Mass Spectrometric Characterization of the MCF7 Cancer Cell Line: Proteome Profile and Cancer Biomarkers

Hetal Abhijeet Sarvaiya

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Biomedical Engineering and Sciences

Iuliana M. Lazar, Ph.D., Committee Chair
Brian J. Love, Ph.D., Committee Member
Yong Woo Lee, Ph.D., Committee Member

April 18, 2006
Blacksburg, VA

Keywords: Mass spectrometry, Cancer, Biomarkers, Proteomics, MCF7, Microfluidics

Mass Spectrometric Characterization of the MCF7 Cancer Cell Line: Proteome Profile and Cancer Biomarkers

Hetal Abhijeet Sarvaiya

Abstract

The discovery of cancer biomarkers is crucial in the clinical setting to facilitate early diagnosis and treatment, thereby increasing survival rates. Proteomic technologies with mass spectrometry detection (MS) have the potential to affect the entire spectrum of cancer research by identifying these biomarkers. Simultaneously, microfabricated devices have evolved into ideal analysis platforms for minute amounts of sample, with promising applications for proteomic investigations and future biomarker screening. This thesis reports on the analysis of the proteomic constituents of the MCF7 breast cancer cell line using a shotgun 2-D strong cationic exchange/reversed phase liquid chromatography electrospray ionization tandem mass spectrometry (SCX/RP-LC-ESI-MS/MS) protocol. A series of optimization strategies were performed to improve the LC-MS experimental set-up, sample preparation, data acquisition and database searching parameters, and to enable the detection and confident identification of a large number of proteins. Over ~4,500 proteins were identified using conventional filtering parameters, and >2000 proteins using a combination of filters and p-value sorting. Of these, ~1,950 proteins had $p<0.001$ (~90%) and more than half were identified by = 2 unique peptides. About 220 proteins were functionally involved in cancer related cellular processes, and over 100

proteins were previously described in the literature as potential cancer markers. Biomarkers such as PCNA, cathepsin D, E-cadherin, 14-3-3-sigma, antigen Ki-67, TP53RK, and calreticulin were identified. These data were generated by subjecting to mass spectrometric analysis ~42 μg of protein digest, analyzing 16 SCX peptide fractions, and interpreting ~55,000 MS^2 spectra. Total MS time required for analysis was 40 h.

Selective SCX fractions were also analyzed by using a microfluidic LC platform. The performance of the microchip LC was comparable to that obtained with bench-top instrumentation when similar experimental conditions were used. The identification of 5 cancer biomarkers was enabled by using the microchip LC platform. Furthermore, this device was also capable to analyze phosphopeptides.

Acknowledgements

I would like to thank my advisor Dr. Iulia Lazar for her invaluable support, countless hours of guidance, and patience throughout my entire research. This study would not have been possible without her direction and encouragement. I am thankful to my committee members Dr. Brian Love and Dr. Yong Woo Lee for their time, help, and support in reviewing my thesis. I appreciate their willingness to serve on my research committee. I would also like to extend my sincere thanks to Dr. Jung Hae Yoon for assistance with cell culture work.

I am grateful to all my friends here at Blacksburg for helping me and keeping me motivated during my graduate career and my entire stay. Finally, I would like to thank my parents, my husband and my entire family for their never-ending love and support. Without their continuous encouragement, I would not be where I am. This work is dedicated to them.

This research was supported by the National Science Foundation (NSF) under Career grant BES-0448840.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 1

Chapter 3

Chapter 4

Abbreviations

2D-DIGE: 2 dimensional-differential image gel electrophoresis

2D-PAGE: 2 dimensional-polyacrylamide gel electrophoresis

AFP: Alpha fetoprotein

Apaf-1: Apoptotic protease activating factor

ATCC: American Type Collection Culture

ATM: Ataxia telangiectasia mutated

ATR: Ataxia telangiectasia and rad3 related

Bax: BCL2-associated X protein

Bcl2: B-cell lymphoma -2 protein

BOE: Buffer oxide etchant

BRCA: Breast cancer genes

CA125: Cancer specific antigen 125

CAD: Computer aided detection

CDC2: Cell division cycle 2

CDK: Cyclin dependent kinase

CE: Capillary electrophoresis

CEC: Capillary electrochromatography

CHK: Checkpoint homolog

CID: Collision induced dissociation

c-Myc: Myelocytomatosis cancer oncogene homolog

CT: Computerized tomography

DC: Direct current

DTT: Dithiothreitol

E2F: Transcription factor

ECD: Electron capture dissociation

EDTA: Ethylendiaminetetraacetic acid

EMEM: Eagle's minimum essential medium

EOF: Electroosmotic flow

ER: Estrogen receptor

ESI: Electrospray ionization

ETD: Electron transfer dissociation

FBS: Fetal bovine serum

FDG: Fluoro-2-deoxy-d-glucose

FISH: Florescence in situ hybridization

FTICR: Fourier transform ion cyclotron resonance

GADD45: Growth arrest and DNA-damage-inducible protein

GO: Gene ontology

hCG: Human chorionic gonadotropin

HER2: Human epidermal growth factor receptor 2

HPLC: High performance liquid chromatography

Hsp: Heat shock proteins

IEF: Isoelectric focusing

IHC: Immunohistochemistry

IMAC: Immobilized metal affinity column

IRMPD: Infrared multiphoton dissociation

LC: Liquid chromatography

LCM: Laser capture microdissection

LIF: Laser-induced fluorescence

LTQ: Linear trap quadrupole

MALDI: Matrix assisted laser desorption ionization

MEMS: Microelectromechanical systems

Mdm2: Murine double minute 2

MRI: Magnetic resonance imaging

MS: Mass spectrometry

MYT1: Myelin transcription factor 1

NCBI: National Centre for Biotechnology Information

p21Cip1: Cyclin-dependent kinase inhibitor 1A

PBS: Phosphate buffer saline

PCAF: p300/CBP(CREB binding protein)-associated factor

PCNA: Proliferating cell nuclear antigen

PCR: Polymerase chain reaction

PET: Positron emission tomography

PK: Protein kinase

PLK1: Polo-like-kinase 1

PR: Progesterone receptor

PSA: Prostate specific antigen

Rb: Retinoblastoma

RF: Radio frequency

RP: Reversed phase

SAGE: Serial analysis of gene expression

SCX: Strong cationic exchange

SEM: Scanning electron microscope

SPEC: Solid phase extraction cartridge

SWISSPROT: Swiss protein database

TFA: Trifluoroacetic acid

TOF: Time-of-flight

TP53RK: Tumor protein 53 regulating kinase

VEGF: Vascular endothelial growth factor

Wee1: WEE1+ homolog

Summary

The objective of this research was to develop effective bioanalytical strategies for the characterization of proteomic extracts from cancerous cells that would enable the detection of trace level biomarkers. The specific aims of this research are: (1) To develop a 2D-strong cationic exchange/reversed phase liquid chromatography-tandem mass spectrometry (SCX/RP-LC-MS/MS) platform for the characterization of MCF-7 breast cancer cell line; (2) To develop a microfluidic liquid chromatography system for proteomic applications and biomarker screening; (3) To detect phosphorylated peptides from a-casein digest before and after dephosphorylation using microfluidic chips.

The content of this thesis is divided into five chapters to address the above mentioned objectives. Chapter one provides the necessary background information on cancer as a disease, its diagnosis and treatment strategies. It also explains the significance of different analytical tools such as mass spectrometry in proteomics for the early detection of cancer and discovery of biomarkers.

The experimental techniques that were developed for this research are provided in chapter 2. Chapter 3 describes various optimization procedures that were performed to improve the analysis of the MCF7 breast cancer cell line, and presents the results of the research. Chapter 4 describes the development and application of microfluidic devices to analyze one of the MCF7 fractions and to determine phosphopeptides. Chapter 5 concludes the thesis providing suggestions for future work.

Chapter 1: Introduction

1.1 Background information

1.1.1 Cancer: disease, detection, and treatment

1.1.1.1 Introduction

Cancer represents an abnormal and uncontrolled growth of cells that can invade and destroy the surrounding healthy tissues (**Figure 1**). Clinically, cancer can be termed as a collection of over 100 diseases that differ in rate of growth, age of onset, metastatic potential, invasiveness, stage of cellular differentiation, diagnostic detectability, and response to treatment and prognosis [1]. It is believed that these diseases occur as a result of a sequence of mutations that are caused by the environment, by a mutagen, or that arise as an eventual manifestation of genetic flaw [2-5]. Ordinarily, normal cells reproduce only as instructed by the body, based upon a perceived need. In humans, overall health and longevity depends on this ordered process [5, 6]. However, cancer cells proliferate in an uncontrolled manner resulting in a mass of excess tissue, more commonly known as a tumor [3]. Tumors can be classified as benign or malignant (**Figure 2**). Benign tumors are noncancerous, and as the cells are immotile, rarely represent a threat to life (except brain tumors). Surgery can be used to remove such growths. On the other hand, malignant tumors are cancerous, and can invade nearby tissues or organs. The cancer cells may also separate and relocate via the bloodstream or lymphatic system. This process, called metastasis, allows the cancer to spread from the primary (original) tumor, to the other parts of the body to form new tumors [7, 8].

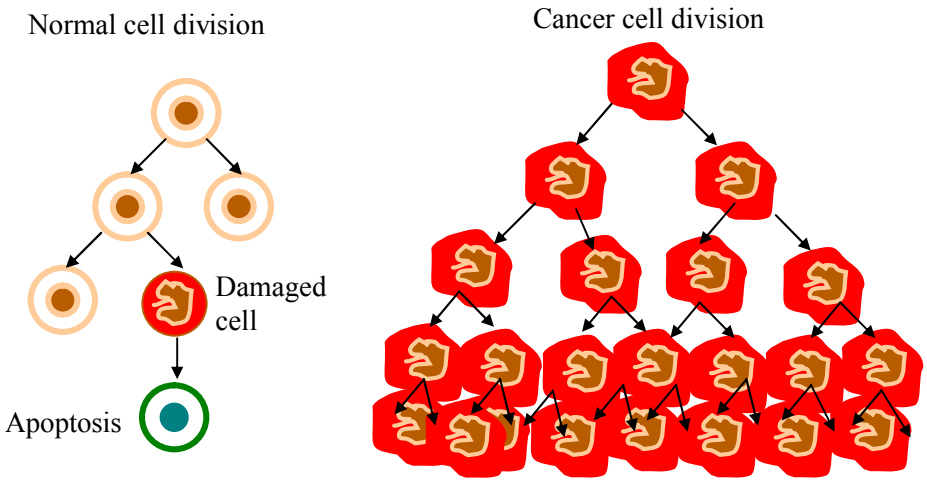


Figure 1. Growth of a normal cell vs. a cancerous cell.

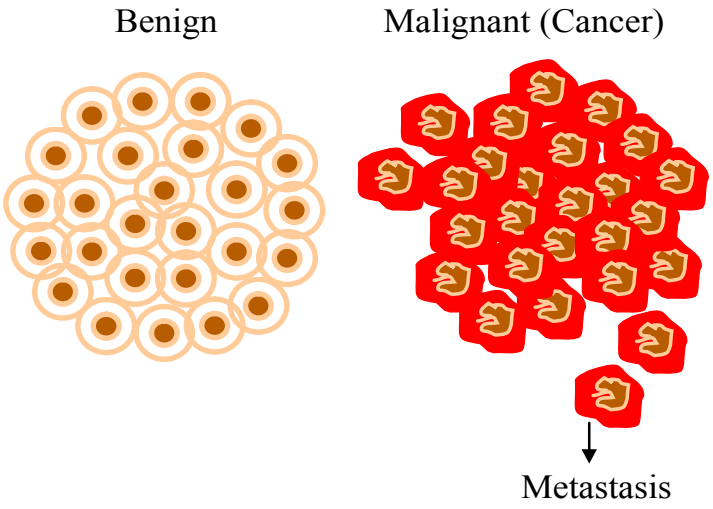


Figure 2. Classification of tumors.

The development of a secondary tumor from a primary carcinoma was first described by Recamier [9]. The term metastasis was introduced as a descriptor for the

migration from the carcinoma in situ, the invasion of new tissue, and the proliferation of secondary tumors [9, 10]. Surgery is far less effective for the treatment of metastatic cancer. Hence, metastasis is one of the more pernicious aspects of cancer. In fact, survival is considered so unlikely upon the discovery of distant metastases that all subsequent treatment is considered merely palliative [5, 6]. Unfortunately, for two-thirds of patients, cancer is already metastatic upon diagnosis [3].

Cancers that affect the bone, muscles, and connective tissues are called sarcomas, those that affect the bloodstream and bone marrow cells are called leukamias, and those that affect the epithelial cells in different organs such as lung, breast, colon, bladder and prostate are called carcinomas. The factors that cause cancer can be internal such as hormones, immune and inherited conditions, or external such as viruses, chemicals, radiation, and lifestyle habits (diet, smoking, and alcohol).

Cancer claims the lives of millions of people worldwide. According to the statistics provided by the American Cancer Society, cancer is the 2nd leading cause of death in the United States after heart diseases. Over 1 million new cases of cancer and >500,000 deaths are expected in 2006. On the average, the diagnosis of the most common types of cancer occurs around the age of 67. Although cancer is relatively rare in children, it is still a leading cause of death between ages 1 and 14. Millions of people who are alive today have had some type of cancer, and only about half are considered cured. Hence it is very important to detect the disease at an early stage to avoid any life threatening risks.

1.1.1.2 Diagnosis

The identification of the disease is largely dependent upon the degree of structural changes. Conventional ways of diagnosing cancer and predicting an outcome involve: clinical examination in combination with pathologic evaluation (examine features of cancer under the microscope), laboratory testing (blood), and imaging studies (X-rays, etc.), as shown in **Figure 3**.

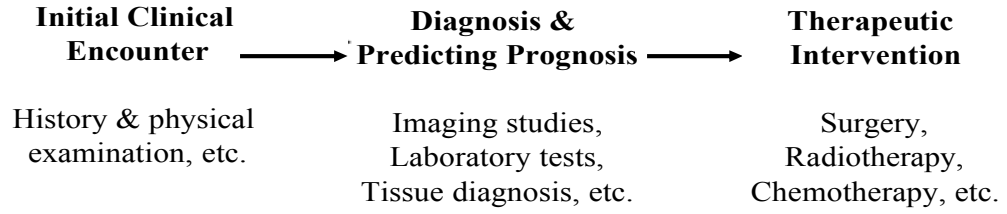


Figure 3. Conventional cancer diagnosis, prognosis & treatment [126].

Various imaging modalities such as computerized tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI), are being used for diagnosis and to track and confirm the effectiveness of a treatment. CT is a diagnostic procedure that uses special X-ray equipment to obtain cross-sectional images of the body, and is used to detect the presence of a tumor, its size, location, and whether it has spread leading to metastasis. There are obvious limitations associated with this technique. For instance, anatomical visualization of the cancerous area is not always possible [11]. The development of new imaging modalities that support the inclusion of metabolic

information do not depend only on anatomical evaluation. PET imaging is based upon the fact that the cancer cells have higher metabolic rates and take up greater amounts of glucose than the surrounding normal tissues. It employs an analogue of glucose, [18F]-2-fluoro-2-deoxy-d-glucose (or FDG), to enter the tumor cell and allow the detection of tumors as high intensity signals [12]. It also allows for the differentiation of benign vs. malignant lesions. The MRI technique uses the magnetic resonance technology to analyze the chemical composition in an area of interest. It is mainly utilized for brain tumors [13]. There is always a risk factor associated with the use of these imaging modalities such as radiation exposure, and allergic reactions to different contrast agents.

Certain laboratory tests utilize cellular and molecular tools for cancer diagnosis. These can be as simple as a blood test, or as complicated as a bone marrow aspiration or biopsy. There are a few diagnostic tools that can be used for the detection of specific cancers. For instance, mammography, ultrasound, and imaging techniques that use computer aided detection (CAD), are used for breast cancer diagnosis [14]. Ultrasound is an imaging technique wherein high frequency sound waves are bounced off the tissues or internal organs to produce an image known as a sonogram. In breast cancer, this technique can help distinguish between solid tumors and fluid filled cysts; the main disadvantage is that it cannot be used consistently to detect early signs of cancer such as microcalcifications [15]. Mammography is the most effective X-ray technique available for breast cancer screening, yet it can be misleading, as the results of mammograms show often false positives that can lead to unnecessary surgery, anxiety and cost [16].

Recently, more accurate laboratory tests that rely on technologies such as polymerase chain reaction (PCR), florescence in situ hybridization (FISH), and flow

cytometry, have been introduced. The sensitivity of the PCR test enables the identification of only a few cells that can be left behind after therapy, or that start to reappear after remission [17]. However, there exists the possibility of contamination, and the quantitation of results is rather difficult. In flow cytometry studies, with the aid of special antibodies against a panel of cell surface markers, doctors can categorize cancers according to the expression patterns of these markers; however, this method lacks sensitivity. Recently, gene mutation has been used for the early detection of cancer [18]. Microarray techniques present unique opportunities to investigate gene function, and provide a versatile platform for utilizing genomic information to benefit human health [19]. Limitations associated with this technique relate to costs and artifacts associated with image and data analysis.

For several decades, physicians and scientists tried to identify specific tumor markers in blood to help diagnose and prognose cancer. The majority of tumors lack specific markers, and up to date, only a handful of tumor markers are used routinely. Some commonly used biomarkers are: alpha fetoprotein (AFP) for hepatocellular carcinoma and germ cell tumors, beta subunit of human chorionic gonadotropin (beta-hCG) for choriocarcinoma, prostate specific antigen (PSA) for prostate cancer, and cancer specific antigen 125 (CA125) for ovarian cancer [20]. Despite their wide clinical application, the lack of sensitivity and specificity generally limits their usefulness. The study of cancer morphology under a microscope and the microenvironment of lymph nodes and/or distant organs still remain some of the most used procedures to predict cancer outcomes.

1.1.1.3 Treatment

Tremendous research is being carried out to discover a new cure for cancer. According to National Cancer Institute the overall budget for research costs in year 2005 will be 6.2 billion dollars. Pharmaceutical companies, biologists, and clinicians are investing considerable effort to improve the prognosis of patients diagnosed with cancer. In order to develop effective treatments it is important to identify the proteins involved in controlling cancer growth. Treatment strategies work differently for different tumors. Even in patients with the same kind of tumor, clinical features such as aggressiveness often differ. In choosing effective treatments with minimal side effects, oncologists rely heavily on biopsy reports that diagnose the tumor type involved.

Surgery remains the most common treatment for cancer. Its purpose is to remove as much of the cancer as possible. Other options include hormone therapy, chemotherapy, radiation therapy, targeted drug therapy, photodynamic therapy, and laser treatment. Hormone therapy is used to keep the cancer cells from getting the hormones they need to grow. This treatment uses different drugs for different cancers, depending upon the related hormonal function. For example, estrogen promotes the growth of about two thirds of breast cancers. As a result, several approaches to block the effect of estrogen, or to lower estrogen levels, are used to treat breast cancer [21]. Tamoxifen is an antiestrogen drug which is used for treating breast cancer. There are many serious side effects associated with the intake of hormone therapy drugs such as blood clots, weight gain, development of some other type of cancer, etc. Radiation therapy is used, at some point, in the treatment of more than half of all cancer cases. High-energy X-rays (ionizing radiation) are delivered using photon beams or particle beams to damage cancer cells, and

stop them from growing and spreading. It can be used to shrink a tumor before surgery, or it can be used after surgery [22]. The disadvantage of radiation therapy is that it can damage normal cells and can cause side effects such as swelling and sunburns in the treated area.

Chemotherapy involves the use of anticancer drugs to kill cancer cells. Unlike surgery and radiation therapy, it is systemic; it works throughout the body. A single drug or a combination of drugs may be used. Chemotherapy is often used after surgery to kill any hidden cancer cells that remain in the body. Chemotherapy reduces the effect of growth hormones, thereby resulting into hair loss, weight loss, loss of blood counts, and premature menopause in women [23]. On the other hand, photodynamic therapy uses a photosensitive agent that is injected in the bloodstream and that can be absorbed by all the cells in the body. This agent stays longer in the cancer cells than in the normal cells, and hence for 24-72 hrs after injection the tumor is exposed to light, and an active form of oxygen is produced that kills the tumor cells [24, 25, 26]. This treatment is limited to the tumors that are about one-third of an inch of tissue. Hence, large tumors as well as metastasized tumors are not treated by this option. In addition, there are many side effects that accompany this treatment, such as sensitivity in the skin and eyes, burns, swelling, pain and scarring of nearby healthy tissue.

Some of the targeted drug therapies for breast cancer include Herceptin (Transtuzumab). About 30% of women with breast cancer have an excess of a protein called HER2, which makes tumors grow quickly. A genetically engineered drug, Herceptin, binds to HER2 and kills the excess cancer cells, theoretically leaving healthy cells unaffected. However, there is a serious side effect related to the use of this drug that

involves weakening of the heart muscle, and possibly congestive heart failure [27]. Though many treatments are available for cancer, technical limitations do apply for each of them. The treatments are all aggressive and result in serious side effects. Moreover, these treatments do not guarantee that cancer will not reoccur. According to the ACS, the best strategy for the successful treatment of cancer is to follow the guidelines for early detection.

1.1.2 Molecular mechanisms and pathways in cancer development

Over 2,500 genes are believed to be involved in the molecular mechanisms and pathways of cancer. These genes are involved in a variety of functions such as cell growth (oncogenes), control of blood supply to the cell (angiogenesis genes), tumor suppression (tumor suppressor genes or antioncogenes), and apoptosis (apoptosis genes) [28,29]. Oncogenes are affected by mutations that result in gain of function, and encode primarily growth factors, growth factor receptors, signal-transduction proteins, transcription factors and cell-cycle control proteins. Tumor suppressor genes are affected by mutations that result in loss of function, and generally encode proteins that inhibit cell proliferation, i.e, cell-cycle control/checkpoint proteins, receptors for secreted hormones, proteins that promote apoptosis, and DNA repair proteins. Other mutations can disable proteins such as p53, which triggers the cell to commit suicide (undergo apoptosis) if its DNA becomes damaged, or if its signaling cascades go out of control [28].

Several genes such as Bax and Bcl2 [30, 31] control the proapoptotic and antiapoptotic signals. Mutations of apoptotic genes result in either abnormal apoptosis, or complete inhibition of the process leading to tumor development. Angiogenic genes control the cell's blood supply, and when mutated, gain function and increase the

production of factors that promote the growth of new blood vessels such as vascular endothelial growth factor (VEGF), fibroblast growth factors, and platelet derived growth factors [32]. When this occurs, hundreds of new capillaries converge on the tumor, and supply it with blood to grow in size. Cancer metastasis results as a deregulation of so called metastasis genes which aid the mobilization of tumor cells from the primary tumor mass to invade the body. These genes encode proteolysis factors (e.g., serine proteases, cathepsins, metallo-proteases), angiogenic factors (VEGF), and factors related to cell adhesion and migration [33].

Quantitative changes in gene expression at mRNA level that may cause cancer are measured routinely using DNA microarray analysis. However, a quantitative correlation between mRNA transcription and protein translation levels does not always exist. Many events under translational or post-translational control interfere in overall protein expression. For example, translation rates and protein synthesis increase in response to stimuli such as growth factors, cytokines, hormones, mitogens, viral infections, etc. [34, 35]. Moreover, selective translation control has been shown to affect a number of genes involved in growth and apoptotic processes (and their protein products). In addition, alternative splicing and post-translational modifications result in a large number of protein isoforms, further increasing the complexity of the proteome. For example, phosphorylation is one of the essential post-translational modifications which regulate protein function in signaling pathways, thus affecting essential cellular processes such as cell division, growth, differentiation and death. Alternatively, protein glycosylation also plays an important role in signaling, as many glycoproteins act as cell surface recognition molecules; alterations in the glycosylation of cell surface proteins has

been shown to be involved with many diseases including cancer. In conclusion, to fully elucidate the molecular mechanisms that govern cancer initiation and progression, the information generated at the DNA and mRNA level must be complemented with a complete and detailed panorama of protein expression levels and their post-translational modifications.

1.1.3 Breast cancer

According to the National Cancer Institute, breast cancer is the most prevalent form of cancer today in the USA. Breast cancer is the leading cause of cancer death in women of age 40-55, and the second leading cause of death, after lung cancer, in women of other age groups. According to the American Cancer Society, at least 211,240 new cases of invasive breast cancer will be diagnosed this year, and approximately 40,410 women will lose their lives to this disease [36]. Every woman is at some risk for breast cancer, and it is estimated that one in eight women has or will develop breast cancer in her lifetime.

Breast cancer develops from the breast tissue in form of a malignant tumor. The human mammary gland is basically composed of ducts, lobules, and stroma, as shown in **Figure 4**. The lobules are milk producing glands, and the ducts are milk passages that connect the lobules to the nipple. The stroma consists mainly of fatty and connective tissue that surrounds the breast, lymphatic vessels, and blood vessels. The lobules and ducts are formed by epithelial cells surrounded by myoepithelial cells. The relative variation in the morphology and metabolism of the mammary gland throughout the life of human is due to the stimulation of epithelial cells by different hormones and growth factors [37].

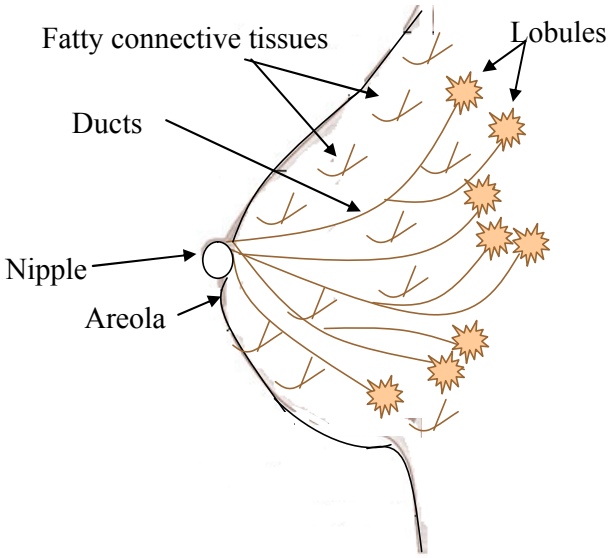


Figure 4. Anatomy of the human mammary gland.

According to the site of occurrence, breast cancer is classified as ductal carcinoma, if it begins in the cells lining the ducts, or lobular carcinoma if it occurs in the cells lining the lobules. The cancer that is confined within the ducts and lobules is also called in situ carcinoma. Alternatively, if the cancer cells invade the surrounding fatty tissue by destroying the basement membrane and become metastasized, the cancer is termed as invasive carcinoma. The seriousness of the invasive carcinoma is strongly influenced by the stage of the disease when it is first diagnosed. According to the American Joint Committee on Cancer (AJCC), tumor size, lymph node involvement, and the presence or absence of distant metastasis, are used in the clinical setting to determine the stage of cancer (I, II, III, or IV). Stage I is a benign or early stage, while stage IV is the most advanced stage of cancer [38]. About 80 % of the invasive cancers are invasive ductal carcinomas, while the invasive lobular carcinomas account for only about 10 %.

Other less frequent invasive cancers include inflammatory breast carcinoma, medullary, and mucinous carcinoma [37]. Unlike in situ and invasive carcinomas, Paget's disease of the nipple spreads to the skin of the nipple and then to the areola.

Most of the breast tumors are benign, i.e., they are not cancerous, but abnormal growths that do not spread outside the breast to other organs. These include fibroadenomas or papillomas. However, some of the benign tumors such as atypical hyperplasia, and a few papillomas, have higher risk of getting converted to breast cancer [37].

