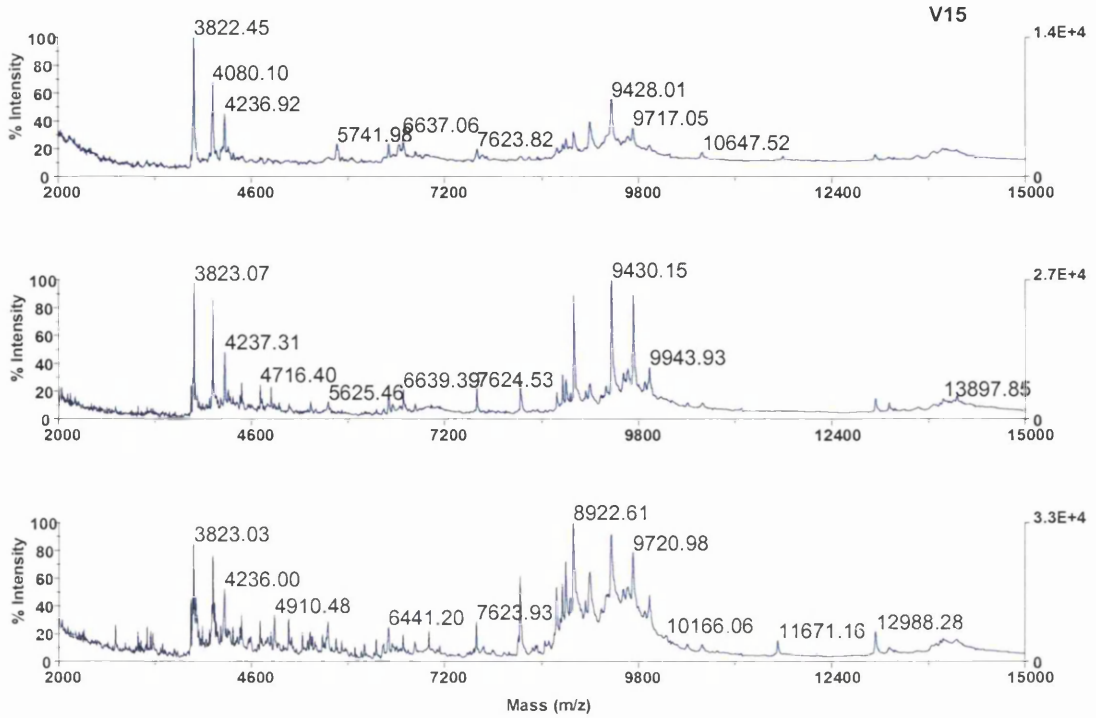


Figure 5.31: Outlier removal. During normalisation, V15a showed up as an outlier, as its normalisation factor (NF) was more than twice the average NF. Hence this spectrum was removed from subsequent analysis.

Consistent with S1, the mass accuracy of the individual peaks across all replicates was found to increase with mass (Figure 5.32); this was even more pronounced for this sample set than for S1. This could be due the manual calibration peaks used (*m/z* 3822, 8920, 13552). In the case of S2, this can be further explained by the use of Cal Mix 3 for external calibration. The smallest peak in that calibrant is insulin at *m/z* 5734.59. Due to this difference, staggered mass tolerances were used during alignment in *mz*Align: <2000 Da, 5000 ppm; 2000-5000 Da, 3300 ppm; and for peaks greater than 5000 Da, 1300 ppm. This greater mass tolerance for the low masses dealt effectively with the low mass accuracy.

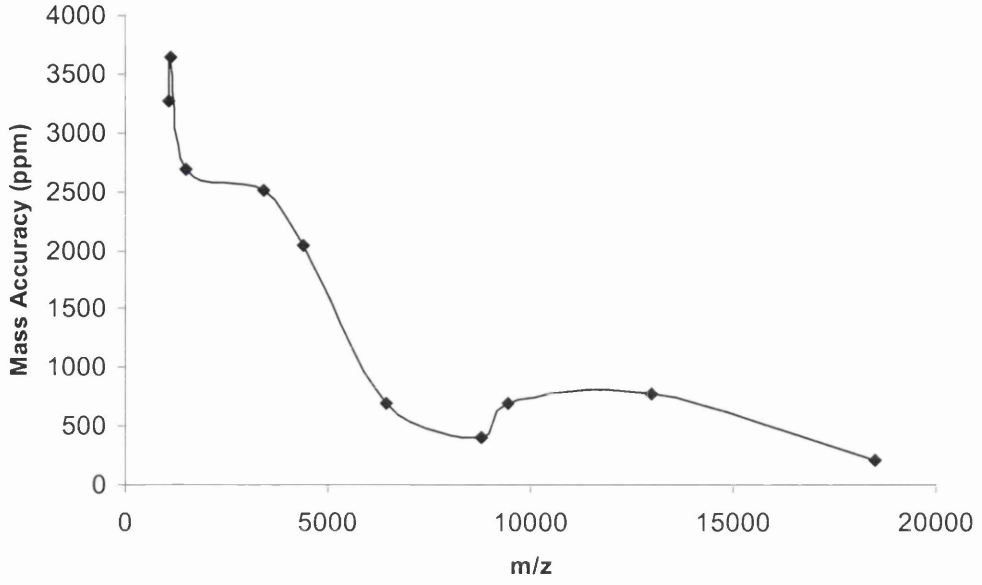


Figure 5.32: Correlation of mass accuracy against *m/z* of individual mass peaks. The mass accuracy again appears to increase with increasing mass.

5.5.1. Biomarkers Discovered in Sample Set S2

Very similarly as for S1, PCA analysis revealed two separate groups with the control and breast cancer samples in separate clusters. Additionally the replicates from each sample are tightly clustered together, as demonstrated by V17 and OP6 (Figure 5.33). OP7c appear to be a bit further away from the rest of the cluster: when inspecting the spectra for OP7 closer in Figure 5.34, it became obvious that the overall peak intensity of the spectrum for OP7c is higher than, for example, OP7b or the spectra from other samples. However it is not significantly different to be flagged up in the outlier identification when calculating the normalisation factors. Nevertheless this shows the usefulness of the PCA analysis in highlighting similarities within the sample groups.

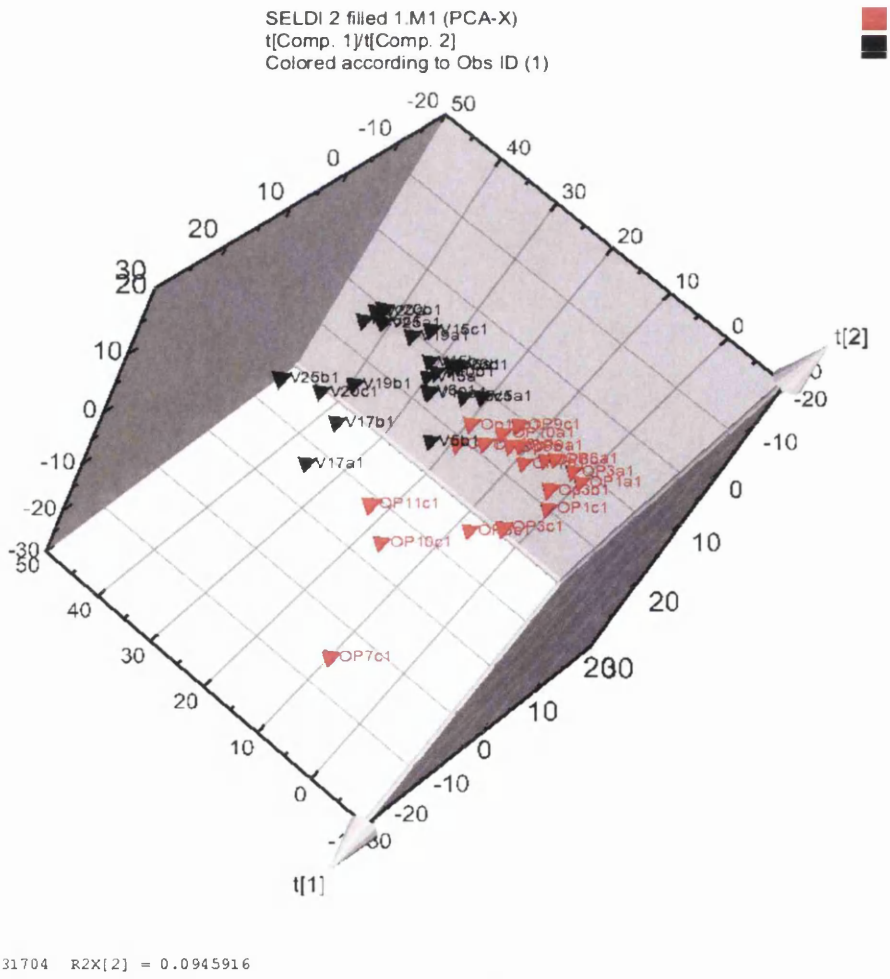


Figure 5.33: Un-supervised principal components analysis (PCA) of MS-based serum protein profiling data derived from healthy controls and metastatic breast cancer patients in S2.

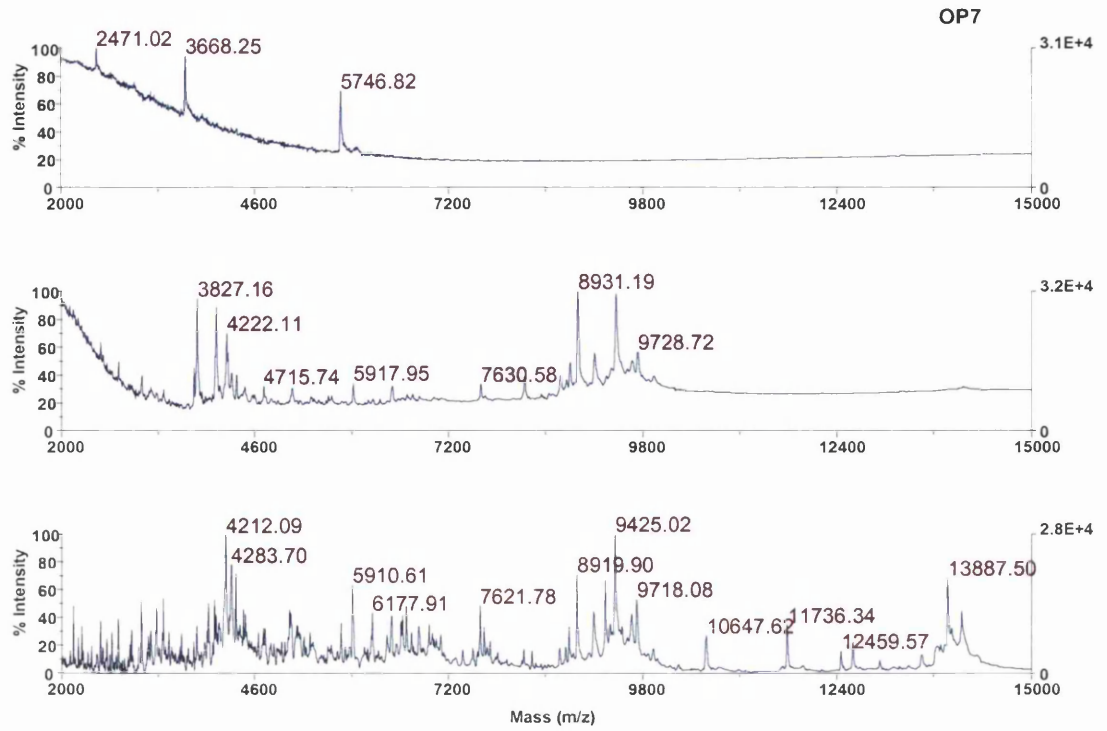


Figure 5.34: MALDI-ToF MS spectra for the three replicates of OP7. The spectrum of OP7a was removed from data analysis since the spectrum is almost empty. Furthermore OP7c appeared to be distant from the other breast cancer in the PCA plot. The spectrum also appeared slightly different from the other samples; however it was not excluded from the sample set.

Using *mz*Align, 505 *m/z* values were aligned in one table, compared to 452 *m/z* values in Markerview. The similarity is not quite as tight as for S1; however in Markerview we were not able to stagger the mass tolerance during import. Nevertheless, the results from both alignment tools are very similar.

After alignment in *mz*Align and Markerview, the *p*-values and fold-changes for each peak in the spectra were calculated in Excel. Peaks with significant *p*-values can be seen in Table 5.8; the table layout is the same as for S1. All discriminant peaks are again shown from the averaged and non-averaged data, also shown, for further confirmation of the results are the results from the log-normalised data. Finally in the last 3 columns the results from the data aligned in Markerview are shown. Ten peaks were significantly different between the two cohorts calculated from the averaged data (better seen in Table 5.9), 5 of those (*m/z* 1184, 2018, 2730, 8771 and 9647) were also found to be significant after normalising the data. These peaks could be considered as more robust.

Table 5.8: Significantly different intensity values of m/z peaks between breast cancer serum samples and control samples. *p*-values were calculated after standardisation with 1. VBA peak alignment in Excel; 2. Markerview peak alignment and t-test in Excel and 3. Alignment and *t*-test in Markerview. The average was also calculated in Excel. The peak numbers in red are those which were detected in less than 25% of samples through alignment.

	Alignment with VBA, t-test with excel								Alignment with Markerview, t-test with excel								t-test with MV	
Centroid mass	n cancer (19)	n control (21)	p-value VBA all	fold-change	p-value average	fold-change	p-value log averages	fold-change	n cancer (19)	n control (21)	p-value MV all	fold-change	p-value averages	fold-change	p-value MV all	fold-change		
1064									10	8	0.016	**4.5**	0.031	**8.1**	0.019	**4.5**		
1184	2	8			0.027	**-6.5**	0.016	**-5.2**										
1225	3	8					0.020	**-3.0**	9	7	0.043	**-2.0**			0.034	**-2.8**		
1272	10	10	0.017	**-2.3**	0.006	**-2.8**			4	10	0.039	**-2.8**			0.034	**-2.8**		
1310									14	12	0.044	-1.6						
1351	15	19					0.046	1.0										
1383									8	5	0.006	**-3.7**						
1495	8	6					0.037	**2.0**										
1562									3	4	0.000	**3.4**						
1603	3	5	0.028	**2.1**														
1749	14	7					0.035	**2.0**										
1838	1	7			0.042	**-9.4**											0.007	**C down**
1932	9	14	0.018	-1.8	0.049	**-2.0**			0	8								
2018	5	6			0.049	**3.5**							0.043	**2.7**				
2397	7	5			0.021	**6.3**	0.015	**2.2**	9	7			0.020	**5.2**	0.039	**5.2**		
2461									6	2								
2535	4	11					0.018	**-3.0**	4	6	0.002	**3.2**						
2714	3	5	0.044	**1.9**					0	4							0.038	**C up**
2730	6	2			0.003	**8.1**	0.002	**6.9**	8	12			0.029	**-3.1**				
2823	0	5			**0.014**	-15985.80	**0.011**	-275.31	6	5			0.023	**3.3**				
2832	8	7			0.029	3.2			4	2	0.003	**-8.4**						
2892									8	3	0.033	**2.1**						
3020									3	2								
3278									6	11			0.013	**-7.2**				
3293									2	7								
3526	4	0					**0.030**	267.49										
3356	6	2					0.019	3.3										
3534	7	9	0.029	**2.7**														
3648	6	6	0.018	**2.0**					2	3	0.008	**6.3**						
4492	2	6	0.032	**2.4**														
4688									6	11	0.023	**-2.6**			0.035	**-2.6**		
5051	9	6	0.011	**3.4**														
5639	0	5					**0.033**	-238.68										
5923	9	5					0.022	**2.1**										
5941									0	6							0.031	**C down**
6565	5	4	0.004	**-4.5**														
6662	8	8	0.019	**-1.9**														
6942	9	11	0.038	**2.0**														
7009	2	3	0.005	**3.2**														
7679	8	4					0.039	**2.0**										
7786	2	3	0.005	**3.9**														
8210	16	16	0.032	**-3.3**														
8771	19	21	0.000	**-1.9**	0.020	**-1.9**	0.017	-1.1	14	16	0.004	**-2.0**	0.015	**-2.2**	0.019	**-2.0**		
9368									15	14	0.015	**1.9**	0.018	**1.9**	0.005	**2.2**		
9647	19	21	0.006	1.6	0.026	1.6	0.012	1.0	15	16	0.020	1.5	0.027	1.6				
10341									0	5					0.046	**C down**		
13550									13	11	0.047	**-2.2**						
18971									2	2	0.024	**-2.9**						

Table 5.9: Statistical analysis of discriminating peaks derived from the average peak intensities from serum protein profiling of breast cancer patients and healthy controls in S2. The first *t*-test was performed on the data aligned in VBA and the second using data that was aligned using Markerview. Both *t*-tests were done in Excel. Peaks that were discovered in less than 25% of spectra are marked in red.

			Alignment VBA		Alignment MV	
Centroid mass	n cancer (19)	n control (21)	p-value average	fold-change	p-value averages	fold-change
1064	10	8			0.031	**8.1**
1184	2	8	0.027	**-6.5**		
1272	10	10	0.006	**-2.8**		
1838	1	7	0.042	**-9.4**		
1932	9	14	0.049	**-2.0**		
2018	5	6	0.049	**3.5**		
2397	7	5	0.021	**6.3**	0.043	**2.7**
2461	6	2			0.020	**5.2**
2730	6	2	0.003	**8.1**	0.023	**3.3**
2823	0	5	0.014	**BC down**		
2832	8	7	0.029	**3.2**		
2892	8	12			0.029	**-3.1**
3293	2	7			0.013	**-7.2**
8771	19	21	0.020	**-1.9**	0.015	**-2.2**
9368	15	14			0.018	**1.9**
9647	19	21	0.026	1.6	0.027	1.6

Furthermore these peaks also had significant *p*-values after performing a *t*-test on the Markerview-aligned data. As discovered during the analysis of S1, visual inspection of the discriminant peaks can prove more valuable than the statistics and therefore spectra for each peak from the averaged data are shown below.

In total, 15 "markers" were discovered from the averaged data, 10 aligned with *mz*Align and 9 with Markerview. Of these, four "markers" were discovered from both datasets. Variation in fold-change results of the VBA and Markerview aligned data can be explained by the difference in peak alignment. The *m/z* values that had many peaks aligned to, such as *m/z* 8771 and 9647 show similar fold-change results for data from both alignments. Shown in order of molecular weight, each differential peak was visually inspected for signal-to-noise ratio and whether there is a true difference visible between the two sample cohorts (Figure 5.34 to Figure 5.39). Of the 15 peaks with significant *p*-values, five peaks were actually large enough to see (*m/z* 1184, 2730, 2832, 8771 and 9647).

As seen in Figure 5.35, the breast cancer spectra showed no peaks for *m/z* 1184 whereas at least 6 samples have a visible peak in the control group Markerview

alignment has also retrieved a significant *p*-value for peak at *m/z* 1064, however due to the poor S/N ratio this cannot be identified as a real peak although the Markerview visualisation looks convincing (Figure 5.36). This was disappointing because m/z 1064 was convincingly different in the S1 results. The reason for this is in part that the S2 spectra were shot with too high levels of laser intensity; however repetition was not possible, because as the samples were re-analysed, it became apparent that they had degraded and the original MALDI target plate had not been preserved.

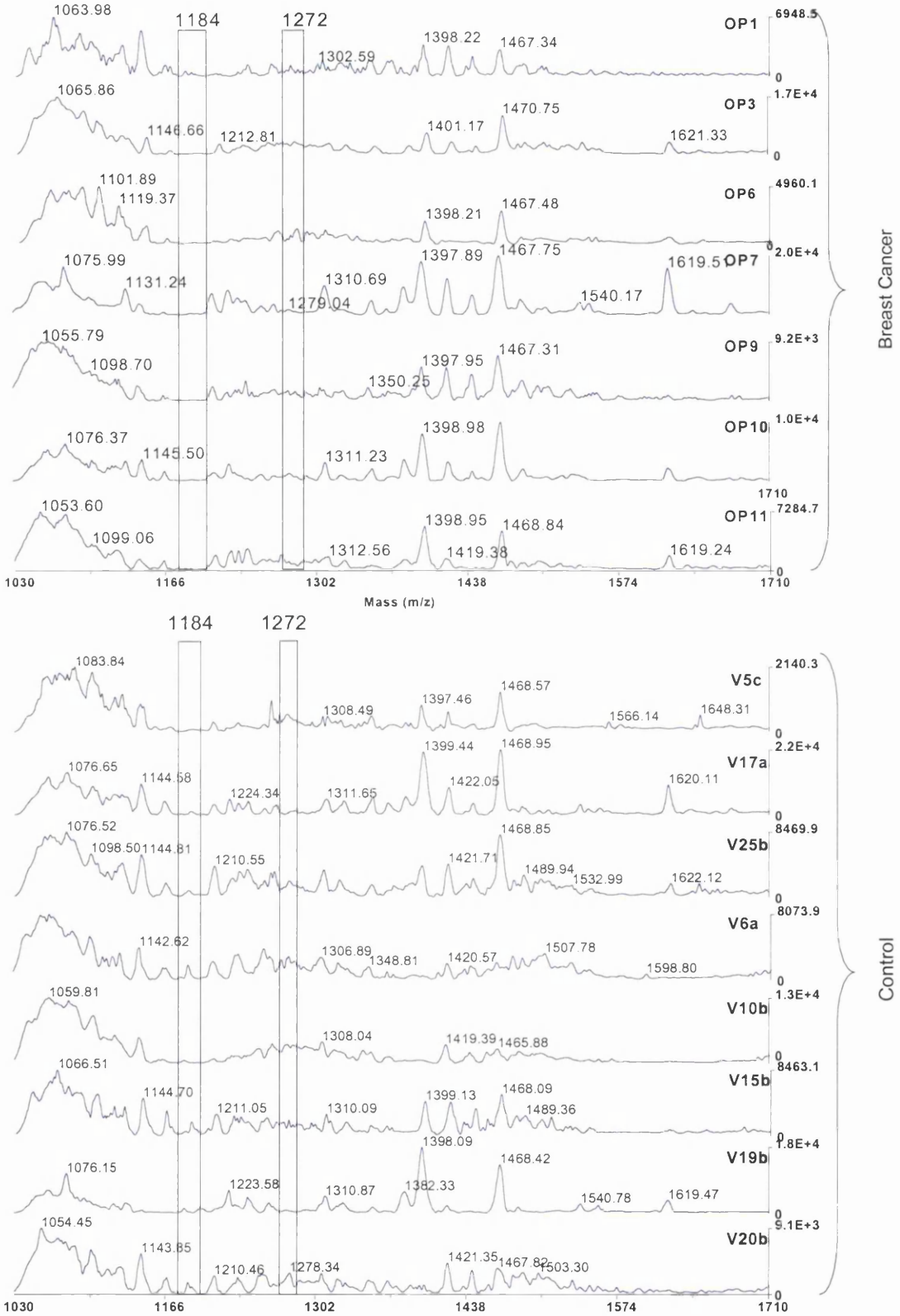
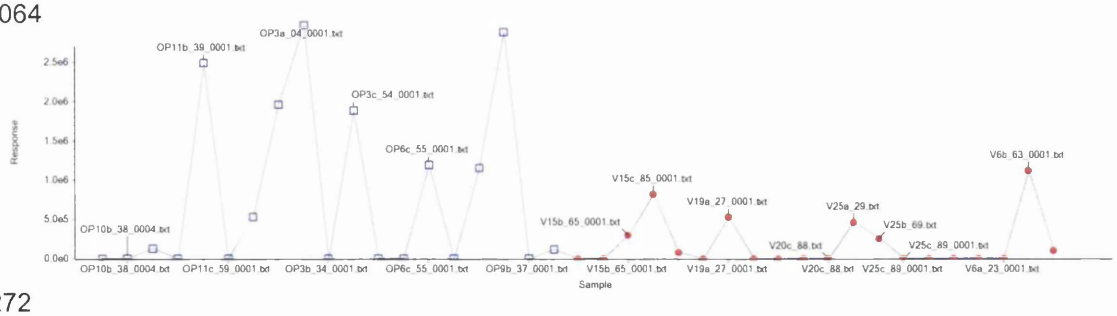


Figure 5.35: MALDI-ToF spectra for the *m/z* range 1000-1700 Da, showing one replicate of each breast cancer sample in the upper panel and one of the control samples in the lower panel. Peaks with significant *p*-values are shown boxed.

m/z 1064



m/z 1272

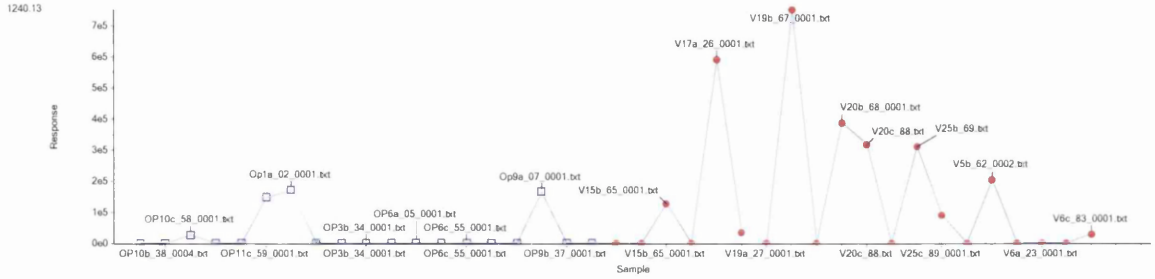


Figure 5.36: Markerview visualisation of *m/z* 1064 and 1272. Peaks from all samples and replicates were aligned using Markerview.

Both *m/z* values 2730 and 2832 have significant *p*-values but show peaks with low S/N ratios (Figure 5.37). No peaks are seen in the control samples for *m/z* 2730 or 2832. The Markerview visualisation also showed that relatively few peaks from across all the samples were detected. Nevertheless both (*m/z* 2737 and 2832) appear to having discriminating *p*-values.

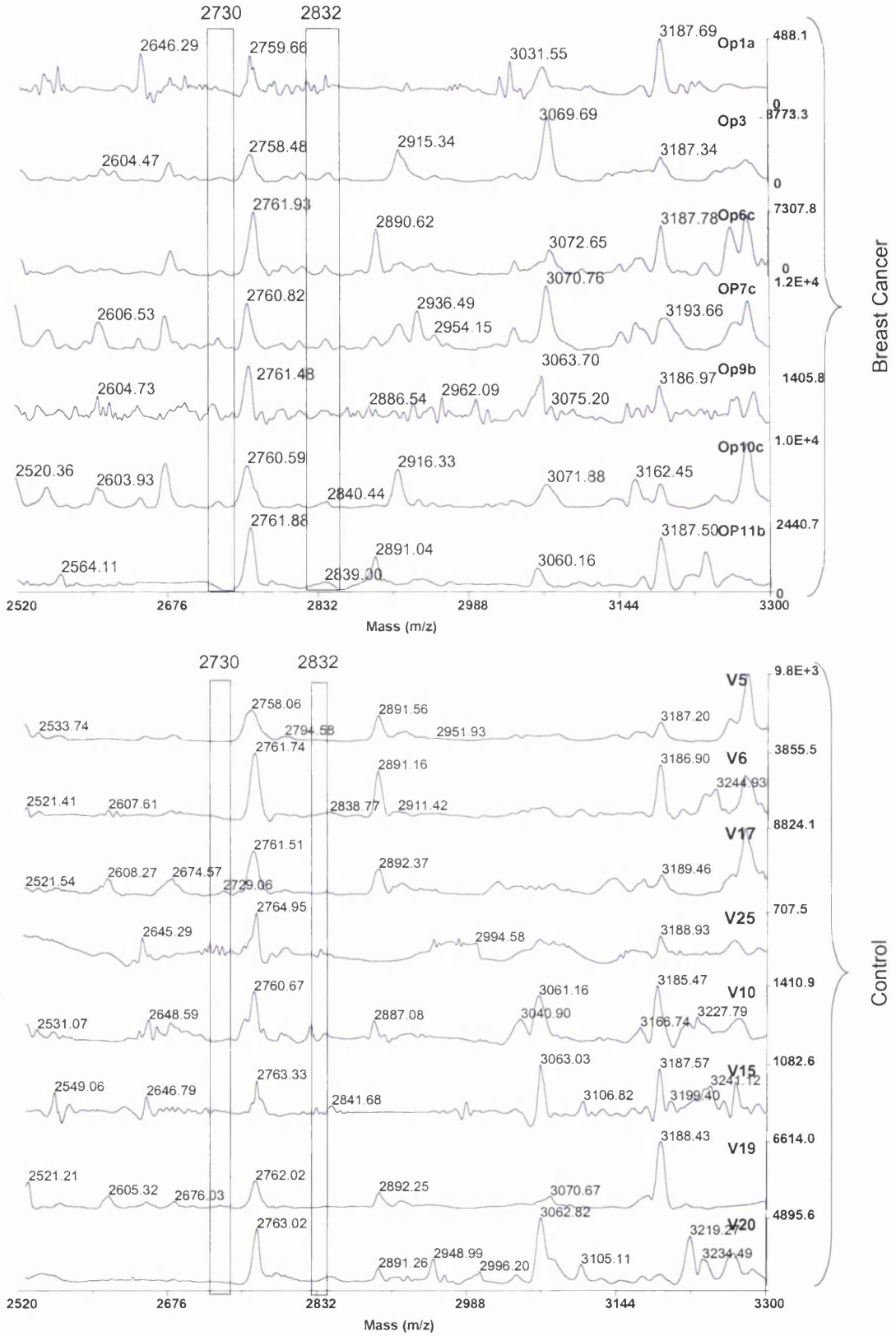


Figure 5.37: MALDI-ToF spectra for the *m/z* range 2500-3300 Da, showing one replicate of each breast cancer sample in the upper panel and one of the control samples in the lower panel. Peaks with significant *p*-values are shown boxed.

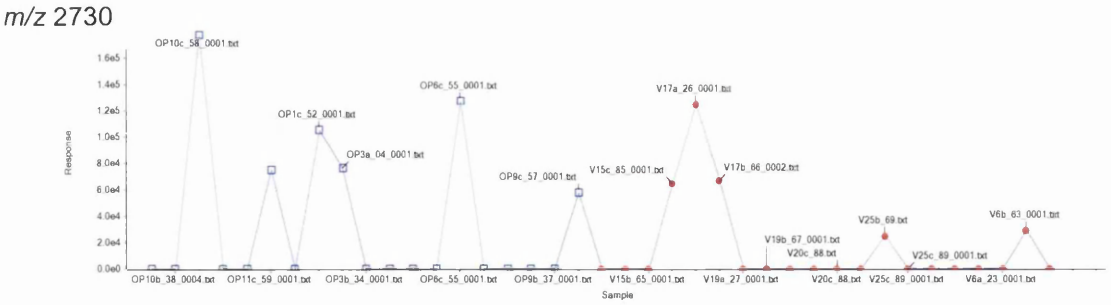


Figure 5.38: Markerview visualisation of *m/z* 2730. Peaks from all samples and replicates were aligned using Markerview.

Only three peaks (*m/z* 1184, 8771 and 9647) from this sample set were convincing enough, statistically and visually, to be considered as potential markers and to be taken forward for identification using tandem MS. Peak *m/z* 1184 was described above (Figure 5.35). Both, *m/z* 8771 and 9647, have good S/N ratios and *m/z* 8771 is consistently down-regulated in the breast cancer cohort whereas *m/z* 9647 has an increased peak intensity across all breast cancer samples compared to the control samples (Figure 5.39). The Markerview visualisation for either is not very compelling but a clear trend is visible (Figure 5.40). It is encouraging that the peaks were changed in opposite directions, whereas *m/z* 9427 remained unchanged between the breast cancer and the control samples.

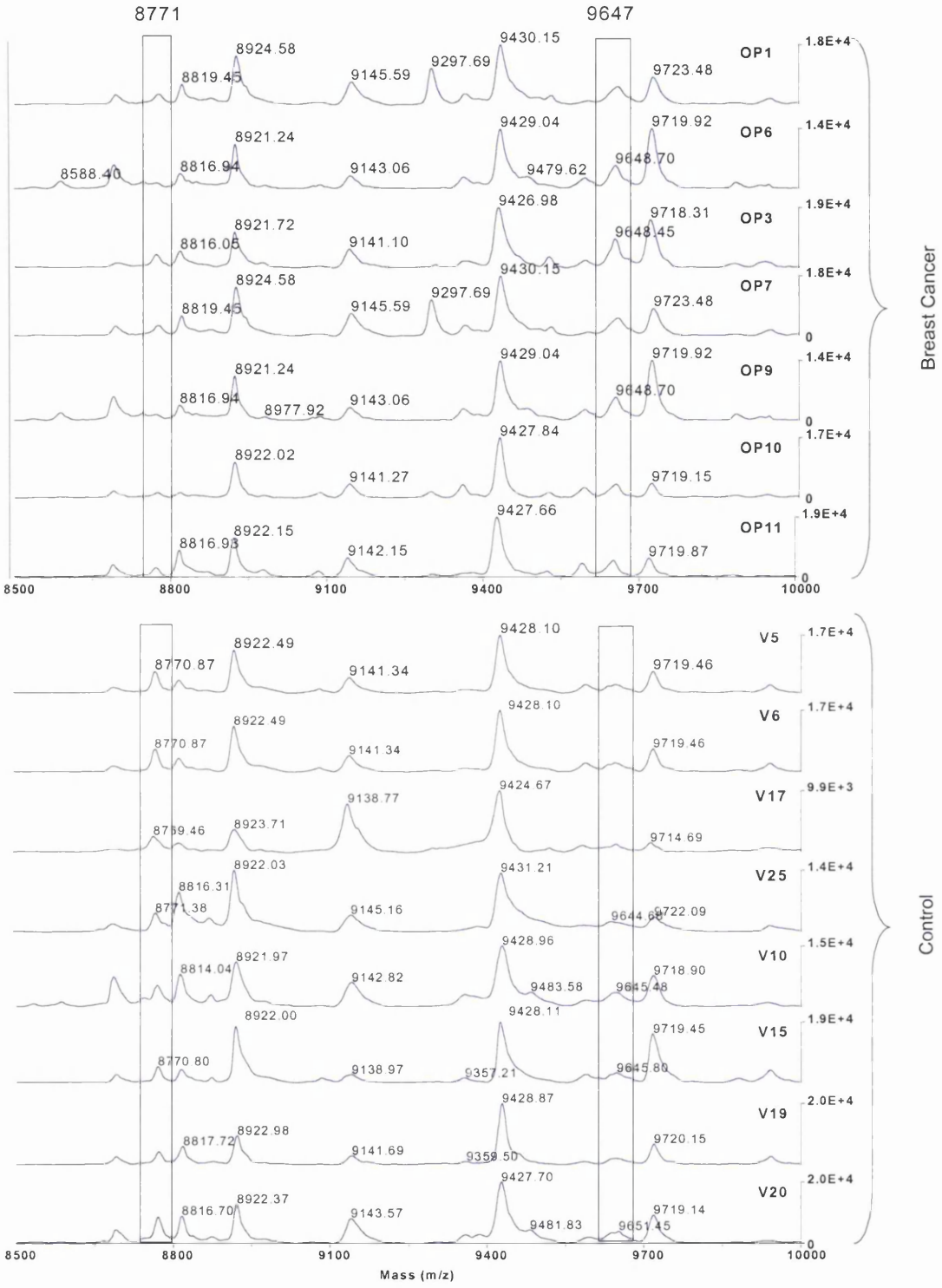
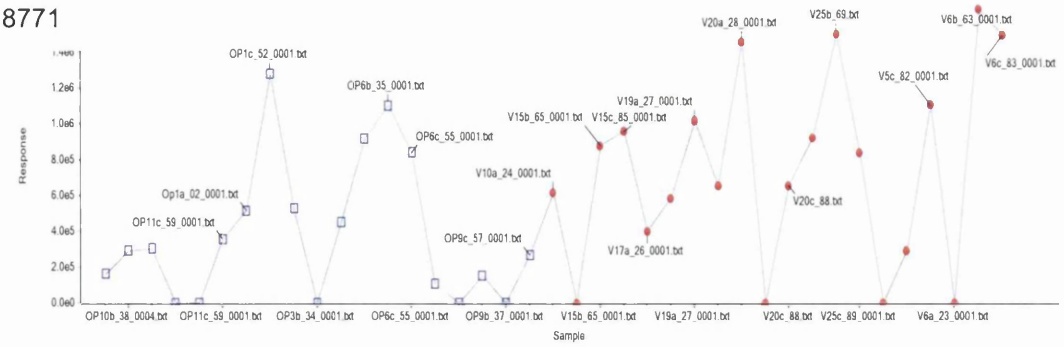


Figure 5.39: MALDI-ToF spectra for the *m/z* range 8500-10000 Da, showing one replicate of each breast cancer sample in the upper panel and one of the control samples in the lower panel. Peaks with significant *p*-values are shown boxed.

m/z 8771



m/z 9647

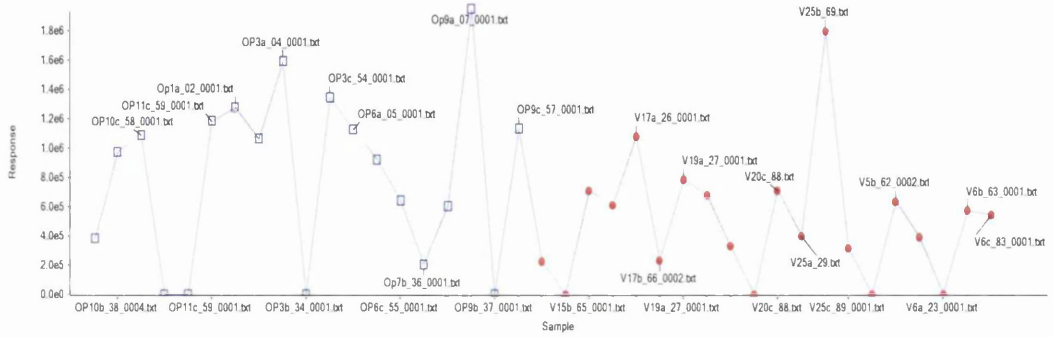


Figure 5.40: Markerview visualisation of *m/z* 8771 and 9647. Peaks from all samples and replicates were aligned using Markerview.

5.5.2. Tandem MS Analysis for Identification of Potential Markers

For marker identification, the peaks with statistically- and visually-significant *p*-values from S1 (*m/z* 1064, 1776, 2556, and 2995) and from S2 (*m/z* 1184, 8771 and 9647) were further analysed. Identification of mass peaks is possible by fragmentation within a mass spectrometer containing a ToF/ToF$^{\spadesuit}$. Here the selected precursor masses were held in the first mass analyser and fragmented using a collision gas before the fragment ions were released into a second ToF tube for detection of their *m/z* ratio. A collaboration with Matthias Glückmann from Applied Biosystems in Germany was set up and some of the samples from the S2 set were Zip-Tipped and spotted with αCHCA matrix on a MALDI target plate for tandem MS analysis. In Germany, a 4700 MALDI-ToF/ToF mass analyser (Applied Biosystems, Darmstadt, Germany) was used for fragment ion identification. The fragmentation spectra were searched against the peptide database using MASCOT (MatrixScience, UK). The significance of the results from MASCOT is graded depending on the number of fragment ions matched to predicted ions from the peptide matched. An ion score of >28 indicates peptides with significant homology and a ion score of >55 indicates identity or extensive homology (p <0.05). Unfortunately, only two (*m/z* 1062 and 2832) MS/MS spectra from the potential markers produced an ion score >28, indicating homology to a protein in the human database. However, one other precursor (*m/z* 1932) for a significantly different peak in the S2 set produced a fragmentation spectrum. Both markers, *m/z* 1932 and 2832, were determined to be significantly different from the averaged data, aligned by *mz*Align and Markerview. The peak for *m/z* 1064 was significantly up-regulated in the breast cancer samples, calculated from the MV aligned data. The fragmentation spectra are shown for *m/z* 1064 (ion score 36, 11/13 *b* and *y* ions matched) with homology to RPPGFSPFR from kininogen precursor (Figure 5.41), *m/z* 1932 (ion score 7, 8/32 *b* and *y* ions matched) with some homology to SLAELGGHLDQQVEEFR from apolipoprotein A-IV precursor (Figure 5.42), and the spectrum for *m/z* 2839 (ion score of 38, 16/52, *b* and *y* ions matched) identifying a peptide sequence with homology to AHYDLRHTFMGVVSLG-SPSGEVSHPR from alpha-2-HS-glycoprotein precursor (Figure 5.43). Although the peak intensity difference for *m/z* 2832 was not as obvious as the others, there was a noticeable increase in intensity in the breast cancer samples (Figure 5.37).

* Tandem MS was performed by our collaborators at Applied Biosystems (Darmstadt, Germany)

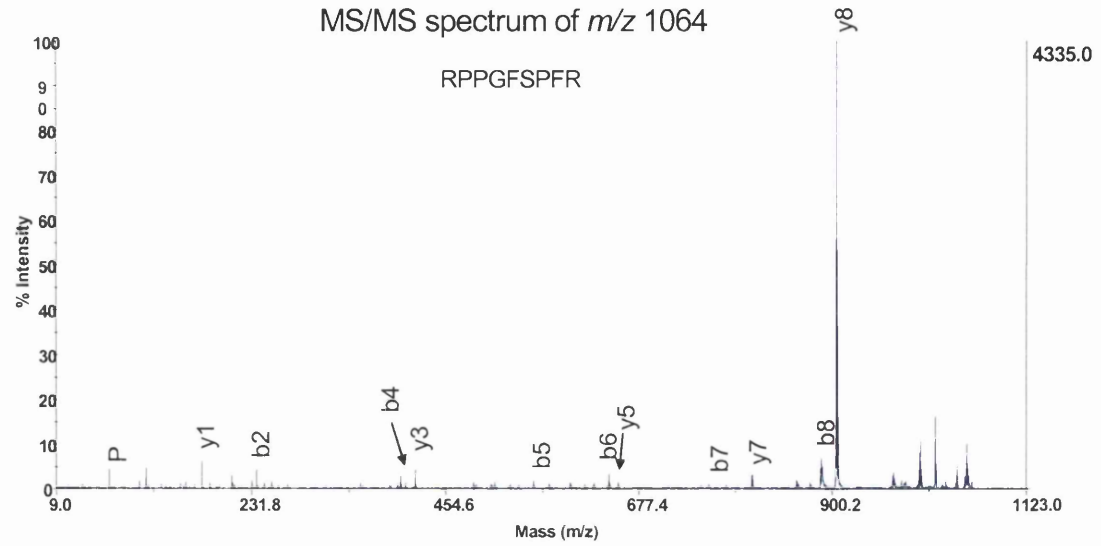


Figure 5.41: Tandem mass spectrum of *m/z* 1064 acquired using MALDI-ToF/ToF MS. Data base searching matched the peptide sequence RPPGFSPFR from kininogen precursor.

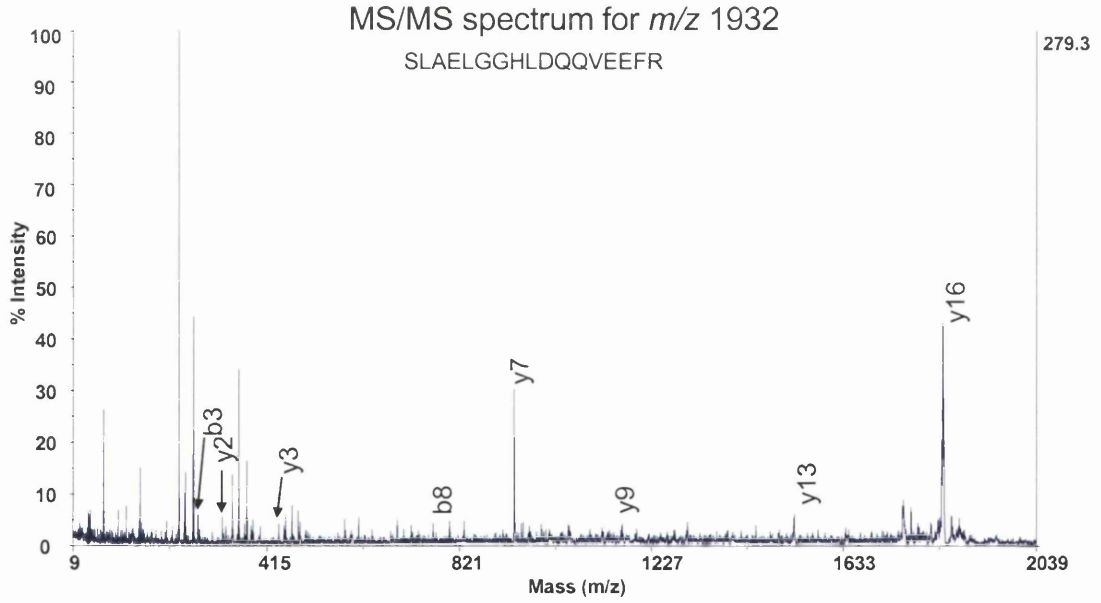


Figure 5.42: Tandem mass spectrum of *m/z* 1930 acquired using MALDI-ToF/ToF MS. Data base searching matched the peptide sequence SLAELGGHLDQQVEEFR from apolipoprotein A-IV precursor.

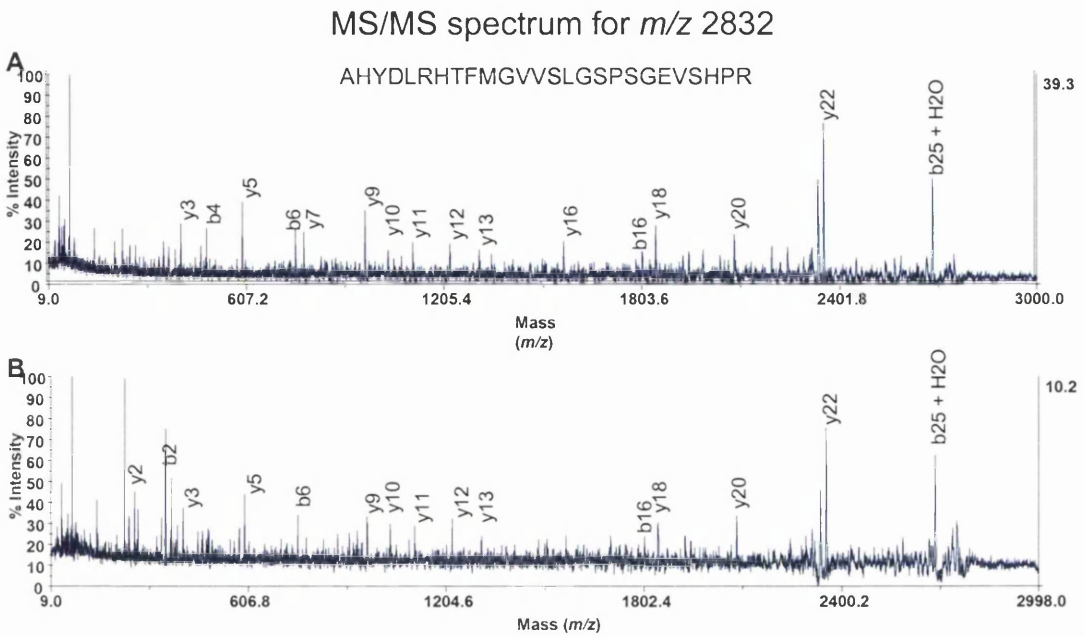


Figure 5.43: Tandem mass spectra of *m/z* 2832 acquired using MALDI-ToF/ToF MS. Data base searching matched the peptide sequence AHYDLRHTFMGVVSLGSPSGEVSHPR from alpha-2-HS-glycoprotein precursor. The precursor was fragmented and identified from two separate spectra (**A** and **B**).

More in depth investigation of the peptide matches revealed that *m/z* 1064 identified the amino acid sequence (from 381 to 389) coding for bradykinin. According to the UniProtKB/Swiss-Prot entry, bradykinin occurs in plasma and is released from kininogen by plasma kallikrein, and shows a variety of physiological effects: (1) influence in smooth muscle contraction, (2) induction of hypotension, (3) natriuresis and diuresis, (4) decrease in blood glucose level, (5) it is a mediator of inflammation and causes (6) increase in vascular permeability, (7) stimulation of nociceptors (8) release of other mediators of inflammation (e.g. prostaglandins), (9) it has a cardio protective effect (directly via bradykinin action, indirectly via endothelium-derived relaxing factor action). The peptide (*m/z* 2832) identified for the alpha-2-HS-glycoprotein precursor was matched to an amino acid sequence coding for the connecting peptide between the A and B chain of the protein. Alpha-2-HS-glycoprotein itself, secreted into plasma, promotes endocytosis, possesses opsonic properties and influences the mineral phase of bone. Furthermore alpha-2-HS-glycoprotein shows affinity for calcium and barium ions. The amino acid sequence contains 3 phosphoserine sites. Finally *m/z* 1932 was matched to an amino acid sequence that codes for repeat 12 within apolipoprotein A-IV. The protein itself is

described in UniProtKB/Swiss-Prot to be secreted into plasma, and responsible for lipid transport and removal of superoxide radicals. Although MASCOT appeared to label the majority of fragment ions, the spectra have relatively low intensity and identification from them is slightly unconvincing.

5.6. Comparison of the S1 and S2 Sample Markers

Very few discriminatory peaks (*m/z* 1064 and 1184) were found in common from the results of the averaged data of both experiments (S1 and S2). Therefore we have chosen to show all the results including *p*-values from un-averaged data and from peak intensities aligned using Markerview. Some protein peaks, *m/z* 1064, 1184, 1272, 1313, 1608, 1776, 1932, 2018, 2717, 2826, 3338, 4457 and 6962, were discriminatory in both S1 and S2. The *m/z* value is not identical, due to the use of different peaks in the calibration file. Furthermore using the larger MWCO membranes, different proteins were recovered in the filtrate. And finally, many of the differential peaks in S1 were of a MW <1700 Da, which was too small to produce good signal-to-noise ratio spectra in the S2 set (Figure 5.37).

5.7. Discussion and Conclusions

Global protein profiling for biomarker discovery from serum using MALDI-ToF MS was widely untested prior to this study. During the course of the analysis a couple of studies were published using MALDI-ToF MS for biomarker discovery [6, 8-10, 21]. Quantitation of intact proteins has been mainly performed using SELDI-ToF MS, which makes use of chemical surfaces to separate proteins; however the mass spectrometry is the same as MALDI-ToF. We hypothesised that MALDI-ToF should therefore be able to perform quantitative analysis of proteins as well. Using the LMW serum fraction reduced the complexity of serum sufficiently to avoid ion suppression. The present type of MALDI-ToF MS instrument can cause irreproducibility of the peak intensity due to the presence of hotspots in the matrix [14]. Furthermore mass accuracy of the individual peaks was not accurate enough for straightforward comparison of individual spectra [22, 23]. And finally for a global proteomic approach using MALDI-ToF MS, comparing multiple spectra, an alignment software tool was required. These problems were addressed in this chapter by mixing of the analyte and matrix prior to application to the target plate to reduce the occurrence of hotspots in addition to accumulating at least 8 spectra for every samples spot on the plate [14]. To further neutralise any variation occurring due to sample preparation of MALDI-ToF MS analysis, three replicates were analysed. In our study we addressed the lack of mass accuracy (despite external calibration) by manually calibrating each spectrum to bring the *m/z* values closer together and by using a comprehensive alignment algorithm coded in VBA to automate alignment. The program was called *mz*Align and proved to be invaluable for automated profiling. A standard protocol was developed to compare 8 breast cancer with 8 control serum samples. The use of replicates and the use of *mz*Align enabled semi-quantitative analysis for marker discovery. We also tested a commercially available software tool called Markerview for peak alignment and marker discovery. In comparison to our *mz*Align, Markerview performed good alignment of protein peaks, very similar to ours, however it was difficult to see how the software actually calculates *p*-values. Moreover, alignment could not be individually adjusted for individual peaks or mass ranges but was fixed for the same mass tolerance across the entire mass region. In *mz*Align the alignment for every mass peak can be individually checked and adjusted if necessary. Markerview has a visualisation option to show alignment of the mass peaks. Visual inspection of the

spectra was important to detect the presence of outliers that may have skewed the data. This was recently addressed by Villanueva *et al.* [22], who published software for viewing and colour-coding of spectral overlays. Ideally *mz*Align could be improved to contain an option for alignment visualisation.

Protein profiling was performed twice to incorporate improvements made to the UF protocol; this resulted in the two sample sets to be quite different and therefore the results had to be analysed separately. The results, as expected, were not directly comparable, due to the different membrane pore size used. However some of the markers were the same between S1 and S2. Those proteins that were significantly different in the S1 and the S2 sample set could be considered as more robust, especially those visually confirmed. We wanted to identify all of them using MS/MS fragmentation of the intact proteins/peptides but were only able to get peptide identifications from three of the potential markers. In fact only two were "identified" with MASCOT scores indicating homology to the suggested protein in the database. For better confidence in the identifications and if we had more time, the larger proteins could be isolated by SDS-PAGE or selective chromatography and identified through tryptic digests and fingerprinting of the peptides. To become a useful marker, each protein would have to be validated by Western blotting or ELISA for the ones where antibodies are available; otherwise antibodies could be raised. Additionally, it would be interesting to see if the markers also exist in tissue by paraffin immunohistochemistry staining or tissue microarrays and whether it is changed there due to cancer growths. Furthermore these markers could then be quantitated in serum from patients with other cancer types to see if they are indicators for illness of the breast cancer specifically and finally if they occur in metastatic cancers only or could be found in early onset patients.

5.8. References

[1] Becker, S., Cazares, L. H., Watson, P., Lynch, H., Semmes, O. J., Drake, R. R. and Laronga, C. (2004) Surfaced-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. *Ann Surg Oncol* **11**, 907-914.

[2] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. and Chan, D. W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* **48**, 1296-1304.

[3] Li, X., Gong, Y., Wang, Y., Wu, S., Cai, Y., He, P., Lu, Z., Ying, W., Zhang, Y., Jiao, L., He, H., Zhang, Z., He, F., Zhao, X. and Qian, X. (2005) Comparison of alternative analytical techniques for the characterisation of the human serum proteome in HUPO Plasma Proteome Project. *Proteomics* **5**, 3423-3441.

[4] Rui, Z., Jian-Guo, J., Yuan-Peng, T., Hai, P. and Bing-Gen, R. (2003) Use of serological proteomic methods to find biomarkers associated with breast cancer. *Proteomics* **3**, 433-439.

[5] Vlahou, A., Laronga, C., Wilson, L., Gregory, B., Fournier, K., McGaughey, D., Perry, R. R., Wright, G. L., Jr. and Semmes, O. J. (2003) A novel approach toward development of a rapid blood test for breast cancer. *Clin Breast Cancer* **4**, 203-209.

[6] Villanueva, J., Martorella, A. J., Lawlor, K., Philip, J., Fleisher, M., Robbins, R. J. and Tempst, P. (2006) Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. *Mol Cell Proteomics* **5**, 1840-1852.

[7] Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J. and Tempst, P. (2005) Correcting common errors in identifying cancer-specific serum peptide signatures. *J Proteome Res* **4**, 1060-1072.

[8] Villanueva, J., Philip, J., Entenberg, D., Chaparro, C. A., Tanwar, M. K., Holland, E. C. and Tempst, P. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal Chem* **76**, 1560-1570.

[9] 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. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* **116**, 271-284.

[10] Callesen, A. K., Mohammed, S., Bunkenborg, J., Kruse, T. A., Cold, S., Mogensen, O., Christensen, R., Vach, W., Jorgensen, P. E. and Jensen, O. N. (2005) Serum protein profiling by miniaturized solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **19**, 1578-1586.

[11] Li, J., Orlandi, R., White, C. N., Rosenzweig, J., Zhao, J., Seregni, E., Morelli, D., Yu, Y., Meng, X. Y., Zhang, Z., Davidson, N. E., Fung, E. T. and Chan, D. W. (2005) Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin Chem* **51**, 2229-2235.

[12] Mathelin, C., Cromer, A., Wendling, C., Tomasetto, C. and Rio, M. C. (2006) Serum biomarkers for detection of breast cancers: A prospective study. *Breast Cancer Res Treat* **96**, 83-90.

[13] Hattan, S. J., Marchese, J., Albertinetti, M., Krishnan, S., Khainovski, N. and Juhasz, P. (2004) Effect of solvent composition on signal intensity in liquid chromatography-matrix-assisted laser desorption ionization experiments. *J Chromatogr A* **1053**, 291-297.

[14] Hattan, S. J. and Parker, K. C. (2006) Methodology utilizing MS signal intensity and LC retention time for quantitative analysis and precursor ion selection in proteomic LC-MALDI analyses. *Anal Chem* **78**, 7986-7996.

[15] Ciphergen Biosystems, I., *ProteinChip® Applications Guide Volume 1: Introductory Guide*, 2004.

[16] Cohen, S. L. and Chait, B. T. (1996) Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal Chem* **68**, 31-37.

[17] Beavis, R. C. and Chait, B. T. (1996) Matrix-assisted laser desorption ionization mass-spectrometry of proteins. *Methods Enzymol* **270**, 519-551.

[18] Richter, R., Schulz-Knappe, P., Schrader, M., Standker, L., Jurgens, M., Tammen, H. and Forssmann, W. G. (1999) Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J Chromatogr B Biomed Sci Appl* **726**, 25-35.

[19] Tammen, H., Mohring, T., Kellmann, M., Pich, A., Kreipe, H. H. and Hess, R. (2004) Mass Spectrometric Phenotyping of Val34Leu Polymorphism of Blood Coagulation Factor XIII by Differential Peptide Display. *Clinical Chemistry* **50**, 545-551.

[20] Tammen, H., Schulte, I., Hess, R., Menzel, C., Kellmann, M., Mohring, T. and Schulz-Knappe, P. (2005) Peptidomic analysis of human blood specimens: comparison between plasma specimens and serum by differential peptide display. *Proteomics* **5**, 3414-3422.

[21] Koomen, J. M., Shih, L. N., Coombes, K. R., Li, D., Xiao, L. C., Fidler, I. J., Abbruzzese, J. L. and Kobayashi, R. (2005) Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin Cancer Res* **11**, 1110-1118.

[22] Villanueva, J., Philip, J., DeNoyer, L. and Tempst, P. (2007) Data analysis of assorted serum peptidome profiles. *Nat Protoc* **2**, 588-602.

[23] Jeffries, N. (2005) Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* **21**, 3066-3073.

CHAPTER 6

Biomarker Discovery using SELDI-ToF MS

LMW markers may be small proteins or fragments of larger proteins arising due to the process of cancer proliferation [1-3]. SELDI-ToF MS is particularly suited for LMW protein analysis [4] and has been widely used to find biomarkers in serum, from patients with many diseases including breast cancer [5-9]. ProteinChip technology is using surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-ToF-MS) analysis and allows purification and concentration of subsets of proteins in a automated fashion prior to MS analysis, enabling the discovery of highly specific biomarkers for early detection of cancer. The mass spectrometry applied is essentially the same as in a MALDI-ToF instrument. The difference lies in the sample preparation and the use of surface-enhancing arrays for protein binding.

In this chapter a proof-of-principle experiment is described to determine whether SELDI-ToF MS could be developed for biomarker discovery from LMW protein serum samples obtained in a clinical setting. To initially investigate the technique and to identify which ProteinChips are most suited for our samples, a pilot study was performed on a small subset of the samples across all array surfaces. Additionally to the different arrays the LMW proteins were pre-fractionated using a weak anion exchange (WAX) resin to determine whether pre-fractionation could increase the number of peaks as well as "biomarkers" discovered even further. For both experiments I spent some time at the Ciphergen research laboratories in Guildford (UK). The LMW serum samples used were exactly the same as in Chapter 5, in fact the SELDI-ToF analysis was performed before MALDI-ToF MS and a different aliquot of the same LMW filtrate was used. The results were encouraging and so the remaining samples (S1) were analysed in a different lab at Cardiff University. However these results were disappointing, so the experiment was repeated on a fresh set of LMW serum samples S2. Unfortunately this experiment failed and the results

were even worse than from S1. Therefore the main part of this chapter will focus on the results gathered during the pilot study in Guildford and some experiments investigating the problems that occurred after.

6.1. Introduction to Chromatographic Chip Surfaces

A number of arrays with different chromatographic chemistries are available which enable the binding of different proteins; additionally each array can be processed using different binding and wash buffers (Figure 6.1). The chemistry on the arrays binds proteins with particular chemical properties (e.g. polar or hydrophobic proteins) but others are washed off. Using washes of different stringencies ensures only proteins with the specific properties favoured by each of the surfaces are retained on the chip. The use of a number of different arrays has the potential to collectively bind and analyse more proteins than uncoated chips (NP20) or MALDI plates alone [10]. This reduces the complexity of the proteome analysed and therefore enables analysis of proteins that occur at low abundance. For this experiment four arrays (hydrophobic (H50), anion exchange (Q10), cation exchange (CM10) and metal affinity capture (IMAC-Cu^{+})) and two different stringencies for Q10 and CM10 chips were tested (Figure 6.2). Effectively 6 different ProteinChips were tested to determine which chromatography binds the most proteins but also which is able to detect the maximum number of potential markers. Within this chapter the term "marker" is used loosely; these markers are merely peaks which were discriminating within our experiment and identification, validation on larger sample sets would be necessary for "identification" and validation of a potential marker.

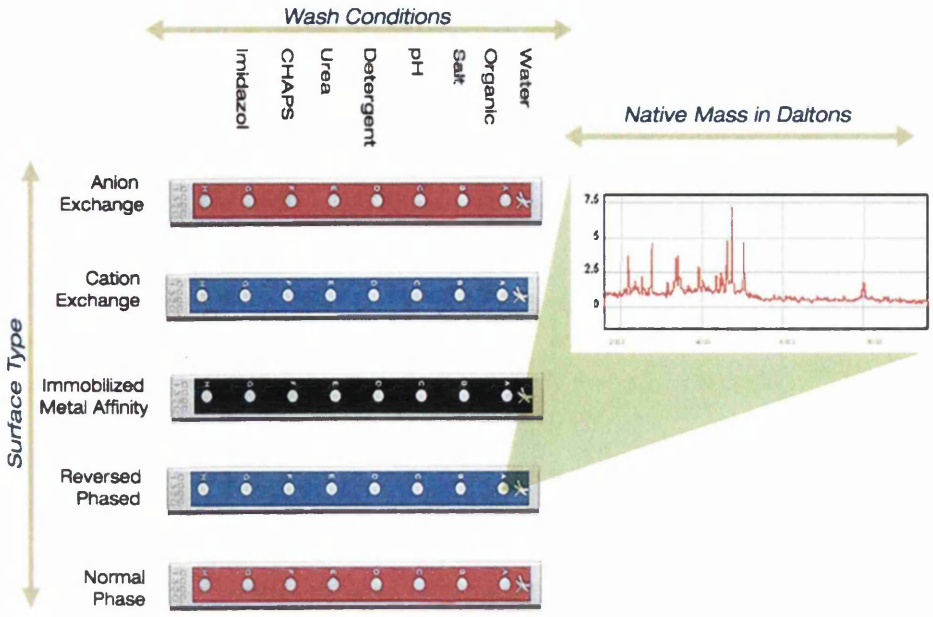


Figure 6.1: The SELDI-ToF MS systems could provides a "3-dimensional" separation system, different ProteinChips bind different proteins, using different binding and wash conditions can increase the magnitude of separation and finally proteins are separated according to their *m/z* ratio within the ToF analyser. (Diagram taken from Ciphergen SELDI-ToF MS Handbook).

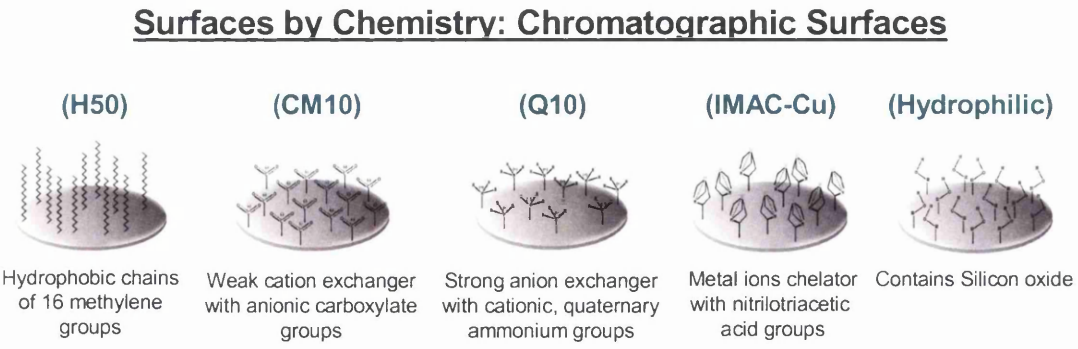


Figure 6.2: ProteinChip chemistries and their binding properties. Reverse-phase binding (H50), weak cation exchange (CM10), strong anion exchange (Q10), immobilised metal affinity chromatography for copper (IMAC-Cu^{+}) and normal phase. (Diagram taken from Ciphergen SELDI-ToF MS Handbook).

Using a bioprocessor (Ciphergen Biosystems Ltd), for preparation of the Protein Chip® arrays, each of the arrays were equilibrated twice with binding buffer (Table 6.1). LMW serum filtrates were diluted 1:10 in binding buffer and added to the arrays. After incubation for 30 min at room temperature the chips were washed 3 times with binding buffer followed by a quick rinse with deionised water. Two 0.6 µl aliquots of a saturated solution of sinapinic acid (SA) matrix (12 mg/ml in 50% ACN and 0.5% TFA in water, v/v) were added to each and air-dried. The ProteinChip® arrays were then analysed, using a ProteinChip® System, Series 4000 in linear mode using the following settings: 150 shots/spectrum collected in positive ion mode, laser intensity 175, detector sensitivity 9 and focus mass 5000 Da. The mass spectrometer was externally calibrated using the "All-in One" peptide mass standard (Ciphergen Biosystems Ltd., Guildford, UK) which contains vasopressin (1084.2 Da), somatostatin (1637.9 Da), bovine insulin β-chain (3495.9 Da), human insulin recombinant (5807.6 Da), and hirudin (7033.6 Da). Data analysis was performed using the Ciphergen ProteinChip® software 3.2 after baseline subtraction, "normalisation" and peak clustering.

Table 6.1: Binding and washing buffers for different chromatographic chip surfaces.

Array Types:	Binding and Washing buffers:
H50	10% acetonitrile, 0.1% TFA
Q10	low: 50mM Tris-HCl, pH 9
	high: 100mM Sodium acetate, pH 6
CM10	low: 100mM Sodium acetate, pH 4
	high: 50mM Tris-HCl, pH 7
IMAC 30	0.1 M sodium phosphate, 0.5 M NaCl pH 7

6.2. Studying the Reproducibility of SELDI-ToF MS Analysis

Reproducibility is very important for protein expression analysis; hence in this experiment the inter- and intra-chip reproducibility was tested. A pooled control sample was applied to two Q10 anion exchange ProteinChip arrays (8 spots per chip) and washed with high stringency buffer for MS analysis. The peaks were labelled and the peak intensity values exported to Excel for coefficient of variance (C.V.) calculation.

$$CV = \frac{StDev \times 100}{Mean}$$

As seen in Figure 6.3 the peaks on both chips look very similar, in fact by eye no differences in different peak intensity could be seen for any of the peaks within or across the arrays. Statistical comparison calculating the C.V. for each peak showed that the majority of peaks had C.V.s of less than 10%; especially for smaller *m/z* values <5500 Da (Figure 6.4). The average C.V. across the entire spectrum was calculated as 13% for all spots. From this it was concluded that the reproducibility is very high and that no variation will be introduced from different chips. Despite this finding, the breast cancer and control samples and their matched pairs were prepared and analysed on the same ProteinChip.

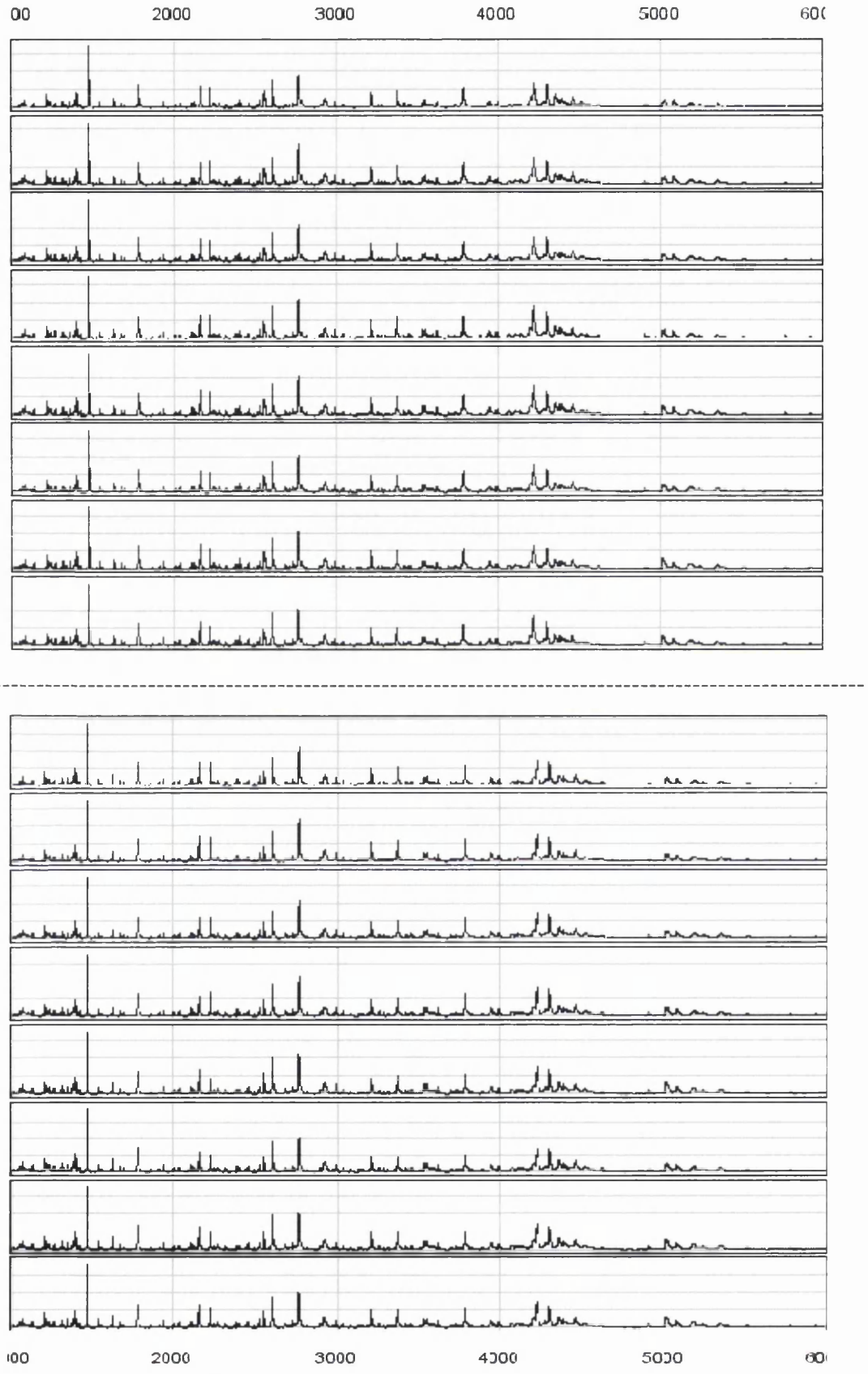


Figure 6.3: The same pooled sample was spotted in two arrays and analysed to investigate the reproducibility within and across arrays.