1.1.4 Biomarkers and technologies for biomarker detection

A biomarker is a molecular entity that is objectively measured and evaluated as an indicator of a normal or diseased biological process, pathogenic process, or pharmacological response to a therapeutic intervention [39, 40, 41]. A biomarker can be a gene, protein, or metabolite that is either found in body fluids such as blood, plasma, and urine, or in tissues. Diagnostic biomarkers can be used for the detection, prognosis, and staging of a disease, as well as for identifying targets for drug development. They are of considerable utility as risk factor indicators for individuals that are susceptible to a disease. The ideal biomarker for assessing the risk of cancer is an early molecular alteration during the premalignant phase of neoplasia, that can predict with high sensitivity and specificity the future progression and invasive potential of malignant cells. The selection of appropriate biomarkers is thus of critical importance to ensure high precision risk assessment to individuals or population sub-groups.

Based on their ability to detect the disease, biomarkers can be classified into detection, high risk, and prognostic biomarkers [39, 40]. One or more of these functions

could be served by a single marker, and hence the marker may fall into more than one category. For example, PSA is used to detect prostate cancer and to monitor its evolution during treatment [42]. CA125 is routinely used for detecting ovarian/cervical cancer [43]. Early detection biomarkers can facilitate a timely intervention in the natural progression of cancer, to inhibit, reduce, or even eliminate the disease. These biomarkers aid the classification of tumors and the staging of the disease, which are both essential steps for choosing a proper treatment strategy [39]. Initially, single markers were used to evaluate the disease; however, in a complex disease such as cancer, high specificity is hard to achieve with a single marker. For instance, PSA has relatively low specificity in prostate cancer diagnosis due to high concentrations in both benign and malignant cancers [44]. a-fetoprotein, which is used as a biomarker for hepatocellular cancer, does not show any change in concentration for patients with small tumors [45].

Biomarkers for risk assessment can provide valuable information to identify individuals who are at risk to develop cancer, before the actual onset of the disease [40]. These biomarkers mainly encode genetic information, i.e., they are mutated genes. For example, mutations in the BRCA1 and BRCA2 genes are used to identify the risk of breast cancer [46]. Similarly, adenomatous polyposis syndrome, a genetic alteration that results in colonic polyps, is used to evaluate the risk of colon cancer [45]. Mutations in the p-53 gene can serve for risk assessment in many types of cancer.

Prognostic markers aid in treatment decisions by providing information about the malignant potential of tumors [39]. Some of the clinically used prognostic markers include hormone receptors such as estrogen (ER) and progesterone receptors (PR), cell proliferation markers such as proliferating cell nuclear antigen (PCNA) and Ki-67,

protein markers of angiogenesis (VEGF), growth factor receptors (HER-2/neu), and tumor suppression protein p53 [47]. Anti-estrogen compounds are used for the treatment of estrogen receptor-positive breast cancers, whereas chemotherapy is used to treat estrogen receptor-negative tumor patients [39].

A list of some potential cancer biomarkers that are reported in the literature is provided in **Table 1,** along with their class/function, method of identification, and type of cancer they are involved in. These biomarkers have a large variety of functions. They can have specific roles in cell cycle regulation, DNA repair, cell differentiation and proliferation, or can have receptor or enzymatic activity. Earlier, most biomarkers were identified using genomic approaches such as DNA microarray analysis and serial analysis of gene expression (SAGE), and other tests such as immunohistochemistry (IHC) and antibody assays; however, with advancements in the field of proteomics, increasing interest is developed towards the discovery of novel protein markers using proteomic technologies [39, 48]. Kolch *et al.* described three main areas of focus in proteomic research for biomarker detection: (1) The analysis of the tumor cell proteome for the discovery of new targets for therapeutic intervention; (2) The analysis of protein markers in the tumor, surrounding tissue, or body fluids for early diagnosis of the disease; (3) The analysis of markers that enable response to the therapy monitoring [49].

Proteomic technologies that are most commonly used for biomarker detection include 2D-gel electrophoresis, 2D-differential image-gel electrophoresis (2D-DIGE), liquid phase separations such as liquid chromatography (LC) and capillary electrophoresis (CE), and protein chip arrays; all these techniques can be coupled with mass spectrometric detection [39]. Mass spectrometry has become an indispensable tool

for protein identifications. Further information about these technologies is provided in the following sections.

Table 1 Cancer biomarkers reported in the literature.

Biomarkers	Class/function	Method of identification	Type of Cancer	Reference
Prostate specific antigen (PSA)	enzymatic activity	IHC	Prostate	20
CA-125	cellular proliferation & apoptosis	IHC	Ovarian (later stage)	20,94
Lysophosphatidic acid (LPA)	growth factor		Ovarian (early stage)	95
Telomerase	cellular inducible enzymes	PCR based assay	Lung	95,96
Heat shock protein 27,60,90	cell cycle regulators	matching with published maps	Breast	97-100
Vascular endothelial growth factor	Angiogenesis factor	protein assay/IHC	Breast	96,101,102
Alpha- fetoprotein (AFP)	apoptosis		Hepatoma; Testicular	20
Chroriogonadotropin (hCG)	human chorionic gonadotropin pathway		Testicular; breast	20,102
Steroid hormone receptors (ER/PR)	hormone receptor pathway	IHC binding assay	Breast	20,96,101,102
Cathepsin D	Oestrogen receptor pathway	antibody/Immunoassay	Breast, colorectal, squamous	96,97,102,103
Insulin like growth factor (IGF)	Insulin and Insulin like growth factors		Breast	102
Epidermal growth factor type 2 (HER2)	Membrane receptor and signal transduction	IHC FISH	Breast	96,101,102,104
Carcinoembryonic antigen (CEA)	Regulation of signal transduction		Colon; breast;lung;pancreatic	20
CA 15.3	cell surface antigen	Monoclonal antibodies	Breast	20
CA19.9	cell surface antigen		Gastrointestinal	20
14-3-3 sigma	molecular chaperone	2D gel/mass spectrometry	Breast	97,105,106
Bcl-2	Apoptosis	IHC	Breast	96,101,102
Cyclin D	cell cycle regulators	IHC	Breast	96,102
E-Cadherin	cell surface receptor	IHC/Methylation-PCR	Breast	107-109
BRCA 1	Regulating pathways controlling cell proliferation &differentiation		Breast, ovarian	20,96,101,102,110
BRCA 2	Regulating pathways controlling cell proliferation &differentiation		Breast, ovarian	20,96,101,102,110

S100 calcium binding protein	Cytoskeleton and Cell adhesion	2DE array	Breast	111,112
Mammary type apomucin MUC-1	Membrane receptor and signal transduction		Breast	102
Matrix metalloproteases MMP-2	cellular inducible enzymes	IHC	Breast	96,102
Cyclooxygenase-2 COX-2	cellular inducible enzymes		Breast	102
Cytokeratins 8, 18, 19, 5	Cytoskeleton and cell adhesion	antibody/mass spectrometry	Breast	97,103
Ki-67	proliferation	Immunohistochemistry	Breast, lung	113,114
p53	nuclear protein	IHC/SSCP sequencing	Breast, ovarian, lung, colorectal	115-117
Calreticulin	molecular chaperone		Breast	97,98,118,119
Breast carbonic anhydrase	metabolic enzyme	antibody	Breast	98
Nuclear matrix proteins	skeleton of nucleus	subcellular fractionation	Breast	120,121
Tropomyosin 1,2, 3	cytoskeleton	comigration	Breast	97,103,122
Inosine-5 monophosphate dehydrogenase	enzyme	comigration antibody	Breast	97,123
Tumor necrosis factor	angiogenesis related		Breast	101
Aldolase A	glucose metabolism		Renal;colorectal;breast	103
Antigens CD9, CD97	Cell migration, adhesion & signaling	2D-DIGE	pancreatic, gastric	124,125

1.2 Mass spectrometry and proteomics

1.2.1 Introduction to mass spectrometry

Mass spectrometry is a powerful analytical technique for analyzing ions in the gas phase. Ions are separated according to their mass to charge ratio (m/z), and are detected in proportion to their abundance. The detection of minute quantities of compounds can be accomplished with this technique. The mass spectrometer generates a mass spectrum, which is a plot of ion abundance versus mass to charge ratio (m/z). Mass spectrometry can be used in a variety of applications to determine the molecular weight, chemical structure, sequence information, isotopic ratio, and the reaction kinetics of compounds, as well as to identify and quantitate various components in complex biological samples. The power of MS detection stems from low detection limits, sensitivity, high mass accuracy, resolution and dynamic range, as well as from the capability to generate a wealth of information. [50].

Prior to mass spectrometry based protein identification, Edman degradation was the method of choice. Edman sequencing is a powerful and simple technique, but is very slow as it identifies only one amino acid from a purified protein at a time. Hence, for a complex sample with thousands of proteins, this technique would be almost impossible to use [51]. The technique of mass spectrometry originated in J.J. Thomson's vacuum tube that demonstrated the existence of electrons and positive rays [52]. Originally a physicist, he proposed that this technique could be very useful to chemists to analyze chemicals. Though for many years the applicability of mass spectrometry remained confined within

the physics labs to determine relative abundances of isotopes and their exact masses, the fundamental knowledge that was gained enabled the further development of mass spectrometry and its acceptance in diverse areas of science. The past decades have witnessed a dramatic evolution of mass spectrometry instrumentation and methodologies with unique capabilities for a variety of applications. The use of MS expands in the area of proteomics for qualitative and quantitative characterizations, and quality control of recombinant proteins. In various biomedical applications, it is used for the identification of proteins, drugs, and metabolites. Moreover, it is the method of choice for the detection and characterization of post translational modifications and covalent modifications that alter the mass of a protein, and that can be a potential cause for the onset of a disease. In addition, the broad applicability of mass spectrometry is complemented by its unparalleled versatility for interfacing with other analytical methods.

The main components of a mass spectrometer include sample delivery system, ion source, analyzer, detector, computer, ion optics elements, vacuum pumps, and power supplies [53]. The sample can be introduced through a direct probe, separation interface, or microfluidic device. Once the sample is introduced, the ion source converts the sample into gas phase ions that are sorted in the mass analyzer according to their mass to charge ratio (m/z). The ions are detected by a detector and the ion flux is converted into an electrical current. The data system records the magnitude of these electrical signals as a function of m/z, and converts this information into a mass spectrum that appears on the computer. Ion optics elements such as lenses, apertures, cylinders, quadrupole, etc., can guide, disperse, or focus the beam of ions with the aid of an electrical field. Turbopumps

are used to create high vacuum and enable efficient detection with electron multipliers or multichannel plates. The block diagram of a mass spectrometer is shown in **Figure 5**.

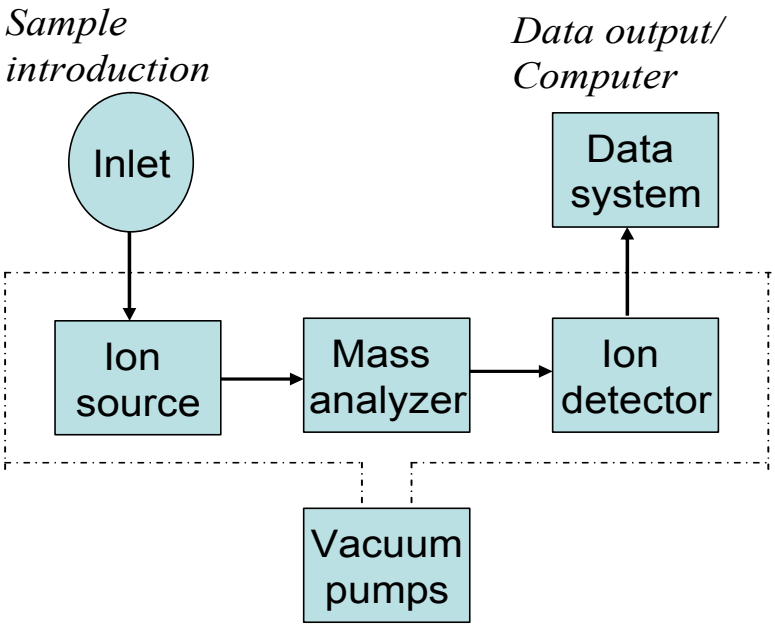


Figure 5. Block diagram of a mass spectrometer with its components.

1.2.2 Ionization methods and mass analyzers

The popularity of mass spectrometry in life sciences has been promoted by the development of two ionization methods, electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). There are many other ionization mechanisms such as electron ionization, atmospheric pressure chemical ionization, desorption ionization, electron capture dissociation, etc., which are appropriate for certain applications, but are not as popular as ESI and MALDI. These techniques are often termed as soft ionization methods, as they are capable to generate ions from large, nonvolatile analytes such as proteins and peptides, with almost negligible analyte fragmentation [54]. ESI can ionize analytes from liquid solutions, and is commonly used for the analysis of complex samples

processed with liquid phase separation methods. Alternatively, MALDI ionizes sample directly from a dry matrix, and is normally used for the analysis of simple peptide mixtures [53, 54].

Fenn was among the first scientists who developed the technique of electrospray ionization MS [55]. A liquid that contains the analyte of interest is pumped at low flow rates through a capillary maintained at high voltage. Under the influence of a strong electric field, the liquid stream emerges into a so called Taylor cone, and then disperses into small charged droplets (a spray). These droplets are desolvated as they pass through the atmospheric pressure region of the source towards a counter electrode. As the size of the droplets decreases, the repulsive forces between the charged analytes on the surface of the droplet increase sufficiently to overcome the cohesive forces of surface tension, and result into the explosion of the droplet. Multiple droplet evaporations and disintegrations continue until gas phase ions are produced. To assist the desolvation process, a stream of dry nebulizer gas is often introduced into the spraying region. Analyte ions are then introduced into the source of the mass spectrometer [50, 56]. A schematic representation of the electrospray ionization process is shown in **Figure 6**. Overall, the ESI signal intensity depends on the flow rate, the concentration and conductivity of the solution, and on the ionization efficiency. The majority of ions produced by electrospray are multiply charged, and as the mass spectrometer measures the m/z ratio of an ion, the mass range of the instrument may be extended by a factor equivalent to the number of charges residing on the analyte molecule [50, 56].

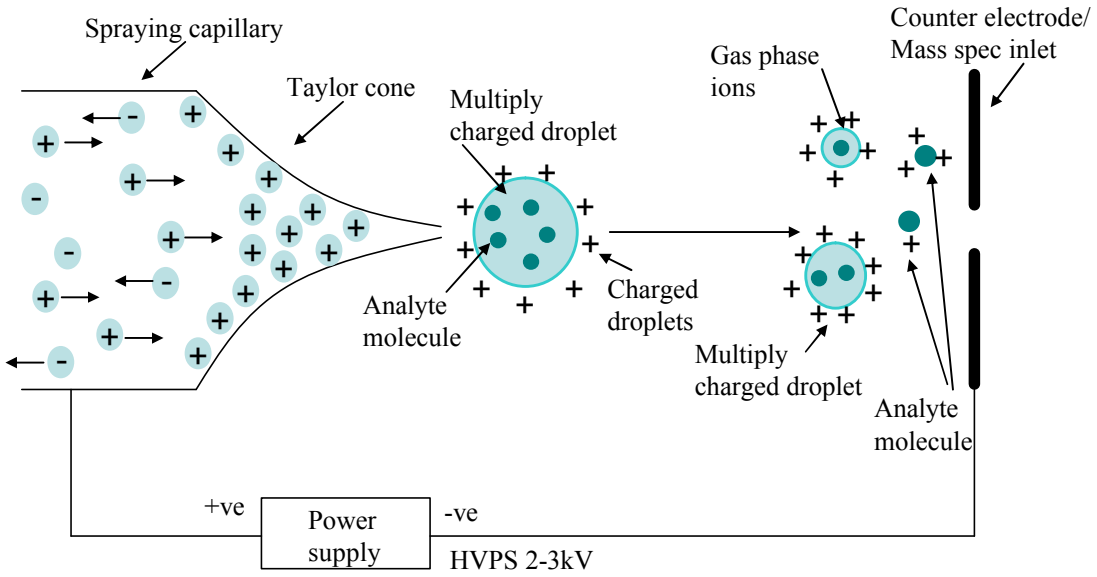


Figure 6. Schematic representation of the ESI process.

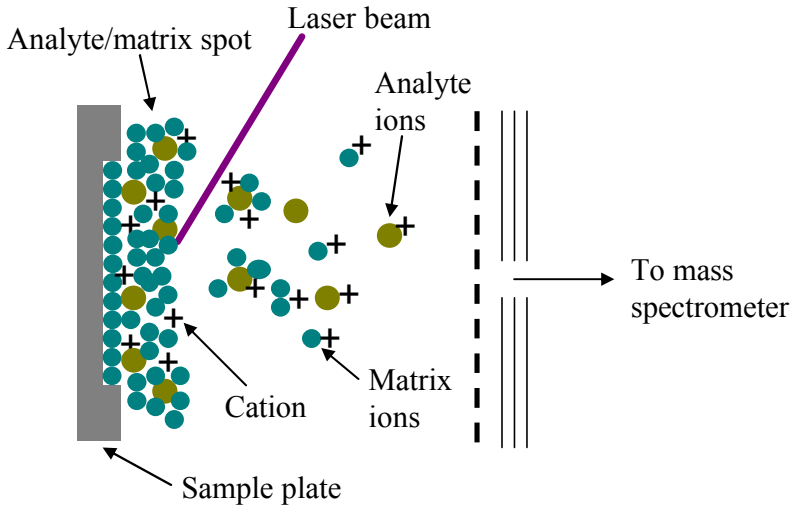


Figure 7. Schematic representation of the MALDI process.

The MALDI process requires the mixing of a sample solution with a volatile matrix solution. This mixture is deposited on a sample plate and allowed to dry to form a homogenous solid solution of sample/matrix, such that the analyte molecules are isolated from one other. Thereafter, a short pulse of a laser beam is focused on the sample/matrix spot at a wavelength which can be absorbed by the matrix, thus resulting into rapid heating of the matrix crystals. This leads to the local sublimation and expansion of the matrix crystals into the gas phase, a process that carries along the analyte molecules, as well. The exact nature of the ionization mechanism in MALDI is not fully understood [50, 57]. The MALDI process is shown in **Figure 7**. Unlike ESI, the majority of ions generated by the MALDI process are singly charged. This ionization technique is mainly used for biopolymers and synthetic polymers (>150,000 Da).

Mass spectrometers are classified based upon the type of mass analyzer. The key quality characteristics of a mass analyzer are sensitivity, resolution, mass accuracy, precision, dynamic range, speed (spectral acquisition/storage rate), duty cycle, and the ability to generate information rich tandem mass spectra (MS/MS) from peptides [53, 58]. There are four basic types of mass analyzers: quadrupole, ion trap, Fourier transform ion cyclotron resonance (FTICR) and time-of-flight (TOF). Each of these has its own strengths and weaknesses. These mass analyzers are discussed in the following sections.

The quadrupole mass analyzer consists of four parallel metal rods (**Figure 8**). Each opposite rod pair is connected electrically, and a radio frequency (RF) voltage (180° out of phase) is applied to each pair. A direct current (DC) voltage is superimposed on the RF voltage. Ions travel down the quadrupole in between the rods. Only ions of a

certain m/z will reach the detector for a given amplitude of the RF and DC voltages; other ions will have unstable oscillations and will collide with the rods. This allows the selection of a particular ion for detection, or scanning the ions by varying the amplitude of the RF and DC voltages. Hence, this analyzer acts as a mass filter. This instrument has good reproducibility but limited resolving power [50, 58].

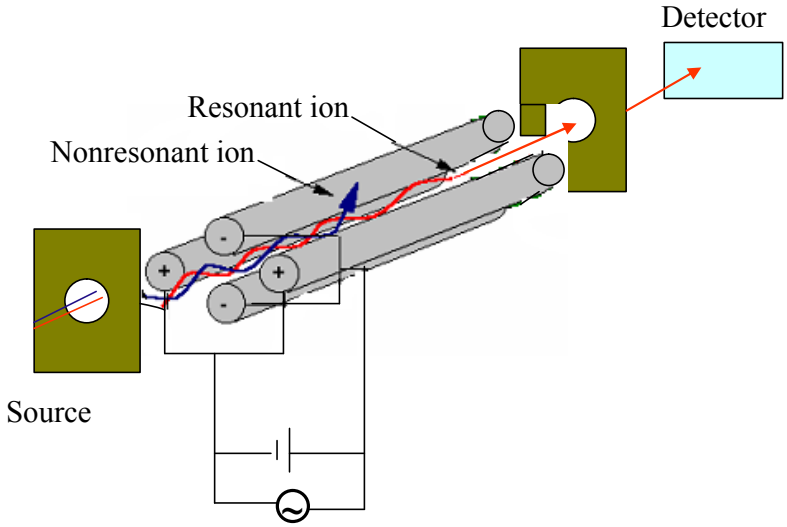


Figure 8. Construction of the quadrupole mass analyzer.

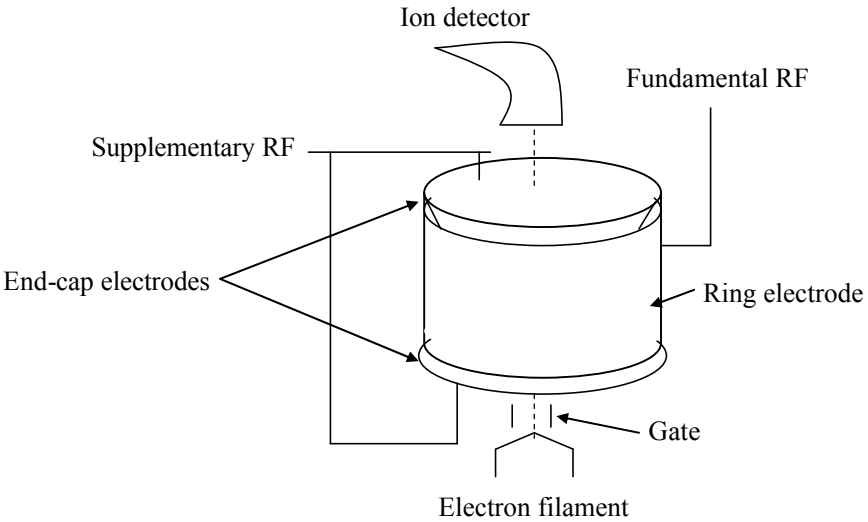


Figure 9. Construction of the ion trap mass analyzer.

There are basically two types of ion trapping mass spectrometers, the quadrupole ion trap often referred as the ion trap MS (dynamic traps) and FTICR MS (static traps). The schematic representation of these instruments is shown in **Figures 9** and **10**. The ion trap mass analyzer is basically a three- dimensional quadrupole that consists of a ring electrode and two end-cap electrodes above and below the ring, forming a trap. The ions are dynamically trapped in a 3D-quadrupole field by a combination of DC and RF potentials. The ion trap is filled with helium gas that promotes fast contraction of ion trajectories towards the center of the trap, as well as enables ejection of ions in dense packets during the mass analysis step. The RF voltage can be manipulated to cause ion excitation and ejection from the trap, or ion isolation and fragmentation [50, 58]. Ion trap mass spectrometers are simple, robust, sensitive, relatively inexpensive, and have the ability to perform multiple stages of MS. The only limitation of this instrument is low mass accuracy and resolution. FTICR MS determines the m/z ratio of the ions based on the cyclotron frequency of ions in a fixed magnetic field. The cubic ICR cell consists of three pairs of parallel plates to which a small potential is applied to keep the ions contained within the ICR cell. Ions move in a circular path perpendicular to the magnetic field. A packet of ions of the same m/z will have the same cyclotron frequency in the cell. The packet of ions, as it is approaching or getting away from the receiver plates in the ICR cell, induces an image current that can be amplified and digitized. The frequency of this current is equivalent to the cyclotron frequency corresponding to a particular m/z ion, and its amplitude is proportional to the abundance of these ions. The image currents induced in the receiver plates will contain frequency components from all mass-to-charge ratio ions. The Fourier transform technique will convert a time-domain

signal (the image currents) to a frequency-domain spectrum (the mass spectrum) [50, 58]. These instruments have the highest resolution, mass accuracy and dynamic range. However, their use in proteomics research is limited due to complex operation, cost, and low peptide-fragmentation efficiency [53].

Time-of-flight mass analyzers measure the time it takes for an ion of a given m/z to fly from the ion source to the detector. Ion packets generated in the ion source are accelerated and expelled within a very short time into a field free region known as the flight tube. During their flight path, the ions are separated according to their m/z. Light ions arrive earlier at the detector relative to heavy ones. As the flight time is proportional to the m/z, the m/z ratios can be determined by measuring the flight time. TOF analyzers have large mass range, high transmission efficiency, and are capable of fast analysis. However, the mass resolution of TOF instruments is poor. Some of the methods employed to increase the resolution include the use of delayed pulse extraction and/or of a reflectron. Nevertheless, high resolution is achieved at the cost of sensitivity and m/z range [50, 58]. A schematic representation of a TOF mass analyzer is shown in **Figure 11**.

Various combinations of analyzers and ionization source are commercially available. Hybrid instruments that take advantage of two mass analyzers such as quadrupole TOF-MS, FT-iontrap MS, and TOF-TOF MS are also available [53]. However, the sensitivity and capability of the ion trap mass analyzer to generate MS^n data, and the high ionization efficiency of ESI, result in a suitable instrumental arrangement for the analysis of complex biological samples.

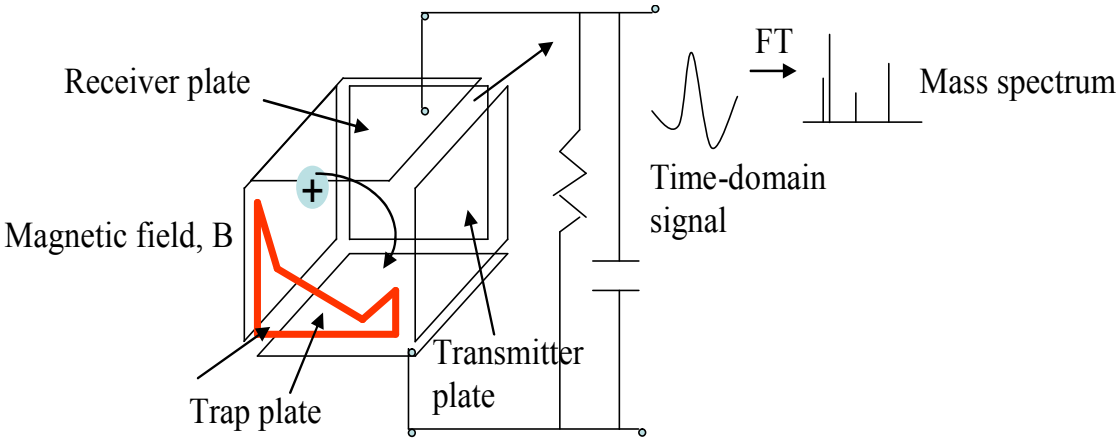


Figure 10. Construction of the FTICR mass analyzer.

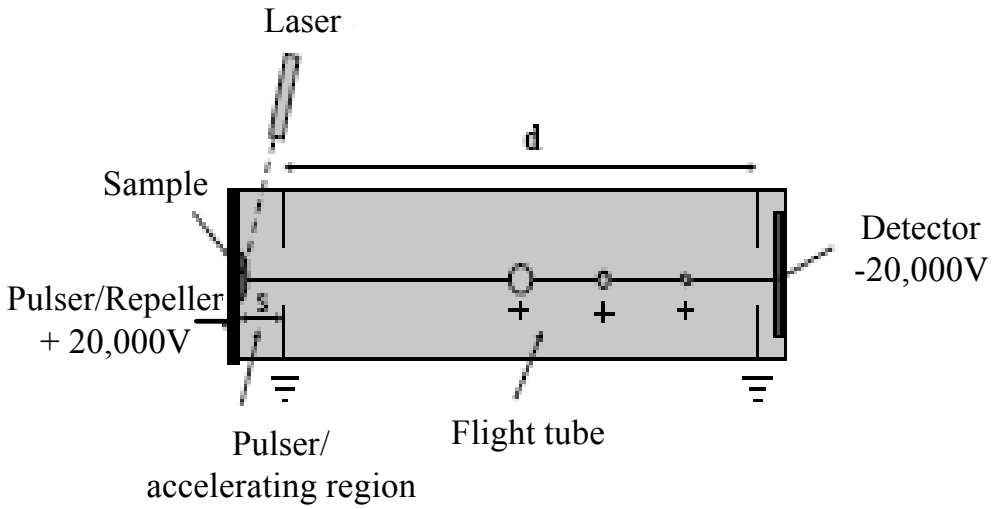


Figure 11. Construction of the TOF mass analyzer.