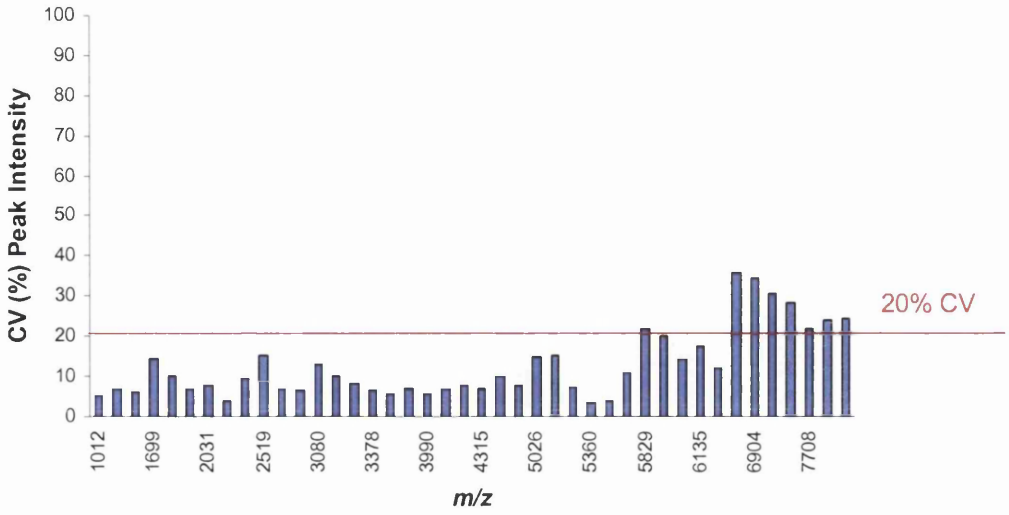


Figure 6.4: The coefficient of variance was calculated for the intensity for 42 detected peaks above signal-to-noise threshold across all spectra from both ProteinChips. The average C.V. peak intensity for all peaks = 12.8%.

6.3. Breast Cancer Marker Discovery from Sample set S1

6.3.1. Optimisation of Array Type – The 4 x 4 Study

Initially only four of the 8 cancer samples were compared to four of the 8 control samples (Figure 6.5). Each of the four ProteinChips was prepared with 4 control and 4 breast cancer LMW serum samples and analysed to determine which arrays would be most suited for biomarker discovery of these samples. Furthermore the Q10 and CM10 arrays were prepare under high and then again under low stringency conditions. As the analysis was time consuming and resources were limited, it was decided to define which arrays were best before embarking on a large study, especially since SELDI-ToF technology was widely untested on the LMW proteome. For the remainder of this chapter this experiment will be referred to as the 4 x 4 study.

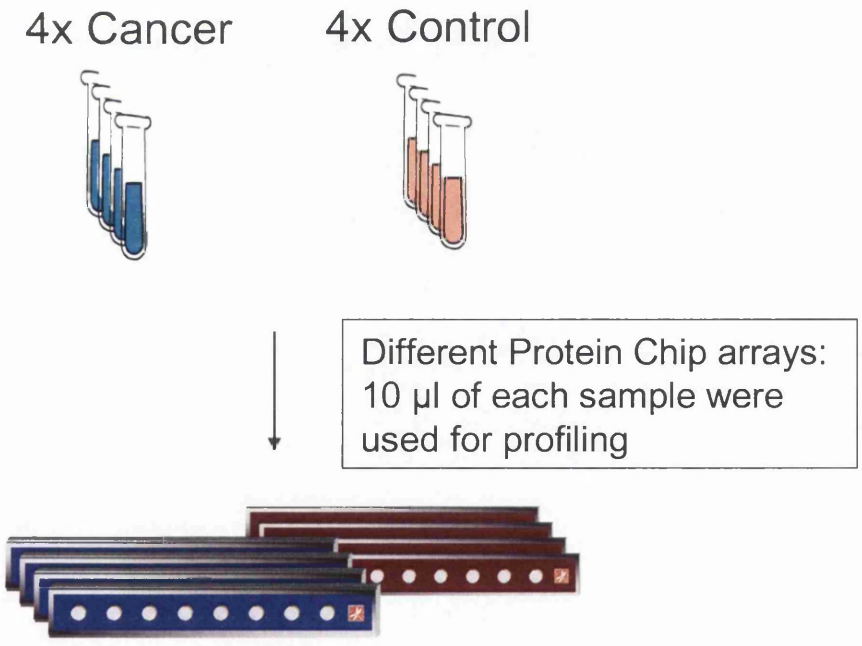


Figure 6.5: Initially only 4 sample from each group were analysed for simplicity. Later, 8 metastatic breast cancer sera were compared with 8 age-matched post-menopausal volunteer samples.

Analysis of the different array types and stringency conditions showed that different proteins bind to the arrays and therefore that using a number of different chips would provide a more complete analysis of the proteome. In Figure 6.6 some of the additional proteins that bind to CM10 chips under low rather than high stringency conditions are boxed in red. However some proteins actually ionize better on the high stringency arrays, it could be assumed that this is due to reduced ion suppression of proteins that bind strongly under low stringency conditions. Similarly Figure 6.7 shows that more proteins bind to the Q10 arrays under low stringency conditions and finally in Figure 6.8 it can be seen that less proteins bind to the IMAC-Cu and H50 arrays. The number of peaks that are bound to each type of array can be compared in Table 6.2. This actually showed that more proteins were detected on the Q10 and CM10 ProteinChips when prepared with high stringency buffers. For analysis all spectra from each type of array were combined in one experiment and peaks were normalised and clustered using the user-defined peaks, peaks with a signal-to noise ratio of 3 were detected and clustered if they were within a mass window of 0.3% of the *m/z*. Peaks that have a minimum S/N ratio of 3 were labelled in the first spectrum, then the other spectra were searched for that peak and labelled if they passed the S/N ratio of 2.

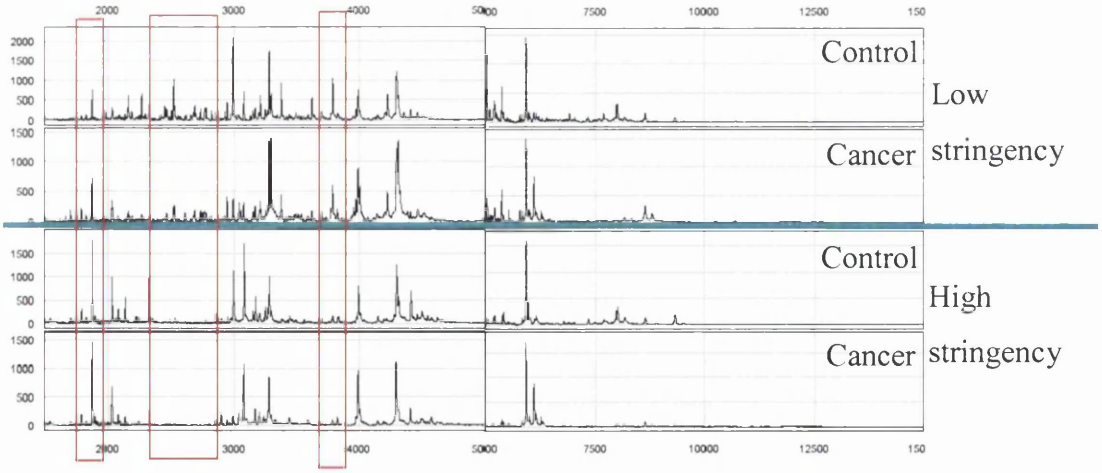


Figure 6.6: Spectra from CM10 arrays prepared at high and low stringency. Different subsets of the proteome retained using different profiling conditions. Obvious differences were boxed in red.

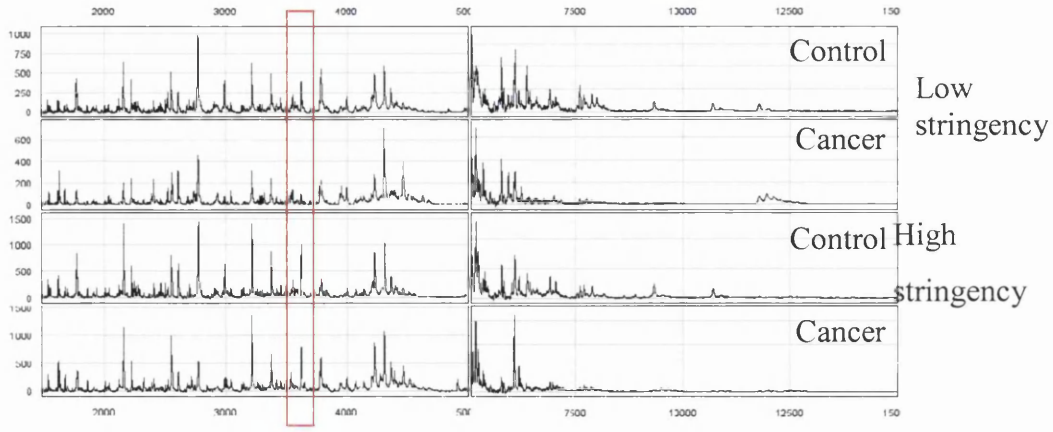


Figure 6.7: Spectra from Q10 arrays prepared at high and low stringency. Different subsets of the proteome retained using different profiling conditions. Obvious differences were boxed in red.

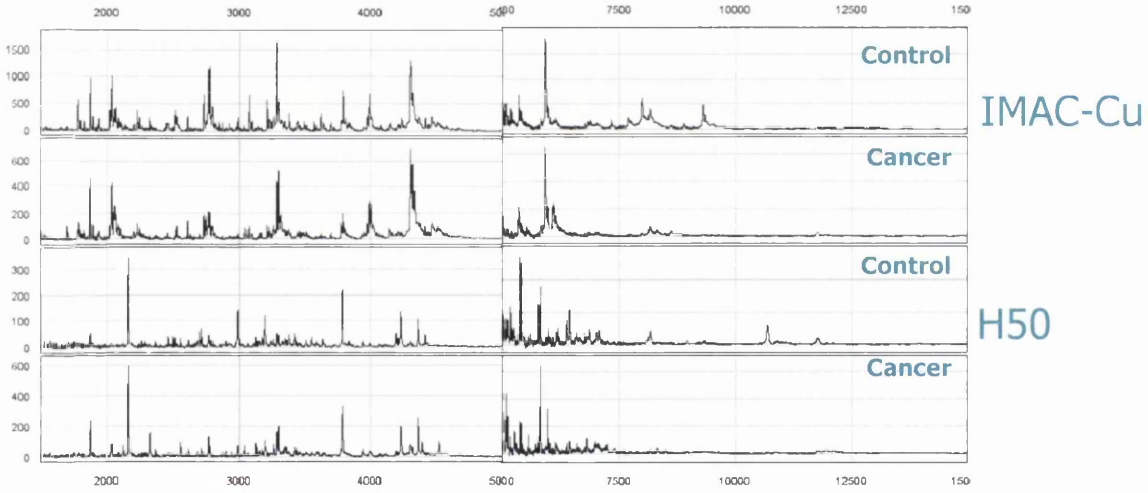


Figure 6.8: Using IMAC-Cu and H50 chips a lower number of peaks were observed. These chips appeared to bind less of the LMW proteins.

From the results of the 4 x 4 study on all the different chip types it appeared that CM10 and Q10 arrays prepared at high stringency detect the most number of proteins although Q10 ProteinChip prepared at low conditions distinguished the most markers (Table 6.2). Although the complexity of the sample needed to be reduced, once the proteins were binding to the chips it is important to use an array that allows maximum number of proteins detected. Binding the maximum number of proteins does not equal maximum number of peaks detected. It is more likely that more peaks are detected on

an array that binds fewer proteins. Hence when using one chip type only, the Q10 chip would be the chip of preference and if 2 types the CM10 would also be useful, preferably prepared at both binding stringencies. It was somewhat surprising that the high stringency binding and wash conditions detected more proteins than the low conditions. It was assumed that proteins have to be more specific to bind to the arrays. However if the proteome binding is less complex, more proteins may reach high enough intensity to be detected.

Table 6.2: Number of peaks and potential markers detected by the SELDI-ToF MS 4 x 4 pilot experiments on all chip types.

Condition	Chip No	Total peak count	*p*-value <0.05
CM10 pH4 (high)	1170202693	73	4
CM10 pH7 (low)	1170202694	55	8
Q10 pH9 (high)	1230154912	73	7
Q10 pH6 (low)	1230154913	64	13
IMAC-Cu	1190177256	66	6
H50	1080151427	51	1
Total		**352**	**35 (28 unique)**

As part of the Ciphergen SELDI-ToF system, sophisticated analysis software is available to perform spectrum processing such as baseline correction and normalisation of the spectra to neutralise spectrum-to-spectrum variation. The spectra can then be clustered; finding peaks in common to be compared using the Biomarker Wizard where Mann-Whitney *p*-values are calculated for each peak in the spectrum to provide statistical evidence for differential peaks. The software also provides visualisation of the protein peaks in the form of peak spectra and gel views that image the peak intensity in form of different intensity bands. Additionally to the Mann-Whitney *U* test results, the peak intensities from each experiment were exported into Excel and the variation was calculated using an unpaired Student's *t*-test. The *t*-test is more powerful than the non-parametric test and moreover in Excel fold-changes of the differences in peak intensity could be calculated [11]. The results from both analyses for all chip types are shown in Table 6.3, the *t*-test appears to be more stringent than the Mann-Whitney *U* test and visual inspection of the peaks confirmed that many of

the *m/z* values that were significant by only the Mann-Whitney *U* test were in fact not actually different or of very low intensity.

Table 6.3: Discriminating peaks from the 4 x 4 study. *p*-values were calculated using the Mann-Whitney *U* test as part of the Biomarker Wizard software and a Student's *t*-test after exporting the data into excel.

m/z	*p*-value (Excel)	fold-change	Mann-Whitney (Ciphergen)	m/z	*p*-value (Excel)	fold-change	Mann-Whitney (Ciphergen)
Q10 high stringency				*CM10 high stringency*			
1078.2	0.021	1.6	0.021	2913.3			0.021
1225.8	0.029	1.7	0.021	2993.3			0.021
1314.9	0.029	1.7	0.043	4498.6	0.016	-1.3	0.043
1402.0	0.005	1.5	0.021	8806.1	0.028	2.7	0.043
1636.5			0.043	*CM10 low stringency*			
1936.5	0.050	-2.1	0.043	2262.8			0.430
2249.1	0.012	-1.9	0.021	2525.0			0.021
2463.5	0.030	-1.6	0.021	2994.9			0.021
2775.8	0.041	-1.9	0.021	3791.6	0.009	-1.5	
2994.8			0.021	3989.8	0.029	1.2	0.021
4207.4			0.043	5026.0	0.009	-1.5	
4400.8			0.021	8803.1	0.035	3.0	0.430
4514.0	0.023	1.5	0.021	14891.2	0.048	1.8	
Q10 low stringency				*IMAC-Cu*			
1780.5	0.003	-2.1	0.021	1780.1	0.011	-2.3	0.021
2994.3			0.021	1802.3	0.001	-2.7	0.021
3537.3	0.010	2.4	0.043	2775.4	0.013	-2.3	0.043
3554.5	0.032	-1.4	0.021	2791.2	0.008	-2.1	0.021
3788.2			0.043	3791.9	0.026	-1.7	0.043
6381.2			0.043	4271.7			0.021
12118.9			0.043	*H50*			
				2992.6	0.037	-3.9	none

The most discriminating peaks from each type of ProteinChip are shown in Figure 6.9 to Figure 6.13. The Spectra from each sample were normalised and overlaid within the ProteinChip Software 3.1; the red lines are from the breast cancer samples and the blue lines from controls. A list of the masses that were statistically and visually significant across all ProteinChip types is shown after all the spectra in Table 6.4.

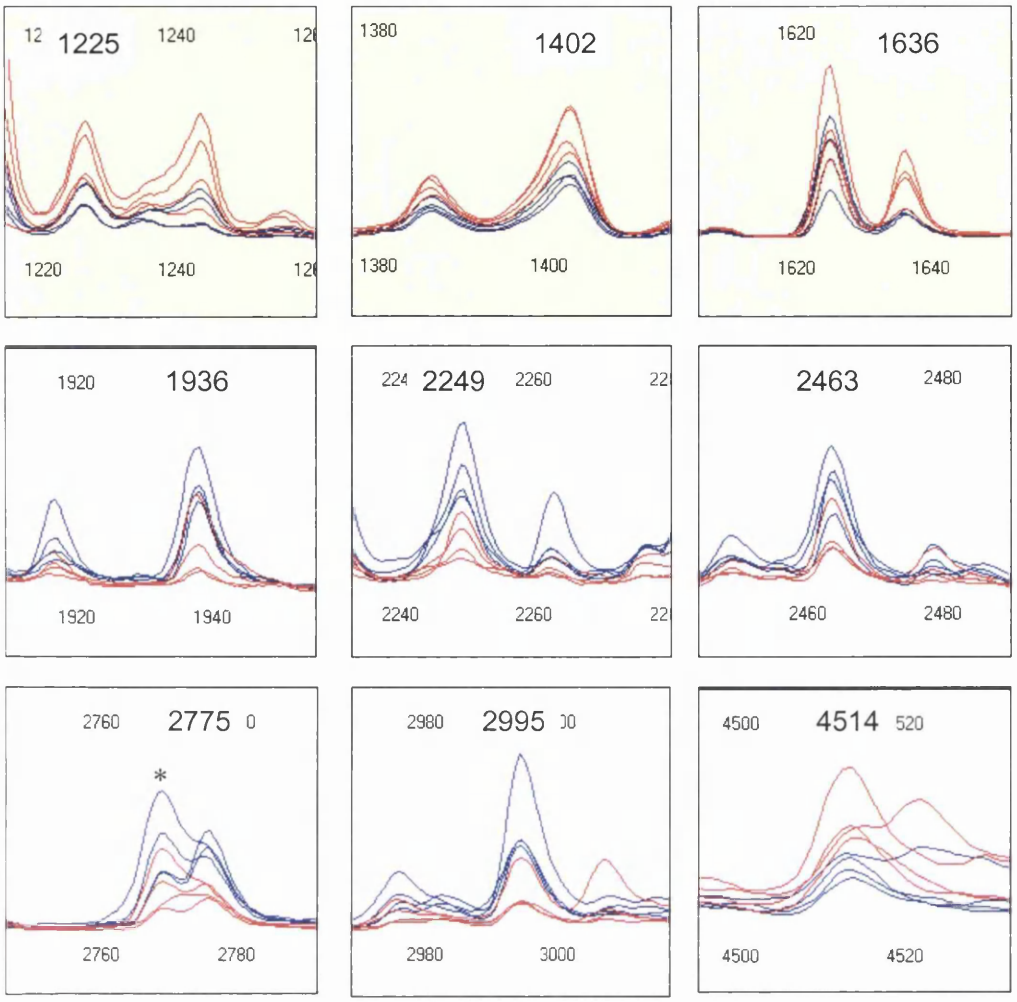


Figure 6.9: The 9 most discriminating peaks recovered from Q10 arrays washed with high stringency (pH 6). The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

Some of the peaks were found to be significant on more than one array. The peak at *m/z* 2775 had a significant *p*-value on the Q10 (high stringency) and IMAC-Cu^{+} arrays (Figure 6.9 and Figure 6.10). However the peak with *m/z* 2765 (marked *) showed no significant distinction between the two sample groups (Figure 6.9). The peak at *m/z* 1780 was also discriminating on two arrays (IMAC-Cu^{+} and Q10 low) and *m/z* 2995 and 3791 on even three arrays. So for IMAC-Cu^{+}, all significant peaks were also detected on other ProteinChip types, except for *m/z* 4271 which was only elevated in one breast cancer sample (Figure 6.10). Therefore the IMAC-Cu^{+} arrays appear to add no further information to the profiling results.

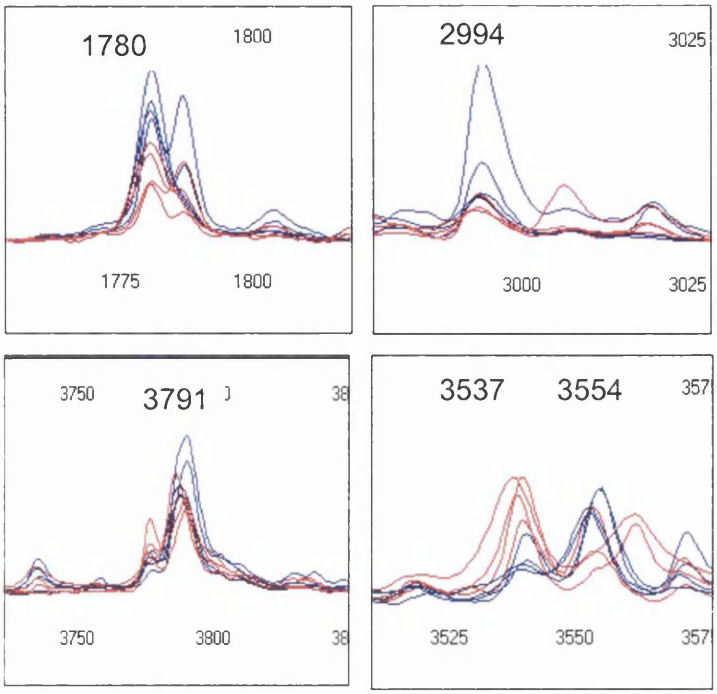


Figure 6.10: The four most discriminating peaks recovered from IMAC-Cu arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

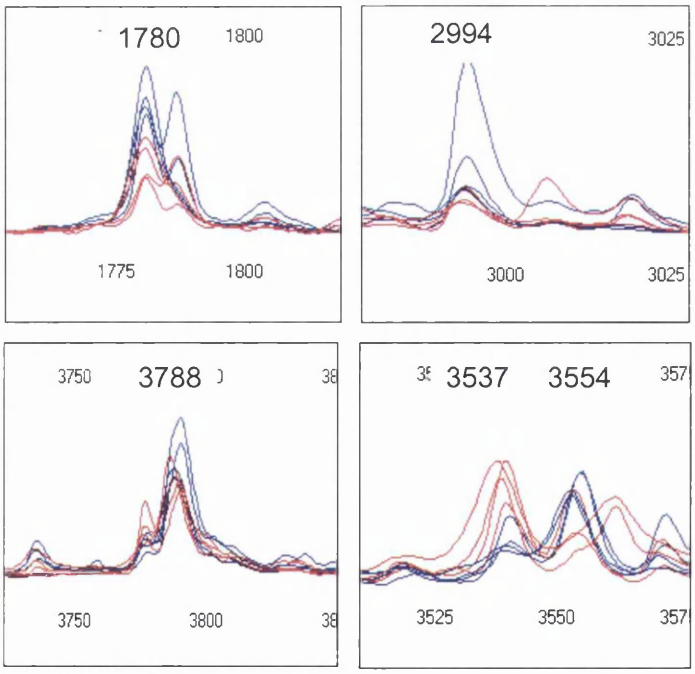


Figure 6.11: The four most discriminating peaks recovered from Q10 arrays washed with low stringency (pH 9). The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

The discriminating peaks at *m/z* 3537 and 3554 in Figure 6.11 were interesting, as there appears to be a mass change in the breast cancer samples. The intensity for the peak at *m/z* 3537 is increased but reduced for the peak at *m/z* 3554 in the breast cancer group. It is possible that due to the cancer this protein was modified, has lost a residue or a fragment. This would be an interesting protein to analyse further. The peaks at *m/z* 3791 in Figure 6.11 present an example of a peak that had a significant *p*-value but was not visually confirmed to be different between the two groups. In fact the same is true for *m/z* 3791 in the CM10 (low stringency) arrays (Figure 6.13). Interestingly this is one of the only peaks that was not significantly different using the Mann-Whitney *U* test. Since the Mann-Whitney *U* statistic is less powerful, it finds less peaks significant that are more likely to be false-positives [11]. On the other hand it can miss significant differences, hence it is justified to use both tests and to verify the results.

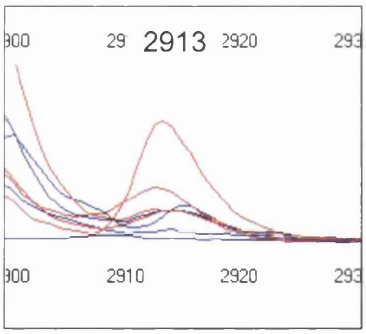


Figure 6.12: The only truly discriminating peak recovered from the CM10 high stringency (pH 7) arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

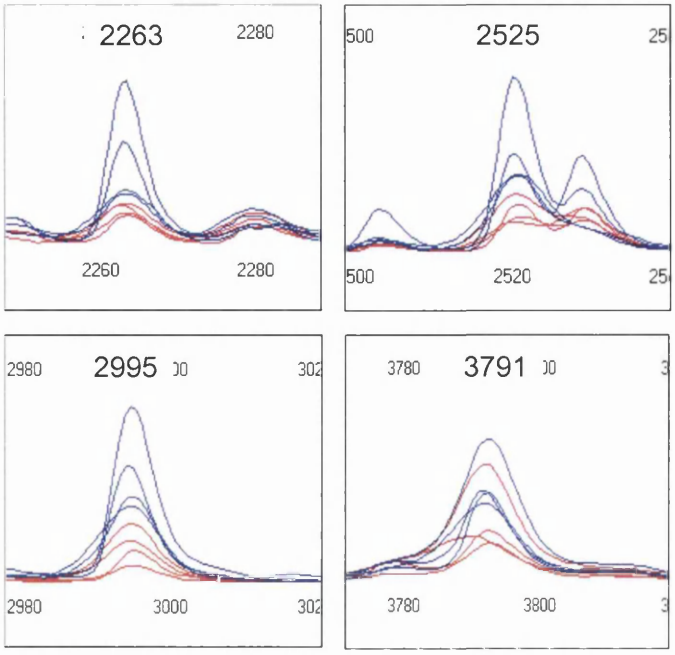


Figure 6.13: The four most discriminating peaks recovered from CM10 arrays washed with low stringency (pH 4). The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

Table 6.4: The 17 most discriminating peaks recovered from all array types, which were also visually confirmed.

m/z	Array type	*p*-value	fold-change
1225.8	Q10 high	0.029	1.7
1402.0	Q10 high	0.005	1.5
1636.5	Q10 high	0.043	2.2
1780.1	Q10 low, IMAC	0.021	-2.3
1802.3	IMAC	0.001	-2.7
1936.5	Q10 high	0.050	-2.1
2249.1	Q10 high	0.012	-1.9
2463.5	Q10 high	0.030	-1.6
2525.0	CM10 low	0.021	-2.5
2775.4	Q10 high, IMAC	0.013	-2.3
2791.2	IMAC	0.008	-2.1
2913.3	CM10 high	0.021	5.3
2994.8	Q10 high, Q10 low, CM10 high, CM10 low, H50	0.021	-2.6
3537.3	Q10 low	0.043	2.3
3554.5	Q10 low	0.021	-1.4
3791.9	IMAC, CM10 low, Q10 low	0.009	-1.7
4514.0	Q10 high	0.023	1.5

6.3.2. Pre-Fractionation of all Samples from S1 using a WAX Separation

To reduce the complexity of the LMW proteins even further, the samples were separated using a weak anion exchange resin. The LMW serum filtrate (30 µl) was fractionated into 6 fractions by weak anion exchange (WAX) separation using a step-pH gradient. A 10 µl aliquot of each fraction was then analysed on weak cation exchange (CM10) arrays under low stringency conditions. The experiment setup is shown in Figure 6.14 and Figure 6.15. For the fractionation all 8 breast cancer and 8 control samples were used. The fractionation was performed in a 96 well plate on a Biomek 2000 robotic sample-processing station (Beckman, UK) to avoid degradation, the plate was cooled.

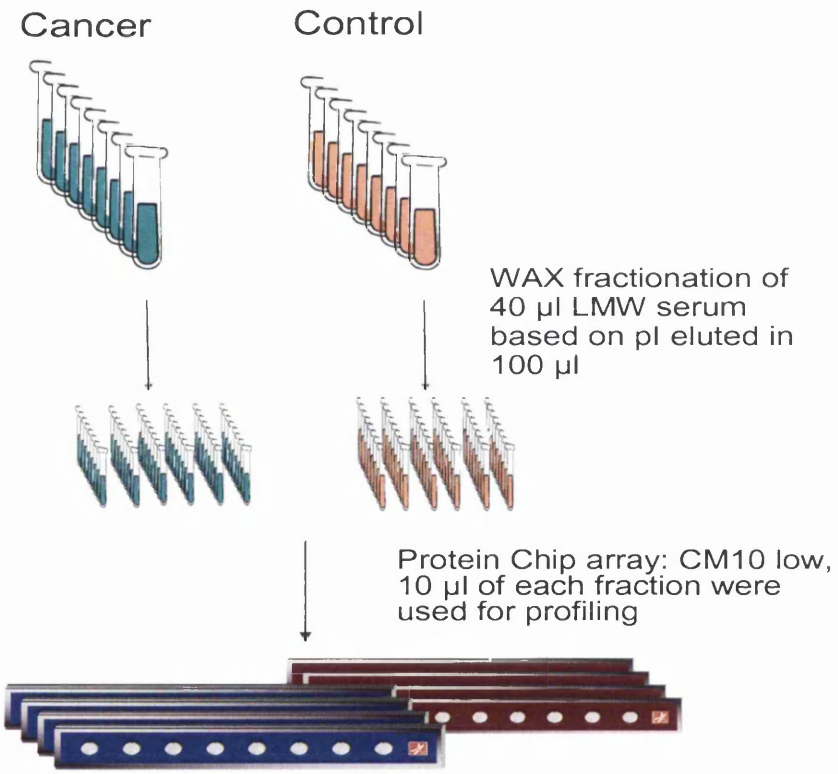


Figure 6.14: Experiment setup of WAX pre-fractionation of LMW serum proteins. Each fraction was then analysed on CM10 ProteinChips prepared at low stringency.

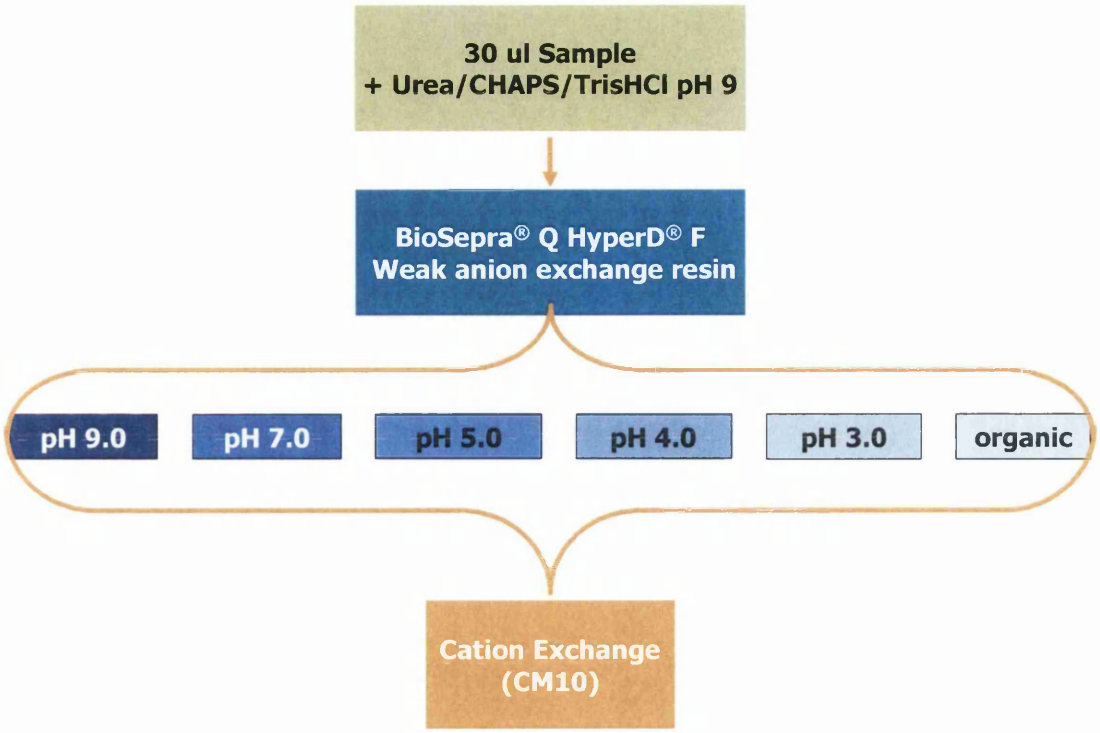


Figure 6.15: Weak anion exchange (WAX) pre-fractionation. The flow-through was collected together with the first fraction (pH 9). Each fraction was analysed on CM10 ProteinChips.

Fractionation of a complex protein mixture prior to MS analysis can increase the number of protein peaks discovered, this is due to the fact that in complex mixtures, more abundant proteins mask lower abundant proteins by ion suppression. Reducing the complexity of the sample has the potential to also increase discovery of possible markers. Figure 6.16 shows that different peaks were present in each of the fractions. Boxed in red are the peaks that were amplified in the fractions but otherwise at low intensity in the un-fractionated sample. In more detail the number of peaks retrieved from each fraction is shown in Table 6.5, from this it appears that many proteins that do not bind to the WAX resin (i.e. cations), bind well to the cation exchange array (CM10). The CM10 chips were prepared at low stringency binding conditions (pH 4), where the high pH fractions (pH 9- pH 4) are positively charged and bind to the cation exchanger. In fraction 5 (pH 3), the proteins are negatively charged in the pH 4 binding buffer and hence bind poorly to the array. This is reflected by the low number of peaks detected in fraction 5 (pH 3) (Figure 6.16).

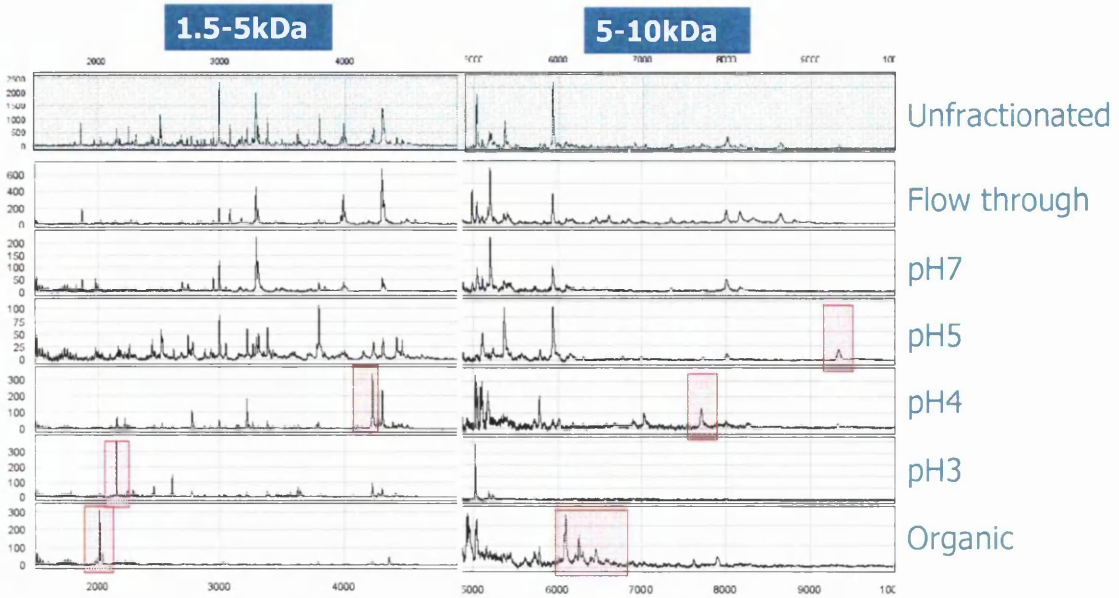


Figure 6.16: CM10 profiling of fractions. Different peaks were eluted in each fraction and therefore peaks that are otherwise masked in the un-fractionated sample were amplified and could be analysed. The flow-through was collected together with the pH 9 fraction.