1.2.3 Tandem mass spectrometry

There are different mass spectrometric techniques used for protein identification. The most common ones are peptide mass fingerprinting (PMF), peptide sequence tagging, and MS/MS peptide sequencing. Peptide mass fingerprinting is a technique wherein a protein is first digested with an enzyme, the absolute mass of the cleaved peptides is measured with MS, and then compared to *in silico* generated theoretical masses. Software programs translate the sequenced genome of an organism into proteins, and then theoretically cut the proteins into peptides with a specific enzyme, to calculate the absolute masses of the peptides. A comparison is made between the experimental peptide mass list of the unknown protein and the theoretical peptide mass list of each protein encoded by the genome. The results are statistically analyzed, and the best protein match to the submitted peptide query is obtained [51, 59]. In sequence tagging, a peptide is fragmented in a mass spectrometer and the sequence of amino acids (tag) is determined. The peptide mass, the sequence of the tag, and the starting and ending masses that flank the tag are used to search a protein database [60]. Tandem mass spectrometry allows for the fragmentation of selected ions to generate structural information. This is used to identify and characterize an unknown biomolecule, to verify that a peptide has a predicted sequence, or to determine posttranslational modifications [54]. Database searches performed with multiple MS/MS spectra are far superior to identifications made with PMF. A single peptide mass can never be confidently correlated to a protein, however, the information-rich fragmentation spectra are reliably matched with a particular protein. High mass accuracy is always required for confident identifications, especially when large databases are used in the search [61].

Tandem mass spectrometry involves multiple steps of mass selection for fragmentation. There are various methods for fragmenting molecules, including collision induced dissociation (CID), electron capture dissociation (ECD), electron transfer dissociation (ETD), and infrared multiphoton dissociation (IRMPD) [58]. The most commonly applied CID method fragments ions by successive collisions with inert gas molecules. MS/MS data generation can be performed in a single mass analyzer over time, such as in the case of ion traps, or in a couple of mass analyzers. For example, one mass analyzer can isolate a peptide (precursor or parent ion) from a mixture of ions entering a mass spectrometer. The second mass analyzer fragments the peptide ions by collisions with a gas (CID) and produces daughter or product ions, and a third mass analyzer catalogs the fragments produced from the parent peptide [62]. Peptide fragmentation typically occurs along the polypeptide backbone to produce y-type (ions that contain the C-terminus) and b-type (ions that contain the N-terminus) product ions. At times, double cleavage of the peptide backbone results in the formation of immonium and internal cleavage (other than N- or C- terminus) ions [62].

Manual sequencing of a fragmentation spectrum is a very difficult task due to the presence of internal fragments, the lack of fragmentation at certain sites, and the difficulty in identifying the fragmentation ion series. Protein identification using tandem mass spectra has been made much easier through the development of computer algorithms that correlate and interpret the CID spectra [63]. The Sequest algorithm selects the top 500 peptide sequences from the database that have the same m/z as the precursor ion selected for fragmentation, and creates a virtual spectrum for each of these peptides that is then cross correlated with the experimental spectrum. Finally, each

comparison is ranked based upon the given scores, and the top score spectrum is displayed [54, 63].

1.2.4 Multidimensional separations – complex samples

The complexity of biological samples requires the development of novel strategies for their complete separation. Various analytical separation techniques have been coupled to mass spectrometric detection; these include two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), and capillary separations such as chromatography (liquid, gas) and capillary electrophoresis. One of the most commonly used techniques for the separation and characterization of proteins is 2D-gel electrophoresis. This technique enacts an isoelectric focusing-IEF (differences in net charge) based separation in one direction, and a size based separation (differences in molecular masses) through polyacrylamide gel electrophoresis in the orthogonal direction. The gels are stained (coomasie blue), and images are taken through a laser scanner or fluorescent imager. Thus, the separation occurs in two spatial dimensions, and the analyst then uses techniques such as mass spectrometry on the resulting spots for the identification of the separated components. This separation system represents the workhorse in many areas of biochemistry and biology, but is particularly valuable when large scale proteomics data need to be generated [64, 65]. However, the technique is time consuming and has many limitations: low abundant and very acidic/basic proteins are lost, several proteins coelute in a single spot, hydrophobic proteins are not soluble in typical gel buffers, and the sample is easily contaminated with keratins and detergents, etc. [54]. Within the past decade new interest in liquid-phase separations has been reported, emphasis being placed on the development of instrumentation, applications, and

theory. Capillary electrophoresis encompasses a family of related separation techniques that use fine fused-silica capillaries (< 100μm i.d) to separate a complex mixture of analytes based on their charge and size, in high electric fields. Although CE provides rapid and efficient separations, it suffers from some limitations such as limited sample loading capacity, and compatibility issues with ESI-MS due to the high concentration buffers that are used in the system [66, 67].

Chromatography is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary while the other is mobile. The sample is loaded on the head of the separation column as a short plug, and is carried down the column by the continuously flowing mobile phase. The sample components distribute themselves between the stationary and liquid phase to various degrees; the ones that interact strongly with the stationary phase are retained longer in the column than the ones that prefer the mobile phase environment and move along with the eluent. Various sample components move down the column with different rates and separate in individual zones [67, 68]. Based on the mechanism of retention, there are different chromatographic methods, such as adsorption, ion-exchange, size-exclusion, and partition.

The quest for a substantial increase in peak capacity, and therefore in the number of compounds that can be separated in a chromatographic run, led to the development of multidimensional techniques. High-performance liquid chromatography (HPLC), based on a partition/adsorption mechanism, is most commonly used in proteomic applications [67]. The potential of two dimensional combinations such as [69] size-exclusion chromatography and reversed phase-LC [70], or reversed phase-LC and

CE [71], combined with MS detection has been reported. High peak capacities were reported in systems using strong cation exchange and reversed phase LC columns [72]. The identification of 80 proteins was made possible by direct analysis of the yeast ribosome complex in a single experiment [72, 73]. Similarly, SCX followed by off-line nano-LC, with long eluent gradients, was able to identify more than 500 proteins from a yeast extract [62]. The overall separation power of these techniques can be significantly extended by using multi-dimensional separations based on orthogonal mechanisms of interaction with the sample components.

1.2.5 Challenges in proteomics research

Proteomics is a relatively new trend in biological sciences that has witnessed tremendous development after the sequencing of the human and several other genomes. Proteomics refers to the comprehensive analysis of protein expression in a cell, tissue, or microorganism; its goal is to identify all the proteins, quantify their expression level, determine protein-protein interactions and post-translational modifications, and finally assign a function to each protein. The challenges of proteomics arise as a result of sample complexity (10^3-10^4proteins/sample), wide range of concentrations (dynamic range of 1: 10^6), low level expression (<1000 copies/cell), limited amount of sample (10^8-10^9 cells), and dynamic composition (different sets of proteins are expressed in various stages of cell development). MS-based detection approaches that provide high sensitivity, throughput, as well as high-confidence protein identifications, are highly desirable. Substantial amount of work is required to develop optimized protocols for the accurate characterization of expressed proteins within a cell. These proteins play an important role in establishing the biological phenotype of a healthy vs. diseased organism. Hence, it is

very important to initially identify these proteins with a reliable approach that will enable further quantitation and differential expression analysis. Proteomic technologies have the power to identify biomarkers specific to a certain disease. However, the gap between what can be measured in a lab, and what can be used effectively in the clinical settings, is broad. Biomarker validation, using complementary proteomic and genomic technologies, is the key factor limiting the migration to routine diagnostics.

The large amount of proteomic data, which are generated by the combination of various analytical technologies with mass spectrometry detection, requires the development of effective bioinformatics tools and high quality databases for accurate interpretation of results. Database integration from multiple sources, and the development of user interfaces that allow data entry, visualization and retrieval, are the key elements in this effort. Inter-laboratory comparisons should be performed for further confirmation and validation of the results.

1.2.5 Proteomic-mass spectrometry methods for cancer cells analysis and biomarker detection

Recent developments in proteomic research and mass spectrometry detection demonstrate the ability of these techniques to provide an alternative choice for the detection of novel disease biomarkers and protein co-expression patterns. While there is a fairly large amount of information regarding the expression of cancer specific protein biomarkers in tissues, blood, cerebrospinal fluid, saliva or urine, relatively few clinical diagnostic tests have been implemented due to the extremely high sensitivities and specificities that are required to justify large scale population screening. The use of a

series of biomarkers, instead of just one, could potentially provide a successful answer to sensitivity and specificity concerns.

A number of analytical platforms such as 1D- and 2D-gel electrophoresis, liquid phase separation techniques (IEF, HPLC, SCX), and protein microchips are being used in tandem with MS detection to identify protein biomarkers. Most of the strategies that are used to study the cancer proteome involve the use of 2D-gel electrophoresis followed by MALDI-MS or micro-LC-ESI-MS [74-83]. Very often the number of protein spots visualized on the gel is relatively large, in the 1,000-1,500 range, however, the number of proteins identified by MS is rather small, only 50-300. Moreover, confident protein identification criteria are not always provided, or the data filtering parameters are set at relatively low values. For example, many users of the ESI ion trap MS instrumentation and of the Sequest/BioWorks algorithm have set cross correlation score values (Xcorr) at 1.5, 2.0 and 3.0 for singly, doubly and triply charged ions, respectively (the Xcorr characterizes the quality of the match between a theoretical and experimental mass spectrum); alternatively, users of MALDI-MS detection, have accepted protein molecular weight values if they were within 150 ppm mass accuracy of the theoretical values. Thus, Somiari has reported the analysis of human infiltrating ductal carcinoma using 2D-DIGE followed by MALDI-MS or ESI-MS/MS; the study resulted in the unambiguous differential identification of ~420 proteins. Differences in protein abundance between cancerous and normal samples ranged between 14-30 % [74].

To overcome some of the limitation of 2D-gel electrophoresis (loss of low abundant, highly hydrophobic, and extreme pI value proteins), alternative liquid phase protocols have been developed. Wang *et al.* used liquid IEF and RPLC followed by ESI-

MS or MALDI-MS detection, and reported the identification of 290 proteins in ES2 human clear cell ovarian carcinoma [84]. Hamler, using a similar approach, has reported the identification of 110 proteins in fractions collected from a limited pH range of an IEF separation [85]. Li has reported the identification of 644 proteins from hepatocellular carcinoma (50,000-100,000 cells) using laser caption microdissection (LCM) and SCX-RPLC-MS/MS. 261 proteins were quantified using the isotope-coded-affinity tag (ICAT) approach [86]. Acceptance criteria for peptide identifications were set at DeltaCn (? Cn)>0.1, and Xcorr vs. charge state at 1.9, 2.2 and 3.7, respectively. The ? Cn value characterizes the difference between the first and second best match proteins. Tomlinson has performed the analysis of KATO III human gastric carcinoma cell line using a combination of methods, i.e., immunoaffinity chromatography, SCX and RPLC, followed by MS/MS detection [87]. The protocol led to the analysis of 1,354 peptide subfractions and resulted in the identification of 1,966 unique proteins by 4,291 peptide sequences. Manual data interpretation was used for the validation of results. Alternatively, Jacobs has used SCX-RPLC-MS/MS to analyze human mammary epithelial cells and reported the identification of 5,838 unique peptides that matched 1,574 proteins [88]. Peptide identifications were accepted if Xcorr vs. charge state values were 1.9, 2.2 and 3.75, and ? Cn>0.1. The analysis of a very small number of cells (10,000) from invasive ductal carcinoma of the breast, using LCM, $^{16}O/^{18}O$ labeling and nano-LC-MS/MS was reported by Li [89]. 76 proteins were identified and about a dozen proteins displayed significant overexpression vs. normal cells. One of the most comprehensive analyses of the MCF7 cancer cell line membrane proteome was performed by Xiang, and the identification of 313 proteins using SCX-LC-MS/MS (Xcorr 1.5, 2, 3.0 and ? Cn>0.1) was reported [90].

As the proteins secreted by a tissue (the secretome) can also reflect the pathological state of an individual, novel technologies have been developed lately that can detect tumor secreted proteins in the blood stream. These proteins could also represent a valuable source of biomarkers. Protein microchips have been introduced for searching for biomarker patterns in serum and tissue samples [91, 92]. While this approach demonstrated relatively good sensitivity and specificity, further work is necessary to resolve issues related to reproducibility and agreement between results reported by various labs [93].

The major part of this thesis is focused on the proteomic characterization of the soluble fraction of the MCF7 cancer cell line. An analytical protocol that consisted of a shotgun 2D SCX-LC separation approach followed by mass spectrometric detection was developed, and resulted in the confident identification of >1,900 proteins. A detailed description of the effect of choosing specific numerical values for various experimental parameters is provided. The list of identified proteins was queried for specific classes of biomarkers that were reported to be associated with breast cancer. To the best knowledge of the author, these results represent the most comprehensive report on the proteomic profile of the MCF7 cell line, and provide an abundant source of reliable data that can be further used in differential protein expression profiling studies. The list of proteins and associated peptides will be made available to public use and peptide MS^2 spectra will be provided upon request.

1.3 References

1. Ruddon, R. W. (1995) Cancer biology 3rd Edition, Oxford University Press: New York

2. Hodgson, L., (2002) Mechanisms of tumor metasasis and cell motility in response to extracellular matrix proteins. The Pennsylvania State University. 1-12

3. McKinnell, R. P., Perantoni, A. Pierce, G. (2000) The biological basis of cancer. Cambridge University Press. 14-310

4. Loeb, L., Loeb, K., Anderson, J. (2003) Multiple mutations and cancer. *Proc. Natnl. Acad. Sci.***100**, 776-781

5. Ames, B., Gold, L., and Willett, W. (1995) The causes and prevention of cancer. *Proc. Natnl. Acad. Sci.* **92**, 5258-5265

6. Fearon, E. (1997) Human cancer syndromes: clues to the origin and nature of cancer. *Science.* **278**, 1043-1050

7. Liotta, L. (1992) Cancer cell invasion and metastasis. *Scientific American.* **266**, 54-63

8. F. Orr, M.B., L. Weiss, ed. (1991) Microcirculation in cancer metastasis: a brief survey of concepts and applications. CRC Press: Boca Raton, FL.

9. Goldberg, I., ed. (1991) Cell motility factors. Basel Co.: Boston. 17-211

10. Becker, W., Kleinsmith, L., and Hardin, J. (2000) The world of the cell, ed. E. Mulligan. Addison Wesley Longman, Inc. 43-786

11. Dobos, N., Rubesin, S. E. (2002) Radiologic imaging modalities in the diagnosis and managemnet of colorectal cancer. *Hematol. Oncol. Clin. North Am.* 16(4), 875-895

12. Zangheri, B., Messa, C., Picchio, M., Gianolli, L., Landoni, C., and Fazio, F. (2004) PET/CT and breast cancer. *Eur. J. Nucl. Med. Mol. Imaging.* **1S1**, 35-42

13. Guo, Y., Sivaramakrishna, R., Lu, C. C., Suri, J. S., and Laxminarayan, S. (2006) Breast image registration techniques: a survey. *Med. Biol. Engg. Computing.* **44**, 15-26

14. Hadjiiski, L., Sahiner, B., Chan, H. P. (2006) Advances in computer-aided diagnosis in breast cancer. *Curr. Opin. Obstet. Gynecol.* **18(1)**, 64-70

15. Smith, A. P., Hall, P. A., and Marcello, D. M. (2004) Emerging technologies in breast cancer detection. *Radiol. Manage.* **26(4)**, 16-24

16. Edell S. L., Eisen, M. D. (1999) Current imaging modalities for the diagnosis of breast cancer. *Del. Med. J.* **71(9)**, 377-382

17. Raj G, Moreno JG, Gomella LG. (1998) Utilization of polymerase chain reaction technology in the detection of solid tumors. *Cancer*. **82**, 1419-1442.

18. Minamoto, T., Ronai, Z. (2001) Gene mutation as a target for early detection in cancer diagnosis. *Critical Rev. Oncology Hematology.* **40**, 195-213

19. Minafra, I. P., Fontana, S., Cancemi, P., Basirico, L., Caricato, S., and Minafra, S. (2002) A contribution to breast cancer cell proteomics: detection of new sequences. *Proteomics.* **2**, 919-927

20. Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and cancer biomarker discovery tool: opportunities and limitations. *Mol. Cell. Proteomics*. **3(4)**, 367-378

21. Verheul, H. A., Coclingh-Bennink, H. J., Kenemans, P., Atsma, W. J., Burger, C. W., Eden, J. A., et al. (2000) Effects of estrogens and hormone replacement therapy on breast cancer risk and on efficacy of breast cancer therapies. *Maturitas.* **36**, 1-17

22. http://www.medicinenet.com/radiation_therapy/article.htm

23. http://www.cancer.gov/

24. Dolmans, DEJGJ., Fukumura, D., and Jain, R. K. (2003) Photodynamic therapy for cancer. *Nature Reviews Cancer.* **3(5)**, 380–387

25. Wilson, B. C. (2002) Photodynamic therapy for cancer: principles. Canadian J. Gastroenterology. **16(6)**, 393–396

26. Vrouenraets, M. B., Visser, G. W. M., Snow, G. B., van Dongen, GAMS. (2003) Basic principles, applications in oncology and improved selectivity of photodynamic therapy. *Anticancer Research.* **23**, 505–522

27. Suter, T. M., Cook-Bruns, N., and Barton, C. (2004) Cardiotoxicity associated with trastuzumab (herceptin) therapy in the treatment of metastatic breast cancer. *The Breast.* **13**, 173-183

28. Kenemans, P., Verstraeten, R. A., Verheijen, R. H. M., (2004) Oncogenic pathways in hereditary and sporadic breast cancer. *Maturitas The Eur. Menopause J.* **49**, 34-43

29. Haber, D. A., Fearon, E. R. (1998) The promise of cancer genetics. *Lancet.* **351**(Suppl II), SII1-8

30. Makin, G., Dive, C. (2003) Recent advances in understanding apoptosis: new therapeutic opportunities in cancer chemotherapy. *Trends Mol. Med.* **9**, 251-255

31. Liu, W, Bulgaru, A., Haigentz, M., Stein, C. A., Perez-Soler, R, and Mani, S. (2003) The BCL2-family of protein ligands as cancer drugs: the next generation of therapeutics. *Curr. Med. Chem. Anticancer Agents.* **3**, 217-223

32. Bourdreau, N., Myers, C. (2003) Breast cancer-induced angiogenesis: multiple mechanisms and the role of the microenvironment. *Breast Cancer Res.* **5**, 140-146
33. Weber, G. F., Ashkar, S. (2000) Stress response genes: the genes that make cancer metastasize. *J. Mol. Medicine.* **78**, 404-408

34. Dua, K., Williams, T. M., and Beretta, L. (2001) Translational control of the proteome: relevance to cancer. *Proteomics*. **1**, 1191-1199

35. Caraglia, M., Budillon, A., Vitale, G., Lupoli, G., Tagliaferri, P., and Abbruzzese, A. (2000) Modulation of molecular mechanisms involved in protein synthesis machinery as a new tool for the control of cell proliferation. *Eur. J. Biochem*. **267**, 3919-3936

36. Jemal, A., Murray, T., Ward, E., *et al.* (2005) Cancer Statistics 2005. *Cancer journal clinical.* Jan-Feb **55(1)**, 10-30

37. Hondermack, H. (2003) Breast cancer: when proteomics challenges biological complexity. *Mol. Cell. Proteomics.* **2**, 281-291

38. AJCC cancer staging manual. 6th ed. New york: Springer-Verlag: 2002

39. Alaiya, A., Mohanna, M. A., and Linder, S. (2005) Clinical cancer proteomics: promises and pitfalls. *J. Proteome Res.* **4**, 1213-1222

40. Srinivas, P. R., Kramer, B. S., and Srivastava, S. (2001) Trends in biomarker research for cancer detection. *The Lancet Oncology.* **2**, 698-704

41. Neuhoff, N. V., Pich, A. (2005) Mass spectrometry-based methods for biomarker detection and analysis. *Drug Discovery Today.* **2(4)**, 361-367

42. Lein, M., Kwiatkowski, M., Semjonow, A., Luboldt, H. J., Hammerer, P., Stephan, C., Klevecka, V., Taymoorian, K., Schnorr, D., *et al.* (2003) A multicenter clinical trial on the use of complexed prostate specific antigen in low prostate specific antigen concentrations. *J. Urol.* **170**, 1175-1179

43. Bast, R. C., Xu, F. J., Yu, Y. H., Barnhill, S., Zhang, Z., and Mills, G. B. (1998) CA125: the past and the future. *Int. J. Biol. Markers.* **13**, 179-187

44. Perrotti, M. (2001) Understanding PSA and prostate cancer risk assessment. *N. J. Med.* **98**, 35-38

45. Bertario, L., Russo, A., Sala, P., Varesco, L., Giarola, M., Mondini, P., Pierotti, M., Spinelli, P., and Radice, P. (2003) Multiple approach to the expectation of genotype-phenotype correlations in familial adenomatous polyposis. *J. Clin. Oncol.* **21**, 1698-1707

46. Nicoletto, M. O., Donach, M., De Nicolo, A., Artioli, G., Banna, G., and Monfardini, S. (2001) BRCA-1 and BRCA-2 mutations as prognostic factors in clinical practice and genetic counseling. *Cancer Treat. Rev.* **27**, 295-304

47. Schnitt, S. J. (2001) Traditional and newer pathologic factors. *J. Natl. Cancer Inst. Monogr.* 22-26

48. Srivastava, S., Srivastava, R. G. (2005) Proteomics in the forefront of cancer biomarker discovery. *J. Proteome Res.* **4**, 1098-1103

49. Kolch, W., et al. (2005) The molecular make-up of a tumor: proteomics in cancer research. *Clin. Sci. (Lond.)* **108**, 369-383

50. Hoffmann, E. de., Stroobant, V. (2001) Mass spectrometry: principles and applications. John Wiley & Sons, Ltd.

51. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci.* USA. **90**, 5011-5015

52. Thomson, J. J., (1913) Rays of positive electricity and their application to chemical analysis, Longmans, Green and Co.: London.

53. Aebersold, R., Mann, M. (2003) Mass spectrometry- based proteomics. *Nature.* **422**, 198-207

54. Aebersold, R., Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* **101**, 269-295

55. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) *Science.* **246**, 64-71

56. Cole, R. B. (1997) Electrospray ionization mass spectrometry. John Wiley & Sons, Inc.

57. Zenobi, R., Knochenmuss, R. (1998) Ion formation in MALDI mass spectrometry. *Mass Spectrom. Rev.* 17(5), 337-366

58. McLuckey, S. A., Wells, J. M. (2001) Mass analysis at the advent of 21st century. *Chem. Rev.* **101**, 571-606

59. Henzel, W. J., Watanabe, C., and Stults, J. T. (2003) Protein identification: the origins of peptide mass fingerprinting. *J.Am. Soc. Mass Spectrom.* **14**, 931-942

60. Mann, M., Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390-4399

61. Zubarev, R. A., Hakansson, P., and Sundqvist, B. (1996) Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. Anal. Chem. **68**, 4060-4063

62. Mann, M., Hendrickson, R. C., and Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437-473

63. Eng, J. K., McCormack, A. L., and Yates III, J. R. (1994) An approach to correlate tandem mass spectral data in peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989

64. Yoshida, M., Loo, J. A., and Lepley, R. A. (2001) Proteomics as a tool in the pharmaceutical drug design process. *Curr. Pharmaceutical Design.* **7**, 293-312

65. Gorg, A., Weiss, W., and Dunn, M. J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics.* **4**, 3665-3685

66. Weinberger, R. (1993) Practical Capillary Electrophoresis. Academic Press Inc.

67. Tomer, K. B. (2001) Separations combined with mass spectrometry. *Chem. Rev.* **101**, 297-328

68. Skoog, D. A., Leary, J. J. Principles of instrumental analysis.

69. Opiteck, C. J., Lewis, K. C., Jorjenson, J. W., and Anderegg, R. J. (1997) *Anal. Chem.* **69**, 1518-1524

70. Opiteck, C. J., Jorjenson, J. W., and Anderegg, R. J. (1997) *Anal. Chem.* **69**, 2283-2291

71. Lewis, K. C., Opiteck, C. J., Jorjenson, J. W., and Sheeley, D. M. J. (1997) *Am. Soc. Mass Spectrom.* **8**, 495-500

72. Link, A. J., Eng, J., Schielt, D. M., Carmack, E., Mize, G. J., *et.al.* (1999) *Nat. Biotechnol.* **17**, 676-682

73. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotech.* **17**, 994-999

74. Somiari, R. I., Sullivan, A., Russell, S., Somiari, S., Hu, H., Jordan, R., George, A., Katenhusen, R., Buchowiecka, A., Arciero, C., Brzeski, H., Hooke, J., and Shriver, C. (2003) High-throughput proteomic analysis of human infiltrating ductal carcinoma of the breast. *Proteomics*. **3**, 1863-1873

75. Oh, J. M. C., Brichory, F., Puravs, E., Kuick, R., Wood, C., Rouillard, J. M., Tra, J., Kardia, S., Beer, D., and Hanash, S. (2001) A database of protein expression in lung cancer. *Proteomics.* **1**, 1303-1319

76. Hathout, Y., Riordan, K., Gehrmann, M., and Fenselau, C. (2002) Differential protein expression in the cytosol fraction of an MCF-7 breast cancer cell line selected for resistance toward melphalan. *J. Proteome Res.* **1**, 435-442

77. Minafra, I. P., Fontana, S., Cancemi, P., Basirico, L., Caricato, S., and Minafra, S. (2002) A contribution to breast cancer cell proteomics: detection of new sequences. *Proteomics.* **2**, 919-927

78. Ying, W., Zhang, K., Qian, X., Xie, L., Wang, J., Xiang, X., Cai, Y., and Wu, D. (2003) Proteome analysis on an early transformed human bronchial epithelial cell line, BEP2D, after a- particle irradiation. *Proteomics.* **3**, 64-72

79. Friedman, D. B., Hill, S., Keller, J.W., Merchant, N. B., Levy, S. E., Coffey, R. J., and Caprioli, R. M. (2004) Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics*. **4**, 793-811

80. Celis, J. E., Gromov, P., Cabezon, T., Moreira, J. M. A., Ambartsumian, N., Sandelin, K., Rank, F. and Gromova, I. (2004) Proteomic characterization of the interstitial fluid perfusing the breast tumor microenvironment: a novel resource for biomarker and therapeutic target discovery. *Mol. Cell. Proteomics*. **3(4)**, 327-344

81. Brown, K. J. and Fenselau, C. (2003) Investigation of doxorubicin resistance in MCF-7 breast cancer cells using shot-gun comparative proteomics with proteolytic ^{18}O labeling. *J. Proteome Res*. **3**, 455-462

82. Tyan, Y. C., Wu, H. Y., Lai, W. W., Su, W. C., and Liao, P. C. (2004) Proteomic profiling of human pleural effusion using two-dimensional nano liquid chromatography tandem mass spectrometry. *J. Proteome Res*. **4**, 1274-1286

83. Zhou, G., Li, H., Gong, Y., Zhao, Y., Cheng, J., Lee, P., and Zhao, Y. (2005) Proteomic analysis of global alteration of protein expression in squamous cell carcinoma of the esophagus. *Proteomics*. **5**, 3814-3821