Table 6.5: The number of peaks retrieved from each fraction. 160 peaks of which 113 unique peaks were detected on CM10 low filters compared to only 55 from the un-fractionated sample. And 11 unique "biomarkers" compared to only 4 on the CM10 low chip before fractionation.

Fraction	pH	Peaks on CM10 low	*p*-value <0.05
F1	FT + pH 9	43	1
F2	pH 7	17	3
F3	pH 5	24	2
F4	pH 4	29	2
F5	pH 3	20	1
F6	organic	27	4
Total	**organic**	**160**	**13 (11 unique)**

As before, the spectra were analysed for expression changes using the Biomarker Wizard and the peak intensities were exported into Excel for calculation of *p*-values using a *t*-test. In this case all 8 samples from each cohort were analysed and so the results could be regarded as a confirmation of the significant *p*-values recovered before. However it is important to keep in mind that these fractions were only prepared on CM10 chips at low stringency conditions. Few markers in the 4 x 4 study

were actually discovered on that chip type and so only two markers (*m/z* 2994 and 3793) from the 4 x 4 experiment where also found amongst the markers in the WAX fractions (Table 6.6). However many markers were seen that are not on any of the chips without fractionation.

In total 160 peaks were detected from all fractions, compared to only 55 from the un-fractionated sample. Of those, 113 peaks were unique; furthermore 11 unique markers were discovered. This is more than double the number from un-fractionated LMW serum on CM10 arrays in the 4 x 4 study. However most of these markers have relatively low fold-change differences between the breast cancer and control samples.

Table 6.6: Discriminating peaks from WAX fractions. Calculated by univariate analysis in Excel and a Mann-Whitney *U* test as part of the Biomarker Wizard.

Markers (*t*-test)				Markers (Biomarker Wizard)		
Fraction	*m/z*	*p*-value	fold-change	Fraction	*m/z*	*p*-value
F2	1013.5	0.031	1.0	F2	1013.7	0.036
F6	1034.4	0.029	1.0	F6	1034.3	0.036
F2, F3, F4	2997.1	0.030	1.1	F2, F3	2998.3	0.046
FT	3172.4	0.030	-1.1	FT	3175.4	0.046
F2	3309.5	0.029	-1.2			
F3	3384.7	0.024	-1.4	F3	3382.1	0.021
F6	3793.7	0.019	-1.4	F6	3792.2	0.016
F5	4234.3	0.043	1.0			
F4	5107.9	0.025	-1.2	F4	5100.5	0.029
F6	7626.6	0.015	-1.0	F6	7620.6	0.009
F6	7911.7	0.025	1.1	F6	7900.4	0.036

Visual inspection of the "markers" in Figure 6.17 to Figure 6.22 showed that most of the peaks appear to be convincingly different between breast cancer (red) and controls (blue) spectra. Some of the higher molecular weight peaks such as *m/z* 5107 and *m/z* 7900 were a bit fuzzy. One of the most convincing markers from the S1 set in the MALDI-ToF MS data was *m/z* 2995. This marker was also shown to be significantly different, using SELDI-ToF MS.

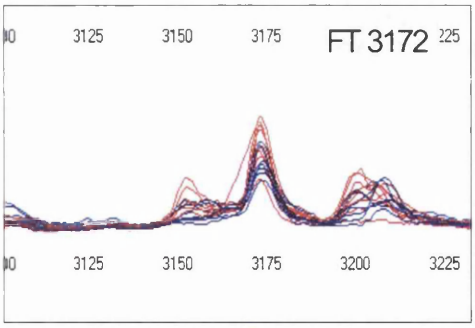


Figure 6.17: The most discriminating peak recovered from fraction 1 (FT and pH 9) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

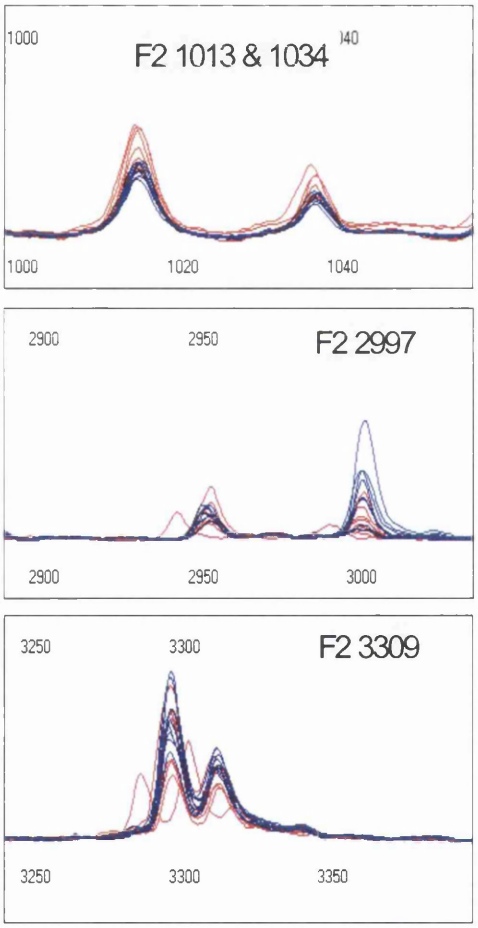


Figure 6.18: The three most discriminating peak recovered from fraction 2 (pH 7) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

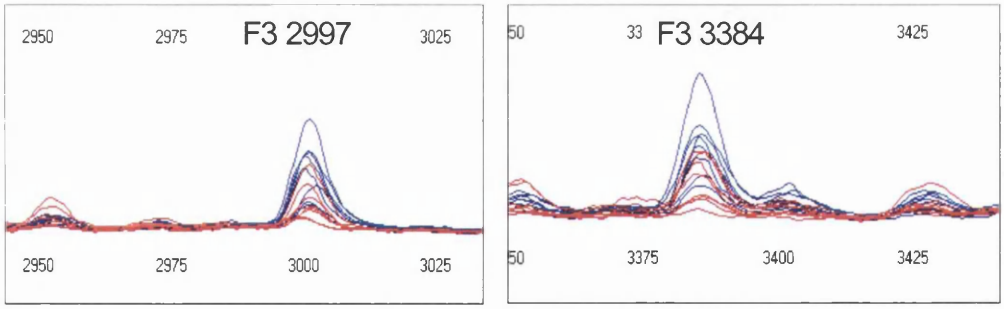


Figure 6.19: The two most discriminating peak recovered from fraction 3 (pH 5) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

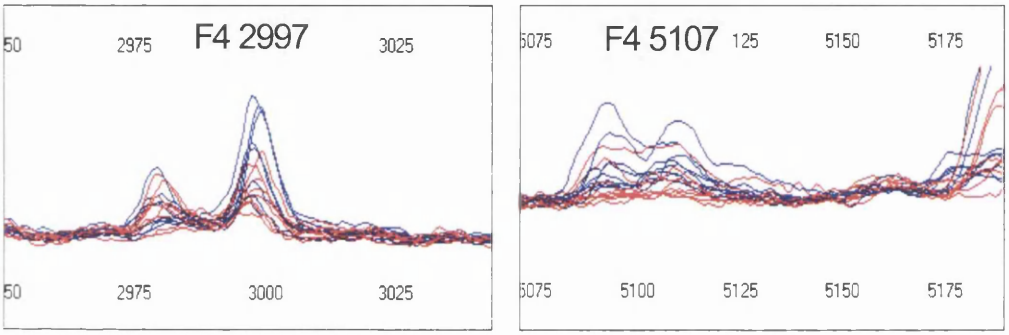


Figure 6.20: The two most discriminating peak recovered from fraction 4 (pH 4) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

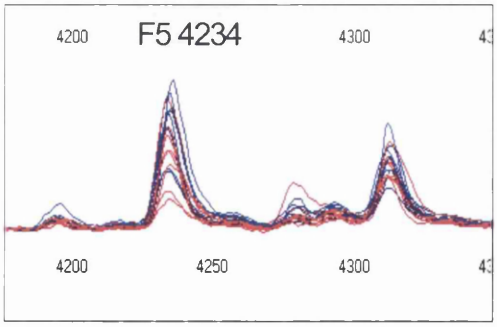


Figure 6.21: The most discriminating peak recovered from fraction 5 (pH 3) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

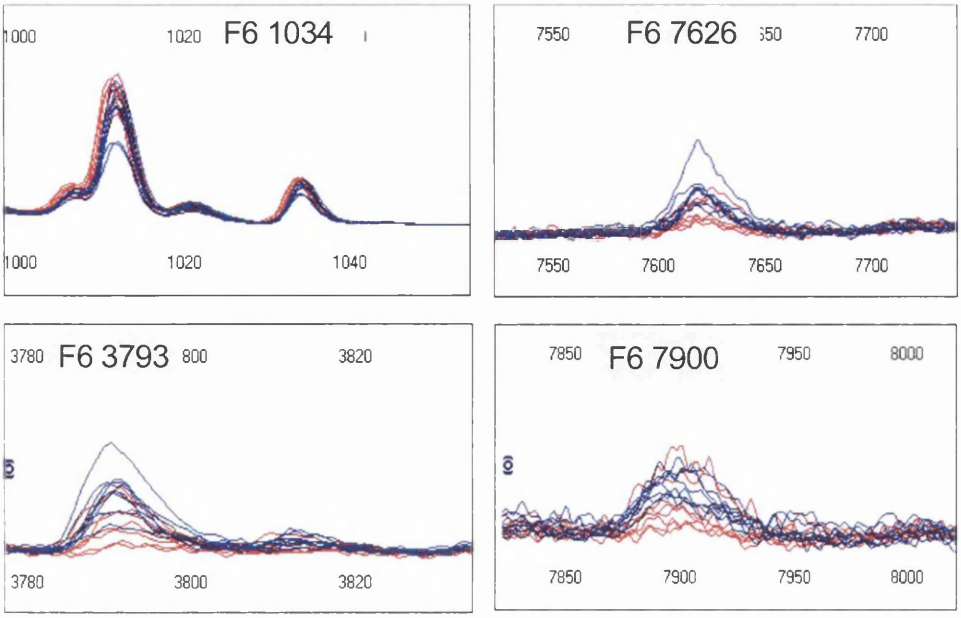


Figure 6.22: The four most discriminating peak recovered from fraction 6 (organic) on CM10 arrays. The spectra from each sample were overlaid, in red the breast cancer samples and in blue control.

It was interesting to see what while the peak intensity for *m/z* 4234 in Figure 6.21 was slightly decreased in most of the breast cancer samples;, it remained unchanged for *m/z* 4310. This added further confidence that the change was due to the cancer and not due to an overall difference in the spectrum intensity. Furthermore there appeared to be more variation among the control samples and the breast cancer peak intensity values were tighter (Figure 6.23). This provided a certain amount of confidence that the discriminating peaks are related to the breast cancer. The control volunteers had very little in common, except for that none of them should have cancer. However the patients all had breast cancer as a common factor.

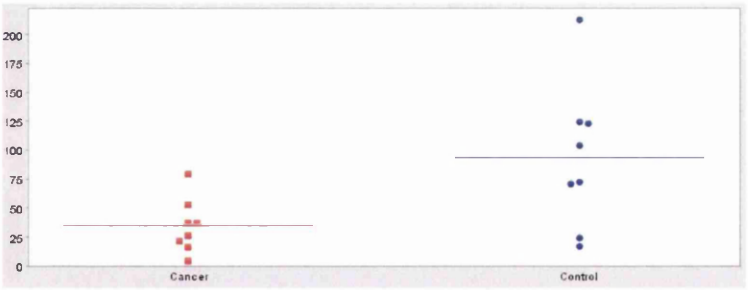


Figure 6.23: Distribution of the peak intensities for *m/z* 2997 in fraction F2, from the breast cancer (red) and control (blue) samples.

6.3.3. The Effects of Sample Pooling on Peak Recovery and Biomarker Discovery

There is an argument for pooling all samples to gain a representative breast cancer sample and control sample and to reduce biological variation to identify makers that are truly different due to the breast cancer status. This would also make analysis faster and reduce the number of SELDI chips to be used. On the other hand there is some argument against this since small changes, possibly due to the cancer, could be masked by the common peaks. Breast cancer is a heterogeneous disease and changes may be present in a certain clinical group, dependent on stage or for example *Her2*-status. Furthermore, markers due to a certain stage in the cancer could be lost when pooling these particular samples with other cancer serum samples. Pooling may be more appropriate for the control sample. However it should be better to treat both sample cohorts the same. In microarray analysis due to the cost of the chips, pooling has been done [12]. However specific statistical analysis and algorithms are necessary to deal with pooling [13, 14].

Here the effect of pooling on peak detection, in comparison to un-pooled samples was investigated. A breast cancer sample (pooled across all 8 patient samples) and a control sample (pooled from all volunteer samples) was created. Both pooled samples were fractionated using the WAX resin and left un-fractionated. The samples and fractions were compared on CM10 arrays at high and low stringency, as well as on an IMAC-Cu^{+} arrays. The number of peaks detected was used as an indication if pooling affects the results. The results revealed that only approximately half the number of peaks was recovered from the pooled samples compared to the individuals on each of the array types (Table 6.7).

Table 6.7: Recovery of the number of peaks is a good measure for suitability of the chip type. Here the results from the analysis of the S1 samples while in Guildford is shown, the number of peaks detected on various array types is compared with the number of peaks detected when analysing the same sample by MALDI-ToF MS.

Method	Profiling conditions/ peak Number						Total Peak Number
MALDI							274
	Arrays	Low stringency	High stringency				
SELDI 4x4	H 50	51	-				51
	Q10	65	73				138
	CM10	55	73				128
	IMAC 30	66	-				66
SELDI pooled	CM10	35	30				65
	IMAC 30	36	-				
WAX resin fractions							
Pre-fractionation + SELDI CM10 low	FT + pH9	pH7	pH5	pH4	pH3	organic	
	43	17	25	28	20	27	160 (113 unique)

Although fractionation enables detection of some proteins that were at very low intensity in the pooled sample, the overall peak intensity was lower (Figure 6.24). This showed further that pooling has a negative effect on peak recovery. However the spectra look very similar, it can be seen that more peaks are present in the un-pooled, un-fractionated sample and F3 and F5 contain more peaks in the un-pooled fractions (Figure 6.24). Both samples were prepared and analysed on CM10 arrays in the same experiment. Table 6.7 shows a summary of the number of peaks detected from each type of array; for comparison the pooled sample and the WAX fractions were also shown. Furthermore the number of peaks recovered by MALDI-ToF MS (274) was also shown. The MALDI-ToF results were exported with a lower threshold for peak detection than the SELDI-ToF spectra, which explains the larger number of peaks discovered. During cluster analysis for SELDI-ToF, if a peak passes the intensity threshold in one or more spectra, the algorithm will find this peak in the other spectra and label it. This is not possible for MALDI-ToF spectra and therefore more peaks had to be exported to ensure complete comparison between all spectra.

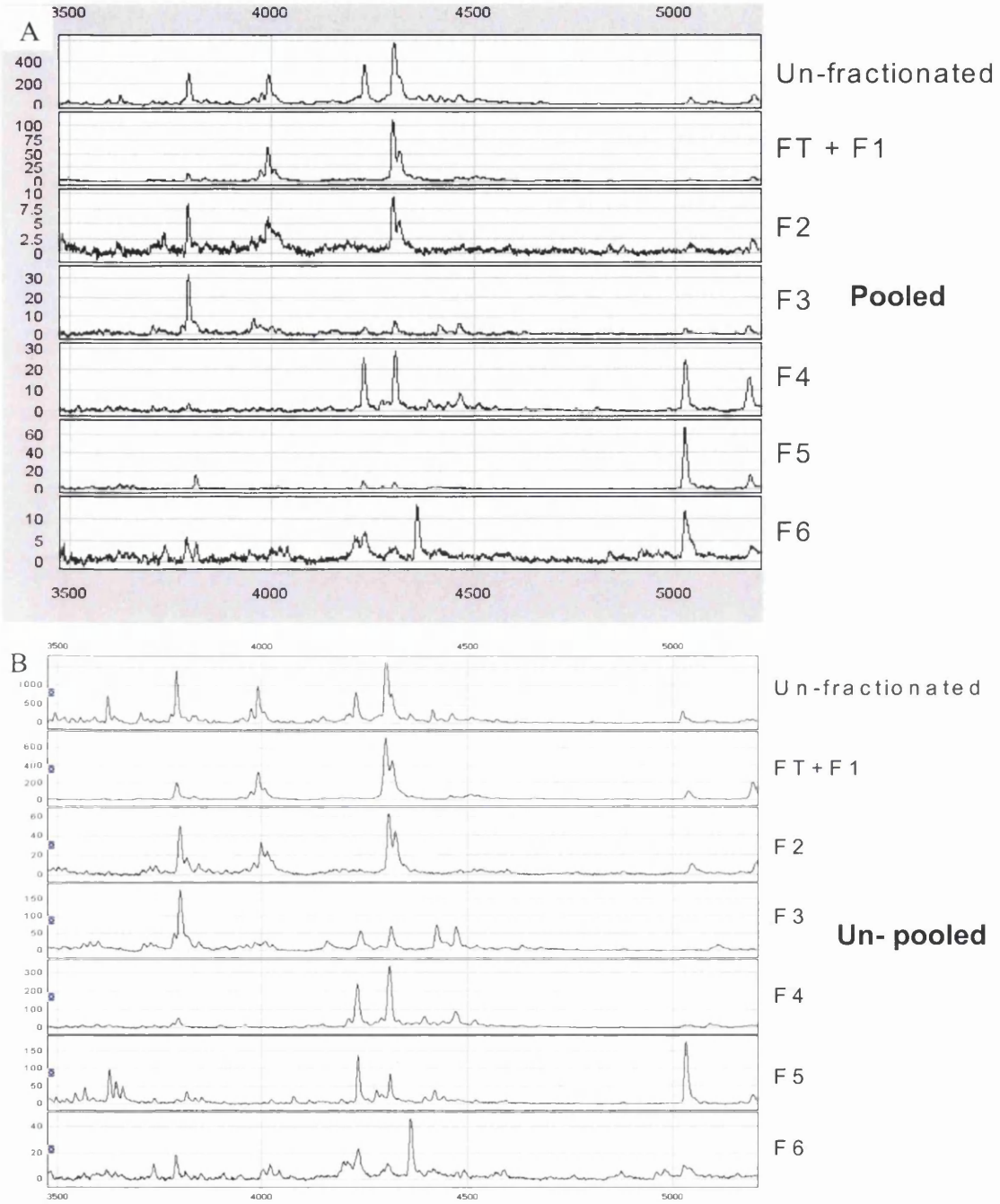


Figure 6.24: WAX fractionation of pooled control serum sample (**A**) and an un-pooled control sample (**B**). A mass range of *m/z* 3500-5300 is shown.

Nevertheless, some of the markers that were discovered in the 4 x 4 study and in the un-pooled fractions were also visibly different comparing the pooled breast cancer and control LMW serum sample.

Due to the lack of replicates no statistical analysis was performed to find discriminating peaks. However, some differences in peak intensity were easily seen by eye. Two examples are shown below (Figure 6.25). It may be argued that the differences between breast cancer and control peaks are more intense in the pooled samples compared to the spectra from un-pooled samples. However since there was only one replicate each, it cannot be said for certain that this has not happened by chance.

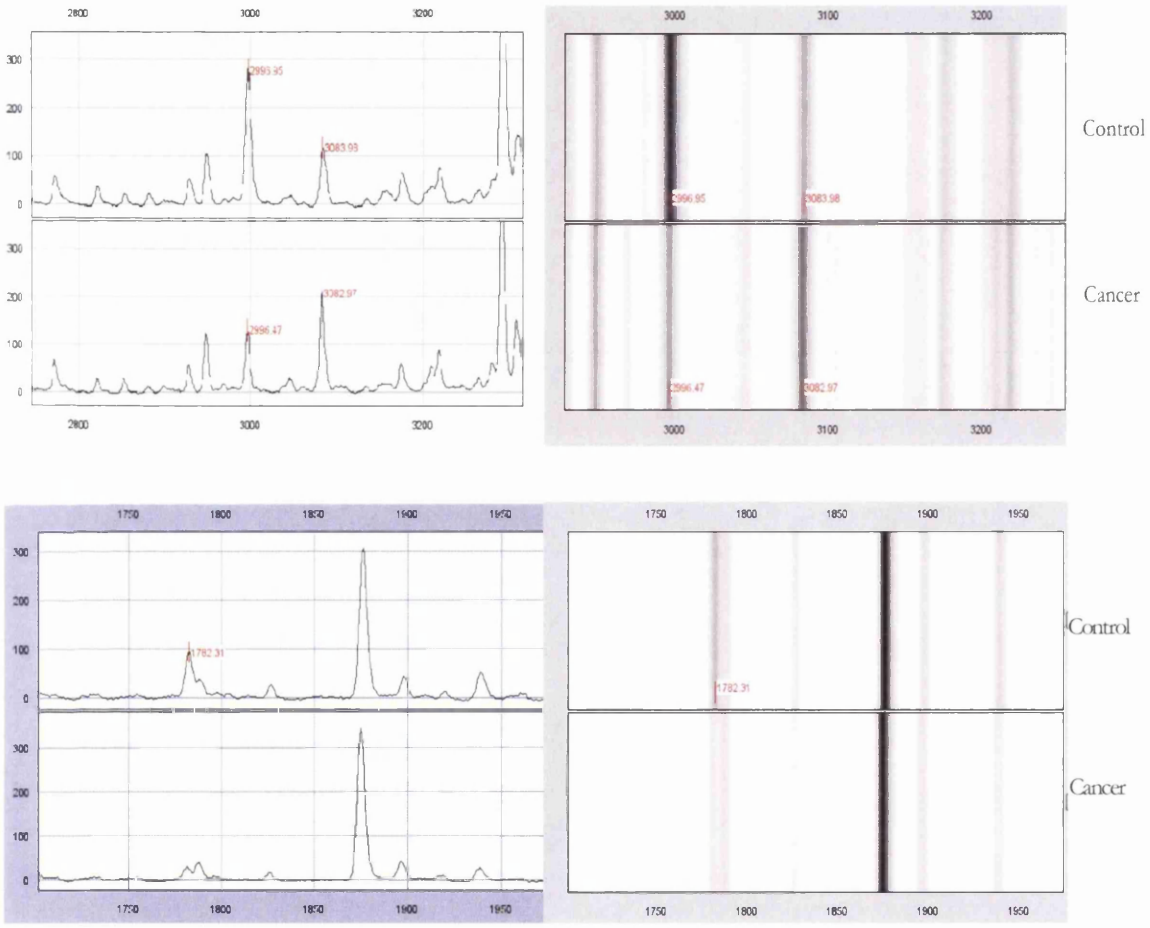


Figure 6.25: Two examples of markers that were retrieved from analysis of the pooled breast cancer and the non-cancer control LMW serum samples. Both examples were chosen because an un-changed peak is visible, giving evidence that the markers are not due to an overall difference in spectral intensity.

Although the peak intensity and resolution was decreased in the pooled compared to the un-pooled samples, markers were still discovered. It might be worth pooling if sample was limited if the cost had to be reduced. It may allow analyses of larger number of technical replicates in an experiment with a large number of biological replicates.

6.3.4. Analysis of the Remaining S1 Samples in Cardiff: The 8 x 8 Study

The 4 x 4 study was continued to analyse all of the 8 breast cancer and 8 control samples (S1 in Chapter 5) on two further chips per sample in Cardiff using a PBS-II ToF ProteinChip® reader (Figure 6.26). This produced three technical replicates from each of the 8 x 8 samples (Table 6.8). Since the mass analyser in Cardiff is an older version of the instrument, the chips previously prepared in Guildford were analysed again alongside the new chips. The arrays were prepared in the same way as before (section 6.1). For the new chips a higher laser intensity (210) was requires to obtain peaks, however the laser intensity was kept at 175 for the chips prepared in Guildford.

Table 6.8: ProteinChip preparation. Four control and four breast cancer samples were bound to the arrays in Guildford. The remaining four of each group were prepared in Cardiff. Furthermore another two replicate chips were prepared of all 8 x 8 samples while in Cardiff. All chips were scanned in Cardiff on a PBS-II ToF ProteinChip® reader.

Replicate Arrays	Samples	Preparation Lab	Samples	Preparation Lab
1	1 - 4 breast cancer 1 - 4 control	Guildford	5 - 8 breast cancer 5 - 8 control	Cardiff
2	1 - 4 breast cancer 1 - 4 control	Cardiff	5 - 8 breast cancer 5 - 8 control	Cardiff
3	1 - 4 breast cancer 1 - 4 control	Cardiff	5 - 8 breast cancer 5 - 8 control	Cardiff

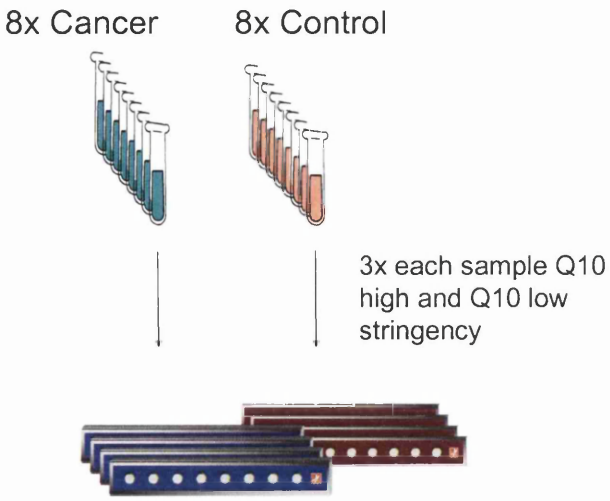


Figure 6.26: ProteinChip preparation for the 8 breast cancer and 8 breast cancer samples on strong anion exchange (Q10) arrays in Cardiff. The chips were prepared in triplicates. However 4 cancer and 4 control chips, already prepared in Guildford, were scanned again in Cardiff.

The spectra from the new ProteinChips look completely different, the peak intensity is very low and the few peaks that were detectable were broad with low resolution and looked asymmetrical (Figure 6.27 and Figure 6.28). For comparison the spectrum from the array prepared in Guildford is shown below the two new replicates. The spectra from the Guildford arrays look good, analysed on the older ProteinChip reader, which suggests that the problem originated from either the sample or the sample preparation during chip binding. This was the same for all samples from both cohorts, two examples from the breast cancer group are show in Figure 6.27 and another example from the control group is shown in Figure 6.28. A high and a low resolution mass range are shown for each sample.

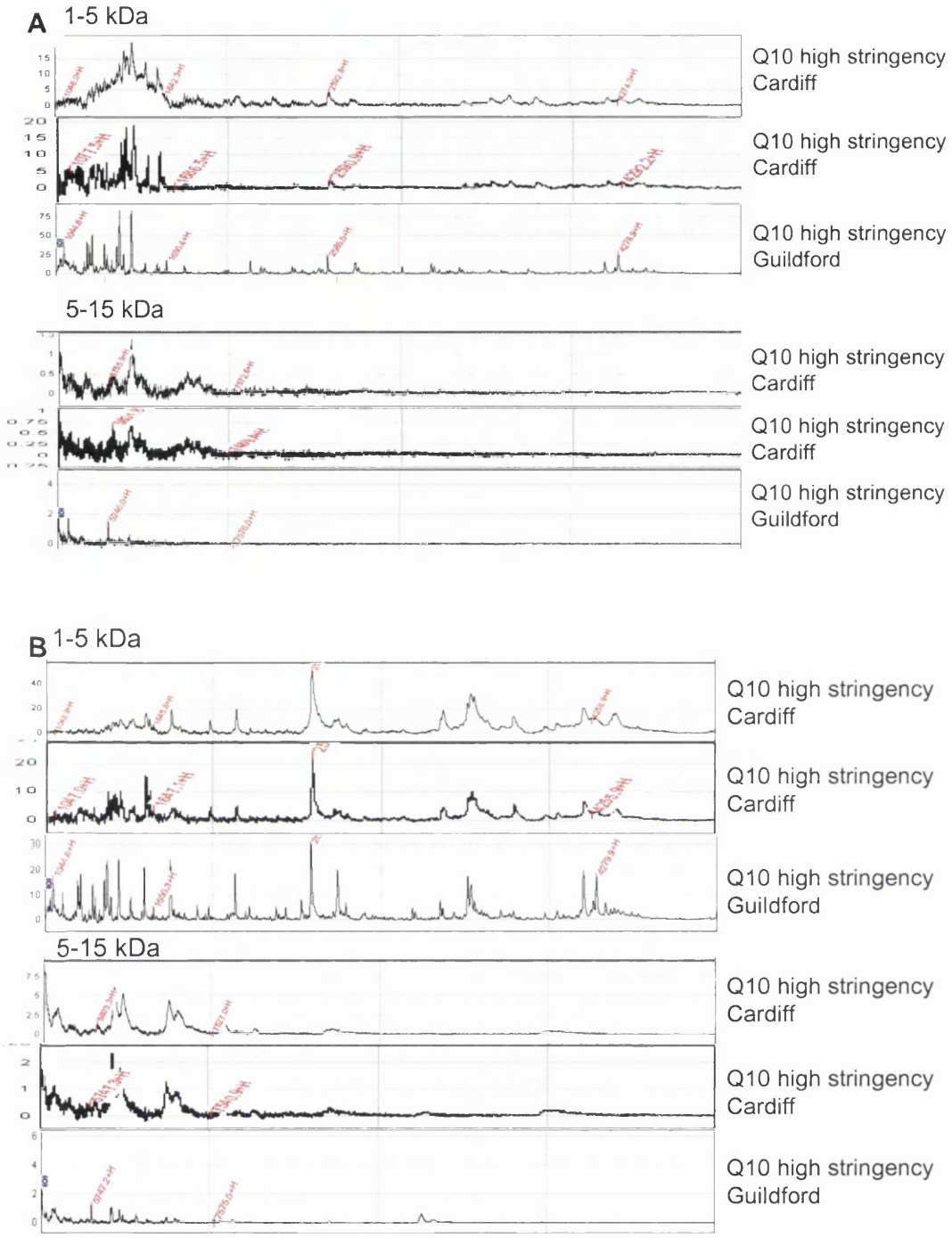


Figure 6.27: SELDI-ToF MS spectra of LMW two breast cancer serum sample, OP4 (**A**) and OP5 (**B**) on Q10 ProteinChips with high stringency wash. Comparison of one LMW serum samples prepared on chip in Guildford and two in Cardiff.

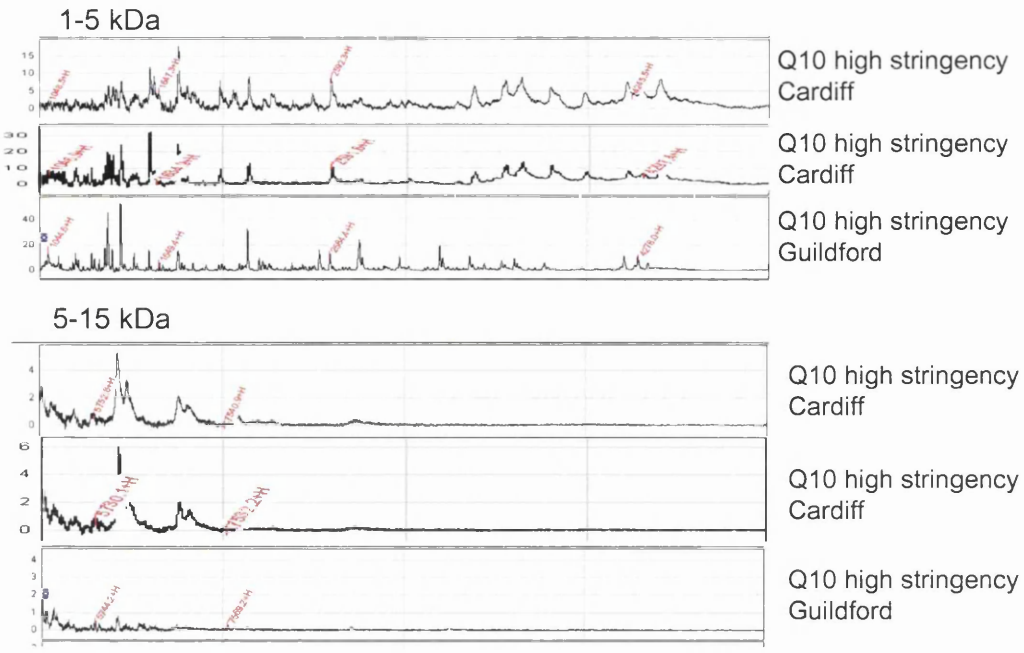


Figure 6.28: SELDI-ToF MS spectra of LMW control serum sample (V6) on Q10 ProteinChips with high stringency wash. Comparison of one LMW serum samples prepared on chip in Guildford and two in Cardiff.

Although the resolution of the chips prepared in Cardiff was of an inferior quality than the ones from Guildford they displayed the same mass peaks (Figure 6.27 and Figure 6.28). Therefore they were analysed using the Biomarker Wizard as described before and the spectra data was exported into Excel (Table 6.9). For statistical analysis the Guildford arrays were removed, to avoid variation. Hence for ½ the samples in each cohort, three replicates and for the other ½ two replicates were analysed. This allowed calculation of sample averages in Excel. The results from a *t*-test and the Mann-Whitney test *U* test are shown in Table 6.9. As before it appears that the *t*-test identifies less significant peaks and even less were discovered from the averaged samples. Each of the significant peaks was inspected visually and the ones that were truly discriminative are shown in Figure 6.30 and Figure 6.29. The low resolution of the peaks can be seen in the low intensity and also in the poor discrimination between the groups.

Table 6.9: Discriminating peaks, of S1 analysed in Cardiff, calculated using a Student's *t*-test in Excel and a Mann-Whitney *U* test as part of the Biomarker Wizard in the ProteinChip software.

All replicates					Sample averages		
m/z	All *p*-value (Excel)	fold-change	Mann-Whitney (Ciphergen)	fold-change	*m/z*	*p*-value (Excel)	fold-change
Q10 high stringency					*Q10 high stringency*		
1139.9	0.046	2.4			1139.9	0.015	2.4
1756.3	0.006	-1.5	0.007	-1.5	1756.3	0.002	-1.6
4292.9	0.032	1.2			4292.9	0.022	1.2
13911.4	0.027	7.1	0.035	7.0	13911.4	0.025	7.0
Q10 low stringency					*Q10 low stringency*		
2096.9			0.024	1.4	2095.9	0.027	1.3
2223.9			0.019	1.3			
2273.0			0.044	1.3			
2371.8			0.048	3.4			
2384.3			0.008	1.3			
2490.0			0.004	1.3	2489.9	0.014	1.4
2591.8	0.023	1.3	0.016	1.2			
2977.8			0.007	-1.4			
3604.0	0.035	-2.0					
5768.5	0.009	-2.0	0.002	-2.5			

As for the MALDI-ToF MS data, if a fold-change was very large, such as 7.0 for *m/z* 13911 on the Q10 (high stringency) arrays, it was usually due to one sample skewing the data with an unusually high intensity. Again this emphasises that visual validation of the markers is crucial, additional to statistical analysis. Four peaks were calculated to have significant *p*-values from the Q10 arrays prepared at high stringency, of those only 2 peaks were visually different between the breast cancer and the control cohort (Figure 6.29).