84. Wang, H., Kachman, M. T., Schwartz, D. R., Cho, K. R., and Lubman, D. M. (2002) A protein molecular weight map of ES2 clear cell ovarian carcinoma cells using a two-dimensional liquid separations/mass mapping technique. *Electrophoresis*. **23**, 3168-3181

85. Hamler, R. L., Zhu, K., Buchanan, N. S., Kreunin, P., Kachman, M. T., Miller, F. R., and Lubman, D. M. (2004) A two-dimensional liquid-phase separation method coupled with mass spectrometry for proteomic studies of breast cancer and biomarker identification. *Proteomics*. **4**, 562-577

86. Li, C., Hong, Y., Tan, Y.X., Zhou, H., Ai, J. H., Li, S. J., Zhang, L., Xia, Q. C., Wu, J. R., Wang, H. Y., and Zeng, R.(2004) Accurate qualitative and quantitative proteomic analysis of clinical hepatocellular carcinoma using laser capture microdissection coupled with isotope-coded affinity tag and two-dimensional liquid chromatography mass spectrometry. *Mol. Cell. Proteomics*. **3**, 399-409

87. Tomlinson, A. J., Hincapie, M., Morris, G. E., and Chicz, R. M. (2002) Global proteome analysis of a human gastric carcinoma. *Electrophoresis*. **23**, 3233-3240

88. Jacobs, J. M., Mottaz, H. M., Yu, L. R., Anderson, D. J., Moore, R. J., Chen, W. N. U., Auberry, K. J., Strittmatter, E. F., Monroe, M. E., Thrall, B. D., Camp, D. G., and Smith, R. D. (2003) Multidimensional proteome analysis of human mammary epithelial cells. *J. Proteome Res.* **3**, 68-75

89. Zang, L., Toy, D. P., Hancock, W. S., Sgroi, D. C., and Karger, B. L. (2003) Proteomic analysis of ductal carcinoma of the breast using laser capture microdissection, LC-MS and $^{16}O/^{18}O$ isotopic labeling. *J. Proteome Res.* **3**, 604-612

90. Xiang, R., Shi, Y., Dillon, D. A., Negin, B., Horvath, C., and Wilkins, J. A. (2004) 2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J. Proteome Res.* **3**, 1278-1283

91. Petricoin III, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572-577

92. Liu, A. Y., Zhang, H., Sorensen, C. M., and Diamond, D. L. (2005) Analysis of prostate cancer by proteomics using tissue specimens. *J. Urology*. **173**, 73-78

93. Veenstra, T. D., Prieto, D. A., and Conrads, T. P. (2004) Proteomic patterns for early cancer detection. *Drug Discovery Today*. **9(20)**, 889-897

94. Moss, E. L., Hollingworth, J., and Reynolds, T. M. (2005) The role of CA125 in clinical practice. *J. Clin. Pathol.* **58**, 308-312

95. Vastag, B. (2000) Some promising biomarkers for cancer. *J. Natl. Cancer Inst.* 92(10), 788

96. Ross, J. S., Linette, G. P., Stec, J., Clark, E., Ayers, M., Leschly, N., Symmans, W. F., Hortobagyi, G. N., and Pusztai, L.(2004) Breast cancer biomarkers and molecular medicine: part II. *Expert Rev. Mol. Diagn*. **4(2)**, 169-188

97. Hondermarck, H., Sophie, A., Edouart, V., Revillion, F., Lemoine, J., Belkoura, I. E. Y., Nurcombe, V., and Peyrat, J. P. (2001) Proteomics of breast cancer for marker discovery and signal pathway profiling. *Proteomics.* **1**, 1216-1232

98. Franzen B, Linder S, Alaiya AA, Eriksson E, *et al.*(1996) *Br. J. Cancer*. **18**, 2832

99. Baselga J. (2004) The science of EGFR inhibition: a roadmap to improved outcomes? *Signal.* **5(3)**, 4-8

100. Ciocca, D. R., Calderwood, S. K. (2005) Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. *Cell Stress Chaperones.* **10(2)**, 86-103

101. Esteva, F. J., and Hortobagyi, G. N. (2004) Prognostic molecular markers in early breast cancer. *Breast Cancer Res*. **6**, 109-118

102. Janssens, J. Ph., Verlinden, I., Gungor, N., Raus, J., and Michiels, L. (2004) Protein biomarkers for breast cancer prevention. *Eur. J. Cancer Prevention.* **13**, 307-317

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

104. Yamashita, H., Nishio, M., Toyama, T., Sugiura, H., Zhang, Z., Kobayashi, S., and Iwase, H. (2004) Coexistence of HER2 over-expression and p53 protein accumulation is a strong prognostic molecular marker in breast cancer. *Breast Cancer Res.* **6(1)**, 24-30

105. Fu, H., Subramanian, R. R., and Masters, S. C. (2000) 14-3-3 proteins: Structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.* **40**, 617-647

106. Vercoutter-Edouart, A. S., Lemoine, J., Le Bourhis, X., Louis, H., *et al*. (2001) Proteomic analysis reveals that 14-3-3 is down-regulated in human breast cancer cells. *Cancer Res.* **61**, 76-80

107. Berx, G. and Roy, F. V. (2001) The E-cadherin/ catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression. *Breast Cancer Res.* **3**, 289-293

108. Leers, M. P. G., Aarts, M. M. J., Theunissen, P. H. M. H. (1998) E-cadherin and calretinin: a useful combination of immunochemical markers for differentiation between mesothelioma and metastatic carcinoma. *Histopathology.* **32**, 209-216

109. Marzo, A. M. D., Knudsen, B., Chan-Tack, K., Epstein, J. I. (1999) E-cadherin as a marker of tumor aggressiveness in routinely processed radical prostatectomy specimens. *Adult Urol.* **53**, 707-713

110. Diamandis, E. P. and Merwe, D. E. (2005) Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin. Cancer Res*. **11**, 963-965

111. Ilg, E. C., Schafer, B. W., Heizmann, C. W. (1996) Expression pattern of S100 calcium-binding proteins in human tumors. *Int. J. Cancer*. **68(3)**, 325-332

112. Hermani, A., Hess, J., Servi, B. D., Medunjanin, S., Grobholz, R., Trojan, L., Angel, P., and Mayer, D. (2005) Calcium-binding proteins S100A8 and S100A9 as novel diagnostic markers in human prostate cancer. *Clin. Cancer Res*.; **11(14)**: 5146

113. Schluter, C., Duchrow, M., Wohlenberg, C., Becker, M. H. G., Key, G., Flad, H. D., and Gerdes, J. (1993) The cell proliferation-associated antigen of antibody Ki-67: a very large, ubiquitous nuclear protein with numerous repeated elements, representing a new kind of cell cycle-maintaining proteins. *J. Cell Biology*. **123**, 513-522

114. Scholzen, T., Gerdes, J. (2000) The Ki-67 protein: From the known and the unknown. *J. Cell. Physiol.* **182**, 311-322

115. Sigal, A., and Rotter, V. (2000) Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome. *Cancer Res.* **60**, 6788-6793

116. Gasco, M., Shami, S., and Crook, T. (2002) The p53 pathway in breast cancer. *Breast Cancer Res.* **4**, 70-76

117. Pharaoh, P. D., Day, N. E., and Caldas, C. (1999) Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br. J. Cancer.* **80**, 1968-1973

118. Giometti, C. S., Williams, K., Tollaksen, S. L. (1997) A two-dimensional electrophoresis database of human breast epithelial cell proteins *Electrophoresis.* **18**, 573-581

119. Kageyama, S., Isono, T., Iwaki, H., Wakabayashi, Y., Okada, Y., Kontani, K., Yoshimura, K., Terai, A., Arai, Y., Yoshiki, T. (2004) Identification by Proteomic Analysis of Calreticulin as a Marker for Bladder Cancer and Evaluation of the Diagnostic Accuracy of Its Detection in Urine. *Clin. Chem.* **50(5)**, 857

120. Khanuja, P.S., Lehr, J. E., Soule, H. D., Gehani, S. K., *et al.,* (1993) Nuclear matrix proteins in normal and breast cancer cells. *Cancer Res.* 53, 3394-3398

121. Samuel, S. K., Minish, T. M., and Davie, J. R. (1997) *J. Cell Biochem.* 66, 9-15

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

123. Williams, K., Chubb, C., Huberman, E., Giometti, C. S.(1998) Analysis of differential protein expression in normal and neoplastic human breast epithelial cell lines. *Electrophoresis.* **19**, 333-343

124. Gronborg, M., Kristiansen, T. Z., Iwahori, A., Chang, R., Reddy, R., Sato, N., Jensen, O. N., Hruban, R. H., Goggins, M. G., Maitra, A., Pandey, A. (2006) Biomarker discovery from pancreatic cancer secretome using a differential proteomics approach. *Mol. Cell. Prot*. **5**, 151

125. Yong, L., Li, C., Shu-you, P., Zhou-xun, C., Vu, C. H.(2005) Role of CD97stalk and CD55 as molecular markers for prognosis and therapy of gastric carcinoma patients. *J. Zhejiang Univ. SCI*. **6B(9)**, 913-918

126. Fu, X. C., Hu, C.-A. A., Chen, J., Wang, J., and Ray Liu, K. J. (2005) Cancer genomics, proteomics and clinical applications. In "Genomics Signal processing and Statistics."

Chapter 2: Experimental Methods of Analysis

2.1 MCF7 cell culture

The MCF7 breast cancer cell line was purchased from ATCC (Manassas, VA). The cells were cultured in Eagle's Minimum Essential Medium (EMEM) supplemented with 10% fetal bovine serum (FBS) and 0.1% bovine insulin. The cells were grown in T75 flasks in an incubator maintained at 37 ºC and 5 % CO_2. After reaching 70% confluence (**Figure 1**), the cell culture medium was removed, the cells were washed twice with phosphate buffer saline (PBS) (pH 7.4), trypsin/EDTA solution (1 mL, 0.25 % trypsin/0.53 mM EDTA) was added for ~5min to the flask for detaching the cells, 4 mL of media were added to stop the tryptic digestion, and the cells were harvested by gentle aspiration with a pipette. Cells were stored at -80 ºC prior to further processing.

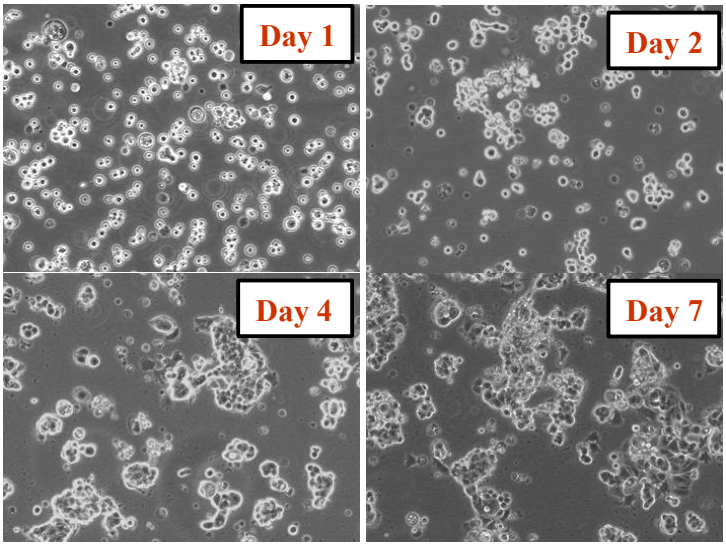


Figure 1. Morphology of MCF7 breast cancer cells in culture.

2.2 Cell lysis and protein extraction

The cell lysis solution was prepared by mixing 1 mL RIPA buffer (500 mM TrisHCl pH 7.4, 1.5 M NaCl, 10 % NP-40, 2.5 % deoxycholic acid, 10 mM EDTA), 100 µL protease inhibitor cocktail (104 mM AEBSF, 0.08 mM aprotinin, 2 mM leupeptin, 4 mM bestatin, 1.5 mM pepstatin A, 1.4 mM E-64), 100 µL NaF (~100mM) and 50 µL Na_3VO_4 (~200 mM) as phosphatase inhibitors, and 8.75 mL of ice cold water. Cells stored at -80 ºC were thawed at room temperature and divided into several Eppendorf tubes. Each of the vials containing cells was added 1 ml of the above prepared lysis buffer, was rocked for 2 h at 4 ºC, and then centrifuged for ~15 min at 13,000 rpm and 4 ºC. The supernatant was collected and the cell pellet was preserved. The protein content in the supernatant (the cytosolic soluble protein cell extract) was measured using the Bradford assay. The concentration of the protein in the soluble extract was ~3 mg/ml. Absorbance measurements were made at 595 nm using a SmartSpec Plus Spectrophotometer (Bio-Rad, Hercules, CA) as per manufacturer's instructions.

2.3 Sample digestion and cleanup

1 ml of the soluble protein cell extract (~3 mg/mL) was treated with urea (8 M) and DTT (4.5 mM) for reducing the disulfide bonds. Additional TrisHCl was not added, as it was present in the RIPA buffer at 50 mM concentration. The mix was heated/denatured for 1 hour at 60 ºC, cooled at room temperature, and diluted 10X with 50 mM NH_4HCO_3. Trypsin, 60 µg, was added to a protein:enzyme ratio of 50:1 w/w, and the sample was digested overnight at 37 ºC. The digestion process was quenched with 10 µL TFA. 1 ml of the MCF7 digest (~3 mg/mL) was further processed with SPEC-PTC18

solid phase extraction pipette tips (Varian Inc., Lake Forest, CA) for desalting. The SPEC cartridge was rinsed with 50 µL of wetting solution CH_3OH/H_2O (50:50) and then with 50 µL of equilibration solution (1 µL TFA / 1 mL H2O). The entire digest solution was passed through the cartridge multiple times, by slowly aspirating and dispensing, to allow for more complete adsorption. Following adsorption, the pipette tip was rinsed with 50 µL wash solution CH_3OH/H_2O/TFA (5:95:0.1), and then the peptides were eluted with 50 µL elution solution I CH_3CN/H_2O/TFA (60:40:0.1), and elution solution II CH_3CN/H_2O/TFA (80:20:0.1). The sample was then concentrated to ~75 µL final volume (~4 mg/mL final concentration) with a vacuum centrifuge, and stored at -20 ºC.

2.4 Experimental setup

A micro liquid chromatography system (Agilent Technologies, Palo Alto, CA) and an LTQ ion trap mass spectrometer (Thermo Electron Corp., San Jose, CA) were used to perform the SCX/RP separation and detection of the protein components in the cellular extract. The interfacing of the LC system to the LTQ-MS was achieved by using an on-column/no-split injection setup. The overview of the experimental arrangement is shown in **Figure 2**. The HPLC pump outlet was connected to the reversed phase separation column through a fused silica capillary (50 µm i.d. x ~50 cm) and two PEEK T-connectors. The first T-connector allowed for eluent splitting, and the second connector for the application of the ESI voltage. The LTQ original nanosprayer was removed from the front end of the instrument, and the source was fitted with an XYZ stage and a home-built fixture that enabled easy alignment of the separation column and its nanosprayer within the LTQ-MS ion source. During sample loading, port 5 on the

LTQ valve was connected to port 6, which was blocked, thus enabling a splitless injection. The entire sample was loaded directly on the reversed phase separation column. During sample analysis, port 5 was connected to port 4, and enabled the splitting of the eluent flow generated by the HPLC pump.

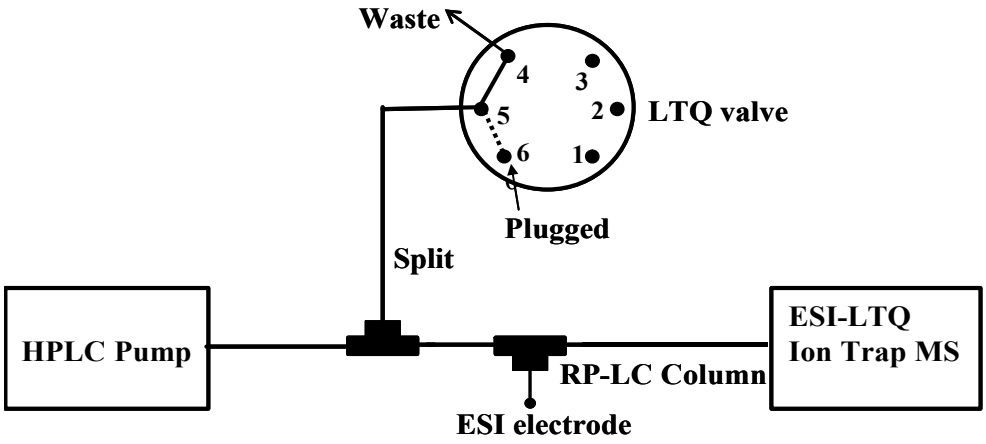


Figure 2. Schematic representation of the experimental arrangement for LC-MS interfacing. Sample load: split closed, port 5 connected to port 6 (plugged) on LTQ valve; Sample analysis: split open, port 5 connected to port 4 (waste) on LTQ valve.

2.5 SCX prefractionation

Sample fractionation was accomplished using a Zorbax Bio SCX Series II column (0.8 mm i.d. x 5 cm) from Agilent Technologies, an SCX column based on silica particles with hydrophilic polymer functionalized with sulphonic acid groups. Solvent A consisted of 0.1% HCOOH in H_2O/CH_3CN (95:5 v/v), and solvent B of 0.1 % HCOOH in H_2O/CH_3CN (95:5 v/v) + 500 mM NaCl. The eluent flow rate was 20 µL/min and the

sample injection volume was 16 µL. At a concentration level of ~4 mg/mL, this is the equivalent of ~64 µg sample injected on the SCX column. The SCX eluent gradient consisted of: 100 % A (0-5 min), 0 to 20 % B (5-35 min), 20 to 100 % B (35-40 min), 100 %B (40-50 min), and 100 % A (50-60 min). A total of 16 fractions were collected. Fraction 1 was collected during the first 5 min wash step, fractions 2-15 (60 µL each) were collected at every 3 min during the salt gradient, and fraction 16 was collected during the last 10 min and consisted mainly of the eluted components at 100 % B.

2.6 RP-HPLC

The 16 SCX subfractions were further analyzed by injecting 40 µL of each fraction on a RPLC-MS system (a total of ~42 µg of peptide mix), while the exact amount injected with each fraction is not known, it is estimated that the average sample amount injected per run was ~1-3 µg. Reversed phase columns (100 µm i.d. x 12 cm) were packed in our laboratory with 5 µm Zorbax SB-C18 packing material (Agilent Technologies) using N_2 pressure at 1800 psi, and were fitted with a 1 cm long (20 µm i.d. x 90 µm o.d.) nanospray emitter. Solvent A consisted of H_2O/CH_3CN (95:5 v/v) + 0.01 %TFA, and solvent B of H_2O/CH_3CN (20:80 v/v) + 0.01 %TFA. Samples were loaded on the column (split closed) at 2 µL/min (100 % A), and eluted at ~170 nL/min (HPLC pump at 10 µL/min, split open). The RP-LC gradient consisted of: 0 to 10 % B (0-1 min), 10-45 % B (1-95 min), 45 to 60 % B (95-110 min), 60 to 100 % B (110-115 min), 100 %B (115-120 min), 100 to 0 % B (120-121 min), and 100 % A (121-150 min).

2.7 ESI-MS/MS

Data dependent MS acquisition conditions were as follows: 1 MS scan (5 microscans averaged) was followed by 1 zoom scan and 1 MS^2 on the top 5 most intense peaks; zoom scan width was ±5 m/z; dynamic exclusion was enabled at repeat count 1, repeat duration 30 s, exclusion list size 200, exclusion duration 60 s, and exclusion mass width ±1.5 m/z; collision induced dissociation (CID) parameters were set at isolation width 3 m/z, normalized collision energy 35, activation Q 0.25, and activation time 30 ms. Protein searching was performed with the BioWorks 3.2 software (Thermo Electron Corp, San Jose, CA) against two human databases. The first database was extracted from the NCBI nr.gz database downloaded on 08/26/05 (included fields were "human" and "sapiens," excluded field was "virus") and contained 131,585 entries. The second database was downloaded from the UniProt website on 03/28/05 (homo sapiens description parameter) and contained 63,973 entries. The database search parameters included: only fully tryptic fragments were considered for peptide matching, missed cleavage sites allowed was 2, peptide tolerance was 2 amu, fragment ion tolerance was 1 amu, and number results scored was 250. Chemical and posttranslational modifications were not allowed, and the capability to match one peptide sequence to multiple references within the database was disabled. Data filtering included 2 sets of filters: filter (1) Xcorr vs. charge state (Xcorr=1.9 for z=1, Xcorr=2.2 for z=2, and Xcorr=3.8 for z= 3), and filter (2) multiple thresholds (Xcorr= 1.9, ?Cn= 0.1, Sp= 500, RSp= 5, Percent ions=30 %; all 5 conditions had to be satisfied).

2.8 Materials and reagents

HPLC grade methanol and acetonitrile were purchased from Fisher Scientific (Fair Lawn, NJ). Cell culturing reagents (EMEM, FBS, insulin, trypsin/EDTA for cell detachment) were purchased from ATCC (Manassas, VA). RIPA lysis buffer was obtained from Upstate (Lake Placid, NY). Protease inhibitors (NaF, Na_3VO_4), NaCl, TFA, HCOOH, TrisHCL, urea, DTT, and all protein standards were purchased from Sigma (St. Louis, MO). Sequencing grade modified trypsin was from Promega Corp. (Madison, WI). NH_4HCO_3 was purchased from Aldrich (Milwaukee, WI). Deionized water (18 MΩ-cm) was generated using a MilliQ ultrapure water system (Millipore, Bedford, MA).

Chapter 3: Results and Discussions

The analysis of complex protein samples represents most often a challenge from both, the qualitative and quantitative point of view. The main objective of this research was to establish a sensitive and reliable LC/MS strategy that will enable the analysis of complex protein samples derived from cancerous cells. A series of optimization strategies were performed with standard protein mixture digests and a set of MCF7 cellular extracts, to enable the detection and confident identification of a large number of proteins. These results assisted in generating data with sensitivities in the high attomole/low femtomole range from pico/nanomolar level solutions. A schematic diagram highlighting the major steps of the analysis is shown in **Figure 1**. MCF7 cancer cells were cultured to 70 % confluence, harvested, lysed, and the soluble protein extract was digested with trypsin and fractionated using a SCX separation column. The sample sub-fractions were further analyzed using RPLC interfaced to ion trap LTQ-MS detection.

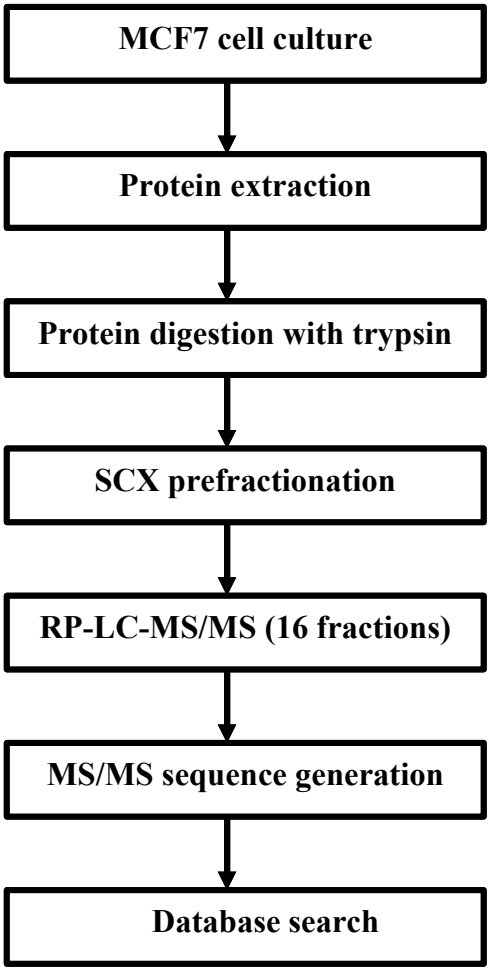


Figure 1. Flowchart including major analysis steps of the MCF7 protein extract.

3.1 Optimization studies

Proteomics is a rapidly developing field in which there are numerous new and improved methodologies used to evaluate the data, making it imperative to use the most appropriate and standardized procedure to assist inter-lab comparisons and improve the reliability of identified proteins. This is crucial particularly for higher organisms and large databases. Keeping this in mind, a broad range of optimizations were performed with a standard mixture of 9 proteins and a set of MCF7 protein digests to enhance parameters related to LC-MS experimental setup, sample preparation, data acquisition, and database searching.

3.1.1 Standard protein mixture

A standard mixture of 9 bovine proteins that included hemoglobin, albumin, carbonic anhydrase, a-lactalbumin, fetuin, a-casein, ß-casein, cytochrome C and insulin, was used for optimization studies and for standardizing the procedures for further MCF7 analysis. The LC-MS interfacing arrangement was evaluated with and without a preconcentrator. The use of an online preconcentrator enabled fast sample loading with high flow rates, however, the sample was retained inside the preconcentrator and was not eluted, worsening thus the detection limits. To avoid this, another setup was made with a backflush preconcentrator arrangement. The LTQ valve positioning with the backflush preconcentrator is shown in **Figure 2A** and **2B,** for sample loading and running conditions. The advantage of this setup is that during loading conditions the sample is retained at the head of the preconcentrator, while under running conditions, the sample is eluted in the opposite direction, such that the sample is flushed out quickly and easily.

Even though this method gave better results than the online preconcentrator, we further compared it with a direct on-column loading setup. Fast loading of the sample was not possible, but better results with ~(3-5) X lower detection limits was achieved. Hence, we adopted the direct sample loading approach for the rest of the experiments. Various LC-MS/MS runs were performed for optimizing other parameters related to sample concentration, data acquisition and database searching. Optimal values for some of these parameters are usually known, but quantitative consequences for deviation from the optimal values are not provided. Experiments were repeated with sample concentrations ranging from 0.5 µM to 0.0005 µM to assess the detection limits and to maximize sequence coverage. A 2-D view of a separation of this protein mixture digest is shown in **Figure 3A,** indicating the high separation efficiencies that were achieved. An inset of the high m/z region of this separation shows the presence of many other components, indicating the capability to detect low intensity ions (**Figure 3B)**.

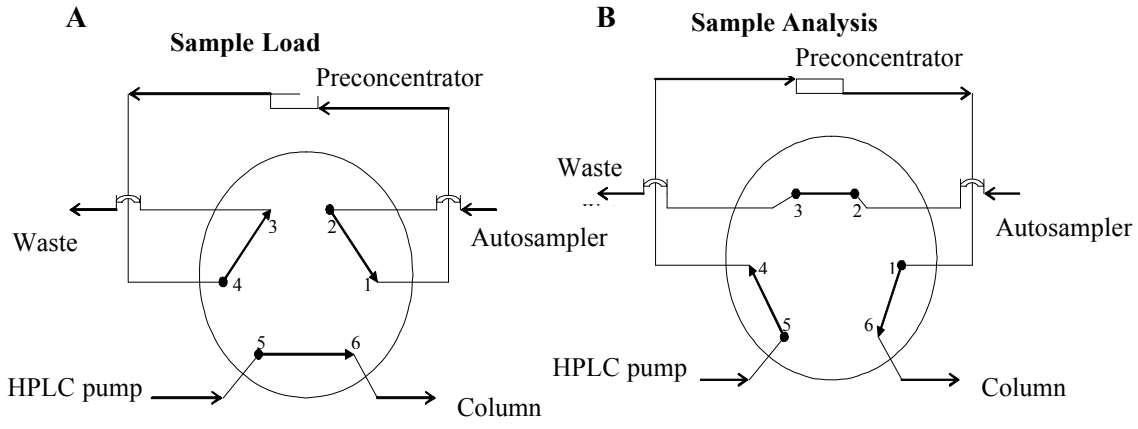


Figure 2. LTQ valve position with backflush preconcentrator for (**A**) sample loading; (**B**) sample running conditions.