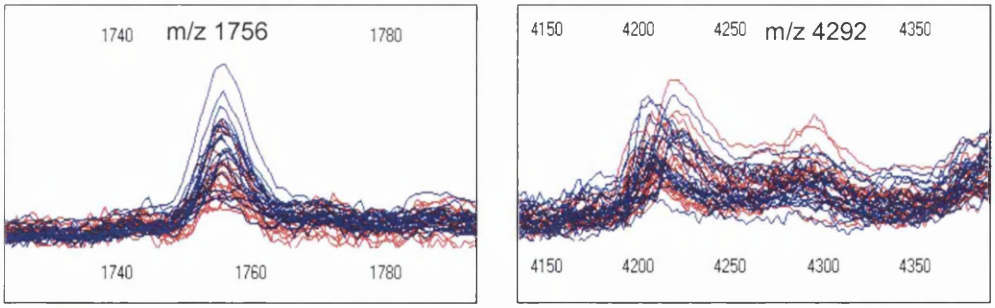


Figure 6.29: The two most discriminating peaks recovered from Q10 ProteinChips prepared at high stringency conditions. All spectra were overlaid, in red the breast cancer and in blue the control samples.

Of the 10 significant peaks on the Q10 (low stringency) arrays seven were visually confirmed (Figure 6.30). Only one (*m/z* 2490) of the significant peaks from the averaged data was actually different, looking at the spectra. Because the spectra processing and data analysis for these arrays was performed in Cardiff, all files were available. For that reason it was possible to get the information on fold-changes calculated from the means obtained form the Biomarker Wizard software.

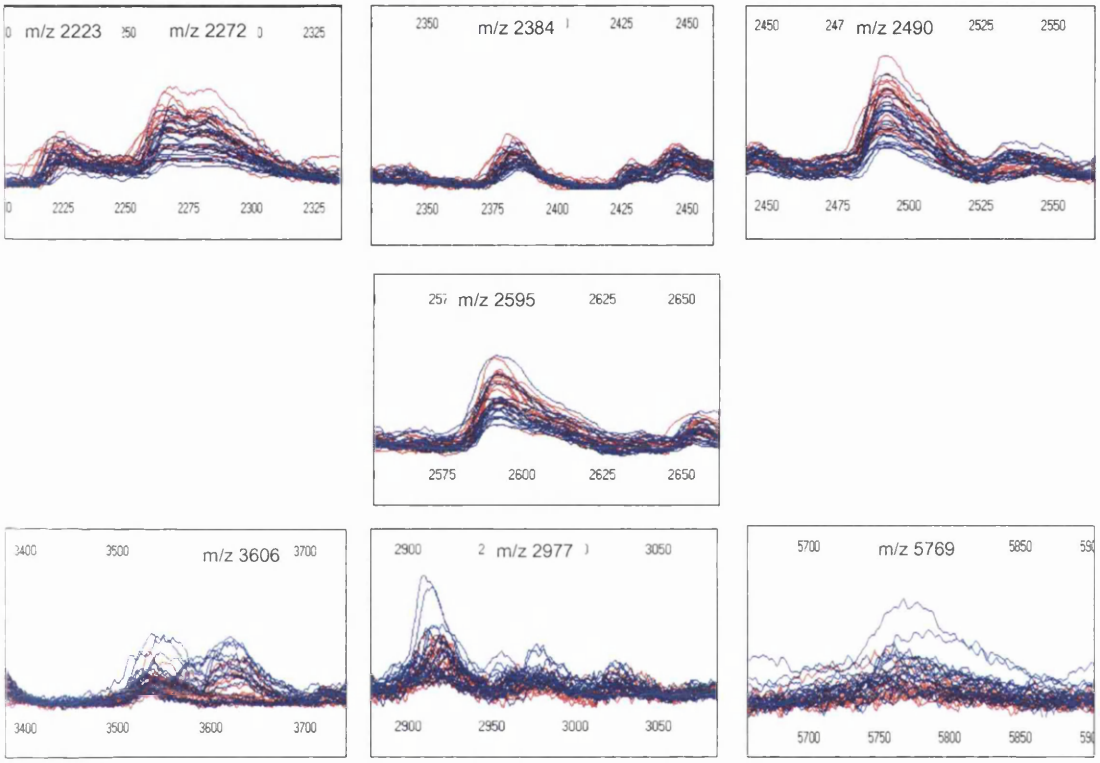


Figure 6.30: The 7 most discriminating peaks recovered from Q10 ProteinChips prepared at low stringency conditions. All spectra were overlaid, in red the breast cancer and in blue the control samples.

Visually the most obvious peak with a significant difference between breast cancer and the control sample was *m/z* 1756 detected from the chips prepared at high stringency. Because of the poor resolution of the peaks this is not visualised as well in the overlaid peak view. Figure 6.31 shows the difference more convincingly. However, here only 3 samples from each cohort were chosen, to not overcrowd the spectra.

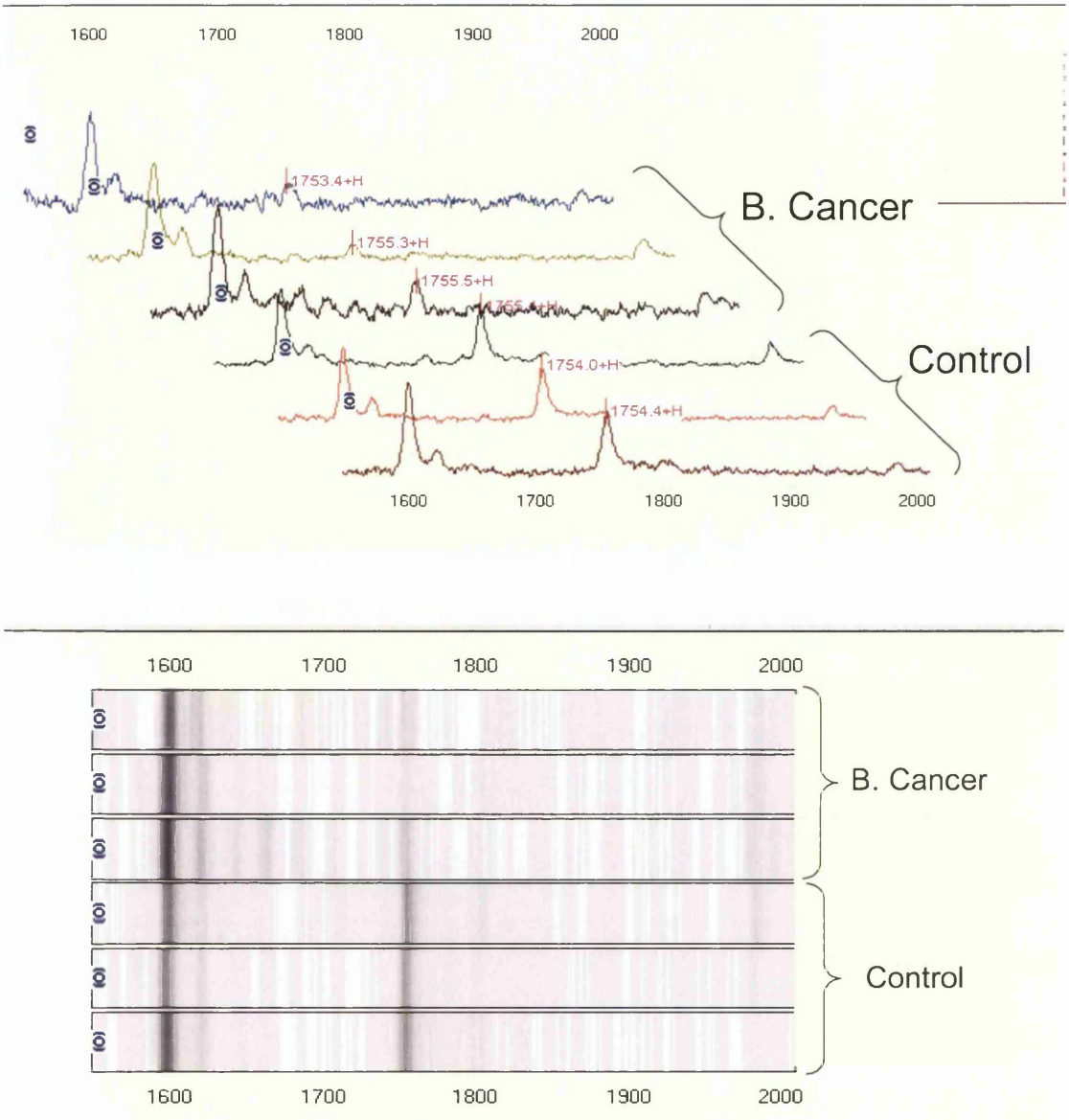


Figure 6.31: Alternative visualisation of the mass spectra from breast cancer and control samples. The peak at *m/z* 1756 is significantly decreased in the breast cancer cohort where as the peak at *m/z* 1600 retained the same peak intensity in both cohorts.

6.4. SELDI-ToF MS Analysis of Sample Set S2

After the failure of the analysis of the S1 sample set with more replicates in Cardiff and discussion with Ciphergen, it was suggested that the ProteinChips had expired and therefore a whole new sample set of LMW serum (S2) was prepared as described in Chapter 5 (section 5.5). Brand new chips and buffers were used for this analysis. The samples were prepared and analysed on Q10 and CM10 ProteinChips. However as can

be seen in Figure 6.32 and Figure 6.33 the resolution of the mass peaks in the SELDI-ToF spectra is very low and the individual peaks were not defined but had merged into broad lumps. These spectra were too poor to be analysed for biomarker discovery. Suggestions that this may be due to degradation were refuted as the MALDI-ToF MS analysis was performed a few weeks after the SELDI-ToF MS analysis on an aliquot prepared at exactly the same time. A number of experiments were preformed to investigate the reason behind this however no explanation could be found and in the interest of saving time and resources SELDI-ToF MS analysis was abandoned.

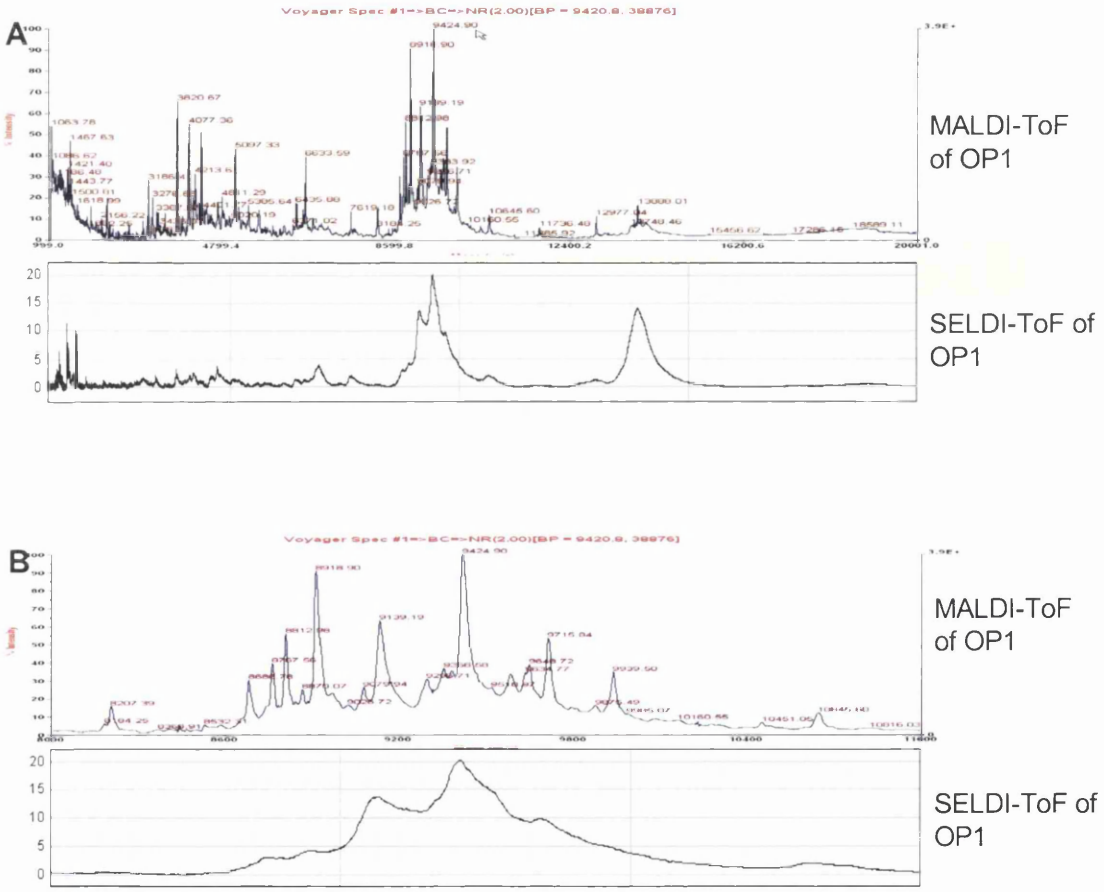


Figure 6.32: Comparison of MALDI-ToF and SELDI-ToF MS of the same sample. The LMW serum sample was Zip-Tipped for MALDI-ToF and prepared on a Q10 array at low stringency. The whole mass range 1000- 20000 Da (**A**) and a narrow range highlighting the peaks between 9000 and 11000 Da (**B**) are shown.

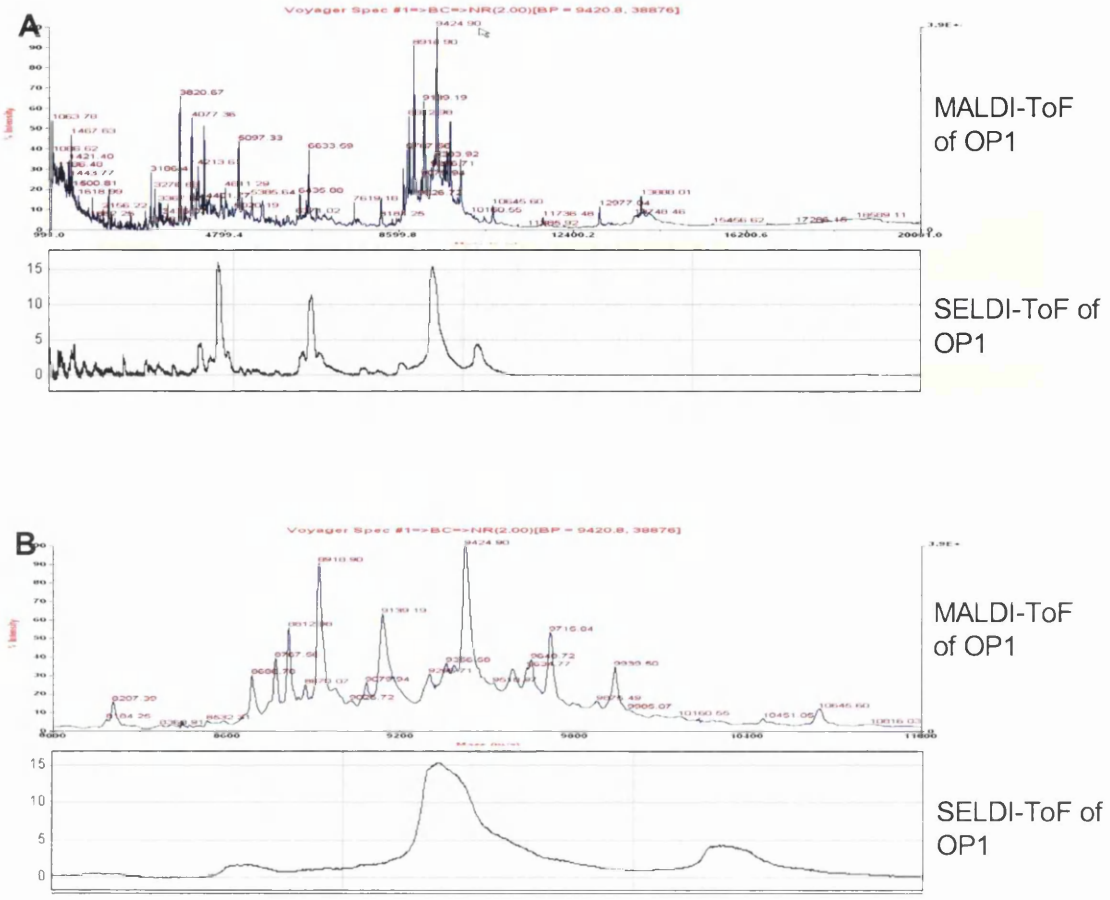


Figure 6.33: Comparison of MALDI-ToF and SELDI-ToF MS of the same sample. The LMW serum sample was Zip-Tipped for MALDI-ToF and prepared on a CM10 array at low stringency. The whole mass range 1000- 20000 Da (**A**) and a narrow range highlighting the peaks between 9000 and 11000 Da (**B**) are shown.

6.4.1. Possible Explanations for the Unsuccessful Experiments

A number of different reasons were explored to investigate why the SELDI-ToF analysis failed and to explain the low spectrum resolution. A fresh aliquot of neat serum was analysed un-fractionated mixed with different amounts of matrix on NP20 arrays (Figure 6.34). For all experiments above 2x 0.6 µl of SA matrix were used per spot; using more matrix (2x 1 µl) did not improve the spectra but made them worse, especially for the low mass range. The number of peaks detected was fewer than expected for a complex mixture such as neat serum. Therefore the serum sample was fractionated using WAX resin and each fraction was analysed on NP20 arrays, this type was chosen as it will bind all proteins similar to a MALDI-ToF target plate. As visible in Figure 6.35 few peaks were present in each of the fractions and the peak widths were broad. This shows that the problems were not related to the ultrafiltration but occurred with all samples tested and furthermore that the complexity of the serum is not the hindering factor since WAX fractionation could not improve the spectra. In any case, the LMW filtrates have a much lower complexity and are free of albumin, which has been a limiting factor in the past.

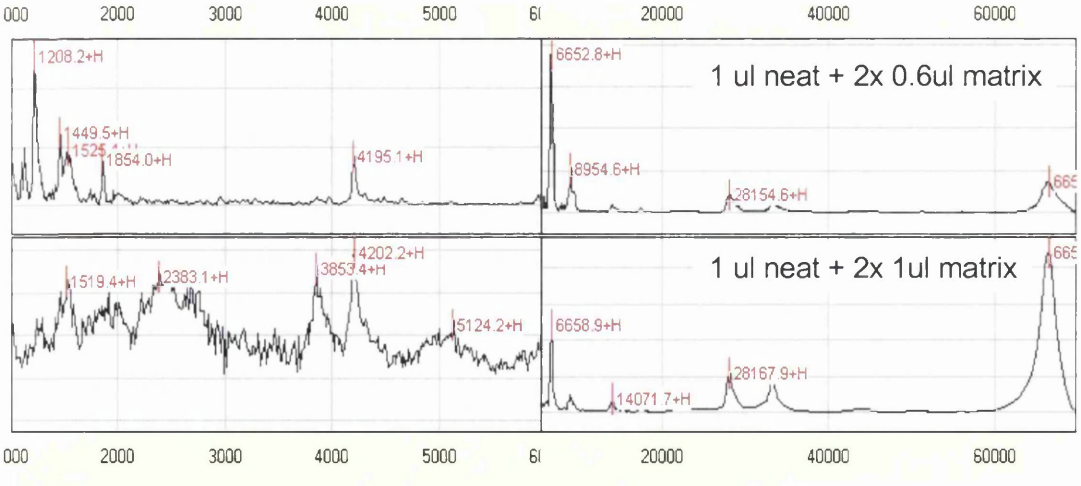


Figure 6.34: Neat serum from a different source was analysed on NP20 chips. The sample was prepared with different volumes of matrix. Very few peaks are present in each spectrum.

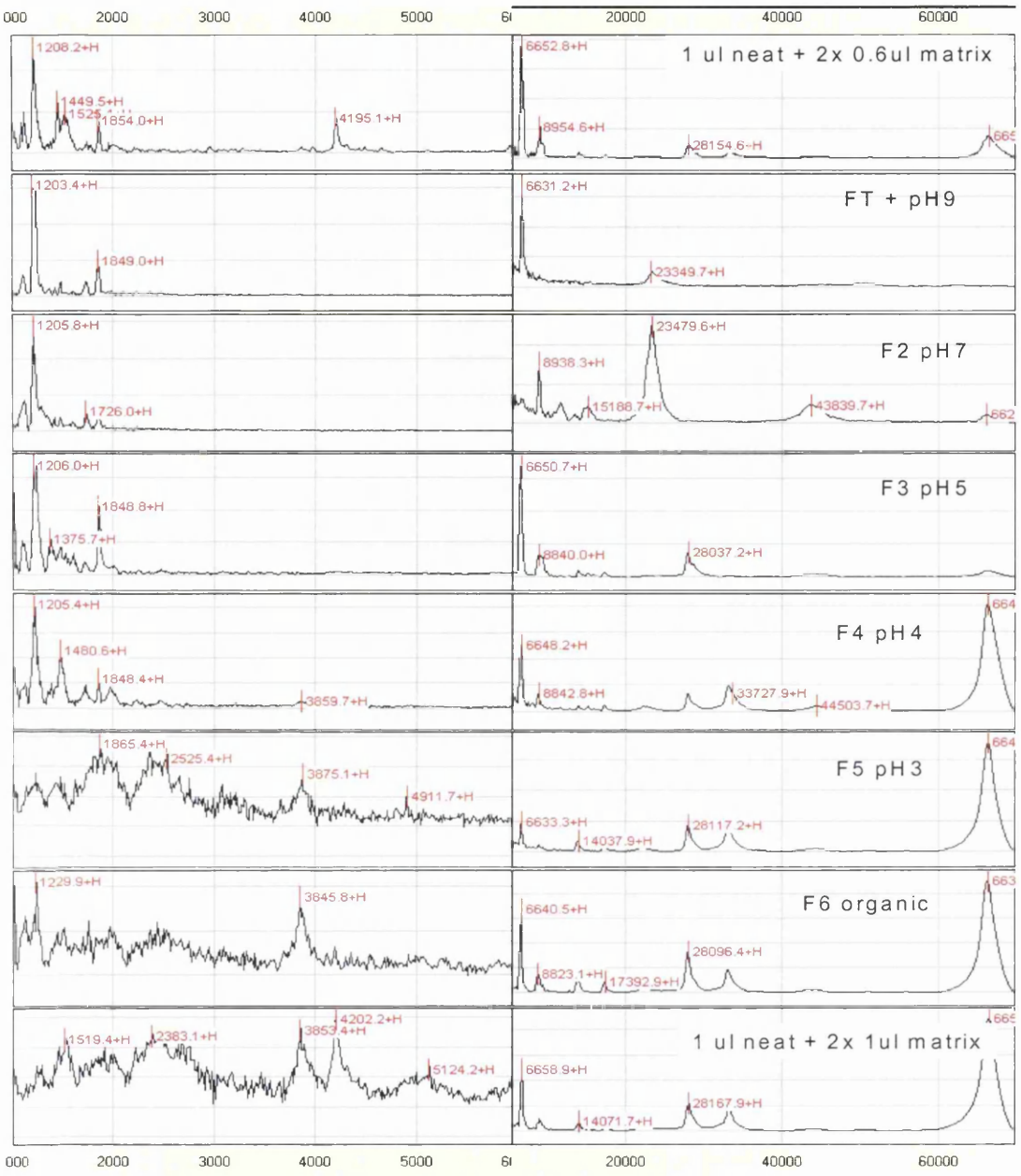


Figure 6.35: Neat serum fractions separated on WAX resin columns. The fractions along with some un-fractionated serum were prepared on NP20 arrays.

We already knew that the ProteinChip reader was functional, since the analysis of the samples prepared in Guildford was successful. New chips and buffers were used for preparation of the S2 set, furthermore these samples had been freshly prepared a day before the SELDI-ToF MS analysis. Also another aliquot of the same filtrates was analysed by MALDI-ToF MS a couple of weeks later and the spectra showed no evidence of degradation or reduced peak resolution. As shown in Figure 6.33 of the

previous section. In conclusion no explanation could be found for the problems and therefore SELDI-ToF analysis was abandoned and all samples from the S2 set were analysed with MALDI-ToF MS only. At the time of the experiments, very little information was available on biomarker discovery using MALDI-ToF MS; however in Chapter 5 we managed to develop a robust protocol for protein profiling. A comparison of the two techniques is described in the next section.

6.5. Comparison of SELDI-ToF with MALDI-ToF MS

Direct analysis of serum is limited by the complexity of the proteome and the great range of protein concentrations within the sample. Analysis of all serum proteins results in detection of the high abundant proteins only (e.g. albumin, immunoglobulins and transferrin). During mass spectrometry, ion suppression causes the highly abundant proteins, that ionize very well, to dominate the spectrum and the lower abundant proteins cannot be seen. The same is true for gel electrophoresis. Therefore serum was fractionated by UF and the LMW proteome, free from albumin and immunoglobulins was analysed. During the UF salts, small enough to pass through the membrane, may accumulate. To remove these the sample is washed with buffers on the SELDI ProteinChips (this is the standard protocol for ProteinChip preparation) and for MALDI-ToF each sample was Zip-Tipped on C18 tips (this is also standard for serum protein preparation for MALDI-ToF MS). Each type of the ProteinChip arrays has a different chromatographic surface, binding specific proteins, in the same way C18 Zip-Tips bind polar proteins. Hence slightly different proteins may be detected on the MALDI-ToF Chips compared to any of the ProteinChips. In Figure 6.36 a spectrum of a H50 array (C8 chromatographic surface) was compared to the MALDI-ToF spectrum of the same sample. The spectra show comparable numbers of peaks and some of the major peaks are the same, some peaks however occur in only one or the other spectrum. In the SELDI-ToF spectrum a peak for *m/z* 1756 is observed and also demonstrated to be a significant marker, this peak is widely absent or of very low intensity from all MALDI-ToF spectra. Another example from a different sample analyzed by MALDI-ToF and on Q10 arrays at high and low stringency is shown in Figure 6.37. A small mass shift was observed, which may be due to different calibration files during manual calibration. For example *m/z* 1441 in the SELDI-ToF spectrum is the same peak as *m/z* 1469 in the MALDI-ToF spectrum. The SELDI-ToF spectra in Figure 6.36 was taken from the good Chips prepared in Guildford, whereas the spectra in Figure 6.37 were taken from the poor S1 chips prepared in Cardiff. However both exhibit the mass shift and the presence of *m/z* 1756.

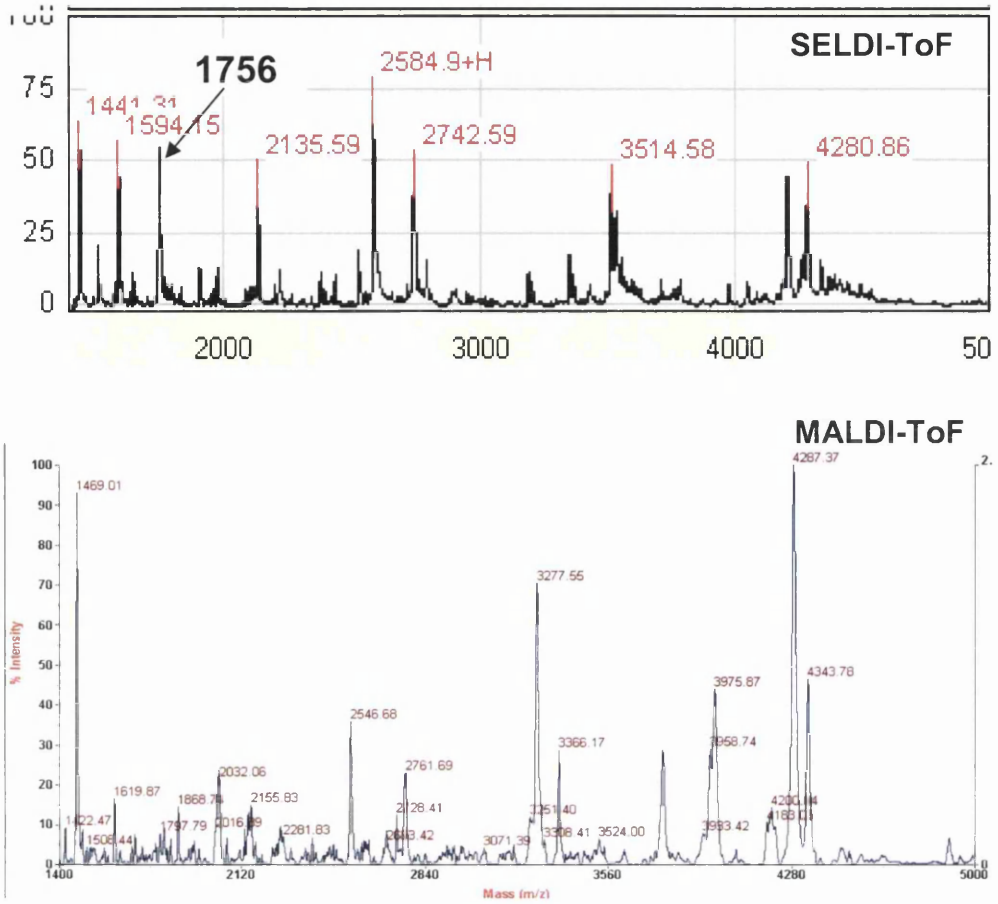


Figure 6.36: Compare MALDI-ToF and SELDI-ToF spectra from the same sample. For MALDI-ToF MS the sample was cleaned using C18 Zip-Tips and for SELDI-ToF MS a H50 array was used.

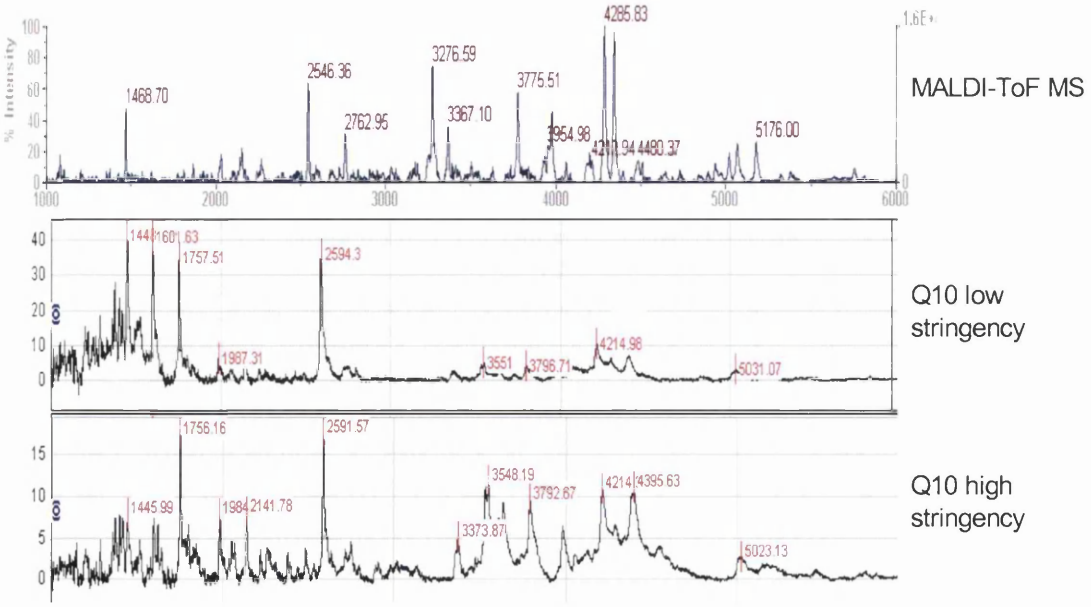


Figure 6.37: Comparison of MALDI-ToF MS spectrum with spectra from Q10 SELDI-ToF arrays.

For reasons described above different peaks were detected by MALDI-ToF and SELDI-ToF MS and therefore different significant markers were detected on each platform. In Table 6.10 *m/z* values that were found to be significantly different on both platforms are shown. These values were taken from across all the SELDI-ToF arrays and from the entire results table including all data (i.e. un-averaged, averaged and Markerview results) of the MALDI-ToF results. However the results from the averaged MALDI-ToF MS results were marked in bold, as these could be considered more robust (Table 6.10). SELDI-ToF "markers", listed in the first column of the table, that were visually confirmed from the spectra above, were also marked in bold. It is worth mentioning again that all the results shown from SELDI-ToF and MALDI-ToF MS analyses in this chapter were generated from the S1 sample set.

Identification of the discriminating peaks was attempted using MALDI-ToF/ToF by our collaborators at Applied Biosystems in Germany. Three peptide identifications were obtained and the proposed amino acid sequences are shown in Table 6.10. This was described in more detail in Chapter 5 (section 5.5.2). Although the database results only suggest homology to the protein matches. It is exciting to know that three potential markers were detected using SELDI-ToF and MALDI-ToF MS, and that they were also identified by MS/MS fragmentation.

Table 6.10: Discriminating *m/z* values retrieved from both SELDI-ToF and MALDI-ToF analysis. The peaks in bold from the SELDI-ToF results are peaks that were visually confirmed to be different and in the MALDI-ToF set the peaks marked in bold are *m/z* values that where significant in the averaged dataset in the Excel data. Three peptides were identified using MALDI-ToF/ToF and the retrieved amino acid sequence is shown.

Markers from SELDI -ToF MS			*Markers from MALDI -ToF MS*			
m/z	***p*-value**	**fold-change**	***m/z***	***p*-value**	**fold-change**	**peptide ID**
1034	**0.036**	**1.0**	**1064**	**0.014**	**1.5**	RPPGFSPFR
1226	**0.029**	**1.7**	1273	0.027	1.8	
1402	**0.005**	**1.5**	1400	0.044	1.8	
1636	**0.043**	**2.2**	1608	0.022	1.6	
1757	**0.021**	**-2.3**	**1776**	**0.008**	**1.0**	
1936	**0.050**	**-2.1**	**1932**	**0.014**	**-1.3**	AHYDLRHTFMGVVSLGS PSGEVSHPR
2464	**0.030**	**-1.6**	2441	0.019	C down	
2525	**0.021**	**-2.5**	**2556**	**0.037**	**-1.6**	
2791	**0.008**	**-2.1**	**2832**	**0.021**	**-2.1**	SLAELGGHLDQQVEEFR
2913	0.021	5.32	2925	0.038	1.5	
2997	**0.007**	**-2.6**	**2995**	**0.002**	**-1.9**	
3309	**0.029**	**-1.2**	3338	0.023	C down	
3555	**0.021**	**-1.7**	3594	0.024	C down	
5108	**0.025**	**-1.2**	5101	0.016	-1.1	

6.6. Discussion and Conclusions

Although the majority of the SELDI-ToF MS analysis was unsuccessful, the results obtained from the experiments in Guildford determined that using multiple array chemistries and binding conditions can increase the number of proteins and potential markers discovered; the best ProteinChip types were Q10 and CM10. The study further showed that pooling of the sample reduces the number of peaks recovered and may therefore mask small changes when comparing two clinical groups. However potential markers were visible. The benefits of fractionation and selective array chemistry, to reduce the sample complexity, were demonstrated; as significantly more proteins and markers were discovered from the WAX fractions.

The preliminary study showed that SELDI-ToF MS technology is reproducible and quantitative, however repeating the experiment in a separate laboratory failed.

Unfortunately it was not possible to repeat the experiment with more replicates to get better statistical confidence. In the same way as for the MALDI-ToF study, analysis of the S1 samples provided some interesting results and could recover markers that were significantly discriminating between breast cancer and control serum samples. However each serum sample was only prepared once by UF; any "outliers" or spectra pushing the data either way may be due to a biological difference but could just as well be due to a difference introduce during the UF process. Although the UF showed to be a robust and reproducible method in Chapter 4, using replicates can minimise any variation and confirm the biological significance of "outliers". Repeating the UF on new serum samples in triplicate yielded good filtrates that were successfully analysed using MALDI-ToF MS (see chapter 5), however the SELDI-ToF analysis, despite brand new ProteinChips and reagents failed completely. Nevertheless the results that were recovered from the S1 samples were compared to the markers retrieved from the MALDI-ToF analysis and encouragingly many of the markers overlap. It can be assumed that markers that are recovered from two completely different forms of analysis are more convincing and maybe regarded as a confirmation of the MALDI-ToF results.

Ideally, for protein identification, all samples would be separated by SDS-PAGE, corresponding molecular weight bands excised and digested with trypsin for identification using LC-MS/MS. However the proteins with significant *p*-values were very small, more like peptides and therefore visualisation on SDS-PAGE is not

straightforward. Alternatively these peaks can be analysed by MALDI-ToF/ToF, as was done with some success, for the markers recovered from MALDI-ToF analysis in the previous chapter. With more time and resources, this would be done for all potential markers. Furthermore each marker, if an antibody is available, should be validated using Western Blotting or ELISA. In the literature three potential markers for breast cancer (*m/z* 4.3, 8.1 and 8.9 kDa) discovered using SELDI-ToF have been described by Li *et al* [7, 15]; and again in a independent study by Mathelin *et al.* [16]. These proteins were not changed in our breast cancer samples. No proteins were discovered in the 8-9 kDa mass range.

6.7. References

[1] Chan, K. C., Lucas, D. A., Hise, D., Schaefer, C. F., Xiao, Z., Conrads, T. P., Janini, G. M., Beutow, K. H., Issaq, H. J. and Veenstra, T. D. (2004) Analysis of the Human Serum Proteome. *Clinical Proteomics* **1**, 101-226.

[2] Chertov, O., Biragyn, A., Kwak, L. W., Simpson, J. T., Boronina, T., Hoang, V. M., Prieto, D. A., Conrads, T. P., Veenstra, T. D. and Fisher, R. J. (2004) Organic solvent extraction of proteins and peptides from serum as an effective sample preparation for detection and identification of biomarkers by mass spectrometry. *Proteomics* **4**, 1195-1203.

[3] Georgiou, H. M., Rice, G. E. and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1**, 1503-1506.

[4] Mian, S., Ugurel, S., Parkinson, E., Schlenzka, I., Dryden, I., Lancashire, L., Ball, G., Creaser, C., Rees, R. and Schadendorf, D. (2005) Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. *J Clin Oncol* **23**, 5088-5093.

[5] Becker, S., Cazares, L. H., Watson, P., Lynch, H., Semmes, O. J., Drake, R. R. and Laronga, C. (2004) Surfaced-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. *Ann Surg Oncol* **11**, 907-914.

[6] Laronga, C., Becker, S., Watson, P., Gregory, B., Cazares, L., Lynch, H., Perry, R. R., Wright, G. L., Jr., Drake, R. R. and Semmes, O. J. (2003) SELDI-TOF serum profiling for prognostic and diagnostic classification of breast cancers. *Dis Markers* **19**, 229-238.

[7] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. and Chan, D. W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* **48**, 1296-1304.

[8] Pawlik, T. M., Fritsche, H., Coombes, K. R., Xiao, L., Krishnamurthy, S., Hunt, K. K., Pusztai, L., Chen, J. N., Clarke, C. H., Arun, B., Hung, M. C. and Kuerer, H. M. (2005) Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry. *Breast Cancer Res Treat* **89**, 149-157.

[9] Vlahou, A., Laronga, C., Wilson, L., Gregory, B., Fournier, K., McGaughey, D., Perry, R. R., Wright, G. L., Jr. and Semmes, O. J. (2003) A novel approach toward development of a rapid blood test for breast cancer. *Clin Breast Cancer* **4**, 203-209.

[10] Ciphergen Biosystems, I., ProteinChip® Applications Guide Volume 1: Introductory Guide, 2004.

[11] Dytham, C., *Choosing and Using Statistics: A Biologist's Guide*, Blackwell Publishing Co, Oxford 1999.

[12] Pletcher, S. D., Macdonald, S. J., Marguerie, R., Certa, U., Stearns, S. C., Goldstein, D. B. and Partridge, L. (2002) Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. *Curr Biol* **12**, 712-723.

[13] Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W. and Stromberg, A. J. (2003) Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* **4**, 26.

[14] Zhang, W., Carriquiry, A., Nettleton, D. and Dekkers, J. C. (2007) Pooling mRNA in Microarray Experiments and its Effect on Power. *Bioinformatics*.

[15] Li, J., Orlandi, R., White, C. N., Rosenzweig, J., Zhao, J., Seregni, E., Morelli, D., Yu, Y., Meng, X. Y., Zhang, Z., Davidson, N. E., Fung, E. T. and Chan, D. W. (2005) Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clin Chem* **51**, 2229-2235.

[16] Mathelin, C., Cromer, A., Wendling, C., Tomasetto, C. and Rio, M. C. (2006) Serum biomarkers for detection of breast cancers: A prospective study. *Breast Cancer Res Treat* **96**, 83-90.

CHAPTER 7

Biomarker Discovery using LC-MS/MS

The ability to quantitatively measure relative protein abundance between different clinical samples is essential for identifying candidate protein biomarkers; however as described above, most of this work has previously been done using intact proteins, by SELDI-ToF MS analysis. The use of LC-MS/MS for human serum or plasma profiling using enzymatically digested proteins was initiated by the first global shotgun proteomics study of human serum published in 2002 by Adkins *et al.* [1]. An explosion of LC-MS/MS-based applications in human serum and plasma soon followed due to the great interest in finding disease-related proteins [2-7]. Shotgun proteomics, where the proteins are digested prior to MS analysis, has now also been used in conjunction with isotopic ($^{16}O/^{18}O$), chemical (ICAT) or metabolic labelling for a global quantitative approach, quantifying relative protein abundance in plasma or serum [6, 8, 9]. This approach has been successfully used for pairwise comparisons where each sample is labelled with a different isotopic tag, however it can often be challenging to compare multiple samples or to use replicates due to the limited number of available labels.

Label-free quantitation instead makes use of the peak area or peak intensity as a measure for quantitation. This approach might be considered to introduce the least variation during sample preparation. Additionally, "label-free" direct quantitation provides greater flexibility for comparing multiple samples and for sample processing. This approach has been published quantitating peak-intensities [9-11]. However, in these papers, experiments were only used on small sample numbers or peptide

standards for method development. We therefore tried to take this approach further and apply it to serum samples, in a comparison of breast cancer and control samples.

In this chapter, our attempts to optimise and develop a LC-MS/MS-based protocol for protein profiling are described. As has been reported in a number of publications, the dynamic range of the sample is a key factor for successful analysis [12-16]. Hence the LMW sub-proteome (containing only 1% of the total serum proteins) appeared to be an ideal source for detection of potential markers. As a first step, the column reproducibility was assessed and the separation gradients and sample concentrations optimised. In a first attempt at quantitation, all samples from the S1 sample set (8 breast cancer and 8 control LMW serum samples) were trypsin-digested and separated by LC-MS/MS twice. The experiment setup is shown in Figure 7.1.

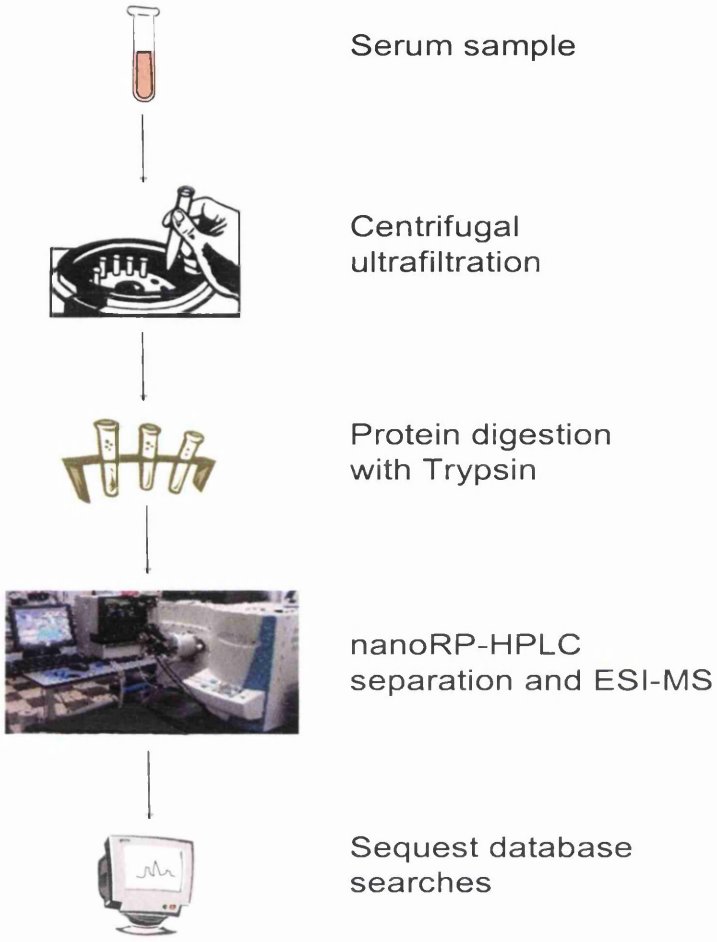


Figure 7.1: Experiment setup. Serum samples were prepared by centrifugal ultra-filtration, the filtrate was digested with trypsin and the peptides separated using a C18 pulled tip reverse-phase column spraying directly into the source of the ion-trap for MS/MS analysis. The fragmentation spectra were matched to peptides in the human FASTA database.

7.1. Optimization of Sample Preparation and HPLC separation

Although LC-MS/MS has been used extensively in the literature for peptide identification, we still thought it important to optimize the technique for quantitative analyses, especially since any experimental irreproducibility could introduce additional variation and influence the results.

For peptide separation using reverse phase C18 columns, fused silica capillaries were pulled into a tapered tip and packed with PepMap C18 reverse-phase stationary phase (Dionex, Camberley, UK) into 10 cm long columns (75 μm ID x 10 cm, 300 Å, 3 μm), as described in the Materials and Methods (section 2.8.1). Initially 20 μm porous particles were used as packing material; however, as seen in Figure 7.2, the peak resolution for these is not as good as when using 3 μm particles. The resolution was tested using a peptide mix of four peptides standards (bradykinin (*m/z* 531), Leu-ENK (*m/z* 556), GluFib (*m/z* 786) and angiotensin II (*m/z* 524)). Only in the separation using the columns with 3 μm porous particles were all four standard peptides eluted in separate peaks.

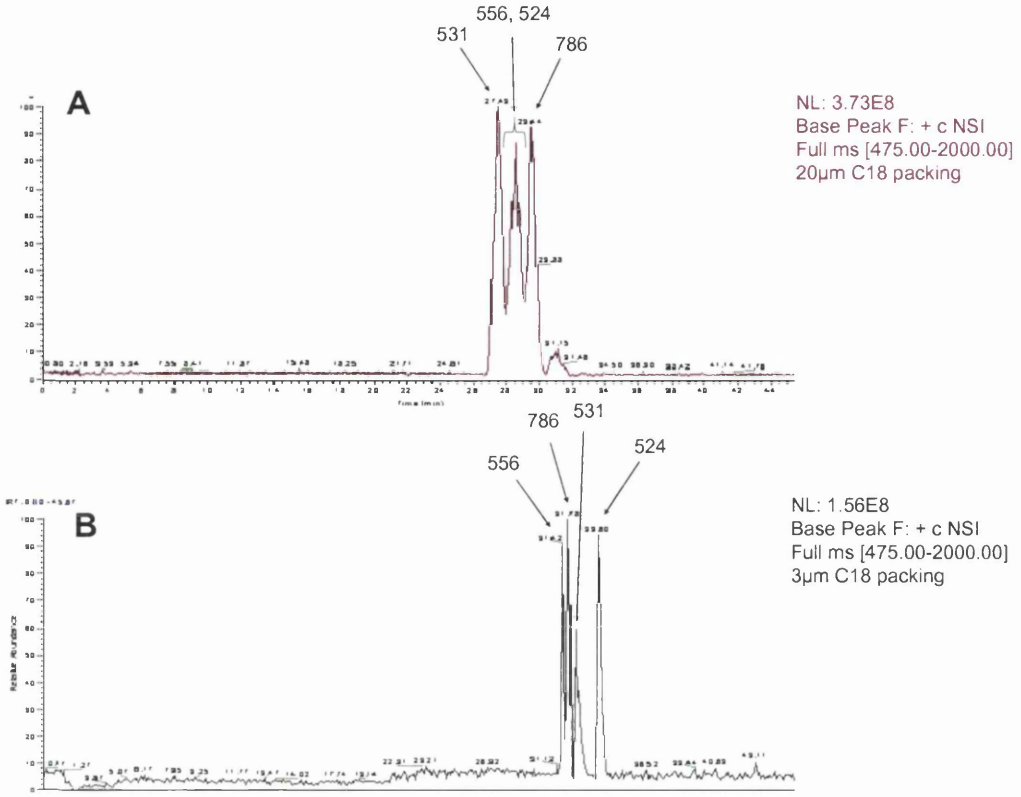


Figure 7.2: RP-LC-MS/MS separation of a test peptide mix: bradykinin (*m/z* 531), Leu-ENK (*m/z* 556), GluFib (*m/z* 786) and angiotensin II (*m/z* 524) using (**A**) 20µm and (**B**) 3µm porous particle C18 packing.

The use of small porous particles, small inner diameter columns and nano-electrospray interfaces has allowed more efficient separations with higher peak capacities [17]. By reducing the inner diameter of the column the flow-rate can be reduced and the analyte is concentrated as it elutes from the column. To reduce costs, all "nano" columns were packed by hand into pulled-tip capillaries. During the packing process, it was noted that although the MetaGuard® C18-packing material (Polaris, Varian Limited, Oxford, UK) produced columns with good chromatography, they were not as robust as the columns packed with C18 PepMap material (Dionex, Camberley, UK). Furthermore a heat source and constant stirring of the packing material was necessary to allow consistent packing of the columns, especially for longer columns and when the ambient temperature was low.

7.1.1. Column Reproducibility for Peptide Elution and Peak Area Detection

To investigate if repeated runs of the same sample on the same column were reproducible, initially a tryptic digest of a mixture of standards (250 fmol of BSA, cytochrome C (CytC), alpha-casein, and beta-casein) was separated and analysed using LC-MS/MS. The digest was injected 7 times and eluted using a continuous gradient from 2% to 98% ACN, 2% formic acid (FA) over 68 minutes (Figure 7.3).

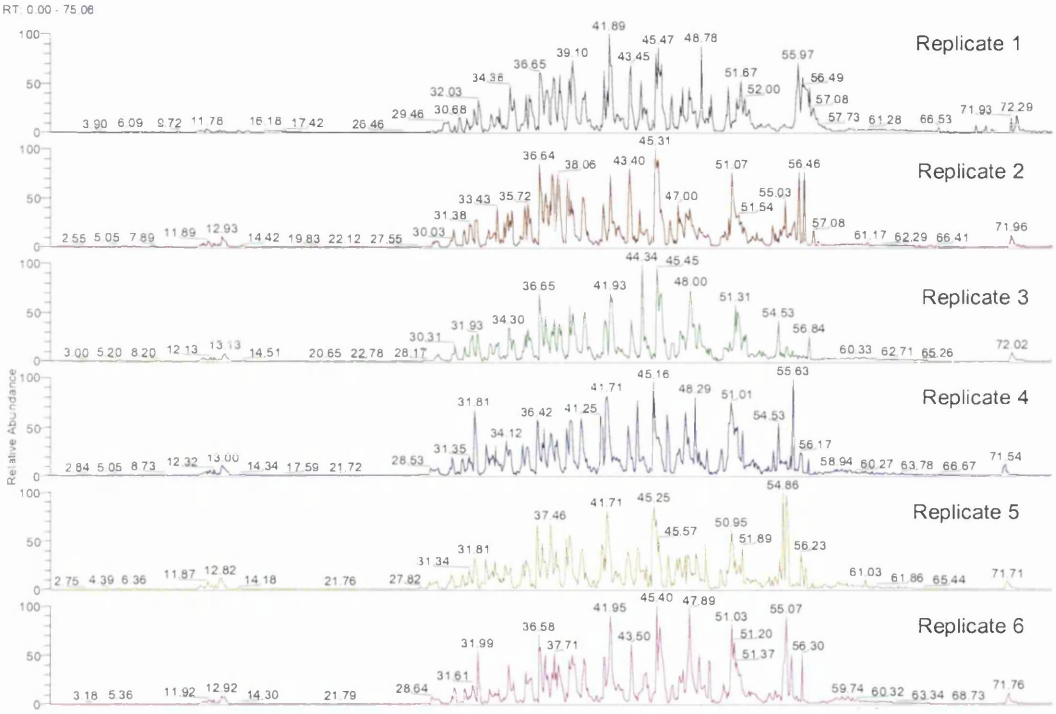


Figure 7.3: nRP-LC-MS/MS separation of LMW serum peptides from replicate separations. The basepeak chromatogram of each of the 6 replicates is shown.

The spectra were searched against the bovine FASTA database using TurboSequest, as part of the Bioworks Browser version 3.1, using the following filter criteria for high stringency cross correlation (X_{corr} 1+, 2+, 3+ = 1.8, 2.5, 3.2 and DeltaCn = 0.08). A number of peaks from the basepeak chromatogram were then selected for comparison of retention time (RT) and peak area (AA). The coefficient of variance (C.V.) was calculated for each of these peak values. A C.V. of < 20% has been shown to be a measure of good reproducibility. The C.V.s for retention time of each peak were

extremely reproducible and the majority of peak area values of the selected peaks showed good reproducibility (Figure 7.4).

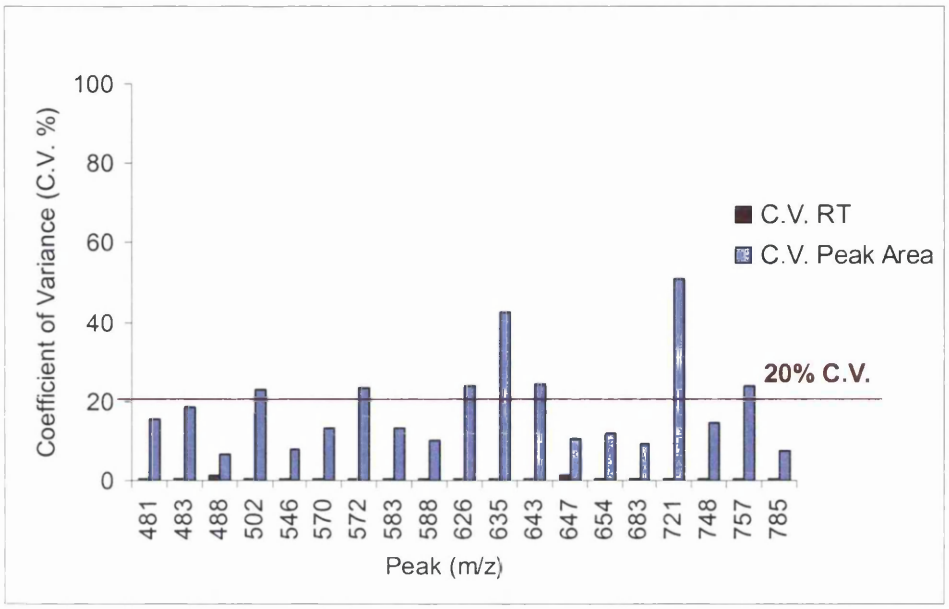


Figure 7.4: Reproducibility of the retention time (RT) and peak area for a number of selected peaks, measured by coefficient of variance (C.V.), across six replicate separations of the peptide digests of BSA, cytochrome C, alpha casein, and beta casein.

In Table 7.1 the amino acid sequence for each of the peptides in Figure 7.4 is shown to demonstrate that they were retrieved from tryptic digests of the standard proteins in the mixture. Additionally the confidence levels of peptide identification are shown in the form of a cross-correlation (X_{corr}) and delta correlation (dCn). Although the peptide identifications for the peaks chosen for evaluation of reproducibility do not show very high levels of confidence, however we know that only these proteins were digested in the mixture of standards. Furthermore the peaks were merely used to assess possible variations in retention time or peak area.

Table 7.1: Peptide identifications of the digest mixture of standards separated by LC-MS/MS and identified in the bovine FASTA database using Sequest. Each *m/z* value was matched to one of the standards. Confidence levels are given in the form of a cross-correlation (X_{corr}) and delta correlation (dCN) value.

m/z	xCorr	dCn	charge	SwissProt ID	Sequence
481.0	3.45	0.25	3	BSA	(R)RHPEYAVSVLLR
483.0	2.30	0.04	2	CytC	(R)EDLIAYLK
488.0	2.50	0.20	2	BSA	(K)DLGEEHFK
501.9	2.14	0.35	2	BSA	(K)LVVSTQTALA
545.5	1.12	----	1	BSA	(K)VASLR
569.7	2.27	0.00	2	beta-casein	
572.2	2.60	0.20	2	BSA	(K)KQTALVELLK
582.7	3.58	0.31	3	BSA	(K)LVNELTEFAK
587.7	3.94	0.02	3	alpha-casein	(K)HQGLPQEVLNENLLR
625.6	3.10	0.09	2	BSA	(R)FKDLGEEHFK
634.5	1.53	----	1	CytC	(K)IFVQK
642.6	1.72	0.32	2	BSA	(R)HPEYAVSVLLR
646.5	1.62	0.16	1	beta-casein	(K)EAMAPK.H
653.7	3.10	0.33	2	BSA	(K)HLVDEPQNLIK
682.7	3.98	0.02	3	BSA	(R)RHPYFYAPELLYYANK
720.5	3.03	0.35	2	BSA	(R)RHPEYAVSVLLR
748.4	1.07	0.16	1	alpha-casein	(K)TTMPLW
756.5	2.88	0.14	2	BSA	(K)VPQVSTPTLVEVSR
785.2	4.67	0.45	2	BSA	(K)DAFLGSFLYEYSR

This experiment was then extended to look at LMW peptides from serum, separated by using the same method as for the peptides. For quality control, the retention time was first checked using bradykinin. The LMW serum peptides were injected and separated 7 times, and the individual spectra are shown in Figure 7.5. In Figure 7.6, the C.V.s for the retention time and the peak area for a number of selected peaks that occur in all spectra, are shown. The majority of peptides elute in a reproducible manner with an average C.V. of 20% for the peak area reproducibility. As for the standards the reproducibility of the retention time of each peak was excellent.

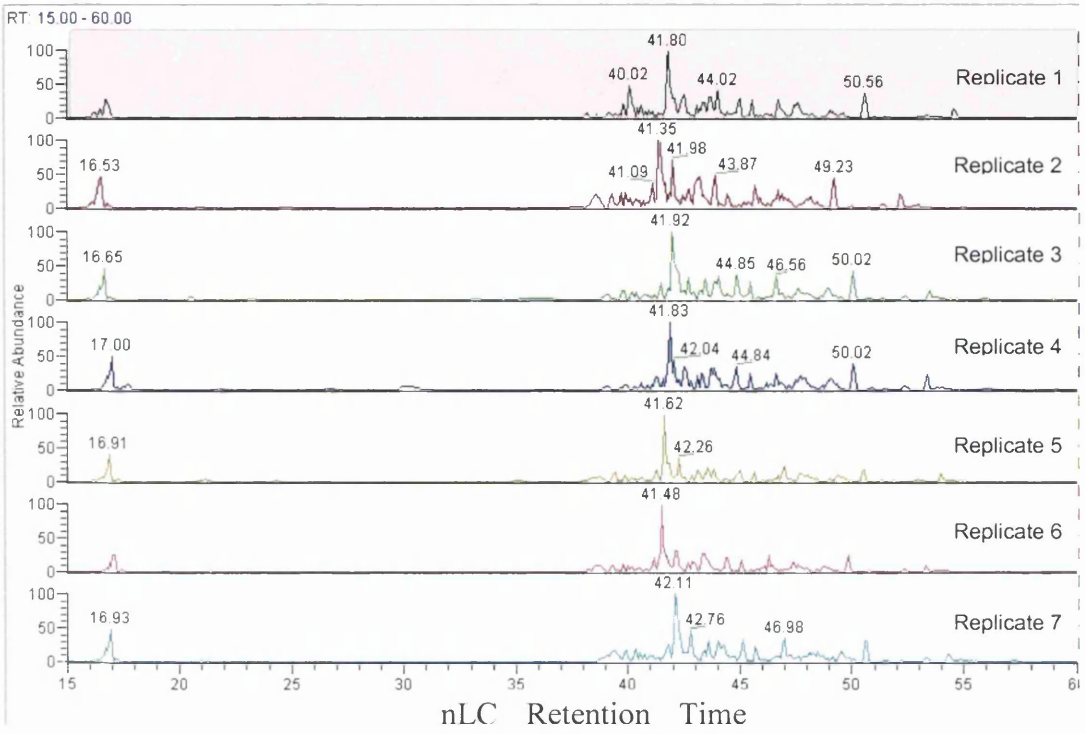


Figure 7.5: nRP-LC-MS/MS separation of LMW serum peptides from replicate separations. The basepeak chromatogram of each of the 7 replicates is shown.

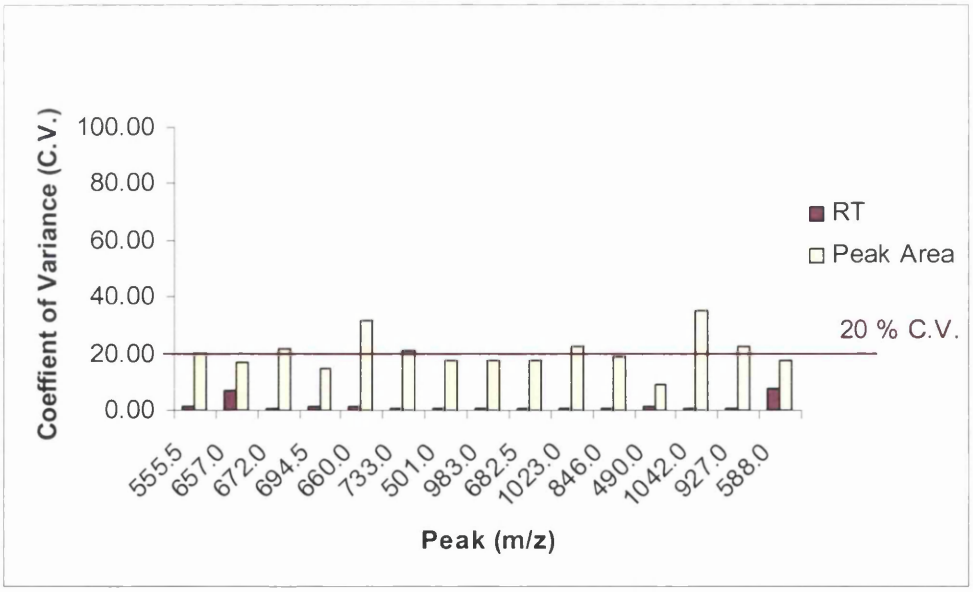


Figure 7.6: Reproducibility of the retention time (RT) and peak area for a number of selected peaks, measured by coefficient of variance (C.V.), across seven replicate separations of the peptide digests of LMW serum.

7.1.2. Elution Gradients for Optimal Peptide Separation

From the results described above, and particularly in Figure 7.5, it became apparent that the chromatography could be improved to allow further separation of the LMW serum peptides and to further improve peptide identification. Enabling greater separation of eluted peptides may also reduce the effect of ion suppression and the exclusion time common to LC separations. Therefore a number of elution gradients were tested, separating LMW serum peptides. It was discovered that the gradient has a strong influence on the way that peptides bind and elute from the column. To assess the differences between the gradients, the chromatograms were visually inspected and the peak resolution for a number of selected peaks was calculated. From previous results, it was observed that most peptides elute from the C18 column during the 30 and 50% organic phase (ACN, 2% FA). Hence it was hypothesised that the gradient in the phase up to 30% ACN could be shorter and steeper to allow more time for the peptides to elute at an organic phase above 30% ACN, without increasing the overall time of the analysis. However as shown in Figure 7.7, most of the peptides actually elute as soon as the steep gradient sets in. A shallower gradient is therefore required from the start to ensure binding to the column and good retention times.

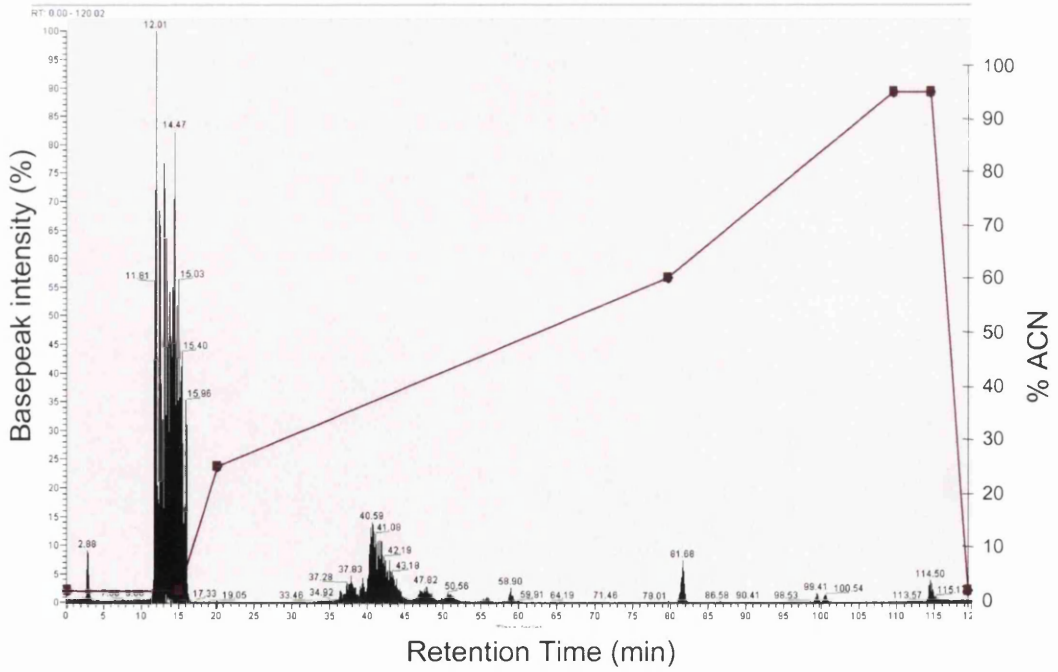


Figure 7.7: Neat serum separated on steep gradient to increase the actual separation time between 30 and 50% ACN.

Increasingly shallow and more "gentle" gradients were therefore tested. As can be seen in Figure 7.8, where the elution gradient itself is shown in red, the peaks became increasingly resolved the shallower the gradient.

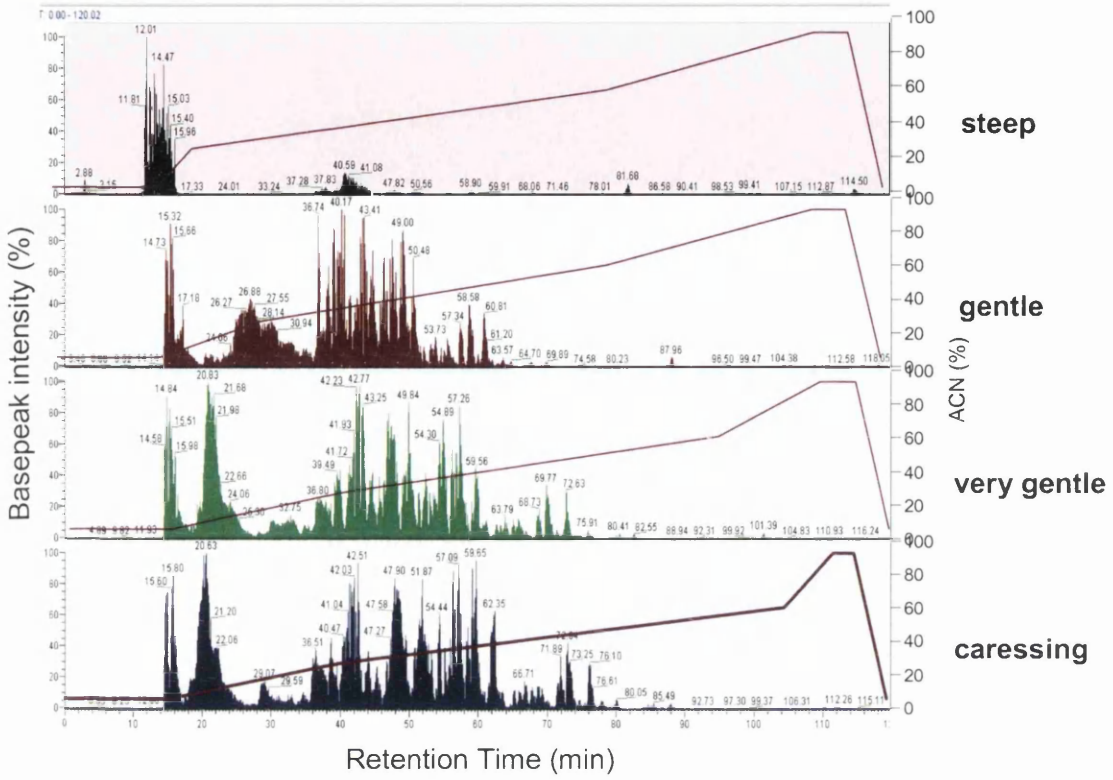


Figure 7.8: Un-diluted LMW serum separated with four different gradients. Peptides were not retained on the column using the steep gradient. However they were retained and separated out using the three more gentle gradients. The separation gradient itself is shown in red.

Additionally the peak resolution of three peak distances for the "gentle", "very gentle" and "caressing" gradient are shown in Table 7.2.

$$R_s = 2\left(\left(t_R\right)_B - \left(t_R\right)_A\right) / W_A + W_B$$

The peak resolution (R_s) can be described by the above equation, where A is the first of two peaks and B the second, which elute at times $(t_R)_A$ and $(t_R)_B$, and W the widths of each peak.

As seen in Table 7.2 no one gradient performed optimally for all peaks analysed. However the very gentle gradient on average resulted in the best peak resolution for

the three peak ranges and produced separation chromatograms with symmetrical, narrow peaks. It is possible that the separation in the caressing gradient is too shallow, resulting in broader peaks. Therefore for future LC-MS/MS experiments, the very gentle gradient was chosen.

Table 7.2: Peak resolution of the three different gradients for three separate areas in the spectrum. In each case, two peaks that elute very close together were chosen. The greater the R_S value, the better the resolution.

Peaks	Gentle (Rs)	V Gentle (Rs)	Caressing (Rs)
m/z 737 (A) - 1101 (B)	1.03	0.79	0.51
m/z 1135 (A) - 1103 (B)	0.46	1.01	0.52
m/z 886 (A) 1006 (B)	0.70	0.77	0.89

7.1.3. Peptide Concentration for Optimal Loading

The chromatograms in Figure 7.8 showed signs of column overloading, as many peptides eluted early at 20 minutes, when the ACN concentration is still low. However, the same peptides were also eluted again later in the gradient. Overloading of the C18 columns can cause early elution of peptides that do not bind strongly to the column, and also peak broadening. Finding the optimal peptide concentration allows binding of as many peptides as possible to the column, while still binding enough sample to analyse as many peptides as possible by tandem MS and finally retain enough sample to perform multiple separations.

A LMW serum filtrate was therefore trypsin-digested and loaded onto the C18 RP-column in a series of dilutions. The initial peptide concentration was determined by BCA assay to be 0.88 mg/ml and a dilution series of 1:5, 1:10, 1:50, 1:100, and 1:500 was created. For each sample, 7 µl were loaded onto the column and separated along the "very gentle" gradient, increasing stepwise from 2% to 98% B (ACN, 2% FA) in 105 min (Figure 7.9). From the concentrated serum sample, 6.16 µg of peptides were injected, whereas only 0.012 µg were loaded onto the column from the 1:500 diluted sample. The retention time of the concentrated sample was slightly shifted compared to the others, with peptides eluting earlier than in the other dilutions, additionally, a

large and wide peak is visible at 20 min (Figure 7.9). These factors together indicate that the column was overloaded and these peptides were eluted at a lower percentage of organic-phase. The spectra from dilutions of 1:5 and 1:10 looked similar to each other and the TIC intensities were $2.3x10^9$, and $1.4x10^9$, respectively. The dilutions 1:50 and 1:100 also looked very similar, with maximum basepeak intensities of $4.3x10^8$ and $5.8x10^8$ respectively, but fewer peaks were observed in these than in the higher concentrations.