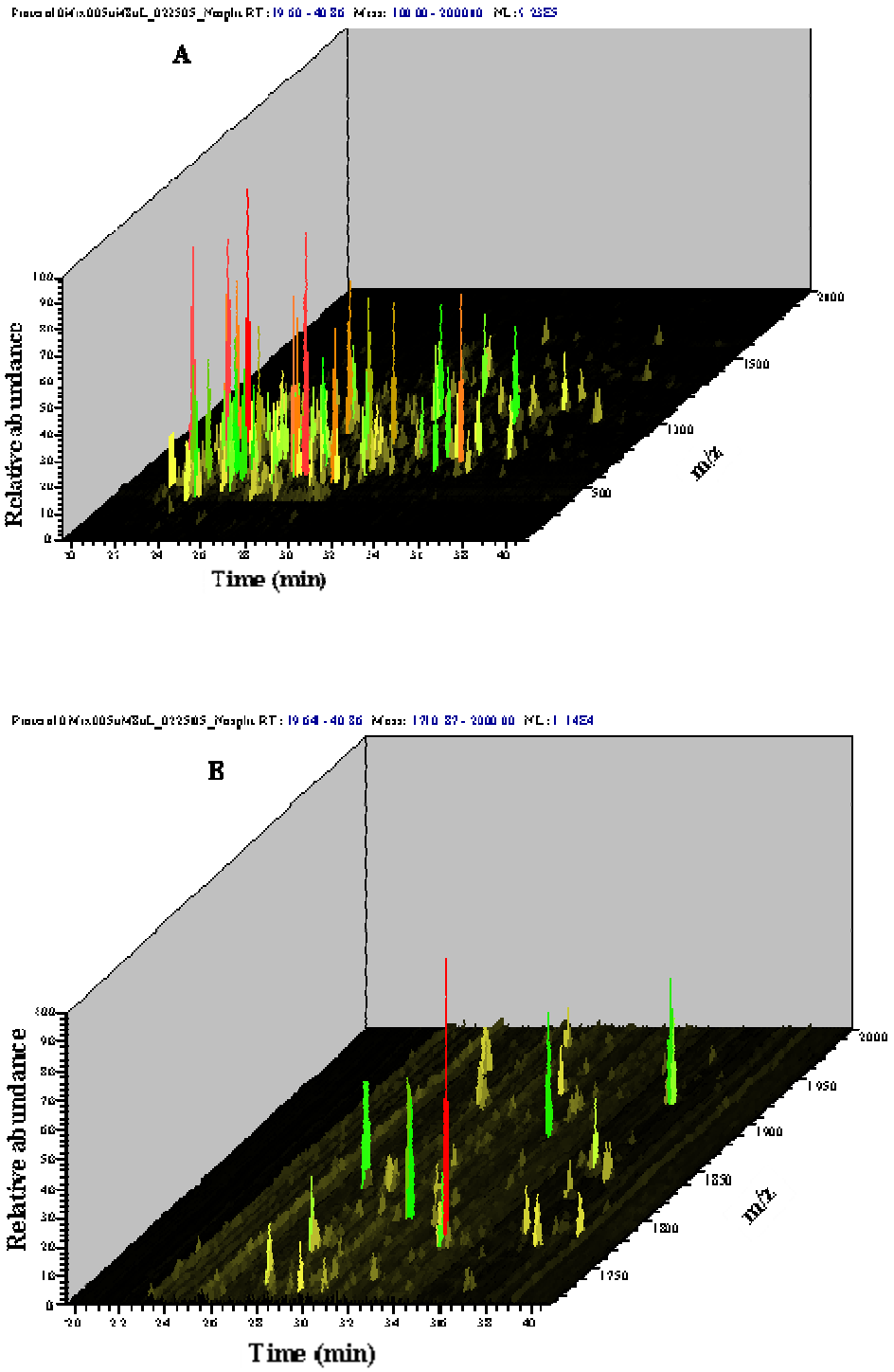


Figure 3. 2D-view chromatogram of a standard protein mix separation: (**A**) m/z 0-2,000; (**B**) inset m/z 1,700-2,000.

3.1.2 MCF7 data sets

3 sets of MCF7 digests were used to perform various optimization studies. The amount of sample injected on the column was critical for identifying a large number of proteins. For example, by lowering the sample injection volume of one of the SCX fractions from 16 μL to 4 μL, the number of identified proteins in that fraction decreased from 354 to 68. An injection volume of 40 μL of sample, the equivalent of about 1-3 μg of peptide mix/fraction, appeared to reach the loading capacity of the reversed phase columns used in this study.

To achieve a good separation of the sample components, a fine tuning of the gradient profile was needed. By increasing the time-window from 55 min to 94 min, for the eluent gradient to proceed from 10 to 45 % B, the number of identified proteins increased from 168 to 220. Obviously, the SCX prefractionation process had a major effect on the number of matched proteins. For one set of MCF7 data, performing LC-MS on the whole cellular digest, without SCX prefractionation, resulted in the identification of only 95 proteins; when the cellular digest was prefractionated by SCX, the number of identified proteins increased to 2,074 (Xcorr were 1.9, 2.2 and 3.8 for z=1, 2, and 3, respectively). A precise setting of the eluent flow rate through the reversed phase column was hard to accomplish, due to the variations in split flow rates that accompany rather small variations in separation column hydraulic resistance. Generally, eluent flow rates <200 nL/min generated 10-20 % more protein matches than flow rates >200 nL/min. The reproducibility of elution times within one set of data was 1-2 % for intra-column and 4-5 % inter-column comparisons, respectively. The overall reproducibility of detecting

overlapping proteins across duplicate runs was ~60%, while the reproducibility of detecting proteins matched by = 2 unique peptides was > 88-90%.

LC peaks at elution were typically 10-40 s wide, however, the use of a maximum sample load and a prolonged eluent gradient resulted in a few peptides with peak widths >1 min; consequently, the number of duplicate peptides reported for the matching proteins increased correspondingly. This is an undesirable outcome, as the mass spectrometer is spending time on performing MS^2 on the same peptides instead of analyzing new ones, and thus structural information is lost. This result was evident mainly for the top proteins on the multiconsensus list that were generally matched by a large number of peptides. For proteins identified by only 2-3 peptides, duplicate entries were, however, a rare event. We refer in this case to duplicate entries as peptides with the same MH^+ and charge state that matched a protein for multiple times. Peptide entries with the same MH^+ but with different charge states were not considered duplicate entries, as the MS^2 was performed at different m/z values for these peptides.

Data acquisition parameters were fine tuned to generate optimum conditions for peptide MS selection and identification. Loss of information due to the MS instrument failure to select certain peptides for CID was observed to happen for several reasons that are discussed in the followings. (1) The greatest contributor to information loss during such an analysis is the complexity of the sample. Even though the experimental setup is always optimized to spread components apart as they elute from the separation system, an ideal situation when only one component elutes at any given time cannot be achieved. The LTQ ion trap instrument enables a fast data acquisition process and the capability to select many ions for MS^2 after each MS event. Some studies report the use of CID on the

top 10 or top 20 most intense ions in each MS. As our experiments involved a triple play data acquisition process, where zoom scans were performed on each ion to determine its charge state, the selection of only 5 top intensity ions for CID resulted in a larger number of protein matches than the choice of 10 top intensity ions. This was a result of the fact that the top 5 selection resulted in 6-8 triple play cycles/min (i.e., 30-40 MS^2/min, a quick updating of the MS panorama and a more comprehensive MS^2 investigation), vs. 3-4 triple play cycles/min with the top 10 selection. To note, however, that these selections are very much dependent on the overall quality of the separation and the peak widths. (2) The quality of the MS^2 spectrum is critical for obtaining a good protein match. The accumulation time before the generation of an MS^2 scan is an essential parameter. At extreme values, by increasing the accumulation time from 10 to 500 ms, the number of the matched proteins increased in one of the SCX fractions from 51 to 138. MS^2 scans were not averaged, as this process would have resulted in lowering the number of triple play cycles/min to a value of 3-4. On the other hand, 5 MS microscans were averaged to generate one single MS scan. This produced a good quality mass spectrum that enabled reliable selections for the MS^2 process while increasing the triple play cycle time by only 0.3 s. (3) If the 400-500 m/z region was selected for data acquisition and CID, about 25-30 % of the MS^2 spectra were generated on uninformative peaks from this m/z region. One set of multiconsensus results for the MCF7 extract revealed, however, that from a total of 4,447 peptide hits, only 131 had a $400<m/z<500$. Consequently, this m/z region was not selected for CID in the final analysis. (4) The time that an ion is sent to the exclusion list (exclusion duration) is tightly related to the peak widths of the components that elute from the separation column. If ions were sent to the exclusion list for only 30 s,

many duplicate peptides were matched for the same protein; the MS instrument time was spent on analyzing the same peptides and not new ones. On the other hand, if ions were sent to the exclusion list for more than 60 s, peptides that had an m/z within ±1.5 of the selected ion (exclusion mass width) and happened to elute in that time window, were lost, as they were not selected for CID. The optimum value for the exclusion duration that resulted in a minimal number of missed peptides was ~60 s. (5) Ions that are selected for CID during the early stages of their elution from the separation column often have low intensity values and do not generate good quality CID spectra. In order to alleviate this problem, experiments with peptide repeat count values of 1 and 2 (within a repeat duration time of 30 s) were conducted, and the results were compared. Each method displayed some peptides that were unique only to that experiment and were not found in the other one. A repeat count value of 2 resulted in more duplicate peptide matches for each protein. Xcorr scores were similar for the two methods. As using a repeat count value of 2 did not seem to improve the overall results, a value of 1 was selected for the final analysis. (6) Peptides with m>1,000 Da often generate spectra where the 2^{nd} or 3^{rd} isotopes are more intense than the 1^{st} one, and these are the isotopes that are selected for CID. As a consequence, the mass range selected for CID (the isolation width) and for sending the ions to the exclusion list (the exclusion mass width) must be large enough to include all the intense isotopes of a specific ion, and to avoid CID on subsequent intense isotopic peaks of the same ion. A range of ±1.5 m/z around the ion of choice was selected in our experiments. (7) It was observed that ions with m/z>1,500 were seldom selected for MS^2, probably due to their low intensity values in spectra that also contained intense, low m/z ions. This problem could be partially resolved by improving the resolution of the

separation system to generate a broader distribution of the sample components; however, for many peptides this seemed to not be a problem, as many of the large m/z ions were the singly charged counterparts of the double/triply charged species that were actually selected for CID.

The parameters selected for database searching had a strong impact on the outcome of the search results. We have chosen parameters that initially enabled the identification of a large set of proteins that were later sorted out with adequate filters. As such, the minimum total ion intensity threshold for database searching was 1,000, and the peptide mass tolerance was +/-2.0. The LTQ mass accuracy would have allowed mass tolerance limits as low as +/-0.5, however, due to the fact that 2^{nd} or 3^{rd} isotopes were often selected for MS^2, the window was maintained rather large to avoid any losses in the peptide-protein matching process.

3.2 Evaluation of mass spectrometric data

The 16 SCX fractions were analyzed using RPLC-MS/MS. Two sets of data, generated by injecting 8 and 40 μL of sample on the RPLC column, were evaluated. The raw files were batch searched with the BioWorks 3.2 software against two human protein databases: one downloaded from NCBI and one from SWISSPROT. The LC-MS/MS experiments summed up to 40 h of mass spectrometric exploration for each data set, and generated a total of 153,472 MS scans and 51,184 MS^2's for the 8 μL injection, and 173,611 MS scans and 54,843 MS^2's for the 40 μL injection. In order to minimize false positive identifications, several peptide acceptance parameters were evaluated. Data were sorted using two sets of filters, Xcorr vs. charge state and multiple thresholds (see experimental section). **Table 1** summarizes the overall findings. The effect of injecting

sufficient sample for analysis is obvious. The 8 μL injection experiment generated 7,196 peptide hits that matched 6,363 entries, of which 2,329 were top match proteins. Alternatively, the 40 μL injection experiment generated 14,981 peptide hits that matched 12,362 entries, of which 4,534 were top match proteins, i.e., approximately twice as many hits as the 8 μL experiment. These data were initially selected by applying only filter 1 (i.e., Xcorr vs. charge state, with values set at 1.9, 2.2, 3.8) that is often used in the reported literature to define high quality data [1]. A close evaluation of the raw MCF7 data indicated, however, that the use of this filtering parameter is appropriate for eliminating poor quality data, but is not sufficient for defining acceptable protein matches with minimum false positives. Moreover, broader efforts to sort large experimental data sets based on these criteria, have demonstrated that if a protein is identified by only one peptide, only 25 % of the peptide hits will result in a reliable protein match [2]. Likewise, our experience in evaluating MS^2 spectra of doubly charged peptides has indicated that only spectra with Xcorr~2.6-3 were of sufficiently good quality to pass a quick visual inspection. Peptides with lower scores often required further examination for validation. Manual evaluation of MS^2 spectra for a few proteins of interest is an achievable or even advisable objective; however, it is not a practical approach for the validation of thousand of spectra.

By increasing the stringency of data acceptance criteria, i.e., by accepting only peptides that passed filter 1 and also had low p values, the number of identified proteins decreased approximately twice. For example, for the 40 μL injection, of the 4,534 top match protein hits, only 2,367 were identified by peptides with p= 0.001 (Note: the p-value represents the probability of a random match, which is 0.1 % for p=0.001; with the

present BioWorks configuration the p-value assignment to proteins is biased, as it is performed by simply applying the p-value of the best scoring peptide to its matching protein). Furthermore, by applying the second data filter in addition to the first one, the number of identified proteins decreased from 2,367 to 1,895 for the same data set. The main reason for this outcome was that the preliminary score (Sp) values from the multiple threshold filter were not meeting the preset criteria of Sp= 500. Most peptides had delta correlation scores ? Cn>0.1. These scores represent the difference between the Xcorr of the top and second best choice peptide. Sp values are computed by taking into consideration the number, abundance and continuity of the fragment ions in a MS^2 spectrum, as well as the presence of certain immonium ions [3]. A number of spectra with Sp<500 were visually inspected, and indeed, either ion intensities were rather low, or some fragments were missing; however, many of the spectra with 400<Sp<500 were of sufficiently good quality to pass manual evaluation. These spectra would have generated a category of false negatives that would have been lost if the combination of filters 1+2 would have been used. It is worth noting that when both filters were applied, as many as 1,089 (~90 %) from 1,207, and 1,895 (~90 %) from 2,107 top match protein hits, were matched by peptides with p<0.001. Alternatively, applying either filter 1, or both filters 1 and 2, in combination with p<0.001, the top match proteins represented 94-97 % of the total protein hits. Similar trends were observed by searching the data against a SWISSPROT database. The total number of identified proteins was somewhat lower, about 80 % of what was reported with the NCBI database; however, the SWISPROT database had only 63,973 FASTA entries (less than half the size of the NCBI database).

Consequently, the confidence of protein identifications can be substantially increased by using a combination of predetermined filter and p-value settings that eliminate false positives comprised of random and second-best matches. Minimizing false positives will inherently maximize, however, false negatives. There will always be a set of data that will meet most, but not all the acceptance criteria, or, that will meet all the acceptance criteria, but only as second best-match results. These data will have to be then manually inspected to confirm their acceptance or rejection, especially when the goal of the project is to search for low level biomarker components. With the experience of analyzing this large set of MS^2 information, we propose the following strategy for evaluating proteomic data: (1) decide for a set of filters that will completely eliminate low quality data, e.g., filter 1; (2) decide for a combination of filters that will pass only very high quality data (no/minimum false positives), e.g., filters 1+2+$p<0.001$; (3) depending on specific needs, manually evaluate the set of data that fall between these two categories. In our case, this would amount to manually evaluate ~400-500 MS^2 spectra. Ideally, the combination of filters should be chosen such that the number of intermediate quality spectra is maintained at a minimum value. The chosen filter values should be used, however, only as guidelines for evaluating the overall quality of the analysis protocol or for inter-lab comparisons; specific values should be set according to the experience of the investigator or the needs of the research project.

Table 1 Number of proteins that were identified in the MCF7 cell line by using different filtering parameters (filter 1: Xcorr vs. charge state; filter 2: multiple thresholds).

	Filter 1			Filters 1 + 2		
	Hits	**p<0.5**	**p<0.001**	**Hits**	**p<0.5**	**p<0.001**
8 µL injection, NCBI database (131,585 entries)						
Total protein	6,363	2,576	1,471	1,887	1,438	1,157
Top match proteins	2,329	1,713	1,383	1,207	1,171	1,089
Total peptides	7,196	6,948	5,940	4,837	4,818	4,445
40 µL injection, NCBI database (131,585 entries)						
Total proteins	12,362	4,806	2,476	3,359	2,502	1,985
Top match proteins	4,534	3,131	2,367	2,107	2,033	1,895
Total peptides	14,981	14,333	11,770	9,677	9,634	8,777
40 µL injection, SWISSPROT database (63,973 entries)						
Total proteins	10,045	3,873	1,927	2,783	1,998	1,569
Top match proteins	3,723	2,525	1,859	1,691	1,632	1,518
Total peptides	14,805	14,217	11,722	9,736	9,695	8,812

Table 2 Search for false positives with the Forward/Reverse NCBI database (filter 1: Xcorr vs. charge state; filter 2: multiple thresholds).

	Filter 1			Filters 1 + 2		
	Hits	**p<0.5**	**p<0.001**	**Hits**	**p<0.5**	**p<0.001**
40 µL injection, NCBI FOR-REV database (263,170 entries)						
Top match proteins	5,285	3,352	2,377	2,133	2,042	1,880
Top match proteins false	1,433	435	24	64	35	2
% Top match proteins false	54.2	25.9	2.0	6.0	3.4	0.2
Total peptides	15,573	14,794	11,733	9,685	9,636	8,715
Total peptides false	1,178	808	36	109	88	6
% peptides false	15.1	10.9	0.6	2.2	1.8	0.1

To estimate the actual false positive identification rates in our study, and the effectiveness of our data selection criteria, a composite database that contained the forward and reverse directions of the protein entries from the NCBI database was created [4, 5]. By applying both filtering parameters and selecting only peptides with $p<0.001$, the 40 μL injection yielded false positive rates of ~0.1 % at the peptide level and ~0.2 % at the protein level (**Table 2**). Should we have used only filter 1 (Xcorr vs. charge state) without applying p-value sorting, the false positive rates would have been much higher, 15.1 % at the peptide level and 54.2 % at the protein level. Peng has analyzed yeast proteins by a 2D-SCX-RPLC protocol and evaluated the data by using a composite database containing yeast ORFs in both forward and reverse direction [5]. He has shown that using Xcorr cut-off values of 1.9, 2.2 and 3.75, equal to the ones that we used for the evaluation of the MCF7 data with filter 1, the false positive peptide and protein rates were 2.6 % and 30.8 %, respectively. The larger false positive rates in our study can be the result of several factors, including the size of the database. A yeast protein database contains ~6,300 protein entries, while the human protein database in our research contained 131,585 entries. Peptide intensity thresholds and mass tolerance values for database searching can also be a contributing factor. For example, by increasing the peptide tolerance from 0.5 to 1.5 amu for searching a human database with data generated by one of the SCX fractions, the number of matched proteins increased by 47 %, from 516 to 759. However, the increase in protein matches was as high as 85 % with other SCX fractions. For the present study, the peptide tolerance for database searching was set at 2 amu, to avoid any losses in possible protein identifications due to peptide selection for MS^2 according to the 2nd or 3rd most intense isotopic peaks. These selections will

clearly affect the search results with reversed databases, as well. Nevertheless, once a preliminary selection of protein matches is performed, by increasing the stringency of data filtering parameters, the rate of false positive peptide/protein matches can be dramatically reduced, from 15.1 % to 0.13 % and from 54.2 % to 0.2 %, respectively.

An additional factor that can be used to increase the confidence of protein identifications is the number of unique peptides that matched a given protein. **Table 3** summarizes data that were selected with a combination of filters and various p-values. The effect of accepting only proteins with low p-values was to eliminate mainly the proteins identified by a single peptide, as most of the proteins that were identified by 2 or more peptides had $p<0.001$.

Table 3 Protein distribution according to the number of unique matching peptides (40 µL injection, NCBI database, filter 1: Xcorr vs. charge state; filter 2: multiple thresholds; filter 3: different peptides; filter 4: top 1 match proteins).

	# of top 1 proteins							
# of unique peptides/protein	**Filter 1 + 3 + 4**				**Filter 1 + 2 + 3 + 4**			
	Hits	**p<0.1**	**p<0.001**	**p<1E-10**	**Hits**	**p<0.1**	**p<0.001**	**p<1E-10**
= 1 peptides	4154	2866	2325	1044	2048	1989	1856	810
= 2 peptides	1432	1381	1328	786	977	977	969	573
= 3 peptides	850	848	843	594	566	566	564	401
= 4 peptides	595	594	592	459	373	373	373	290
= 5 peptides	422	421	420	346	270	270	270	228

The data within each of the 16 SCX fractions was also analyzed in detail. **Figure 4A** is a schematic representation of the protein and peptide distributions across the SCX fractions, indicating a fairly uniform distribution, which is a highly desirable outcome when the sample is very complex. The ratio of unique peptides/total peptide hits in each fraction was >80%, indicating that the rate of duplicate hits was relatively small, and that the MS instrument time was efficiently utilized on identifying new sequences. The distribution of proteins as a function of their p-values across all fractions is shown in **Figure 4B**. These represent data that were selected with the aid of filters 1 and 2; 85-90 % of top match proteins had $p<0.001$. The number of peptide hits per each of these fractions was fairly high, culminating with more than 1,000 total hits in fraction 5. Base-peak chromatograms of this fraction with 8 and 40 μL sample injections are shown in **Figure 5A** and **B**. The 8 and 40 μL experiments were performed a few weeks apart, using 2 different separation columns. The samples came from two different sets of SCX fractions. The similarity between the two chromatograms confirms the reliability of the overall 2D-SCX-RPLC separation method. The difference in elution times between the two chromatograms is a result of using 8 and 40 μL sample loops for the two experiments. The use of the 40 μL loop resulted in approximately ~15 min delay in the onset of the gradient through the RPLC column. Most of the proteins identified in the 8 μL injection were present in the 40 μL injection, as well. If all the proteins were counted, only ~ 5-9% of the proteins were unique to the 8 μL injection, while if only proteins matched by 2 peptides were counted, there were no unique proteins identified in the 8 μL injections. A comparison between the data generated with the 8 and 40 μL injections for the SCX fractions 5, 6, and 7 is shown in **Table 4**. The distribution of the sample

components within the sample elution window, the separation efficiency, and the complexity of the mixture is highlighted by a 2D-view of this (5^{th} fraction) separation (**Figure 6A**). An inset of the high m/z mass region of the chromatogram is shown in **Figure 6B**. Most of the ions in this region were not selected for MS^2 as they did not make it to the top 5 most intense peak list. This resulted in some loss of structural information for ions that were not the singly charged counterparts of the more intense multiple charged species in the spectrum that were selected for CID.

Table 4 Protein comparison between the 8 and the 40 µL injections for the SCX fractions 5, 6, and 7 (filter 1: Xcorr vs. charge state; filter 2: multiple thresholds).

Top 1 proteins						
	Fraction 5		**Fraction 6**		**Fraction 7**	
	=1peptide/ protein	**=2peptide/ protein**	**=1peptide/ protein**	**=2peptide/ protein**	**=1peptide/ protein**	**=2peptide/ protein**
Total proteins	434	90	567	141	633	146
Overlapped proteins	151	67	224	107	256	116
Proteins (only 8 µL)	25	0	40	0	55	1
Proteins (only 40 µL)	258	23	303	34	322	29
% Overlap	35	74	40	76	40	80

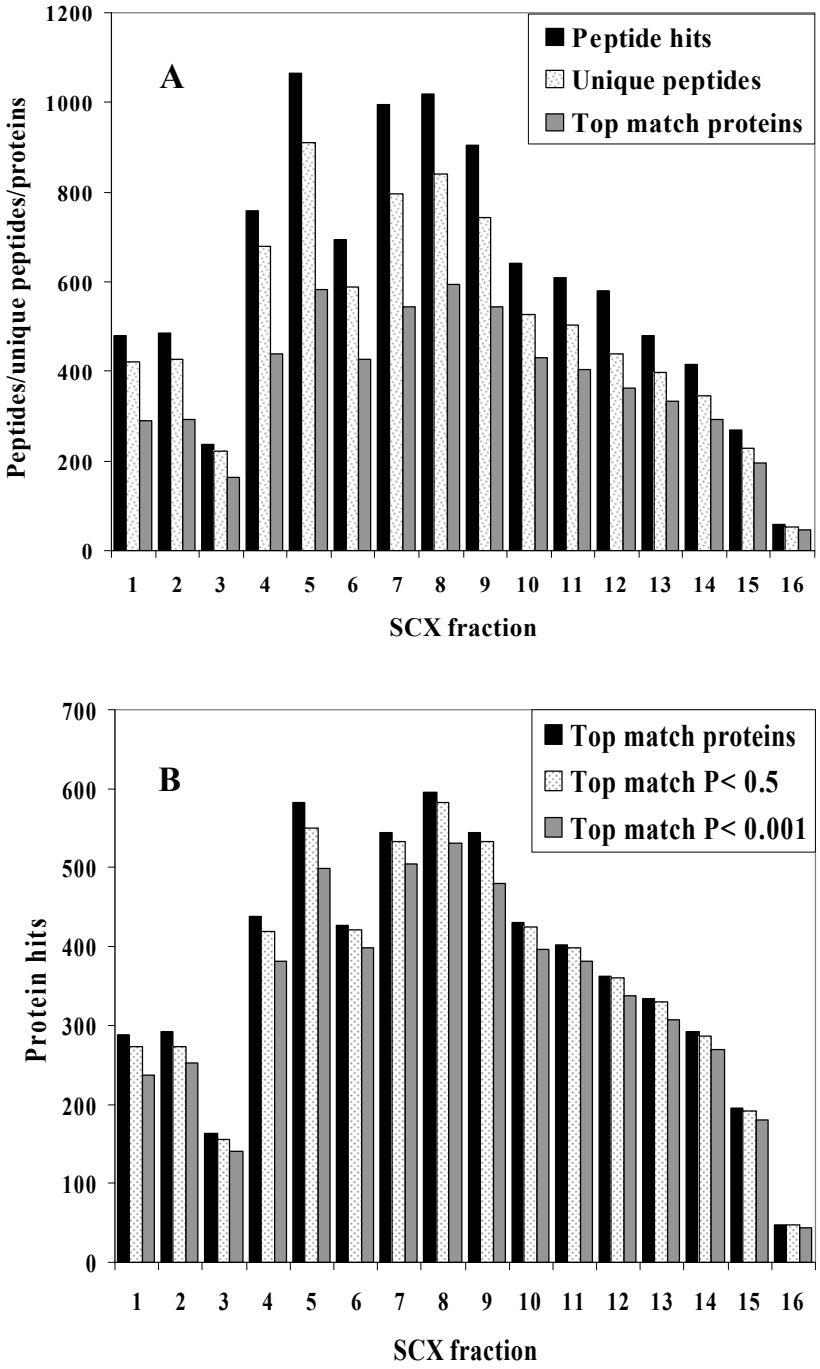


Figure 4. Number of peptide and protein identifications in each of the SCX fractions (40 µL injection). (**A**) Peptide/protein distribution across the SCX fractions; (**B**) p-value distribution of first choice proteins across the SCX fractions. Data were selected with the Xcorr vs. charge state and multiple threshold filters.

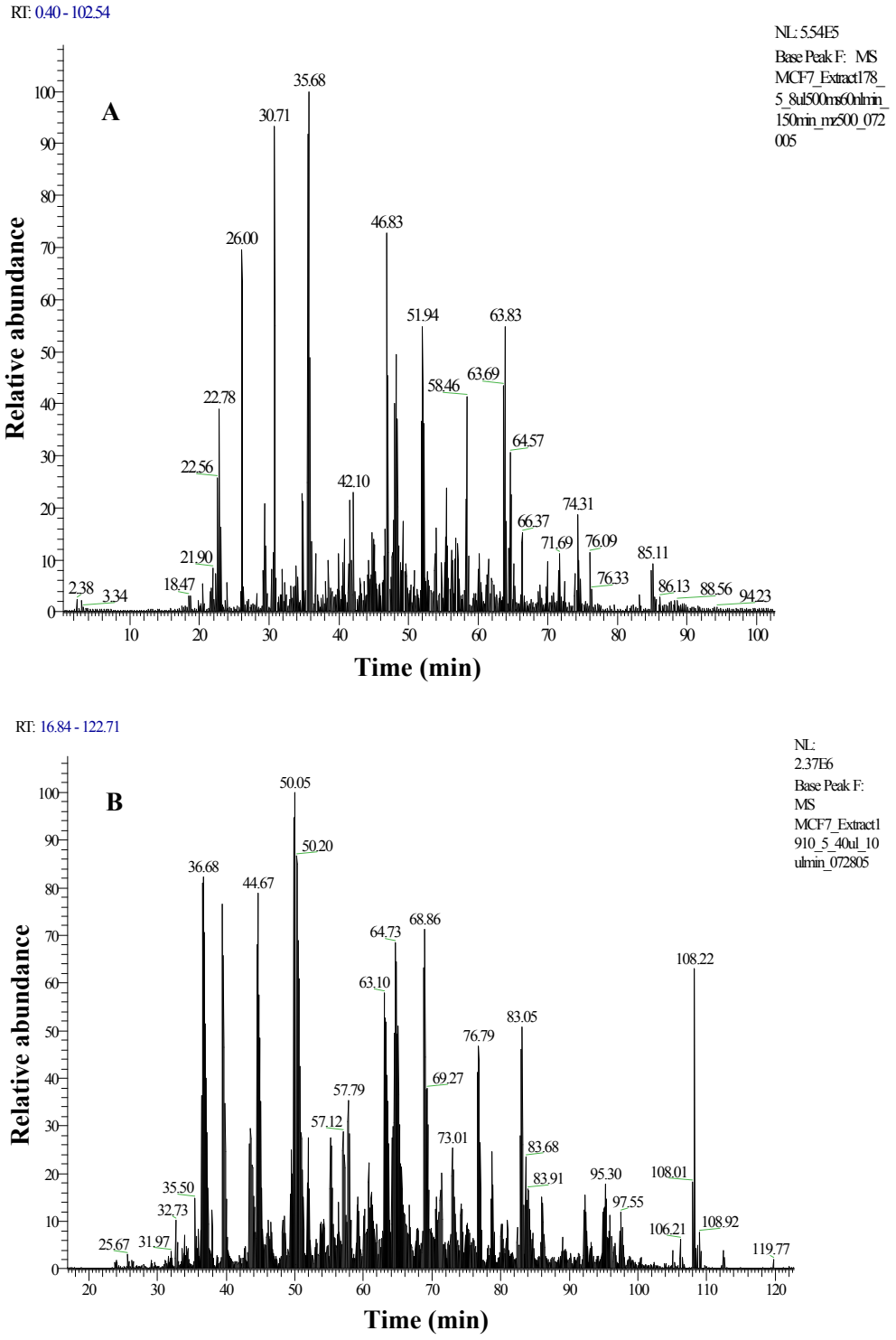


Figure 5. Representative chromatograms of complex LC-MS/MS separations. (**A**) Base peak chromatogram of SCX fraction 5 (8 µL injection); (**B**) Base peak chromatogram of SCX fraction 5 (40 µL injection).

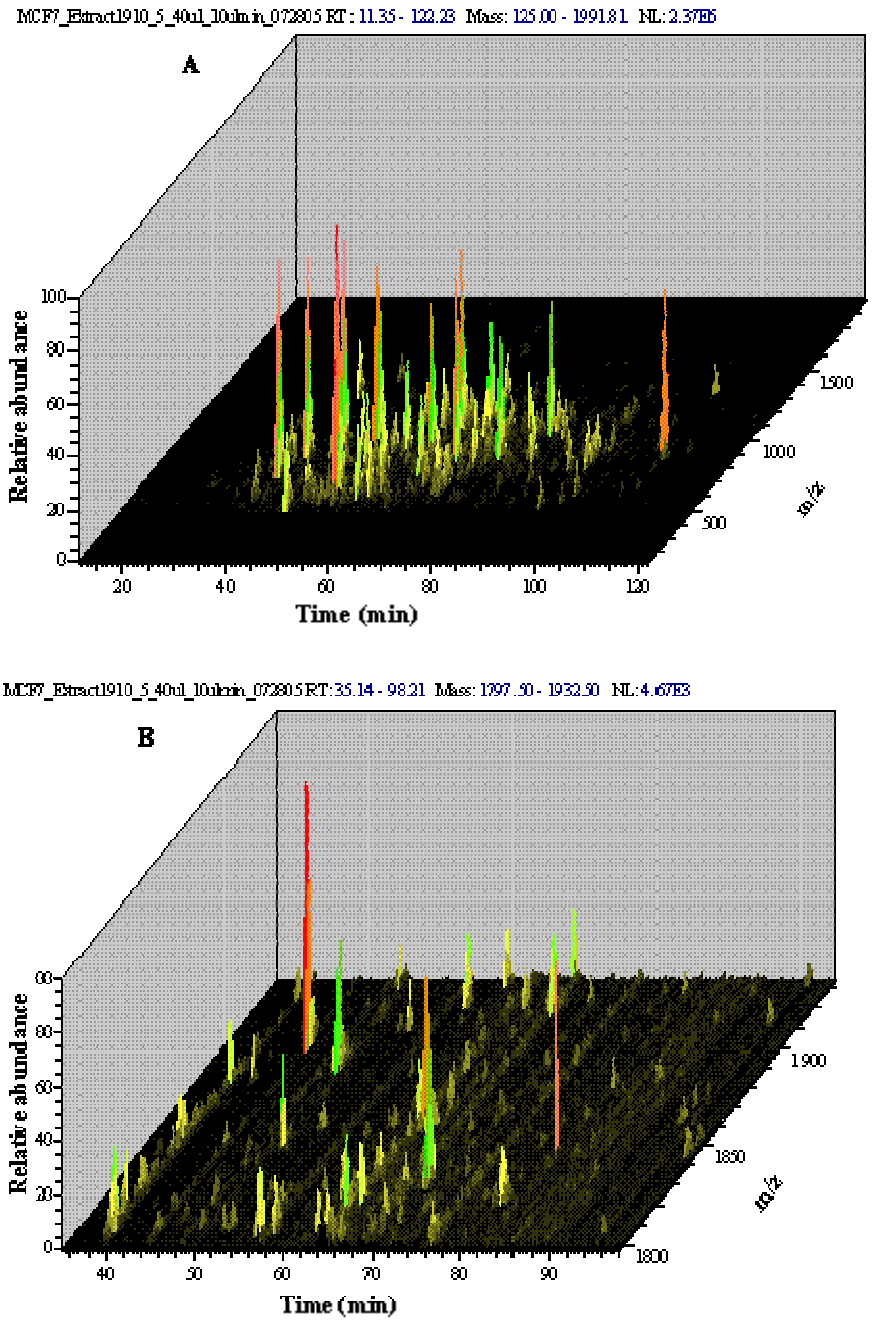


Figure 6. Representative 2D-chromatograms of complex LC-MS/MS separations. (**A**) 2D-view chromatogram of SCX fraction 5 (40 µL injection); (**B**) Inset from 5A, showing the 1,800-2,000 m/z region. Conditions are given in experimental section.

3.3 Protein categorization and pathway profiling

A total of 1,859 proteins with $p<0.001$ from the SWISSPROT database were categorized using the Gene Ontology (GO) identification tool (geneontology.org). The proteins were classified based on cellular localization and biological process, as illustrated in **Figures 7A** and **B**. It was not possible to classify all proteins, as some did not have a GO assignment. The graphical display for cellular location and biological process covers only 78% and 82 %, respectively, of the total number of identified proteins. As it is seen from **Figure 7A**, the larger compartments comprise proteins from the cytoplasm (17.19 %), nucleus (14.62 %), and cell membrane (13.74 %). **Figure 7B** illustrates a variety of biological processes associated with the list of identified proteins. We searched for specific processes known to be essential for the onset and development of cancer; approximately 218 proteins were identified under these categories: cell differentiation (17), cell growth and proliferation (62), cell cycle regulation (61), cell adhesion (19), apoptosis (42) and DNA repair (17).

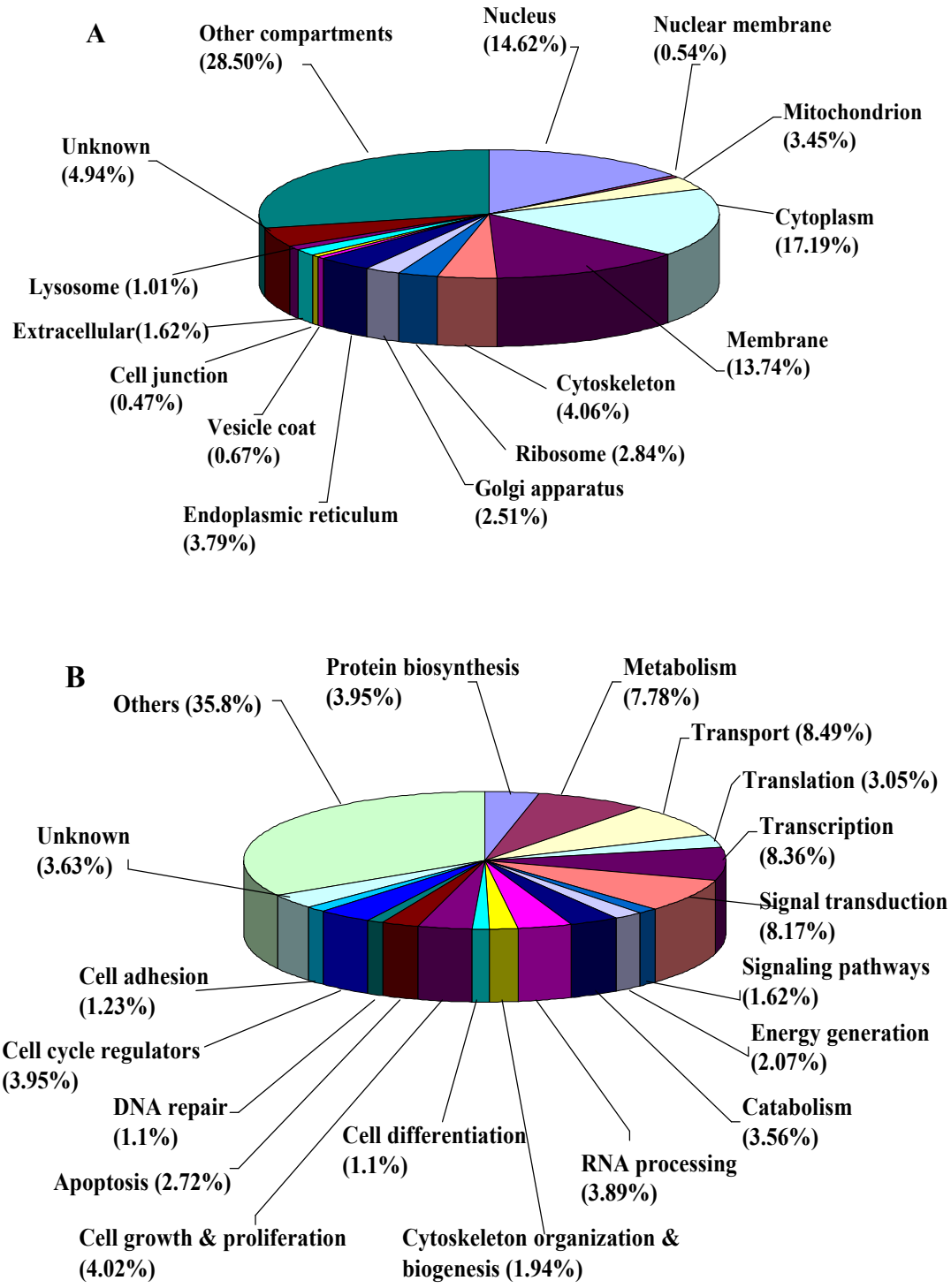


Figure 7. Protein categorization of 1,859 proteins identified in SWISSPROT. (**A**) Cellular location; (**B**) Biological process.

Many of the proteins that are present in some of the major cancer related pathways such as p53 signaling, programmed cell death or apoptosis signaling, and cell cycle regulation were identified in our results. Mutation of the p53 gene is very common in various types of human cancer. There is abundant data supporting the significance of p53 tumor protein in carcinogenesis [6, 7]. The main role of p53 is to eradicate and hinder the proliferation of abnormal cells, thereby preventing neoplastic development; the p53 signaling pathway is activated under conditions of cellular or genotoxic stresses generated by UV irradiation or DNA damage [8]. The mechanism of p53 activation is under a complex control: it can induce cell cycle arrest to eliminate damaged cells in response to DNA damage, or apoptosis if the damage cannot be repaired. The major inter playing factors that control p53 activation are protein interactions, post-translational modifications (mainly phosphorylation), and modification of subcellular protein localizations [8]. The pathway highlighting various activation and degradation mechanisms of p53 is shown in **Figure 8**. The transcription of p21 is turned on upon p53 activation caused by ?-irradiation, which further leads to binding and inhibition of cyclin dependent kinases (CDK). The immediate response to this binding is hypophosphorylation of retinoblastoma (Rb) protein that prevents the release of E2F and blocks the G1-S transition [9]. Deregulated expression of many proteins like c-Myc, Bcl-2, E2F, and Apaf-1 are recognized as blocking agents to block the cellular effects of p53. The human phosphoprotein homologue of the murine double minute 2 (Mdm2) gene, also known as ubiquitin-ligase, forms a complex with p53 leading to degradation, and hinders p53-induced cell cycle arrest and apoptosis [10]. Hence, the autoregulatory loop of p53 with Mdm2 is known to control the activity of p53.

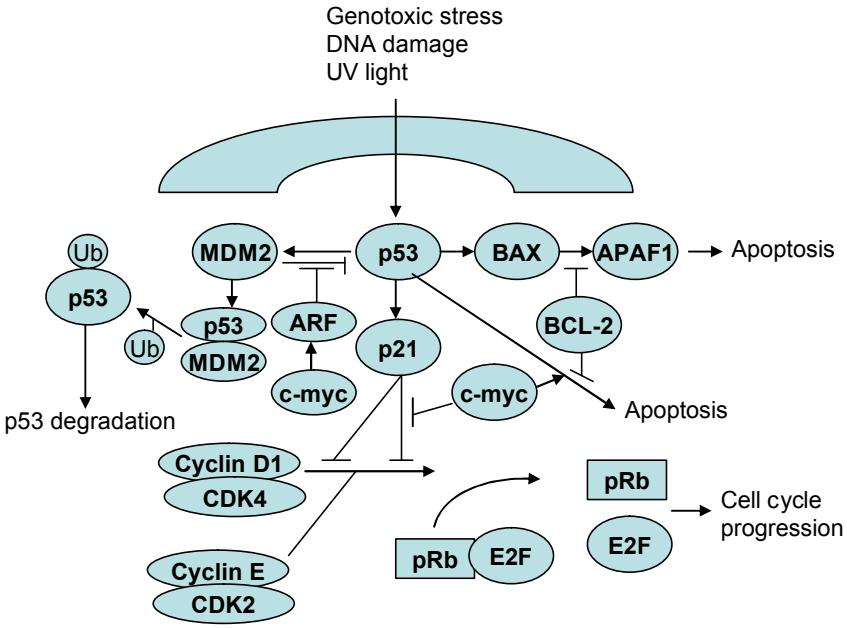


Figure 8. p53 signaling pathway highlighting activation and degradation of p53.

Table 5 Proteins involved in the p53 signaling pathway and identified in our results.

Protein	Accession #	p-value	of peptides	Function
TP53RK	gi|14714958|	3.71E-7	1	Tumor suppressor
p21-activated kinase2	gi|32483399|	1E-30	3	Inhibition of CDK
p21	sp|P38936|	2.34E-3	1	Inhibition of CDK
Cyclin dependent kinase 2 (CDK2)	sp|O14519|	1E-30	1	regulation of the cell cycle by binding with cyclins
Cyclin E	sp|Q9UII4|	0.0148	3	regulates the cell cycle transition from G1 phase to S phase
Retinoblastoma (pRb-1)	sp|P28749|	0.0128	3	Regulates cell cycle, facilitates differentiation & resrains apoptosis
c-Myc	sp|Q99417|	1.2E-4	1	DNA binding transcription factor that regulates expression of specific target genes
Mdm2	tr|Q96DY7|	3.23E-3	1	Binds to p53 and degrades its activity
Bax	sp|Q07812|	4.46E-7	1	proapoptotic functions
ARF1	sp|P84077|	1.58E-13	2	GTPase activity

The proteins involved in the p53 signaling pathway that were identified in our results are summarized in **Table 5**, along with their functions; a few proteins such as Bcl-2, E2F, and CDK4, were present in our data with a p-value >0.5; thus their identification was not reliable enough to be included in the final results.

Another interesting pathway involved in cancer is the apoptosis signaling pathway, induced in response to DNA damage. Apoptosis is basically a mechanism of removal of unwanted, aged and damaged cells. There are a wide range of stimuli such as cell surface receptors, and processes including growth factor withdrawal and exposure to chemotoxins, that are considered to be responsible for triggering programmed cell death [11]. The interactions of proapoptotic and antiapoptotic proteins of the B-cell lymphoma 2 (Bcl-2) families is responsible for the regulation of this process [12]. The relative abundance of proapoptotic and antiapoptotic proteins decides the susceptibility of the cell to programmed death. These proapoptotic proteins act at the surface of the mitochondrial membrane to reduce the mitochondrial trans-membrane potential and increase the permeability of the membrane, thereby releasing cytochrome C into the cytoplasm [12]. Cytochrome C can bind with, and activate the apoptotic protease activating factor (Apaf-1), also released from mitochondria. Apaf-1 further binds to caspapses such as caspase-9 activating a caspase cascade that finally leads to apoptosis [11]. The apoptosis signaling pathway is illustrated in **Figure 9,** and the proteins involved in the process that were identified in our results are summarized in **Table 6,** along with their p-values, number of peptides that identified the protein, and their function.

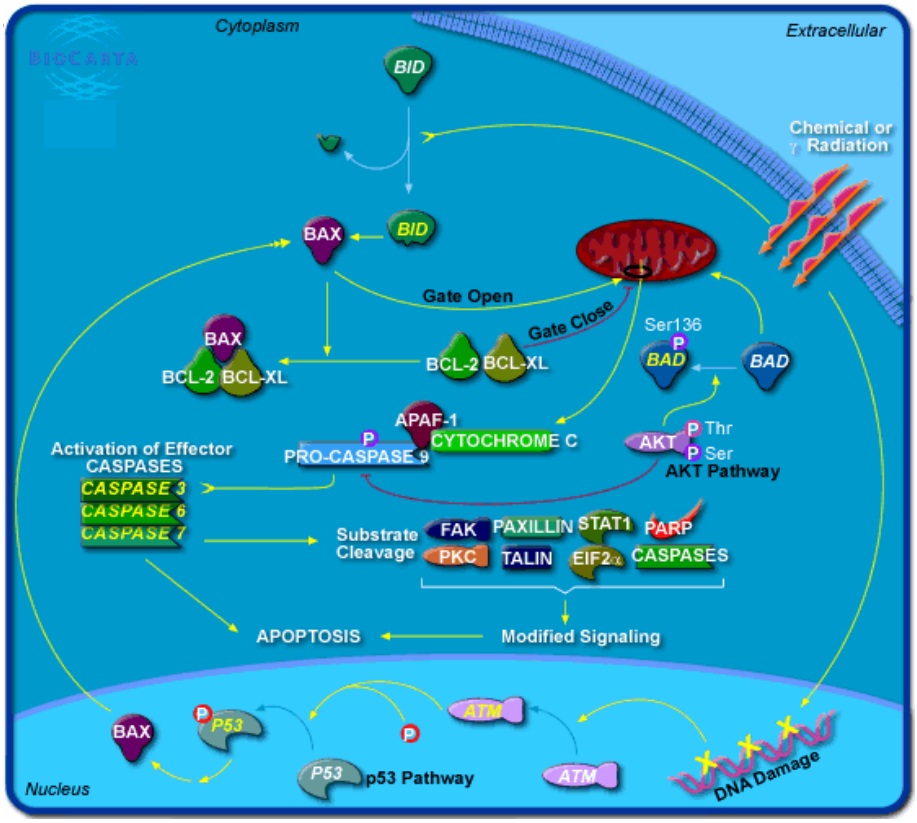


Figure 9. Apoptotic signaling pathway (www.biocarta.com).

Table 6 Proteins involved in the apoptosis signaling pathway and identified in our results.

Protein	Accession #	p-value	of peptides	Function
ADP-ribosyltransferase	gi|31415697|	7.76E-7	3	DNA repair
ATM	gi|28144171|	0.0112	3	Control the cell growth rate
Bax	sp|Q07812|	4.46E-7	1	proapoptotic functions
Caspase 9	sp|P55211|	7.08E-3	2	Initiatir caspase that activates effector caspase for apoptosis
Cytochrome C	sp|P99999|	1.26E-11	3	Released to create cascade of reactions leading to apoptosis
Eukaryotic translation initiation factor2 subunit1	sp|P05198|	1.62E-7	3	Blocks the cellular effects of p53
Talin 1	sp|Q9Y490|	5.01E-15	15	Cell motility
STAT-1	sp|P42224|	2.24E-9	6	Mediates growth arrest & apoptosis
TP53RK	gi|14714958|	3.71E-7	1	Tumor suppressor

One of the important signaling pathways in the human body is related to cell cycle regulation. Cyclins and cyclin dependent kinases play a key role in regulating the cell cycle by forming activated kinases that can phosphorylate the targets. The breakdown of the cell cycle regulation mechanism results in tumor formation, i.e., uncontrolled growth of cells. As a result of DNA damage, cell cycle progression is stopped by some key proteins such as p53, CDK inhibitors (p21), and retinoblastoma, until the damage is repaired [13]. If DNA damage occurs at the G2/M checkpoint of the cell cycle, than the transition of the cell to the mitosis (M) phase is prevented. The cell cycle regulation pathway (G2/M checkpoint) is shown in **Figure 10**. The activation of cyclinB/CDC2 complex is very important for the transition from the G2 phase to the M phase of the cell cycle [14]. This complex is deactivated by the Wee1 and MYT1 kinases during the G2 phase, and activated by the phosphatase CDC25 which is initially activated by the polo-like-kinase (PLK1) during the M phase [14]. Further more, the cyclinB/CDC2 complex is inactivated by two parallel pathways initiated by the DNA activated-PK/ATM/ATR kinases. The first pathway inactivates CDC25 by phosphorylation of CHK kinase and inhibits CDC2 activation, thereby hindering the progression into the M-phase. The second pathway involves the p53/Mdm2 complex, and it proceeds slowly. The DNA-binding and transcriptional activity of p53 is activated by the dissociation of Mdm2 from p53 due to phosphorylation and acetylation of p53 by p300/PCAF [13]. A few genes are turned on by p53; for instance, 14-3-3 sigma exports the phosphorylated CDC-2/cyclin B kinase to the nucleus by binding to it, GADD45 dissociates the cyclinB/CDC2 complex by binding to it, and p21Cip1 inhibits CDK's [13]. Proteins involved in the cell cycle regulation mechanism that were identified in our results are summarized in **Table 7**.

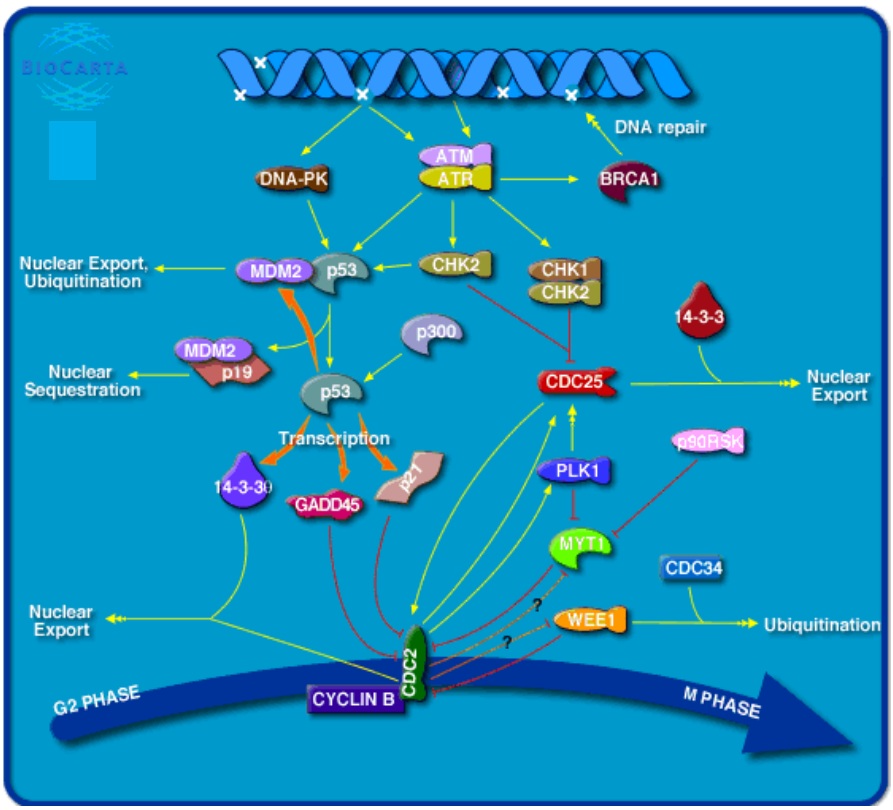


Figure 10. Cell cycle regulation pathway (www.biocarta.com).

Table 7 Proteins involved in cell cycle regulation and identified in our results.

Protein	Accession #	p-value	of peptides	Function
ATM	gi|28144171|	0.0112	3	Control the cell growth rate
TP53RK	gi|14714958|	3.71E-7	1	Tumor suppressor
Mdm2	tr|Q96DY7|	3.23E-3	1	Binds to p53 and degrades its activity
cyclin dependent kinase inhibitor 1A (p21)	sp|P38936|	2.34E-3	1	Inhibition of CDK
CHK1 checkpoint homolog	gi|30584865|	0.0204	2	During S phase, protects DNA breakage
PLK-1 polo like kinase 1	sp|Q9NYY3|	0.0112	1	Cell cycle progression & cell division
14-3-3 sigma	sp|P31947|	2.89E-14	16	p53-regulated inhibitor of G2/M progression
Ribosomal protein S6 kinase	sp|P62753|	6.31E-14	9	Cell growth & proliferation

3.4 Biomarkers in cancer research

The discovery of cancer biomarkers is crucial in the clinical setting to facilitate early diagnosis, treatment, and increase survival rates. There is tremendous research being performed to discover new biomarkers and targets for drug development; however, very few markers are currently used in clinical practice [15]. Marker detection and correlation with tumor growth can be relatively easily performed for advanced tumors, but the true value of novel methods lies rather in early tumor diagnosis, where successful therapy is still possible. The main aim should be to enhance the methodology to obtain higher specificity and prognostic value of a wide range of biomarkers that could be used collectively towards early detection of cancer.