Injection of 1.24 µg (1:5 dilutions) of LMW peptides appeared to overcome the effect of overloading and the peaks are more spread out and narrower. The dilutions of 1:50 to 1:500 showed a peak cluster (RT: 56 - 60 min) that has similarity to a polymer common in LC-MS/MS spectra, and the intensity of the serum peptides is too low to be visible compared to the polymer. Hence approximately 0.62 µg (1:10 dilutions) were necessary for separation of the LMW peptides without interference of background contaminants or elution problems. Shen *et al.* [18] also reported that overloading of the column can reduce the number of identified proteins. They further stated that 0.5 – 2.5 µg of sample was optimal for peptide identification; which is very similar to our findings.

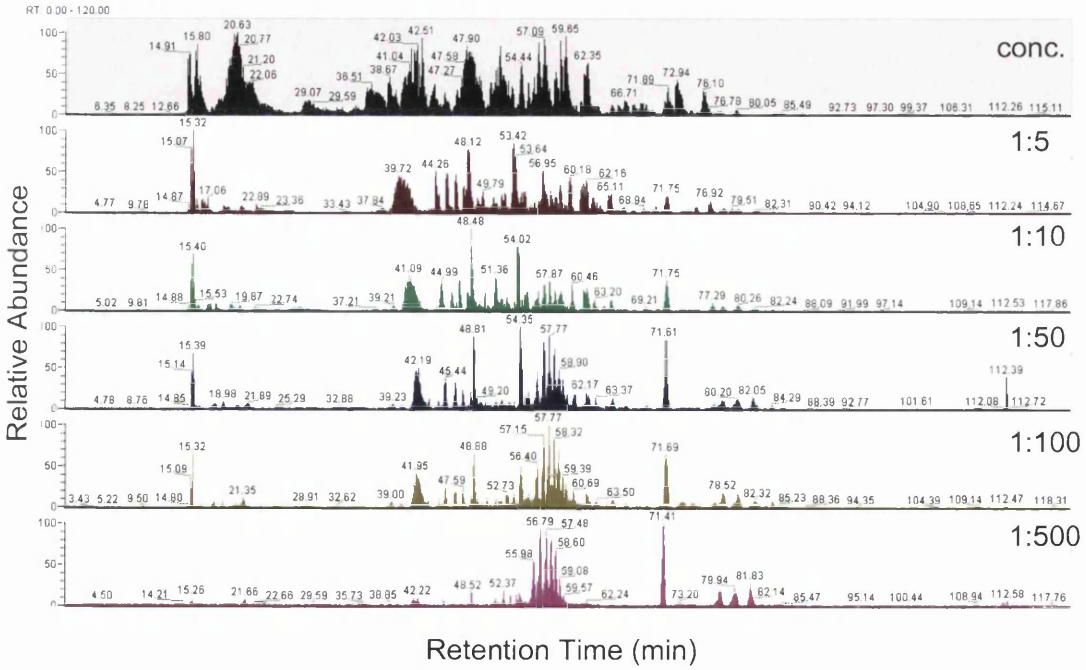


Figure 7.9: Whole basepeak chromatogram of a dilution series of a concentrated LMW serum peptide sample (conc.) separated over 105 min from 2% to 98% ACN.

After LC-MS/MS separation, the number of .dta files (hence the number of peaks selected for MS/MS) was used to gain a measure of the efficiency of peptide recovery (Figure 7.10). Since the MS/MS analysis was performed in data-dependent mode, excluding precursors that had been fragmented three times within a given time window, it can be assumed that most .dta files correspond to different peptides.

Also seen in Figure 7.10 the number of peaks selected for MS/MS analysis decreased with the amount of protein loaded onto the column: a sudden drop is observed between 1.24 µg and 0.62 µg of protein. We tried to detect any peaks that were present in the neat LMW serum but not in the 1:500 dilution but could not find any. This indicates that the same peak information is present in the low concentration samples as in the very concentrated samples.

The number of MS/MS scans increased with every replicate separation (Figure 7.10). This cannot be readily explained, however the spray may have stabilized during the course of the experiment or the background noise could have reduced during the course of the experiment. Nevertheless, the same trend of increasing MS/MS scans with increasing amount of peptides was observed in every run.

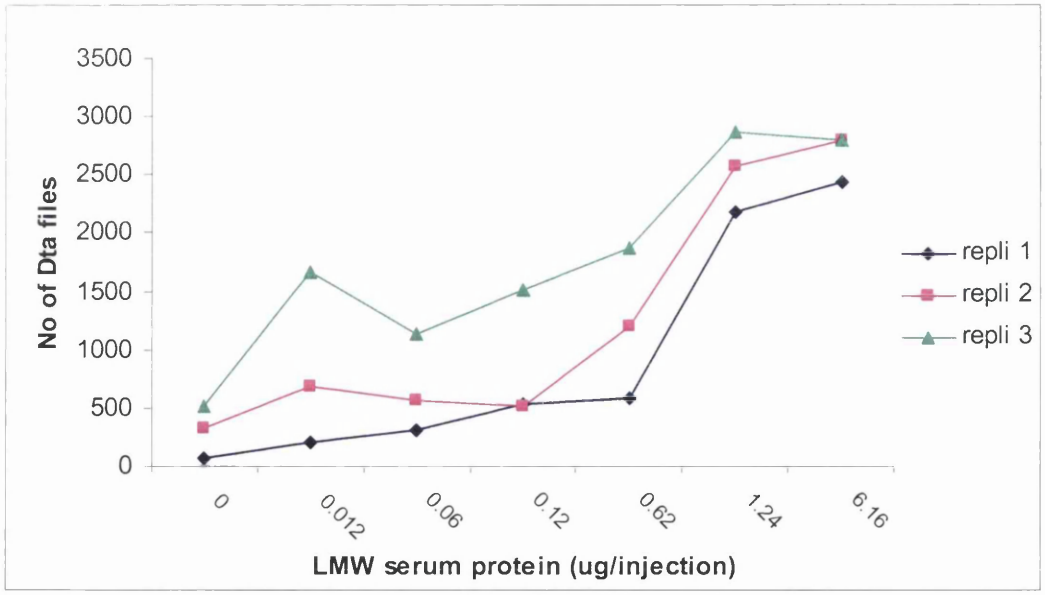


Figure 7.10: The number of peaks selected for MS/MS fragmentations across three replicate separations. The number of .dta files increases with every replicate run, however the trend of the slope is the same.

To investigate whether the differences between the three replicate runs had an influence on retention time and peak area, a number of selected peaks from the 1:500 dilution were observed for retention time and peak area (Figure 7.11 a and b). It appears that the peaks eluted earlier in the third replicate for the majority of peaks (Figure 7.11 a), however this can be compensated for in the spectra by adjusting the baseline. However, although the number of .dta files showed a trend, increasing across the replicate separations, the peak area of the chosen peaks varied randomly across the three replicates and no trend is visible (Figure 7.11 b).

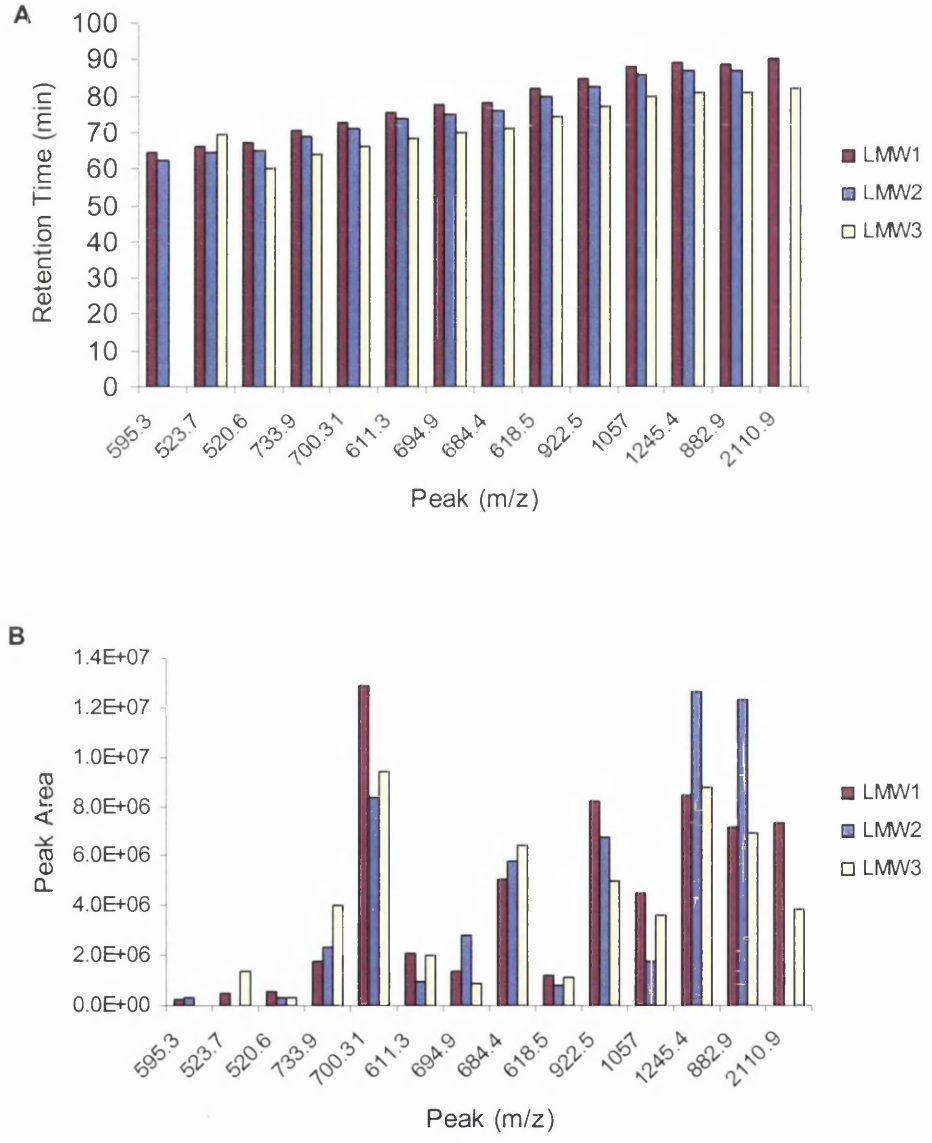


Figure 7.11: Replicate separation of LMW serum peptides (12 ng/injection). A number of selected peptides were analysed with regards to retention time (RT) (**A**) and peak area (**B**).

Additionally, we looked at various individual peaks and observed how they perform in the dilutions: a decrease in peak intensity and area with increasing dilution is visible in Figure 7.12. The peak intensity was adjusted to an absolute intensity of 6 $x10^8$ for the un-diluted LMW serum and 1 $x10^8$ for the dilutions to visualise the decreasing peak height of *m/z* 668.4.

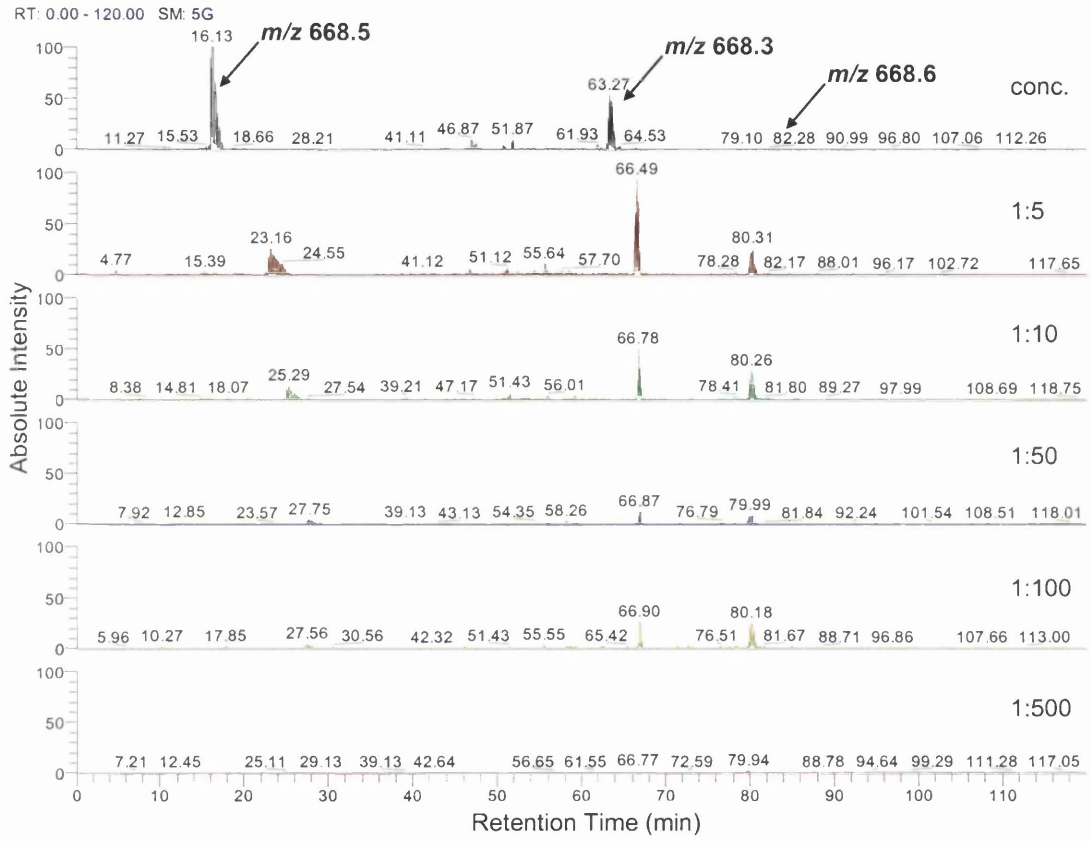


Figure 7.12: The extracted ion chromatogram of *m/z* 668.4 for the dilution series from 6.16 – 0.01 µg of LMW serum. The peak intensity was adjusted to an absolute intensity of 6 $x10^8$ for the undiluted serum and 1 $x10^8$ for the dilutions to enable visualisation of the decrease in peak height.

Furthermore, in Figure 7.13, the relative intensity for each concentration is shown in the extracted ion chromatogram (XIC) and in this way the change in area of the three peaks that elute at different times in the mass range *m/z* 667.9-668.9 can be seen. Three peaks are visible in each chromatogram, at 24 min (*m/z* 668.5), at 67 min (*m/z* 668.3) and at 80 min (*m/z* 668.6). The intensity of each of these in relation to one another changed with decreasing peptide concentration: a shift in the retention time

was visible between the concentrated and the diluted samples (Figure 7.13). A change in the relative peak height can also be seen in the spectra as the samples became increasingly more dilute.

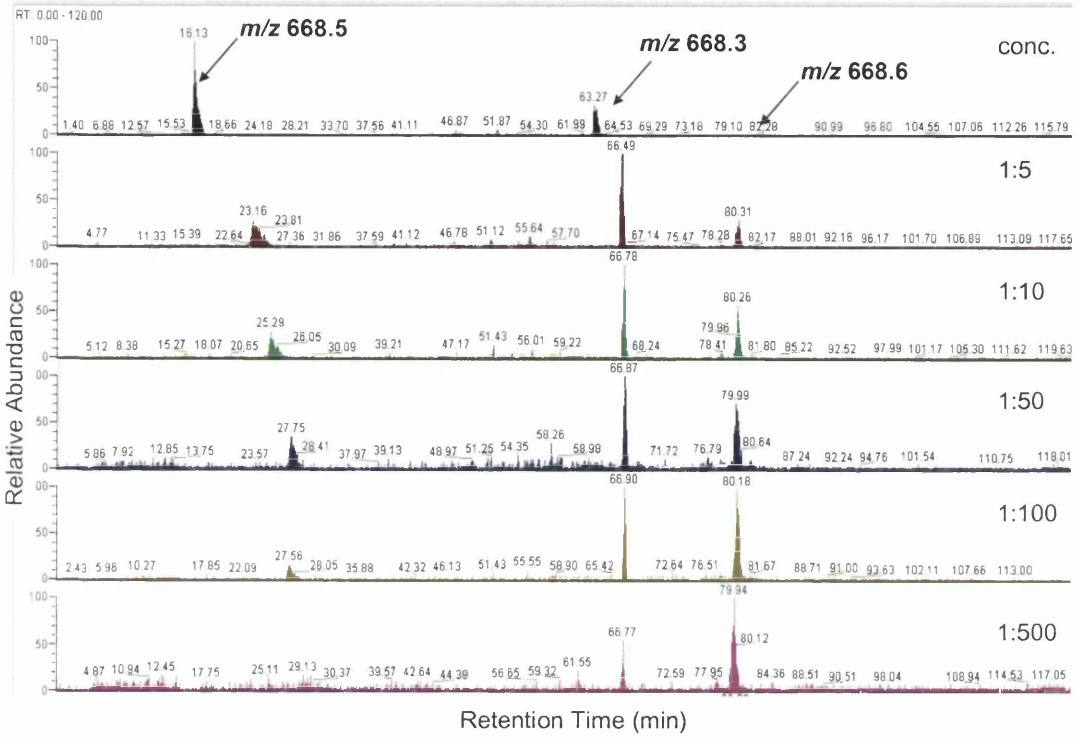


Figure 7.13: The extracted ion chromatogram (XIC) of *m/z* 668.4 for the dilution series from 6.16 – 0.01 µg of LMW serum. The peak intensity was adjusted to the relative basepeak, showing the highest peak in the spectrum at 100%. This enables visualization of the different peaks in the XIC of *m/z* 667.9-668.9.

The results showed that finding a balance between overloading of the column and recovery of a sufficient number of MS scans is crucial: this was achieved by injecting 1.2 µg of peptides onto the column. For separation of larger amounts of protein, a larger column, either in terms of length or inner diameter, would be required. However larger columns may cause the peak resolution to suffer.

7.2. Label-Free Quantitation of Peptides for Biomarker Discovery

The effectiveness of label-free peptide quantitation was tested using peptides (bradykinin, LeuENK, ß-glufibrino-peptide and angiotensin II) spiked into LMW serum samples. From the XICs, the mass peaks were detected and quantitated by manual integration of the peak area, as well as by using the Bioworks Browser version 3.2 (Thermo Finnigan, UK) to identify the MS/MS spectra and for automated quantitation of the peak areas. An aliquot of 3 µg of LMW serum peptides was spiked with a series of concentrations of each of these standard peptides (0.5, 50, 100, 150 and 200 fmol/ injection). Each sample was injected and separated twice and the results were combined. In the Bioworks Browser, the area for each peptide identified from the FASTA database using TurboSequest, was computed. For the standard peptides, since the *m/z* for each was known, each peak was actively searched for in the XICs for manual peak integration. The results using both approaches are shown in Figure 7.14. As shown, some of the values from Bioworks Browser are missing due to lack of identifications from the MS/MS spectra. All peptides show good linearity except for angiotensin II, this may ionize slightly better. Similar results have been shown by Wang G. *et al.* [19], where a special computer program was developed to implement label-free quantitation of bovine serotransferrin spiked into albumin-depleted plasma. In the same study, protein changes in knock-out $p53^{-/-}$ compared to wild-type cells were determined and four peptides present only in wild-type and a further 12 only in $p53^{-/-}$ cells were detected.

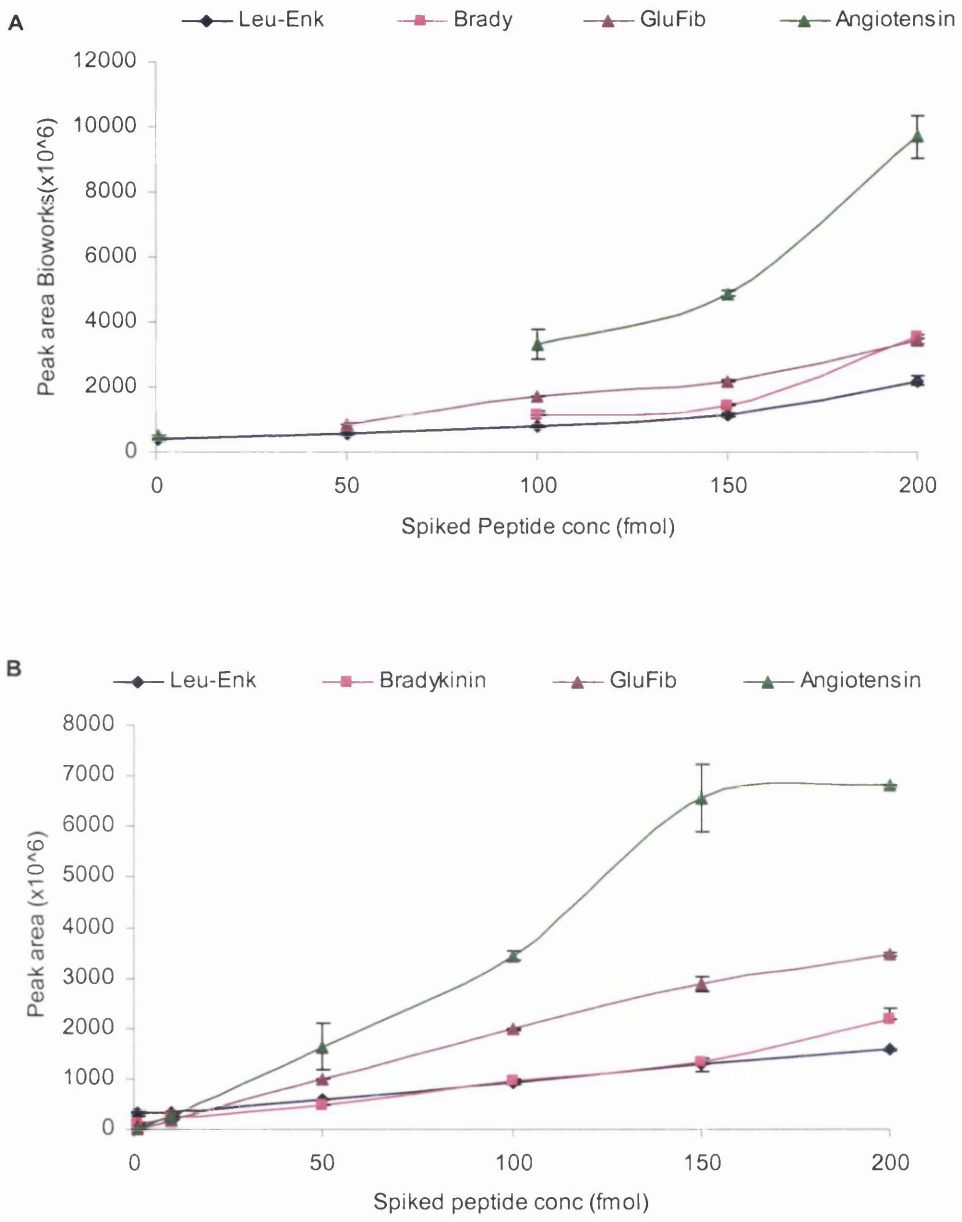


Figure 7.14: Quantitation of peptides spiked into LMW serum peptides by area integration from LC-MS/MS analysis. In **A,** the area quantitation method within the Bioworks Browser was used to automatically detect the area for peptides that were identified by Sequest. In **B,** the area for each peptide was manually searched and integrated in the XIC within Qual Browser of Xcalibur™ version 1.4 and the area for that peak recorded.

The results from Figure 7.14 showed that quantitation using the Bioworks Browser has limitations. A low number of MS/MS scans are performed due to the slow scan rate of the LCQ Deca (Thermo Finnigan, UK) compared to a newer instrument such as the LTQ. Additional to the limited number of MS/MS spectra, fragmentation

spectra that are not present in the database, such as modified sequences or post-transcriptional changes cannot be quantitated. In the spiked standards experiment we were able to directly search for the peptide peaks using the XIC, as their mass was known, and integrate the peak area for quantitation. However in a complex biological mixture of unknown peptides, such as serum, this is not possible.

Nevertheless we did try to compare the 8 breast cancer and 8 control samples from the S1 sample set (see previous chapters). Each LMW serum sample was trypsin-digested and analysed by LC-MS/MS twice before the MS/MS scans were searched through the human FASTA database for peptide identification. The spectra could not be analysed with the Bioworks Browser version 3.2, as only two files can be compared with each other and, as described above and seen in Figure 7.14 a, the analysis is limited to identifiable MS/MS spectra. Spectra were therefore compared visually for different sections and mass ranges highlighted in XICs across all samples with the intention to detect (more or less by chance) any peptides that had a different peak area or intensity in the breast cancer samples compared to the controls. No formal statistical analysis was performed.

We managed to detect one peptide that was only present in the breast cancer samples, in every single one, but was not present in the control samples. As seen in Figure 7.15, this peptide at *m/z* 790.2 is present and at relatively high intensity (the y-axis was fixed to 2.0 $x10^8$) only in the breast cancer samples. The same data is shown, using a slightly different view, in Figure 7.16: this visualization shows an even more convincing difference between the two sample groups.

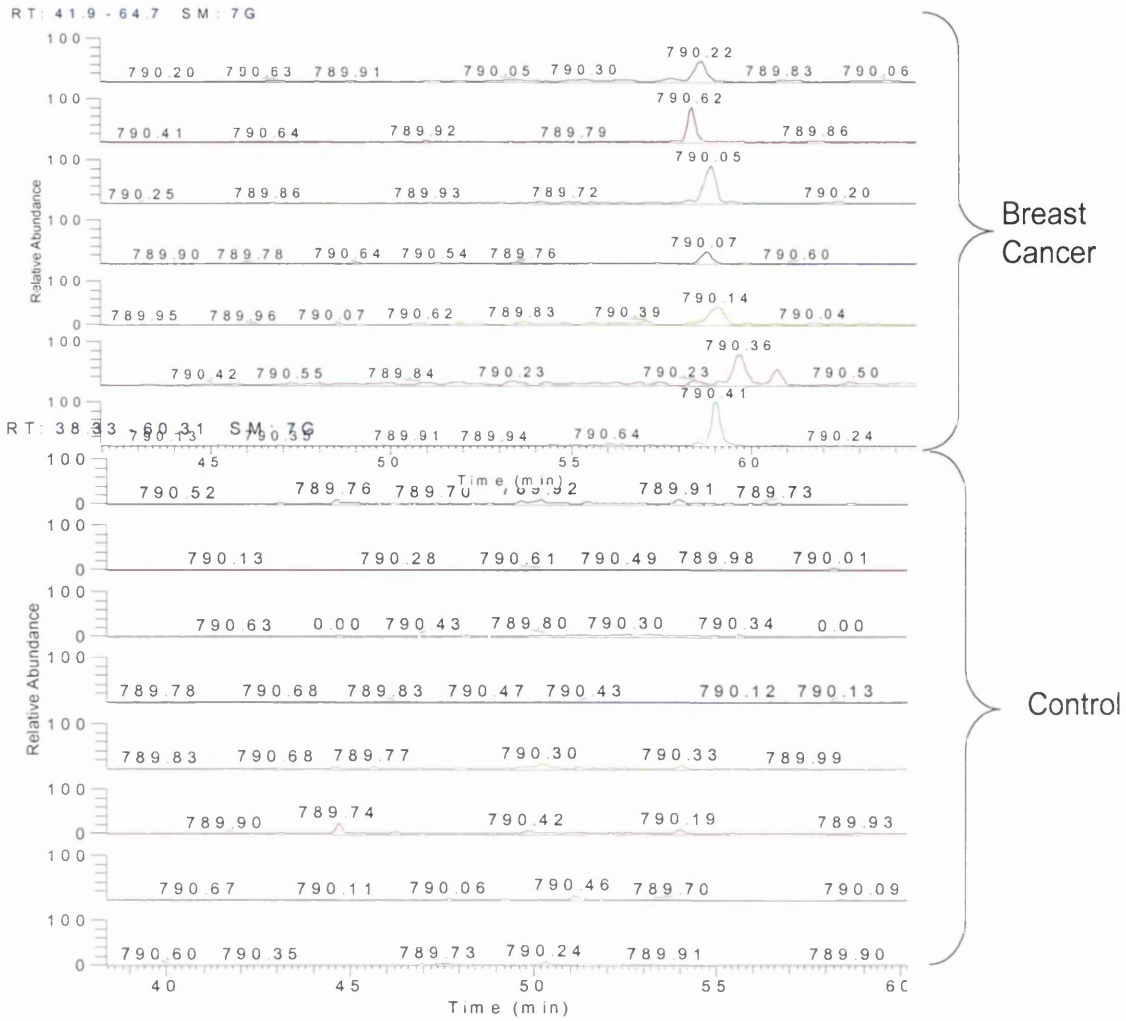


Figure 7.15: Extracted ion chromatogram for *m/z* 790.2 for all breast cancer and control LC-MS/MS separations. The y-axis of the spectra was fixed to 2.0 x10^8 for the breast cancer and 2.0 x10^7 for the controls (The lower intensity for the control spectra was chosen to ensure that the peptide is really not present.)

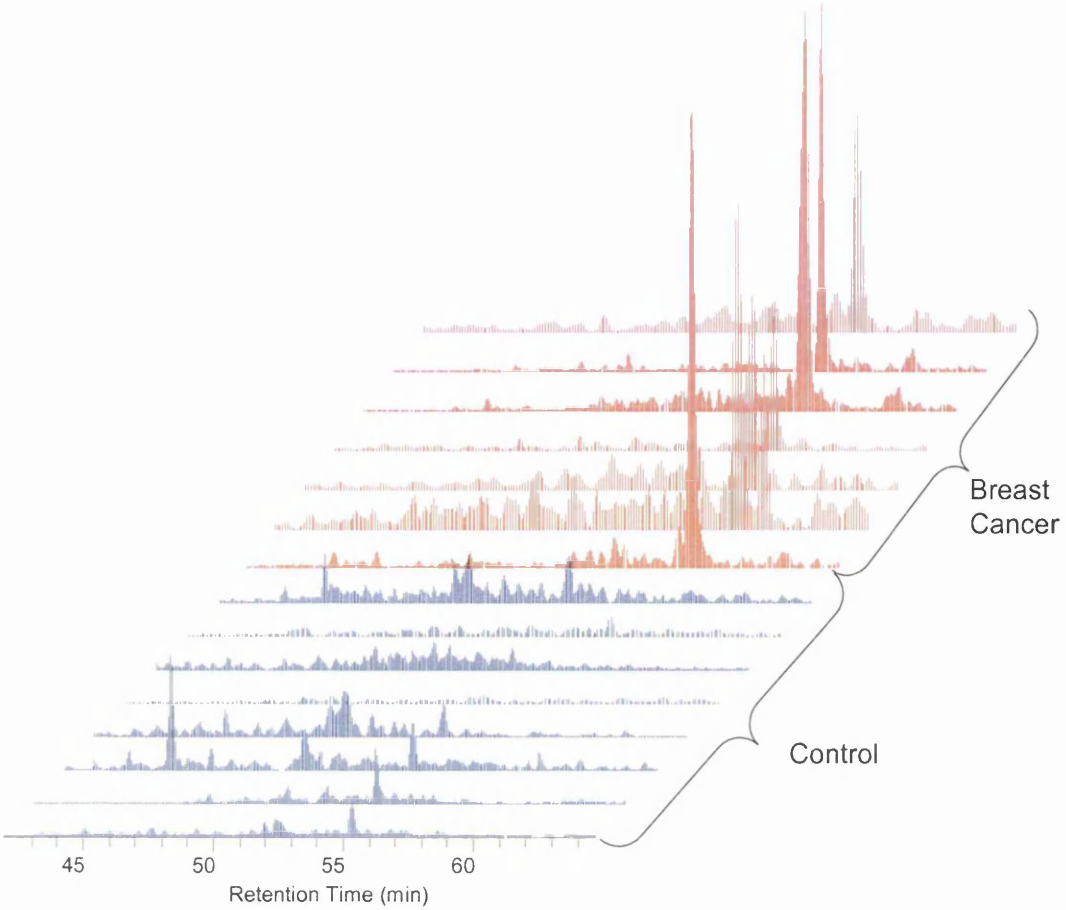


Figure 7.16: A different view of the XIC *m/z* 790.2 of all breast cancer (red) and control samples the different angle and the overlaying of the spectra show the difference between the two clinical cohorts for *m/z* 790 even more clearly.

A search of the literature revealed no published data identifying differentially expressed proteins or peptides detected using label-free quantitation from serum. Hence this is a novel discovery. Ideally, we would have thoroughly analysed this dataset to identify any further potential marker peptides, however, due to the lack of available software programs for more thorough quantitation, no relative quantitation of the other peptides in the spectra could be performed. It is likely that other peptides were up- or down-regulated in the breast cancer samples. Since *m/z* 790.2 is absent in the control group, no relative quantitation can be calculated, however, this only makes it more valuable, it represents a true absent/present marker. To detect a peptide that is only present in one of two sample groups is a real treasure. Identification of this peptide could lay a path for the development of a clinical tool or diagnostic test, given thorough validation. However since the peptide occurred in all breast cancer samples,

despite the heterogeneity of the patient group, this is an important first step towards identifying a diagnostic marker.

7.3. Tandem MS Analysis for Identification *m/z* 790.2 Da

All spectra were searched against the human FASTA database; however no identification for any of the scans containing *m/z* 790.2 was obtained. Therefore the remnants of all the breast cancer samples were pooled and analysed again by selective ion fragmentation for *m/z* 790.2, specifically. Here MS/MS fragmentation was performed only on scans containing *m/z* 790.2, reducing any effect of ion suppression or the slow scan rate of the ion trap. The peak intensity of the resulting MS/MS spectra was very low (NL: 1.06 $x10^3$) and almost indistinguishable from noise (Figure 7.17). Furthermore, although the zoom scan in Figure 7.18 indicates that the peptide is singly charged, the fragment ions exceed the *m/z* of the precursor ion. The peaks in the zoom scan were not well defined and another peak may be hidden underneath. Hence further analyses to determine the charge-state of this peptide was required.

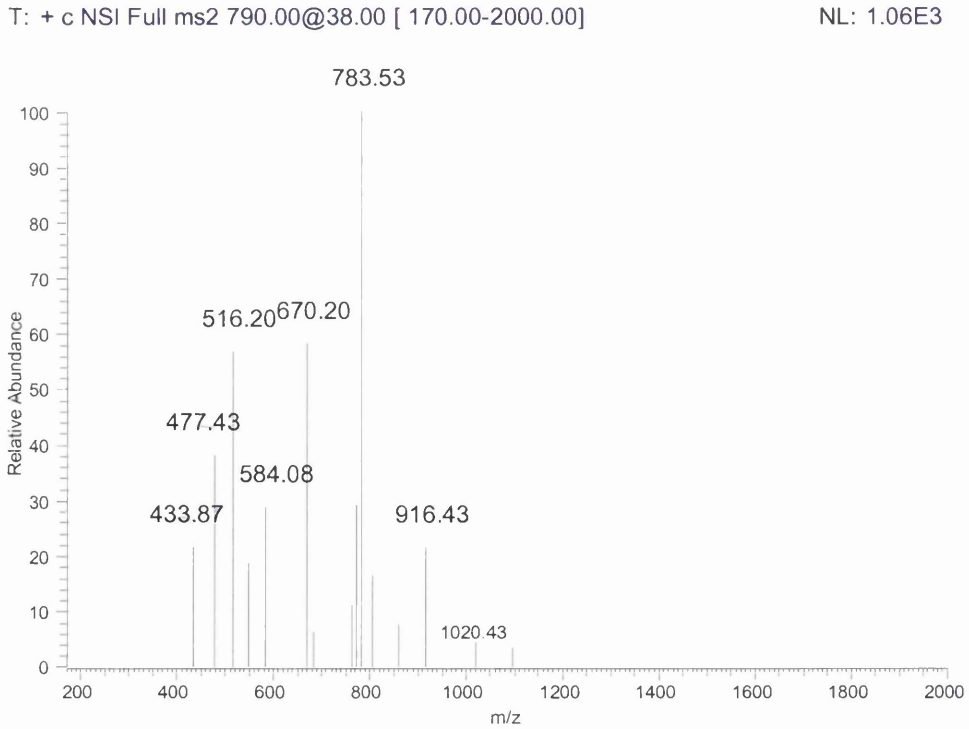


Figure 7.17: Fragmentation pattern of *m/z* 790, the peak intensity is low and the fragment ions are larger than the precursor mass, although the ion is supposed to be singly charged.

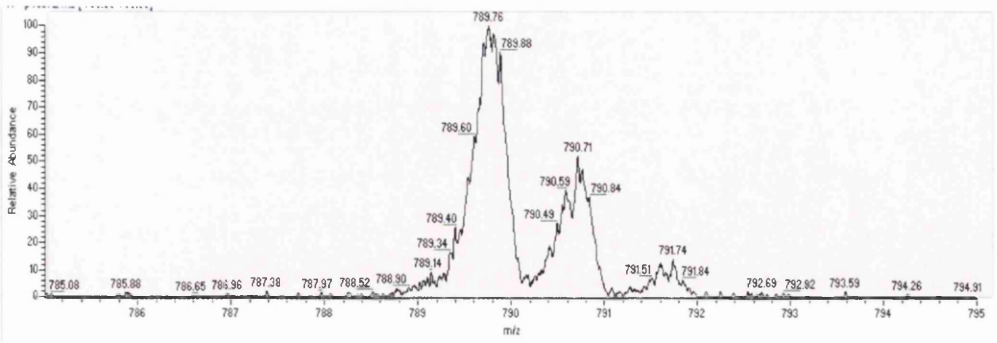


Figure 7.18: A zoom scan of the peak at *m/z* 790.2. The spectrum indicates that the peptide has a single charge.

To further attempt to identify our potential marker peptide, the sample was separated on a C18 reverse-phase column and analysed online by MS/MS using a Q-ToF MS (Micromass, Waters, UK) as described in the Materials and Methods (section 2.8.4). The Q-ToF has greater resolving power for individual mass peaks and may be more sensitive than the LCQ. Directly scanning for *m/z* 790 again resulted in very low intensity fragmentation spectra, and although MASCOT searching (version 1.9; MatrixScience, London, UK) and *denovo* sequencing was attempted the fragment ions were too small and too few to get any useful identification (Figure 7.19). The precursor ion was still present in the spectrum; from this at least the single charge state of the ion could be confirmed (Figure 7.20).

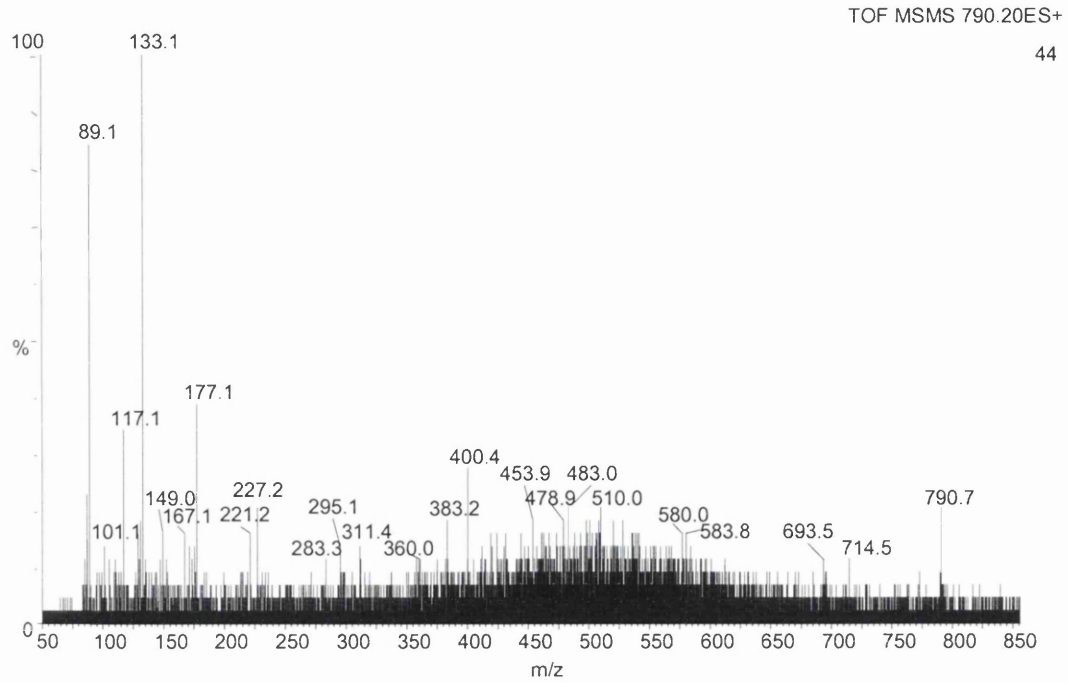


Figure 7.19: Tandem MS spectrum of the peak at *m/z* 790.2. The intensity of the spectrum is low and few fragment ions are visible, sequencing was not successful.

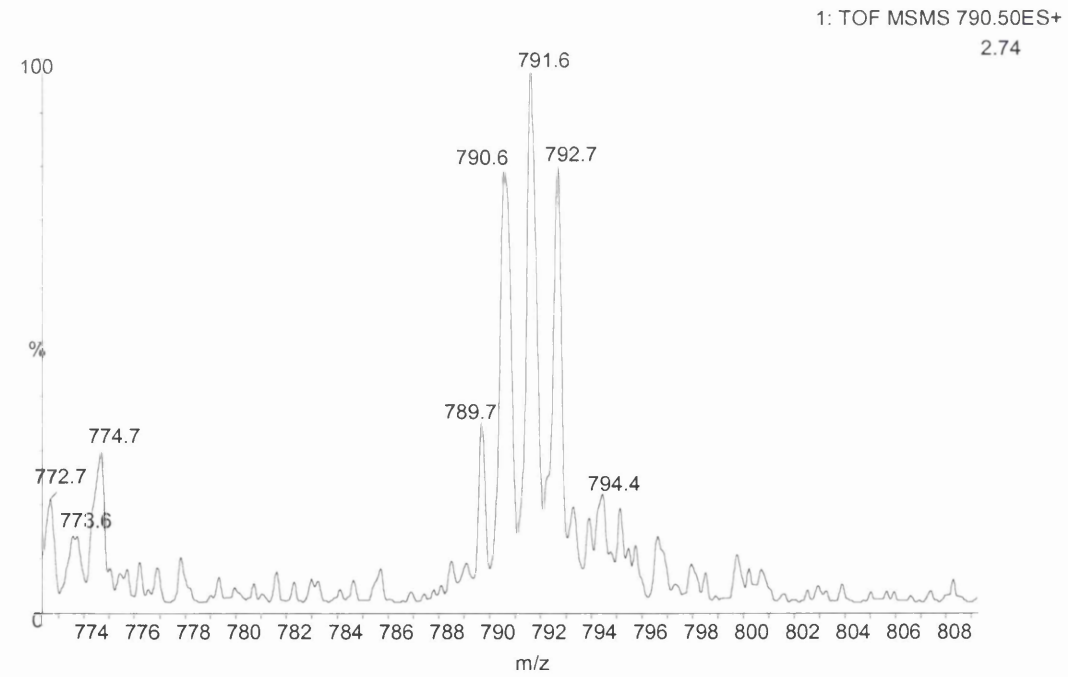


Figure 7.20: Tandem MS spectrum of peak at *m/z* 790.2. Zooming in over the precursor confirmed that the peptide is singly changed.

7.4. Discussion and Conclusions

In this final results chapter the use of LC-MS/MS based analysis of peptides was tested for biomarker detection from serum. Label-free quantitation of peptides from LC-MS/MS separation relies on the reproducible elution of the peptides and the reproducible determination of their peak areas. Using LMW serum protein digests the complexity of the sample was reduced and the reproducibility of the retention time and peak area detection was demonstrated on a number of selected peptide peaks. Furthermore to identify a maximum number of peptides an optimal elution gradient and concentration of peptides was determined to resolve peptides into separate peaks. Finally for high throughput analysis including data analysis, sophisticated software tools are required to analyse the vast amount of data produced, in a reliable and rapid manner. Initially we demonstrated that standard peptides spiked into LMW serum can be semi-quantitated, from the XICs, using peak area integration. However quantitation of those standards using the Bioworks Browser showed to be more complicated. The quantitation algorithm within Bioworks makes use of the MS/MS scans, and only peaks that produce a positive peptide identification are quantitated. Due to the limited number of MS/MS scans produced with the LCQ Deca, even from the standards some peaks were not identified and hence not quantitated. However a small profiling study was still undertaken using the S1 sample set (see Chapter 5 and 6) for label-free quantitation. Using the optimized elution gradient and the optimal concentration of peptides we were able to detect a peptide that was only present in the breast cancer samples. Although breast cancer is a heterogeneous disease, this potential "marker" was present in all breast cancer samples but in none of the controls. Discovering a peptide present in only one sample cohort even with the use of statistical software, is rare. Even in the absence of a positive identification for the peptide at *m/z* 790.2, this could, if validated in repeat experiments and larger sample groups, become a potential marker for detection of breast cancer patients. If we were able to isolate a peak at *m/z* 790.2 it may be possible to identify the peak, although it appeared to be difficult to fragment this peptide.

For a more global search for peptides of different intensity within a complex sample of unknowns the use of a high-speed 2D linear ion trap such as an LTQ or LTQ-

Orbitrap could increase the protein coverage greatly compared to an LCQ with a 3D ion trap [20]. Using the LCQ with data-dependent MS/MS scans the analysis suffered from a limited number of MS/MS fragmentations due to ion suppression and loss in the exclusion time, as only a proportion of species observed in the survey MS scan was selected for MS/MS fragmentation [21]. To overcome this all samples were analysed in MS mode only and then once differences are detected these could be specifically identified in a separate MS/MS experiment. However automated quantitation using the identification-based algorithm in Bioworks was not possible. It has been reported [22] that multiple sources of variation affecting the peak intensity may be introduced during the analysis, such as differences in electrospray ionization efficiencies among different peptides and samples as well as differences in separation (as seen in Figure 7.10 for replicate runs and in Figure 7.14 for angiotensin II). These issues are often peptide dependent, resulting in differences in relative abundance in peptides from the same protein, making an automated approach using all peptides identified from one protein such as using Bioworks Browser quantitation more difficult. For these collected reasons we decided that the algorithm in Bioworks is not suitable for our data collected from the LCQ Deca analysis, and the experiment was not extended to analysing the S2 sample set.

7.5. References

[1] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L. and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* **1**, 947-955.

[2] Ornstein, D. K., Rayford, W., Fusaro, V. A., Conrads, T. P., Ross, S. J., Hitt, B. A., Wiggins, W. W., Veenstra, T. D., Liotta, L. A. and Petricoin, E. F., 3rd (2004) Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml. *J Urol* **172**, 1302-1305.

[3] Qian, W. J., Jacobs, J. M., Camp, D. G., 2nd, Monroe, M. E., Moore, R. J., Gritsenko, M. A., Calvano, S. E., Lowry, S. F., Xiao, W., Moldawer, L. L., Davis, R. W., Tompkins, R. G. and Smith, R. D. (2005) Comparative proteome analyses of human plasma following in vivo lipopolysaccharide administration using multidimensional separations coupled with tandem mass spectrometry. *Proteomics* **5**, 572-584.

[4] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J., Petricoin, E. F., 3rd, Liotta, L. A., Veenstra, T. D. and Conrads, T. P. (2004) An investigation into the human serum "interactome". *Electrophoresis* **25**, 1289-1298.

[5] Deutsch, E. W., Eng, J. K., Zhang, H., King, N. L., Nesvizhskii, A. I., Lin, B., Lee, H., Yi, E. C., Ossola, R. and Aebersold, R. (2005) Human Plasma PeptideAtlas. *Proteomics* **5**, 3497-3500.

[6] Zang, L., Palmer Toy, D., Hancock, W. S., Sgroi, D. C. and Karger, B. L. (2004) Proteomic analysis of ductal carcinoma of the breast using laser capture microdissection, LC-MS, and 16O/18O isotopic labeling. *J Proteome Res* **3**, 604-612.

[7] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R. and Lobley, A. (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* **3**, 311-326.

[8] Johnson, K. L. and Muddiman, D. C. (2004) A method for calculating 16O/18O peptide ion ratios for the relative quantification of proteomes. *J Am Soc Mass Spectrom* **15**, 437-445.

[9] Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M. and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **75**, 4818-4826.

[10] Chelius, D. and Bondarenko, P. V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* **1**, 317-323.

[11] Fang, R., Elias, D. A., Monroe, M. E., Shen, Y., McIntosh, M., Wang, P., Goddard, C. D., Callister, S. J., Moore, R. J., Gorby, Y. A., Adkins, J. N., Fredrickson, J. K., Lipton, M. S. and Smith, R. D. (2006) Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol Cell Proteomics* **5**, 714-725.

[12] Qian, W. J., Jacobs, J. M., Liu, T., Camp, D. G., 2nd and Smith, R. D. (2006) Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol Cell Proteomics* **5**, 1727-1744.

[13] Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198-207.

[14] Lee, H. J., Lee, E. Y., Kwon, M. S. and Paik, Y. K. (2006) Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics. *Curr Opin Chem Biol* **10**, 42-49.

[15] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P. and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* **2**, 1096-1103.

[16] Bjorhall, K., Miliotis, T. and Davidsson, P. (2005) Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **5**, 307-317.

[17] Smith, R. D., Loo, J. A., Edmonds, C. G., Barinaga, C. J. and Udseth, H. R. (1990) Sensitivity considerations for large molecule detection by capillary electrophoresis-electrospray ionization mass spectrometry. *J Chromatogr* **516**, 157-165.

[18] Shen, Y., Jacobs, J. M., Camp, D. G., 2nd, Fang, R., Moore, R. J., Smith, R. D., Xiao, W., Davis, R. W. and Tompkins, R. G. (2004) Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal Chem* **76**, 1134-1144.

[19] Wang, G., Wu, W. W., Zeng, W., Chou, C. L. and Shen, R. F. (2006) Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J Proteome Res* **5**, 1214-1223.

[20] Mayya, V., Rezaul, K., Cong, Y. S. and Han, D. (2005) Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics. *Mol Cell Proteomics* **4**, 214-223.

[21] Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D. and Yates, J. R., 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* **75**, 2470-2477.

[22] Tang, K., Page, J. S. and Smith, R. D. (2004) Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J Am Soc Mass Spectrom* **15**, 1416-1423.

CHAPTER 8

Final Discussion and Conclusions

During the three years that this study has taken, little has changed in terms of biomarker discovery for breast cancer. In fact very few markers have been reported for any diseases, or discovered using proteomics. [1]. A review by Hu *et. al.* [2] summarizes all the markers discovered using proteomics and mass spectrometry for any disease from serum and they reported less than 15 for breast cancer.

In biomarker research, the focus has changed from pattern profiling and the use of SELDI-ToF to attempts to develop and optimise other types of mass spectrometry as well as studies of the capabilities and reproducibility of current protocols and techniques.

In particular, in the case of serum it has become apparent that storage and handling is crucial for quantitative analyses, as much variation can be introduced to LMW proteins during these first steps.

8.1. Sample Preparation

There was surprisingly little in the literature about methods of standardising sample preparation and handling at the time this project was started. This may be due to the fact that researchers were caught up in the excitement of the fast development of mass spectrometry and new applications, discovering biomarkers. In fact, factors such as sample preparation and handling, for instance clotting time, storage temperature and time, freeze/thaw cycles and the time taken for further sample preparation are all very important. This is especially important when each candidate marker will have to be validated on large independent sample sets [3]. Recently, the

Human Proteome Organization (HUPO) has published a meta-analysis of information gathered over the last 3 years on these issues [4], and a number of other papers have been published on the back of the data collected [5-7].

At the beginning of this study it was thought that serum samples could be stored in small aliquots at -80°C for a prolonged time. Since then there have been reports that serum samples may in fact be best stored in liquid nitrogen and that even here they may only be stable for a few months [1, 8]. Furthermore it may in fact be the case that serum may be most stable in dry form, after precipitation. In my own research, I have discovered that the serum samples were not stable in the -80°C freezer for long periods, such as 18 months, even in the absence of freeze/thawing. The LMW samples, which were stored in NH_4HCO_3 following centrifugal ultrafiltration (UF) were even less stable and appeared to degrade within 6 months. A possible explanation for this may be the lack of albumin; albumin has been reported to have protective properties against protease activity in serum [3]. Protease inhibitors could be added to the sample to prevent protein degradation during storage and handling, however these may interfere with subsequent MS analysis or lead to formation of covalent bonds with other proteins [4, 9-12]. Furthermore, Villanueva *et al.* [13] reported that the use of protease inhibitors could mask the presence of disease-related proteases that produce the diagnostic signature of endogenous peptides.

From the start we were very careful to standardise all sample preparation steps and so the type of serum collection tubes, clotting time, centrifugation and storage were kept the same for all samples. And this decision is supported by the most current literature. Freeze/thawing, on the other hand appears to cause less obvious changes to LMW proteins and peptides [3] than previously thought as long as the samples are kept on ice [14]. The only factor that we did not have power over was the storage time and, in light of the studies published recently [3, 7, 9, 14], this may be a major problem. In our experiment, storage time has presented as the major limiting factor, as we were not able to repeat any of the experiments described in this thesis later on the same samples. The only possibility was to use those LMW samples that were stored completely lyophilized, as they were more stable than the neat serum or LMW samples in solution. If I was to repeat this project, I would use all serum samples within 6 months of freezing, and ensure that all processed samples were lyophilized to complete dryness. I would not use protease inhibitors or precipitation prior to

freezing, as these may introduce additional variation or complicate future analysis. However this sort of change is not possible for large-scale sample collection projects such as serum banks started before this project. More research is clearly required into the effects of storage.

8.2. Albumin Depletion

In this thesis, we set out to develop a protocol for detection of serum markers for breast cancer. It quickly became obvious that serum is too complex to analyse and that albumin and other proteins of high abundance present in serum interfere with the analysis of other serum proteins, of low abundance. These low abundance proteins may have not been studied in great detail, but may be clinically most interesting. To study low-abundance proteins, the high-abundance species are often eliminated as the first step in the analytical protocol [15-21]. Removal can be achieved by affinity depletion through LC columns or cartridges, protein precipitation or centrifugal ultrafiltration.

We have tested the above methods and shown that UF is the best technique for removal of albumin and the enrichment of low molecular weight (LMW) proteins. Using UF, recovery of the LMW sub-proteome was possible in one single step without the use of high salt buffers that would later interfere with MS analysis. The reproducibility of this step is paramount and therefore an in-depths study was performed. The results confirmed the reproducibility of UF and that the complexity of the serum proteome was sufficiently reduced by this method for quantitative studies. During the extensive optimisation of the UF procedure, a new protocol was designed, different to those previously described in the literature [22-25]. Using this optimised method we were able to recover a larger number of LMW proteins in a more reproducible fashion.

The use of UF may be further justified by the fact that the only reported marker discovered in the last 4-6 years, before the beginning of this study, had been peptides of low molecular weight [1, 26, 27]. Markers of higher molecular weight have often turned out to be unrelated to the actual disease itself [13].

In summary, we have shown that UF is a robust sample pre-processing method to enrich the LMW proteome for subsequent biomarker discovery in serum. Furthermore, UF has been shown here to be more selective at removing albumin and immunoglobulin from serum than affinity chromatography and precipitation without the loss of other proteins.

8.3. Biomarker Discovery from Intact Proteins

The aim of this project was to discover a biomarker or a protein pattern as a signature of breast cancer. At the start of the thesis, very little literature was available on the use of mass spectrometry for quantitative proteome analysis. SELDI-ToF MS was the most cited technology and a number of convincing studies had been published, using this technique for protein quantitation from serum. We therefore started off using SELDI-ToF analysis, which appeared to be optimal for the LMW sub-proteome, since SELDI-ToF has an optimal mass range smaller than 30,000 Da. As there was no SELDI-ToF mass spectrometer available on-site, I formed a collaboration with a group at Cardiff University. MALDI-ToF MS, on the other hand, was available in Swansea. Although quantitation of all proteins in the complex mixture had not been reported using MALDI-ToF we also analysed all our samples with MALDI-ToF MS. The two sample sets (S1 and S2), of 8 breast cancer and 8 control samples, were each analysed using both techniques. For the S1 sample set an initial "4 x 4 study" was set up of a smaller subset of samples analysed on all SELDI chip types. Also on this sample set, pre-fractionation using WAX resin was tested and this was termed the "8 x 8 study". The two ToF instruments work relatively similarly; hence intact protein analysis should be directly comparable. The SELDI-ToF analysis of the first sample set (S1) produced some very interesting results, hinting at some potential markers. From the 4 x 4 study, 17 discriminating protein peaks were discovered and 11 peaks identified from the fractions of the 8 x 8 study. After the first experiment, it became apparent that it is essential to use multiple replicates through-out the whole experiment to allow for any variation occurring during the sample preparation or MS analysis. Therefore we could not be sure about the authenticity of the markers because no replicates of the actual samples were prepared. For this reason the experiment was repeated using new serum samples,

prepared by UF in triplicate before MS analysis. Unfortunately, the second experiment using SELDI-ToF MS failed; the samples produced very low resolution spectra and could not be used for marker discovery. Despite extensive research into the cause for the low resolution spectra we were not able to resolve the problems and SELDI-ToF analysis was therefore abandoned. We know that the samples were of good quality as they were subsequently analysed by MALDI-ToF and produced good resolution spectra. Analysis of these S1 samples on triplicate MALDI spots by MALDI-ToF MS resulted in 9 potential markers, of which 4 were visually convincing enough to be taken forward.

The results recovered from the S1 samples by SELDI-ToF were compared to the markers retrieved from the MALDI-ToF MS analysis and encouragingly many of the markers overlap. It can be assumed that markers that are recovered from two completely different forms of analysis should be more convincing and may be regarded as a confirmation of the MALDI-ToF results.

The second sample set, S2, provided results with more confidence as three replicate UF from each serum sample were carried out before analysis by MALDI-ToF. Using alignment with *mz*Align and Markerview, 16 proteins with significant *p*-values were detected. Of these, three markers were visually convincing enough to be further analysed in addition to four markers from the S1 results using tandem MS for protein identification. Three markers produced MS/MS spectra with enough fragment ions to be matched to a peptide sequence using MASCOT. Of those, two results had an ion score with enough confidence to indicate homology to bradykinin and the connecting peptide between the A and B chain of α-2-HS-glycoprotein. However, unfortunately no peptide identification was found for the most convincing peptides in the S2 sample set. The peaks at *m/z* 8771 and 9647 were statistically the most significant but since these peptides were relatively large, no fragmentation spectra were obtained from the MS/MS analysis for identification. The most convincing peak from the S1 data set is *m/z* 2995 but this peak was not present in the spectra from the S2 samples, hence no identification by MS/MS was possible.

8.4. Data Analysis

Semi-quantitative protein profiling from serum samples had not been published when the project was started and so the experiment and in particular the data analysis had to be developed from basic principles. Peak alignment proved to be the most challenging aspect of the analysis, although this later also proved to be the most important factor when comparing multiple spectra from different samples, including replicates. The mass accuracy of the present MALDI-ToF instrument was relatively poor and so a wide range of masses was produced for the same peak from the spectra of individual samples. These first had to be aligned in order to be recognized as the same peak before they could then be used to calculate intensity changes. A program, later called *mz*Align, was written in VBA for Excel to combine the masses from the peaks across all spectra into one reference mass list. A second sequence of the program aligned the peak intensities for the same peak across all spectra. To avoid misalignment, the mass tolerance was checked and could be adjusted in the program. This proved to be invaluable and one of the greatest advantages of *mz*Align over other commercially-available options. For example Markerview, supplied by Applied Biosystems, that became available later in the study (2006), was tested and compared to *mz*Align. The application is very similar, creating a reference mass list and then aligning the peak intensities from across the spectra in one datasheet. Markerview provides more sophisticated visualization options for control of peak alignment than *mz*Align, but the mass tolerance for individual peaks cannot be altered, which was a disadvantage, resulting in some peaks being wrongly aligned from distinct peaks in the spectra or false peaks created. Additionally, statistical analysis using a *t*-test and principal component analysis (PCA) are possible. However using the *t*-test, *p*-values were found to be calculated for peaks that were not present in one of the cohorts. This is statistically impossible; hence we did not use this function, but instead exported the aligned results into Excel to perform statistical analysis there. Furthermore, because of the great variation in the results with a large number of missing values for each peak across all spectra, PCA analysis was not possible. In Markerview the scatter plots showed too much variation across principal component 1. In Excel, however, the average of all replicates (biological and technical) for each patient sample could be calculated and in this way we were able to reduce the variation greatly. This average was then used to calculate *p*-values in Excel and to perform PCA analysis using SIMCA-P from Umetrics (Windsor, UK). The scatter

plot here showed that the breast cancer and control samples clustered in distinct groups although a lot of variation was still present.

Visual inspection of the discriminant peaks also proved to be paramount, as the peak intensity variation was great and significantly different results could result from single outliers. This was also described by Villanueva *et al.* [28, 29]; this group was the first to publish quantitative data of serum profiling by MALDI-ToF MS. For peak alignment this group used the GeneSpring analysis platform (Agilent Technologies UK Limited, Stockport, UK), which was originally designed for microarray analysis. In one of their more recent papers, a new algorithm for peak alignment was also proposed [28]. This program enables the user to visualise aligned peaks by overlaying them, very similar to what was possible in the SELDI-ToF software and the overlaid spectra look like those shown in Chapter 6. Ideally *mz*Align could be improved to contain an option for alignment visualisation in this way.

Furthermore, our analysis method could be generalized for many diagnostic and predictive purposes, as an *in vitro* phenotypic readout of catalytic and metabolic activities in body fluids or tissues, utilizing either endogenous substrates or measured quantities of externally assed isotopic labelled substrates followed by quantitative analysis.

8.5. LC-MS/MS

Shot-gun proteomics has been practiced successfully for a while now for identification of many proteins from serum, and is being used to identify more and more protein from biofluids, cells and tissue. However quantitative profiling studies have mainly been carried out using isotopically-labelled samples. Nevertheless label-free quantitation using peak intensities has been proposed in a number of studies which tested the application using standards [30, 31]. We therefore attempted label-free quantitative analysis using tryptic digest of the S1 samples. The new version of Bioworks 3.2 has been designed to be able to quantitate peptides using the peak area of each peptide from LC-MS spectra. An experiment using standard peptides spiked into serum quickly revealed that a high resolution iontrap MS would be necessary for

such analysis, as well as an instrument with a fast scan rate to minimise the effect of the exclusion time. Due to the exclusion time of the LCQ, very few peptides were analysed by MS/MS, despite a long elution gradient. Better results were achieved using manual integration of the peptide peaks in the extracted ion chromatogram (XIC) for each of the standard peptides. However this is not possible for complex mixtures without the prior knowledge of the *m/z* of the peptides.

In an attempt to find a differentially expressed peptide, small mass ranges were selected in XICs and inspected for peaks that had different expression levels between the breast cancer and the control cohorts. One peptide was discovered that is exclusively present in the breast cancer samples, and is completely absent in the controls.

It was decided that for relative quantitation, more sophisticated algorithms are required for normalization and to calculate differences while taking into account variation. Especially where a large number of replicates per sample is required. A number of algorithms are available commercially, such as the DeCyder MS Differential Analysis Software (GE Healthcare, UK) or Progenesis from Nonlinear Dynamics (Newcastle upon Tyne, UK), to name only a few. Johansson *et al.* [32] published a study successfully comparing levels of integrase in *E.coli* using DeCyder MS for automated detection and relative quantitation of unlabelled peptides in LC-MS/MS data. DeCyder MS generates 2D representations of the peptide patterns from individual LC-MS/MS analyses which can then be matched and compared. The use of other algorithms has also been published for peak area comparison, such as *Q-MEND* [33].

For future research it would be useful to develop a software tool, similar to *mz*Align, to automatically integrate all peaks in the spectra, normalize them and adjust the mass alignment. However, unless all samples were analysed in full scan mode, an instrument with a faster scan rate would be required to obtain reproducible peptide identifications from tandem MS spectra.

8.6. Future Prospects

If more time had been available to take the project further there are various routes that could have been followed. For protein/peptide identification of potential markers, proteins could be isolated by SDS-PAGE or selective chromatography and identified through tryptic digests and peptides fingerprinting. To qualify as a useful marker, each protein would then have to be validated by Western blotting or ELISA for these where antibodies are available; otherwise antibodies could be raised. Additionally, it would be interesting to see if the markers are also present in tissue rather than serum by paraffin immunohistochemistry staining or tissue microarrays and if the identified proteins were differentially present in tumour and normal tissue. Furthermore these markers could then be quantitated in serum from patients with other cancer types to see if they are indicators for breast cancer specifically, and finally, if they occur in metastatic cancers only or could be found in serum from patients with early stage disease.

In conclusion the use of multiple MS platforms for biomarker discovery from breast cancer serum samples has highlighted the advantages and limitations of the current technology. The minimum requirements for a protein profiling platform suitable for biomarker discovery in a clinical setting, as summarized by Qian *et al.* [34] are (i) a high dynamic range to detect low abundance proteins (ii) high confidence protein identifications; (iii) accurate quantitation for relative protein abundances across different sample groups and between replicates; (iv) high-throughput that allows analysis of large sample numbers to provide sufficient statistical power to address biological variation and, finally, (v) comprehensive informatics software for data mining and statistical analysis. At the moment, no single one of the platforms in use can meet all of these requirements for effective biomarker discovery. This therefore re-emphasises the need for the use of multiple platforms on the same sample set as was carried out in this thesis.

Development of a standardised method for sample collection, preparation and handling is also crucial for serum proteome analysis. In our study, UF appeared to be an effective means to unmask low abundant proteins in serum. Quantitative MS analysis was achieved using a combination of SELDI-ToF, MALDI-ToF and LC-MS. However for identification of potential markers, the proteins need to be isolated

in order to do peptide fingerprinting. Further improvements in technology will be necessary for biomarker discovery, especially to allow more high-throughput analysis. Unlike genomic analysis using microarrays, proteomics is far more complicated and encompasses too many options, as evidenced by the results from this thesis. It is critical to increase the ease of use, throughput, speed and accuracy of currently available proteomics techniques to allow their more general use in a clinical setting. For example for more high-throughput analysis during identification of markers, and for label-free LC-MS quantitation, the samples were trypsin digested, which required a 24 hour protocol. To increase the speed of analysis this step could be optimized.

The use of microwave-assisted trypsin digestion as reported by Juan *et al.* [35] and Sun *et al.* [36], can improve enzymatic digestion and enable a more high-throughput protocol for peptide profiling experiments. Although not reported in the results, the efficiency of trypsin digestion at different times in the microwave was compared with 6 hour-, and the most commonly used 16 hour-digestion at 37°C, without the use of the microwave. The results showed that overnight digestion at 37°C was inferior to the other methods, but a 30 min treatment of the proteins in the microwave produced superior results. This experiment was performed after all other experiments were completed. However, in any future experiments, I would perform a microwave-assisted trypsin digestion, to achieve better peptide coverage and increased experimental throughput.

As a take home message, quantitation of intact proteins using MALDI-ToF MS was successful using the LMW sub-proteome, however a larger number of technical replicates is required for biomarker discovery. In the future I would prepare each serum sample in triplicate by UF and then analyse each replicate again in triplicate using MALDI-ToF MS analysis. For data mining, the use of *mz*Align or Markerview allows fast and effective alignment and analysis of a large dataset. Markers were detected easily but still had to be confirmed by visual inspection of the spectra. A visualisation tool such as the software described by Villanueva *et al.* [28] would be beneficial, to overlay and inspect all spectra for a particular peak at once. For small peptides, the MALDI-ToF target plate should be saved to perform further MS/MS analysis on the spots containing discriminating peaks directly for identification. The

use of LC-MS/MS for label-free quantitation of peptides appears to be promising but is limited by the type of instrument used and more sophisticated alignment algorithms are necessary to compare multiple samples and their replicates.

8.7. References

[1] West-Norager, M., Kelstrup, C. D., Schou, C., Hogdall, E. V., Hogdall, C. K. and Heegaard, N. H. (2007) 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**, 30-37.

[2] Hu, S., Loo, J. A. and Wong, D. T. (2006) Human body fluid proteome analysis. *Proteomics* **6**, 6326-6353.

[3] Hsieh, S. Y., Chen, R. K., Pan, Y. H. and Lee, H. L. (2006) Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. *Proteomics* **6**, 3189-3198.

[4] 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. (2005) HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* **5**, 3262-3277.

[5] Traum, A. Z., Wells, M. P., Aivado, M., Libermann, T. A., Ramoni, M. F. and Schachter, A. D. (2006) SELDI-TOF MS of quadruplicate urine and serum samples to evaluate changes related to storage conditions. *Proteomics* **6**, 1676-1680.

[6] Govorukhina, N. I., Reijmers, T. H., Nyangoma, S. O., van der Zee, A. G., Jansen, R. C. and Bischoff, R. (2006) Analysis of human serum by liquid chromatography-mass spectrometry: improved sample preparation and data analysis. *J Chromatogr A* **1120**, 142-150.

[7] Luque-Garcia, J. L. and Neubert, T. A. (2006) Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry. *J Chromatogr A*.

[8] Schrader, M. and Schulz-Knappe, P. (2001) Peptidomics technologies for human body fluids. *Trends Biotechnol* **19**, S55-60.

[9] West-Nielsen, M., Hogdall, E. V., Marchiori, E., Hogdall, C. K., Schou, C. and Heegaard, N. H. (2005) Sample handling for mass spectrometric proteomic investigations of human sera. *Anal Chem* **77**, 5114-5123.

[10] Banks, R. E., Stanley, A. J., Cairns, D. A., Barrett, J. H., Clarke, P., Thompson, D. and Selby, P. J. (2005) Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clin Chem* **51**, 1637-1649.

[11] 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. (2005) 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**, 3226-3245.

[12] Tammen, H., Schulte, I., Hess, R., Menzel, C., Kellmann, M., Mohring, T. and Schulz-Knappe, P. (2005) Peptidomic analysis of human blood specimens: comparison between plasma specimens and serum by differential peptide display. *Proteomics* **5**, 3414-3422.

[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. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* **116**, 271-284.

[14] Baumann, S., Ceglarek, U., Fiedler, G. M., Lembcke, J., Leichtle, A. and Thiery, J. (2005) Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin Chem* **51**, 973-980.

[15] Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W. and Guild, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics* **4**, 1175-1186.

[16] Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L. and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422-432.

[17] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L. and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* **1**, 947-955.

[18] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., Schatz, C. R., Miller, S. S., Su, Q., McGrath, A. M., Estock, M. A., Parmar, P. P., Zhao, M., Huang, S. T., Zhou, J., Wang, F., Esquer-Blasco, R., Anderson, N. L., Taylor, J. and Steiner, S. (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics* **3**, 1345-1364.

[19] Hinerfeld, D., Innamorati, D., Pirro, J. and Tam, S. W. (2004) Serum/Plasma depletion with chicken immunoglobulin Y antibodies for proteomic analysis from multiple Mammalian species. *J Biomol Tech* **15**, 184-190.

[20] Quero, C., Colome, N., Prieto, M. R., Carrascal, M., Posada, M., Gelpi, E. and Abian, J. (2004) Determination of protein markers in human serum: Analysis of protein expression in toxic oil syndrome studies. *Proteomics* **4**, 303-315.

[21] Bjorhall, K., Miliotis, T. and Davidsson, P. (2005) Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **5**, 307-317.

[22] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P. and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* **2**, 1096-1103.

[23] Harper, R. G., Workman, S. R., Schuetzner, S., Timperman, A. T. and Sutton, J. N. (2004) Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. *Electrophoresis* **25**, 1299-1306.

[24] Georgiou, H. M., Rice, G. E. and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1**, 1503-1506.

[25] Merrell, K., Southwick, K., Graves, S. W., Esplin, M. S., Lewis, N. E. and Thulin, C. D. (2004) Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J Biomol Tech* **15**, 238-248.