A list of identified cancer specific proteins, along with their probability values, sequence coverage, molecular weight, function, and number of peptides that matched each protein, is given in **Table 8**. This table includes a series of proteins that were searched with descriptors such as tumor protein, cancer/carcinoma, oncogene, growth factor, receptor, antigen, and some membrane proteins that are, or could be, related to cancer. The list includes a wide range of proteins that could be potential markers, or are known to be established markers for cancer, as they were determined to be differentially expressed between normal and cancerous cell states. Cancer biomarkers that include cathepsin D, E-cadherin, proliferating cell nuclear antigen (PCNA), Ki-67, TP53RK, CA125, and 14-3-3 sigma, were identified in our results from the MCF7 cellular extracts. Not all proteins were identified in both databases.

Table 8 List of potential biomarkers identified in the MCF7 cell line (reference of origin for each biomarker is provided).

Protein Name and Reference	p-value	Coverage	MW	Peptide (Hits)	Function
sp|P31947|1433S_HUMAN 14-3-3 protein sigma (Stratifin) [35,36,38]	2.89E-14	46.8	27756.7	16 (16 0 0 0 0)	Regulation of signal transduction pathways
gi|30584113|gb|AAP36305.1| Homo sapiens cathepsin D [16-19,35]	1.00E-30	40.4	44636.8	29 (28 0 0 0 1)	Promotes cancer cell migration
gi|6682961|dbj|BAA88956.1| E-cadherin [Homo sapiens] [22-24]	1.05E-06	3.53	90885.7	2 (2 0 0 0 0)	Cellular adhesion & signal transduction
sp|P46013|KI67_HUMAN Antigen KI-67 - Homo sapiens [30]	3.29E-08	6	358523.8	15 (4 4 3 3 1)	Cell proliferation
sp|P12004|PCNA_HUMAN Proliferating cell nuclear antigen (PCNA) [25-29,35]	1.66E-12	42.1	28750.3	17 (17 0 0 0 0)	Cell proliferation
gi|24419041|gb|AAL65133.2| ovarian cancer related tumor marker CA125 [15,32]	0.0655706	0.9	2352946	7 (1 2 1 0 3)	Cellular proliferation & apoptosis
gi|913148|gb|AAB33281.1| calreticulin=calcium binding protein [29,35,54,57]	3.25E-13	61.29	3740.9	5 (5 0 0 0 0)	Molecular chaperone
gi|14714958|gb|AAH10637.1| TP53RK protein [Homo sapiens] [6,7,31]	3.75E-07	8.33	27968.8	1 (1 0 0 0 0)	Tumor suppressor
gi|12653819|gb|AAH00698.1| Keratin 18 [Homo sapiens] [35,43,48,49]	1.00E-30	62.6	48002.6	116 (113 3 0 0 0)	Cytoskeleton and cell adhesion
gi|7594732|dbj|BAA94607.1| keratin 19 [Homo sapiens] [35,43,48,49]	1.00E-30	60.1	23330.9	40 (37 1 1 1 0)	Cytoskeleton and cell adhesion
gi|66933016|ref|NP_000875.2| inosine monophosphate dehydrogenase 2 [35,66]	9.88E-14	24.3	55769.7	12 (10 1 0 0 1)	Regulation of cell growth
gi|63252896|ref|NP_001018004.1| tropomyosin 1 alpha chain isoform 3 [35,67]	2.56E-06	4.9	32716.7	2 (2 0 0 0 0)	Cytoskeletal protein in cell growth
gi|3132833|gb|AAC16450.1| vascular endothelial growth factor receptor 2 [40-42]	0.0836	1.1	151431.4	1 (1 0 0 0 0)	Angiogenesis factor
gi|1620018|dbj|BAA13431.1| heat shock protein 90 [Homo sapiens] [29,35,44,45]	7.52E-13	37.1	16808.3	11 (11 0 0 0 0)	Cell cycle regulators

gi|14326412|gb|AAK60261.1| heat shock protein 60 Hsp60s2 [29,35,44,45]	2.12E-08	6.2	27078.9	1 (1 0 0 0 0)	Cell cycle regulators
gi|136378|sp|P02786|TFR1_HUMAN Transferrin receptor protein 1[68]	9.73E-10	4.9	84848	2 (2 0 0 0 0)	Provides iron for DNA synthesis
gi|6650599|gb|AAF21930.1| epidermal growth factor receptor [45,69]	1.89E-14	3.9	94197.1	2 (2 0 0 0 0)	Phosphorylation of tyrosine to initiate cell proliferation
gi|60823739|gb|AAX36654.1| tumor protein D52 [70]	3.13E-07	31.5	19835.2	5 (5 0 0 0 0)	DNA binding protein
gi|61679634|pdb|1Y41|A ChainA, human translationally controlled tumor protein [71,72]]	8.66E-14	18.9	20647.1	5 (5 0 0 0 0)	Calcium binding & apoptosis
gi|47606203|sp|Q99973|TEP1_HUMAN Telomerase protein component 1[42,73]	0.7943282	1	290229.2	2 (1 0 0 1 0)	Cellular inducible enzyme
gi|15080490|gb|AAH11988.1| CD9 antigen [74]	0.0001844	4.4	25369	1 (1 0 0 0 0)	Cell migration
gi|38305346|gb|AAR16191.1| antigen MLAA-42 [Homo sapiens] [75]	8.88E-15	64.2	16860.5	8 (8 0 0 0 0)	Elongation factor
gi|42560541|sp|P48960|CD97_HUMAN CD97 antigen precursor [76]	4.24E-10	1.9	91781.3	1 (1 0 0 0 0)	Adhesion and signaling
gi|14714785|gb|AAH10541.1| S100 calcium binding protein A16 [46,47]	1.24E-11	26.2	11794	2 (2 0 0 0 0)	Calcium binding & cell adhesion
gi|13477125|gb|AAH05019.1| S100 calcium binding protein A14 [46,47]	1.16E-08	25	11654.8	3 (3 0 0 0 0)	Calcium binding & cell adhesion

As annotations in diverse databases can vary, the search for specific proteins by their name in a report must be performed diligently, to avoid confusions or potential misinterpretation of the results. Moreover, as the "reporting for duplicate references" feature was not enabled during the database search, only the first protein match to a specific amino acid sequence was recognized. Additional proteins (often from the same family), that have large identical amino acid sequences to the first entry, were not identified in our study. For example, in the NCBI database there were 4 proteins with the name of "E-cadherin" (different accession numbers and slightly different amino acid sequences) and one entry with the name "E-cadherin epithelial." Only the first E-cadherin protein was reported in the Sequest report. In the SWISSPROT database, on the other hand, there were 2 proteins with the same name of "E-cadherin" (different accession numbers and slightly different amino acid sequences) that none were identified in our search, as the two peptide sequences that matched "E-cadherin," also matched another entry annotated "Epithelial-cadherin precursor (E-cadherin)", that was queried first in the database and reported in the final list.

MS^2 spectra for some of these proteins are given in **Figures 11-17**. Only the relevant ions (b, y and a few others) were marked in the spectra, however, many additional ions such as (a), (b-H_2O), and (b-NH_3), were also assigned. The peptide that identified Ki-67 (with ~25 % matched ions in the spectrum, while the required threshold was 30 %) and the peptide that identified TP53RK (with Sp value of 347, while the required threshold was 500) passed only filter 1, but not 2. This is a relevant example of the limitations associated with the selection of proteins based solely on cut-off values of filtering parameters.

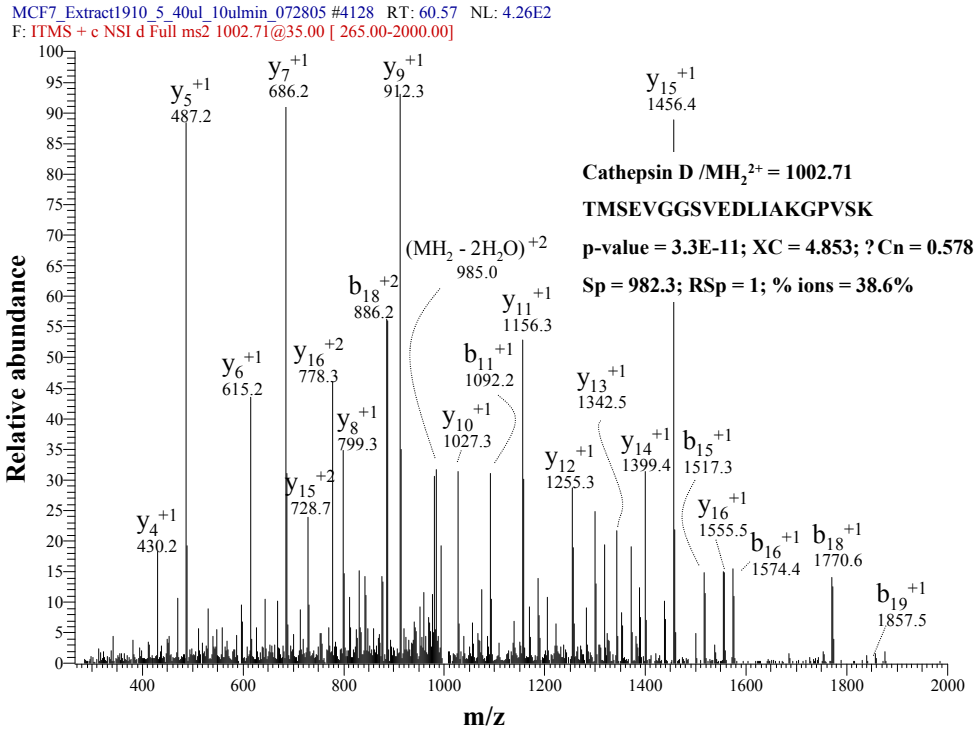


Figure 11. Mass spectrum of Cathepsin D.

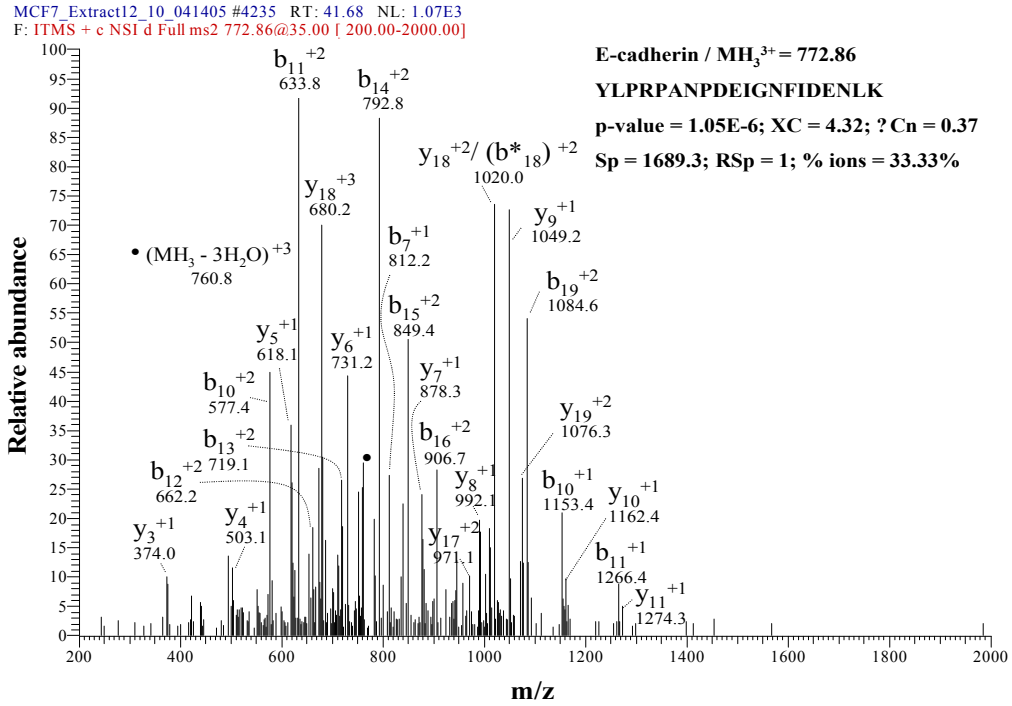


Figure 12. Mass spectrum of E-cadherin; Note: "o" represents ions that lost one molecule of H2O. "*" represents ions that lost one molecule of NH3.

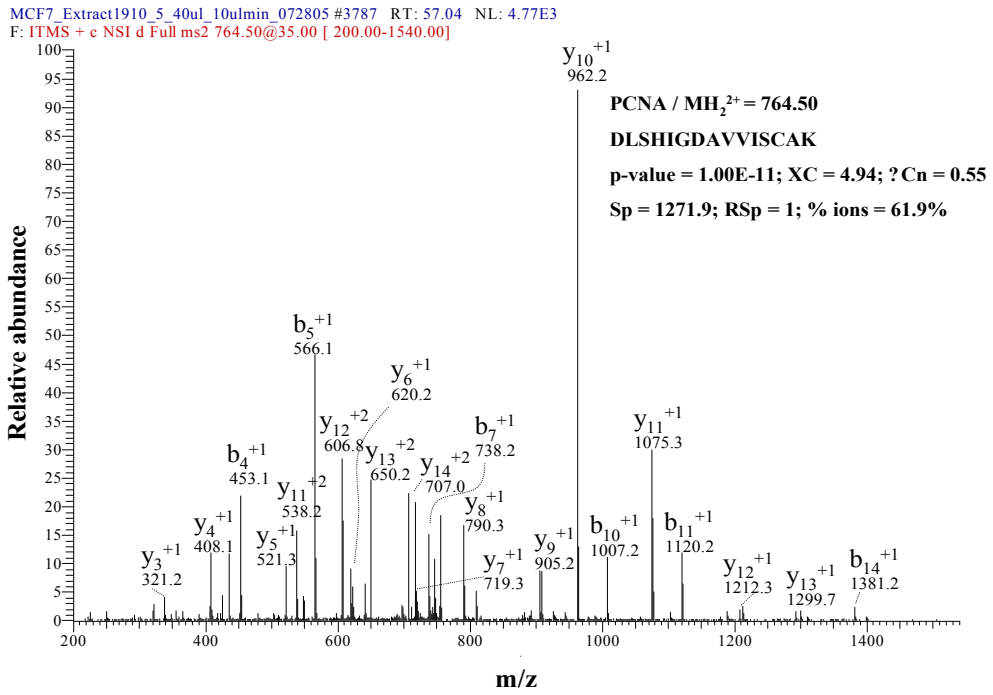


Figure 13. Mass spectrum of proliferating cell nuclear antigen (PCNA).

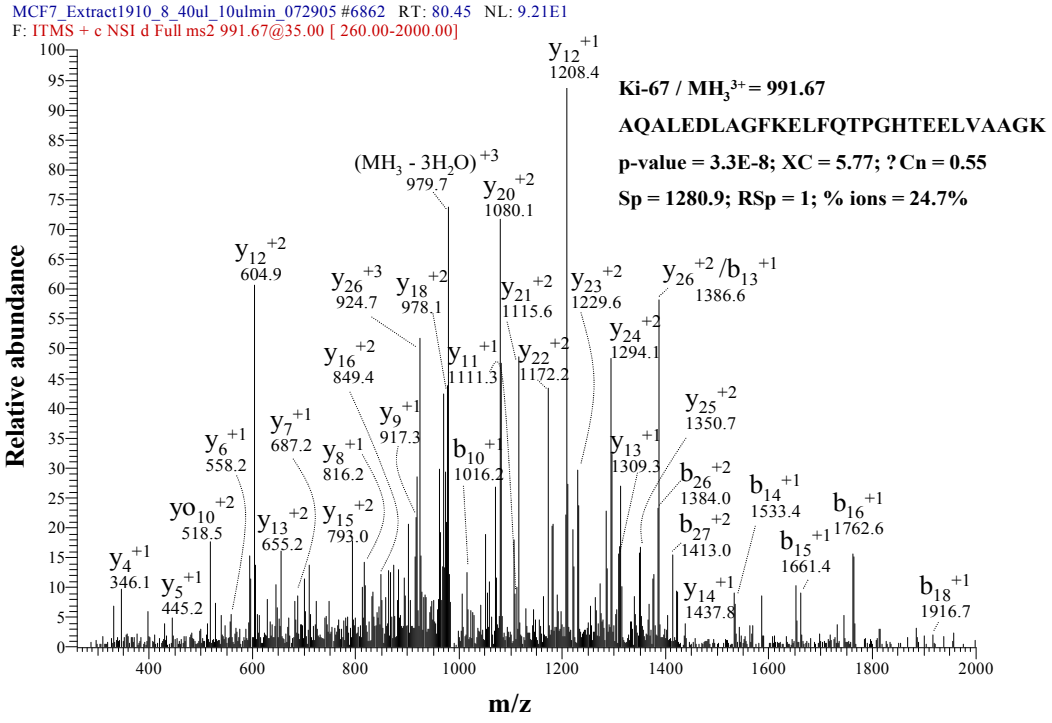


Figure 14. Mass spectrum of cell proliferation antigen Ki-67; Note: "o" represents ions that lost one molecule of H2O. "*" represents ions that lost one molecule of NH3.

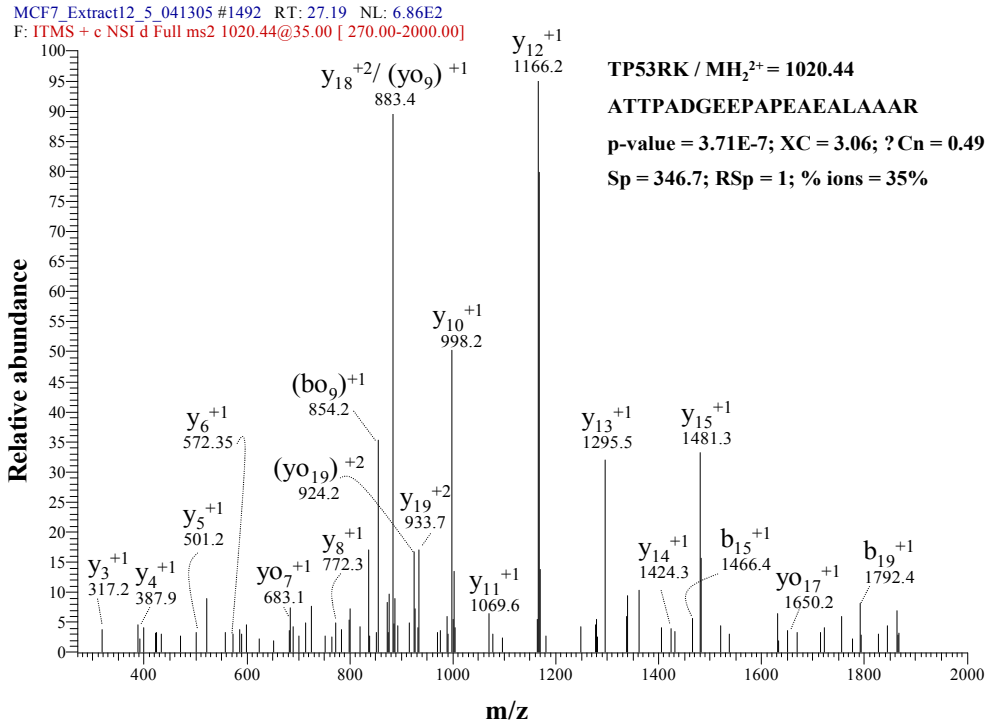


Figure 15. Mass spectrum of TP53RK; Note: "o" represents ions that lost one molecule of H2O. "*" represents ions that lost one molecule of NH3.

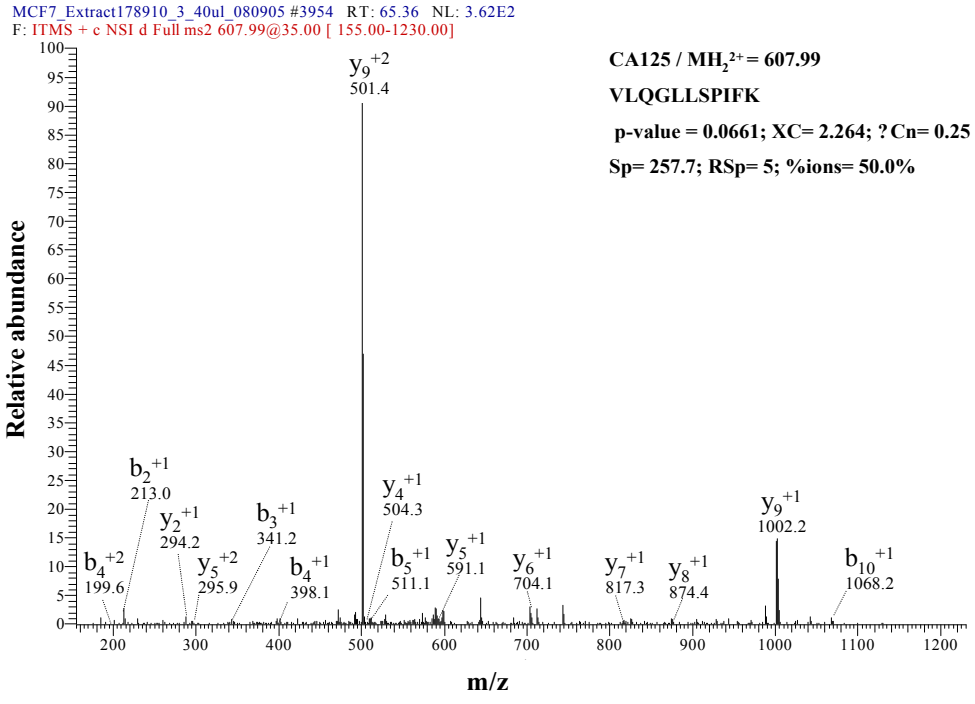


Figure 16. Mass spectrum of CA125; Note: "o" represents ions that lost one molecule of H2O. "*" represents ions that lost one molecule of NH3.

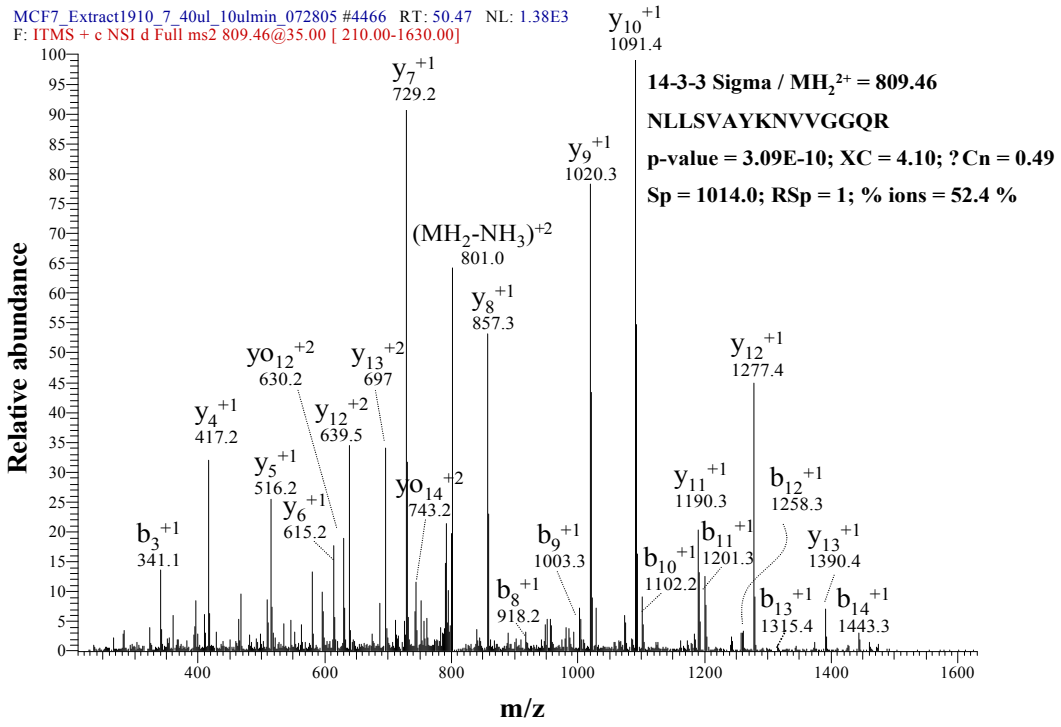


Figure 17. Mass spectrum of 14-3-3 sigma; Note: "o" represents ions that lost one molecule of H2O. "*" represents ions that lost one molecule of NH3.

Cathepsin D is an aspartyl lysosomal protease that is involved in protein metabolism, tissue remodeling, and cancer cell proliferation [16, 17, 18, 19]. An increased expression (2-50 fold) of cathepsin D has been found in estrogen positive breast cancer cells like MCF7, using a variety of techniques such as immunohistochemistry, cytosolic immunoassay, in situ hybridization and northern and western blot analyses [20]. The MS^2 spectrum of a peptide that matched cathepsin D is shown in **Figure 11** with a p-value of 1.00E-11. Its presence was identified by 28 peptide hits.

E-cadherin was identified in both the NCBI and SWISSPROT database by 2 peptide hits and a p-value of 1.05E-06 (**Figure 12**). It is believed that E-cadherin mediated cell-adhesion suppresses the tumor in breast cancer [21, 22, 23]. Apart from its involvement in the cell adhesion process, E-cadherin also plays an important role in signal transduction. It has been shown in a study by Berx et al. that the E-cadherin gene is mutated in lobular breast cancer, and hence its reduced expression is associated with invasiveness and unfavorable prognosis of the disease [24].

PCNA is involved in cell proliferation, cell cycle progression, and DNA replication. It is an acidic nuclear polypeptide with molecular weight of 36 kDa that is involved in nucleic acid metabolism [25]. **Figure 13** shows the MS^2 spectrum for a PCNA peptide. It was identified by 6 unique peptides and a p-value of 1.00E-11. Studies have revealed its association with many different types of cancers such as lung, pancreatic, and breast [25-28, 29].

Another protein involved in cell proliferation is the nuclear antigen Ki-67. It is found in all the phases of the cell cycle except the resting phase G_0, and hence Ki-67 is being used as a proliferation marker to measure the growth fraction of cells in human tumors [30]. Ki-67 was identified by 3 peptide hits and a p-value of 3.31E-8 (**Figure 14**). Immunohistochemical staining tests are typically used to determine the level of this protein.

Tumor protein p53 regulating kinase (TP53RK) was identified by one peptide with a p-value of 3.71E-7. The MS^2 spectrum of the peptide indicating the assignment of fragment ions is shown in **Figure 15**. Mutations of p53 are very common in different types of human cancers. On the average, 20% of p53 mutations are associated with breast

cancer [31]. P53 prevents the neoplastic development of cancer by blocking the proliferation of abnormal cells [6, 7].

CA125 is a glycoprotein with high molecular weight and is expressed mainly by epithelial ovarian cancers. Research has shown poor sensitivity and specificity associated with CA125 as an ovarian tumor marker, as it leads to many false positive results [32]. Investigations have shown CA125 to be present in other diseases, as well [33]. In one of the studies by Bast *et al.*, a monoclonal antibody test was used to detect CA125 [34]. CA125 was identified in our results by one peptide with a p-value of 0.0661. The MS^2 spectrum of the peptide that identified CA125 is shown in **Figure 16**.

Reports have shown the identification of 14-3-3 sigma, as a strong marker for the non-cancerous state of breast epithelial cells, by 2D gel and mass spectrometry techniques [35, 36]. The key role of this protein is in regulation of signal transduction pathways that control cell proliferation and differentiation [37, 38]. 14-3-3 sigma has been shown to negatively regulate cell growth by associating with cyclin-dependent kinases [39]. Moreover, a study by Hubert *et al.* has shown the clinical value of this protein by considering its down regulation in breast cancer biopsies, as compared to normal epithelial cells [38]. The presence of 14-3-3 sigma was indicated by 6 unique peptides and a p-value of 3.09E-10 (**Figure 17**).

Protein markers such as calreticulin, cytokeratin 18 and 19, heat shock proteins Hsp60 and Hsp90, and S100 calcium binding proteins, that are known to be over expressed in cancerous cells, were also identified in our results [32, 35, 40-47]. Cytokeratins belong to the intermediate filament family of proteins and when released from proliferating or dying tumor cells during apoptosis, they provide a useful marker for

epithelial malignancies, as evidenced by the number of available immunochemical assays for cytokeratins [43, 48, 49]. Heat shock proteins play an important role in cell differentiation and proliferation, invasion, metastasis, and apoptosis. Circulating levels of Hsp could be useful biomarkers for tumor diagnosis and carcinogenesis [44, 45]. Up-regulation of heat shock proteins is common in several cancers [50-53]. A study by Susumu *et al.* has shown results indicating the usefulness of calreticulin as a urinary tumor marker for bladder cancer [54]. Over expression of calreticulin has also been reported in different cancerous tissues such as breast, liver, and prostate cancer [39, 54-57]. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and are involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. Chromosomal rearrangements and altered expression of this gene have been implicated in breast and colorectal tumor metastasis [46, 47, 58, 59].

Established biomarkers such as BRCA1, BRCA2, carcinoembryonic antigen, and PSA, were also identified in our results with the use of filter 1 alone, however the probability values were very low, and after extensive visual inspection and validation they were not included in the final results. Genetic mutations linked to breast and ovarian cancer are often linked to the BRCA1 and BRCA2 genes. An estimated 10-15% of breast cancer cases are due to BRCA1 and BRCA2 mutations [NCBI/Medscape]. BRCA2 was identified in our results with filter 1 by 5 unique peptides, all with XCorr=2.21-3.24. From these, four peptides had a ? Cn value between ~0.11-0.52, and one peptide a ? Cn value of ~0.07. The preliminary score (Sp) and rank of preliminary score (Rsp) ranged from 151-324 and 7-41, respectively. Not all the peptides matched BRCA2 as a top match protein, except one; the other peptides were considered better matches for other

proteins. This peptide had Xcorr=2.269, ? Cn=0.332, Sp=306.6, Rsp=14, and %ions=33%. Even though the peptides had good Xcorr and ? Cn values, they did not pass validation due to poor spectral quality, as a result of insufficient Sp and Rsp values. This information on BRCA2 can be considered as one of the challenges associated with the proper selection of predetermined filtering parameters.

Similarly, PSA, a biomarker for prostate cancer, was identified by a single peptide with Xcorr=3.52, ? Cn=0.21, Sp=225.6, Rsp=1, and %ions=24%. However, due to inadequate p-value and low preliminary score, the spectral quality was not good, and failed manual validation. Although a biomarker for prostate cancer, studies have shown its presence in many nonprostatic sources. One of the studies by Ferdinando *et al.* has shown the presence of PSA in breast secretions and tissues of diseased females [60]. Likewise, ErbB2 (HER2/neu) was also identified in our results but with an inadequate p-value. It is believed that HER2 gene encodes a membrane glycoprotein with tyrosine kinase activity that belongs to a growth factor receptor family [61]. Immunohistochemistry staining and fluorescence in situ hybridization (FISH) tests are used to determine the quantity and expression of the HER2. HER2 gene amplification and overexpression plays a crucial role in tumorigenesis and metastasis. Amplification of HER2 gene causes over expression of ErbB2 receptor protein in the cell. An increase in the rate of cell division followed by cancerous cell formation is due to excess ErbB2 formation. About 25% of breast cancers have ErbB2 over expression [62]. Expression of HER2 was not found in the MCF7 cell line, as shown in a study by Xiang *et al.*, instead it was found in BT474 cell line [63].

3.5 References

1. Wolters, D. A., Washburn, M. P., and Yates III, J. R. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683-5690

2. Ommen, G. S. (2004) AACR Conference on Advances in Proteomics and Cancer Research, Key Biscane, FL.

3. Eng, J. K., McCormack, A. L., and Yates III, J. R. (1994) An approach to correlate tandem mass spectral data in peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989

4. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: An algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13**, 378-386

5. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2002) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2**, 43-50

6. Sigal, A., and Rotter, V. (2000) Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome. *Cancer Res.* **60**, 6788-6793

7. Gasco, M., Shami, S., and Crook, T. (2002) The p53 pathway in breast cancer. *Breast Cancer Res.* **4**, 70-76

8. Jimenez, G. S., Khan, S. H., Stommel, J. M., and Wahl, G. M. (1999) p53 regulation by post-translational modification and nuclear retention in response to diverse stresses. *Oncogene*. **18**, 7656-7665

9. Gostissa, M., Hofmann, T. G., Will, H., and Sal, G. D. (2003) Regulation of p53 functions: let's meet at the nuclear bodies. *Curr. Opin. Cell Biol.* **15**, 351-357

10. Gu, J., Chen, D., Rosenblum J., Rubin, R. M., and Yuan, Z. M. (2000) Identification of sequence element from p53 that signals for Mdm2-targeted degradation. *Mol. Cell. Biol.* **20(4)**, 1243-1253

11. Kelekar, A., Tompson, C. B. (1998) Bcl-2-family proteins: the role of the BH3 domain in apoptosis. Trends Cell Biol. **8**, 324-330

12. Kuwana, T., Newmeyer, D. D. (2003) Bcl-2-family proteins and the role of mitochondria in apoptosis. *Curr. Opin. Cell Biol.* **15**, 691-699

13. Taylor, W. R., Stark, G. R. (2001) Regulation of the G2/M transition by p53. *Oncogene*. **20**, 1803-1815

14. Smits, V. A. J., Medema, R. H. (2001) Checking out the G2/M transition. Biochimica Biophysica Acta. **1519**, 1-12

15. Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and cancer biomarker discovery tool: opportunities and limitations. *Mol. Cell. Proteomics*. **3(4)**, 367-378

16. Garcia, M., Platet, N., Liaudet, E., Laurent, V., Derocq, D., Brouillet, J. P., and Rochefort, H. (1996) Biological and clinical significance of cathepsin D in breast cancer metastasis. *Stem Cells.* **14**, 642-650

17. Fusek, M., Vetvicka, V. (2005) Dual role of cathepsin D: ligand and protease. *Biomed. Papers.* **149(1)**, 43-50

18. Westley B, Rochefort H. (1980) A secreted glycoprotein induced by estrogen in human breast cancer cell lines. Cell. **20**, 353

19. Augereu P, Garcia M, Mattei MG, Cavailles V. *et al*. (1988) Cloning and sequencing of the 52K cathepsin D complementary deoxyribonucleic acid of MCF7 breast cancer cells and mapping on chromosome 11 .*Mol. Endocrinol.* **2**, 186

20. Henry, J. A., McCarthy, A. I., Angus, B. *et al*. (1990) Prognostic significance of the estrogen regulated protein, cathepsin D, in breast cancer. An immunohistochemical study. *Cancer.* **65**, 265-271

21. Frixen, U. H., Behrens, J., Sachs, M., Erbele, G., Voss, B., Warda, A., Lochner, D., Birchmeier, W. (1991) E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells. *J. Cell Biol.* **113**, 173-185

22. Leers, M. P. G., Aarts, M. M. J., Theunissen, P. H. M. H. (1998) E-cadherin and calretinin: a useful combination of immunochemical markers for differentiation between mesothelioma and metastatic carcinoma. *Histopathology.* **32**, 209-216

23. Marzo, A. M. D., Knudsen, B., Chan-Tack, K., Epstein, J. I. (1999) E-cadherin as a marker of tumor aggressiveness in routinely processed radical prostatectomy specimens. *Adult Urol.* **53**, 707-713

24. Berx, G. and Roy, F. V. (2001) The E-cadherin/ catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression. *Breast Cancer Res.* **3**, 289-293

25. Chu, J. S., Huang, C. S., and Chang, K. J. (1998) Proliferating cell nuclear antigen (PCNA) immunolabeling as a prognostic factor in invasive ductal carcinoma of the breast in Taiwan. *Cancer Lett.* **131(2)**, 145-152

26. Caputi, M., Esposito, V., Groger, A. M., Pacilio, C., Murabito, M., Dekan, G., Baldi, F., Wolner, E., and Giordano, A. (1998) Prognostic role of proliferating cell nuclear antigen in lung cancer: an immunohistochemical analysis. *In Vivo.* **12(1)**, 85-88

27. Yue, H., Na, Y. L., Feng, X. L., Ma, S. R., Song, F. L., and Yang, B. (2003) Expression of p57kip2, Rb protein and PCNA and their relationships with clinicopathology in human pancreatic cancer. *World J. Gastroenterol.* **9(2)**, 377-380

28. Horiguchi, J., Iino, Y., Takei, H., Maemura, M., Takeyoshi, I., Yokoe, T., Ohwada, S., Oyama, T., Nakajima, T., and Morishita, Y. (1998) Long-term prognostic value of PCNA labeling index in primary operable breast cancer. *Oncol. Rep.* **5(3)**, 641-644

29. Franzen B, Linder S, Alaiya AA, Eriksson E, *et al. Br. J. Cancer* 1996; **18**: 2832

30. Scholzen, T., Gerdes, J. (2000) The ki-67 protein: from the known and the unknown. *J. Cell. Physiol.* **182**, 311-322

31. Pharaoh, P. D., Day, N. E., and Caldas, C. (1999) Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br. J. Cancer.* **80**, 1968-1973

32. Moss, E. L., Hollingworth, J., and Reynolds, T. M. (2005) The role of CA125 in clinical practice. *J. Clin. Pathol.* **58**, 308-312

33. Daoud, E., Bodor, G. (1991) CA-125 concentrations in malignant and non-malignant disease. *Clin. Chem.* **37**, 1968-1974

34. Bast, R. C., Feeney, M., Lazarus, H., et al. (1981) Reactivity of a monoclonal antibody with human ovarian carcinoma. *J. Clin. Invest.* **68**, 1331-1337

35. Hondermarck, H., Sophie, A., Edouart, V., Revillion, F., Lemoine, J., Belkoura, I. E. Y., Nurcombe, V., and Peyrat, J. P. (2001) Proteomics of breast cancer for marker discovery and signal pathway profiling. *Proteomics.* **1**, 1216-1232

36. Vercoutter-Edouart, A. S., Lemoine, J., Le Bourhis, X., Louis, H., *et al*. (2001) Proteomic analysis reveals that 14-3-3 is down-regulated in human breast cancer cells. *Cancer Res.* **61**, 76-80

37. Hondermarck, H., Dolle, l., Belkoura, I. E. Y., Edouart, A. S. V., Adriaenssens, E., and Lemoine, J. (2002) Functional proteomics of breast cancer for signal pathway profiling and target discovery. *J. Mammary Gland Biol. Neoplasia*. **7(4)**, 395-405

38. Fu, H., Subramanian, R. R., and Masters, S. C. (2000) 14-3-3 proteins: Structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.* **40**, 617-647

39. Bini, L., Magi, B., Marzocchi, B., Arcuri, F., Tripodi, S., Cintorino, M., et al. (1997) Protein expression profiles in human breast ductal carcinoma and histologically normal tissue. *Electrophoresis.* **18**, 2832-2841

40. Esteva, F. J., and Hortobagyi, G. N. (2004) Prognostic molecular markers in early breast cancer. *Breast Cancer Res*. **6**, 109-118

41. Janssens, J. Ph., Verlinden, I., Gungor, N., Raus, J., and Michiels, L. (2004) Protein biomarkers for breast cancer prevention. *Eur. J. Cancer Prevention.* **13**, 307-317

42. Ross, J. S., Linette, G. P., Stec, J., Clark, E., Ayers, M., Leschly, N., Symmans, W. F., Hortobagyi, G. N., and Pusztai, L.(2004) Breast cancer biomarkers and molecular medicine: part II. *Expert Rev. Mol. Diagn*. **4(2)**, 169-188

43. Barak, V., Goike, H., Panaretakis, K. W., and Einarsson, R. (2004) Clinical utility of cytokeratins as tumor markers. *Clin. Biochemistry.* **37**, 529-540

44. Ciocca, D. R., Calderwood, S. K. (2005) Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. *Cell Stress Chaperones.* **10(2)**, 86-103

45. Baselga J. (2004) The science of EGFR inhibition: a roadmap to improved outcomes? *Signal.* **5(3)**, 4-8

46. Ilg, E. C., Schafer, B. W., Heizmann, C. W. (1996) Expression pattern of S100 calcium-binding proteins in human tumors. *Int. J. Cancer*. **68(3)**, 325-332

47. Hermani, A., Hess, J., Servi, B. D., Medunjanin, S., Grobholz, R., Trojan, L., Angel, P., and Mayer, D. (2005) Calcium-binding proteins S100A8 and S100A9 as novel diagnostic markers in human prostate cancer. *Clin. Cancer Res*.; **11(14)**: 5146

48. Trask, D. K., Band, V., Zajchowski, D. A., Yaswen, P., *et al.* (1990) Keratins as Markers that Distinguish Normal and Tumor-Derived Mammary Epithelial Cells *Proc. Natl. Acad. Sci. USA.* **87**, 2319-2323

49. Moll, R., Franke, W. W., Schiller, D. L., Geiger, B., Krepler, R. (1982) The catalog of human cytokeratins: Patterns of expression in normal epithelia, tumors and cultured cells *Cell.* **31**, 11-24

50. Lebret, T., Watson, R.W., Molinie, V., O'Neill, A., Gabriel, C., Fitzpatrick, J. M., Botto, H. (2003) Heat shock proteins HSP27, HSP60, HSP70, and HSP90: expression in bladder carcinoma. *Cancer* **98**, 970-977

51. Helmbrecht, K., Zeise, E., Rensing, L. (2000) Chaperones in cell cycle regulation and mitogenic signal transduction: a review. *Cell Prolif.* **33**, 341-365

52. Jaattela, M. (1999) Escaping cell death: survival proteins in cancer. *Exp. Cell Res.* **248**, 30-43

53. Jolly, C., Morimoto, R. I. (2000) Role of the heat shock response and molecular chaperones in oncogenesis and cell death. *J. Natl. Cancer Inst.* **92**, 1564-1572

54. Kageyama, S., Isono, T., Iwaki, H., Wakabayashi, Y., Okada, Y., Kontani, K., Yoshimura, K., Terai, A., Arai, Y., Yoshiki, T. (2004) Identification by Proteomic Analysis of Calreticulin as a Marker for Bladder Cancer and Evaluation of the Diagnostic Accuracy of Its Detection in Urine. *Clin. Chem.* **50(5),** 857-866

55. Yu, L. R., Zeng, R., Shao, X. X., Wnag, N., Xu, Y. H., Xia, Q. C. (2000) Identification of differentially expressed proteins between human hepatoma and normal liver cell lines by two-dimensional electrophoresis and liquid chromatography-ion trap mass spectrometry. *Electrophoresis.* **21**, 3058-3068

56. Alaiya, A., Roblick, U., Egevad, l., Carlsson, A., Franzen, B., Volz, D., et al. (2000) Polypeptide expression in prostate hyperplasia and prostate adenocarcinoma. *Anal. Cell. Pathol.* **21**, 1-9

57. Giometti, C. S., Williams, K., Tollaksen, S. L. (1997) A two-dimensional electrophoresis database of human breast epithelial cell proteins *Electrophoresis.* **18**, 573-581

58. Moog-Lutz, C., Bouillet, P., Regnier, C. H., Tomasetto, C., Mattei, M. G., Chenard, M. P., Anglard, P., Rio, M. C., Basset, P.(1995) Comparative expression of the psoriasin (S100A7) and S100C genes in breast carcinoma and co-localization to human chromosome 1q21-q22. *Int. J. Cancer* **63**, 297-303

59. Tanaka, M. Adzuma, K., Iwami, M., Yoshimoto, K., Monden, Y., Itakura, M. (1995) Human calgizzarin: one colorectal cancer-related gene selected by a large scale random cDNA sequencing and northern blot analysis. Cancer Lett. **89**, 195-200

60. Mannello, F., and Gazzanelli, G. (2001) Prostate-specific antigen (PSA/hk3): a further player in the field of breast cancer diagnostics? *Breast Cancer Res*. **3**, 238-243

61. Akiyama, T., Sudo, C., Ogawara, H., Toyoshima, K., Yamamoto, T. (1986) The product of the human c-erbB-2 gene: a 185-kDa glycoprotein with tyrosine kinase activity. *Science.* **232**, 1644-1646

62. Yamashita, H., Nishio, M., Toyama, T., Sugiura, H., Zhang, Z., Kobayashi, S., and Iwase, H. (2004) Coexistence of HER2 over-expression and p53 protein accumulation is a strong prognostic molecular marker in breast cancer. *Breast Cancer Res.* **6(1)**, 24-30

63. Xiang, R., Shi, Y., Dillon, D. A., Negin, B., Horvath, C., and Wilkins, J. A. (2004) 2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J. Proteome Res.* **3**, 1278-1283

64. Giometti, C. S., Williams, K., Tollaksen, S. L. (1997) A two-dimensional electrophoresis database of human breast epithelial cell proteins *Electrophoresis.* **18**, 573-581

65. Kageyama, S., Isono, T., Iwaki, H., Wakabayashi, Y., Okada, Y., Kontani, K., Yoshimura, K., Terai, A., Arai, Y., Yoshiki, T. (2004) Identification by Proteomic Analysis of Calreticulin as a Marker for Bladder Cancer and Evaluation of the Diagnostic Accuracy of Its Detection in Urine. *Clin. Chem.* **50(5)**, 857

66. Williams, K., Chubb, C., Huberman, E., Giometti, C. S.(1998) Analysis of differential protein expression in normal and neoplastic human breast epithelial cell lines. *Electrophoresis.* **19**, 333-343

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

68. Savelleno, D. H., Boss, E., Blondet, C., Sato, F., Abe, T., Josephson, L., Weissleder, R., Gaudet, J., Sgroi, D., Peters, P. J., and Basillion, J. P. (2003) The transferrin receptor: a potential molecular imaging marker for human cancer. *Neoplasia.* **5(6)**, 495-506

69. Kosaka, T., Yatabe, Y., Endoh, H., Kuwano, H., Takahashi, T., and Mitsudomi, T. (2004) Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications. *Cancer Res.* **64**, 8919-8923

70. Byrne, J. A., Balleine, R. L., Fejzo, M. S., Mercieca, J., *et al.,* (2005) Tumor protein D52 (TPD52) is overexpressed and a gene amplification target in ovarian cancer. *Int. J. Cancer.* **117**, 1049-1054

71. Tuynder, M., Fiucci, G., Prieur, S., Lespagnol, A., *et al.,* (2004) Translationally controlled tumor protein is a target of tumor reversion. *Proc. Natl. Acad. Sci. USA.* **101(43)**, 15364–15369

72. Arcuri, F., Papa, S., Carducci, A., Romagnoli, R., Liberatori, S., *et al.,* (2004) Translationally controlled tumor protein (TCTP) in the human prostate and prostate cancer cells: expression, distribution, and calcium binding activity. *The Prostate.* **60**, 130-140

73. Vastag, B. (2000) Some promising biomarkers for cancer. *J. Natl. Cancer Inst.* 92(10), 788

74. Gronborg, M., Kristiansen, T. Z., Iwahori, A., Chang, R., Reddy, R., Sato, N., Jensen, O. N., Hruban, R. H., Goggins, M. G., Maitra, A., Pandey, A. (2006) Biomarker discovery from pancreatic cancer secretome using a differential proteomics approach. *Mol. Cell. Prot*. **5**, 151

75. Chen, G., Zhang, W., Cao, X., Li, F., Liu, X., Yao, L. (2005) *Leukamia Res.* **29**, 503

76. Yong, L., Li, C., Shu-you, P., Zhou-xun, C., Vu, C. H.(2005) Role of CD97stalk and CD55 as molecular markers for prognosis and therapy of gastric carcinoma patients. *J. Zhejiang Univ. SCI.* **6B(9)**, 913-918

Chapter 4: Microfluidic Devices

4.1 Introduction

Microfluidics refers to a set of technologies that control the flow of minute amounts of sample in a miniaturized system. These microfabricated architectures integrate an array of functional elements that include separation channels, microreactors, mixers, sample valving components, fluid propulsion elements, and MS interfaces. They offer several advantages over conventionally sized systems such as compact size, high speed analysis, increased functionality, high throughput, reduced costs, as well as integration and multiplexing capabilities. The ability to perform precise and accurate sample handling operations enables process control, automation, and the generation of reliable and high quality data. Moreover, this technique allows the implementation of operational principles that are not feasible in the macro-scale setting. The miniature format allows the fabrication of contamination-free, disposable devices with wide applications in biomedical and biotechnology fields. These microfabricated structures represent promising analytical platforms for proteomic investigations, and have the potential to become future point-of- care devices.

The recent past has witnessed significant progress in the field of microfluidics and its integration with mass spectrometry detection [1, 2, 3]. Although microfluidic devices were coupled with various optical and electrochemical detectors, for instance laser-induced fluorescence (LIF) [1, 4], the interfacing of microchips to mass spectrometry provides fast and reliable detection, with no need for sample derivatization. A wide range

of separation techniques integrated on a microchip, such as capillary electrophoresis, capillary electrochromatography (CEC), and micro-liquid chromatography [5-10], have been demonstrated.

A microfluidic device that integrates an LC system that was used for the analysis of an MCF7 extract is reported in this part of the work. The microchip contained all the functional elements necessary for stand-alone operation of the liquid chromatography system. Two experiments were conducted: one to demonstrate the microfluidic-LC platform for the analysis of one of the SCX fractions of the MCF7 cell line, and the other to demonstrate the applicability of the microfluidic chip for the detection of phosphopeptides from an a-casein digest, before and after dephosphorylation with alkaline phosphatase. The ultimate goal of these efforts relates to the development of microfluidic chips for high-throughput proteomic research and biomarker discovery and screening.

4.2 Microfabrication techniques

Microfabrication refers to a process that involves a set of techniques and equipment commonly used to manufacture integrated circuits and microelectromechanical systems (MEMS). Microchips are fabricated using a range of materials such as glass, silicon, and polymeric substrates. In earlier days, the silicon substrate was very popular due to high stiffness and heat conductivity, but it has limited optical, electrical and chemical properties. Today, polymeric materials have acquired popularity due to their low manufacturing costs, as well as low-temperature sealing capability. Glass substrates are used most commonly due to their good optical properties, well-known surface characteristics, and well-developed fabrication procedures adapted

from the microelectronics industry. The most important factors that are considered for choosing an appropriate material relate to surface chemistry, ease of fabrication, price, and disposability. However, there are some other aspects of the material that need to be considered while making a suitable selection. Some practical aspects related to surface reactivity, adsorption and electroosmosis, must be considered. To prevent the formation of analyte adducts from the chip itself, in the ESI process, the selected material must have sufficient chemical stability. Sample adsorption on the surface of the chip can result in loss of analytes and overall sensitivity of detection; however, the total surface that comes in contact with the sample can be minimized by integrating some of the functional elements on the chip [11]. Moreover, the techniques developed for CE to reduce sample adsorption (the use of low-pH buffers and charged/neutral hydrophilic surface coatings), can help reduce adsorption and electroosmosis in microchips [12], as well.

Basic fabrication procedures currently used in the manufacturing of microfluidic devices from glass, quartz, or silicon include: (1) deposition of thin films using various chemical or physical techniques; (2) photolithography, to transfer the desired pattern onto the substrate; (3) etching with different chemicals in the liquid or gas phase to generate the microfluidic channels; and (4) sealing of the substrate to a cover plate to enclose the microfluidic network of channels [13, 14].

In our research, microfluidic devices were fabricated from glass using previously described photolithography and wet chemical etching protocol [15, 16]. The design of the photomask was prepared using the AutoCAD software. Microchips were prepared from soda lime glass slides sputtered with chrome and positive photoresist (Nanofilm, West lake Village, CA). The substrate was exposed through the photomask to UV radiation

(360 nm) for microchannel pattern imprinting. Chemical development of the exposed chips was performed using MF-319 developer. Next, the chrome was removed using a chrome mask etchant. Sample handling and micropump channels were etched in the substrate and cover plate to a depth of 50 μm and 1.5-2 μm, respectively, using buffer oxide etchant (BOE) solution. The etch depth was measured using a Dektak 6M stylus Profilometer (Veeco, Tucson, AZ). To access the pump and channels, holes with 0.8-1 mm diameter were drilled in the chip. The substrate and coverplate were cleaned with acetone and methanol to strip the photoresist. Prior to bonding, the chips were soaked in detergent and activated with a solution of NH_4OH, H_2O_2 and H_2O. Finally, the cover plate was thermally bonded to the substrate by gradually raising the temperature to 550 ºC. Glass reservoirs were glued to the chip using epoxy glue (Epotek, Epoxy Technology, Billerica, MA).

Reagents: MF-319 developer was purchased from Microchem (Newton, MA). Buffer oxide etchant and chrome etchant were obtained from Transene Co. (Danvers, MA). Hydrogen peroxide and ammonium hydroxide were obtained from Mallinckrodt Baker Inc. (Philipsburg, NJ).

4.3 MCF7 analysis and biomarker detection on a chip

A microchip liquid chromatography system (0.5” x 2.5”) that integrates a multichannel electroosmotic flow (EOF) pumping technique [17], and combines a separation channel, micropump, valve, mixer, and ESI interface on a single unit to perform pressure driven separations, is reported in this part of the research. Under identical conditions, the performance of this device is similar to the benchtop LC system.

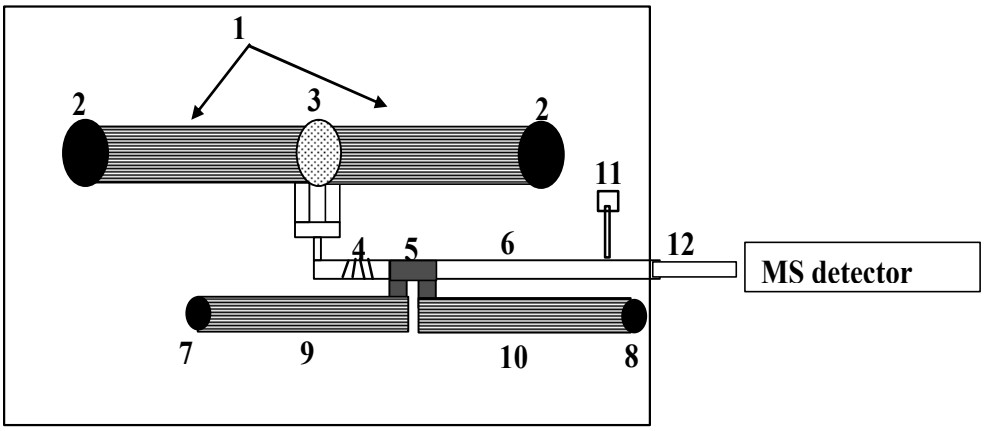


Figure 1. Schematic representation of the microfluidic LC system.

4.3.1 Experimental section

The microchip integrated LC system (**Figure 1)** comprises 2 EOF pumps, a valving component, a separation channel with an on-column preconcentrator, and an ESI interface. The separation channel (6) was 2 cm long with a depth of ~50 μm. Reversed phase packing material, Zorbax SB-C18, d_p=5 μm (Agilent Technologies) was loaded manually in the channel from the LC waste reservoir (11) with the aid of a 250 μL syringe. The packing material was retained in the separation channel or the preconcentrator with the aid of short (~100 μm), multiple channel structures, similar to the pump or to commonly used filter elements. The two EOF pumps (1) consisted each of 200 nanochannels (2 cm long, ~1.5 μm deep), and had different inlet reservoirs (2) and a

common outlet reservoir (3). The voltage for EOF generation in the pumps was applied to reservoirs (2) and (3). The voltage applied to reservoir (3) represents also the voltage for electrospray generation. EOF leakage in the outlet reservoir (3) was prevented by a porous glass disc (5 mm diameter, 0.8-1 mm width, 40-50 Å pore size) purchased from Chand Associates (Worcester, MA). The disc was secured to the bottom of reservoir (3) and enabled only the exchange of ions but not of bulk flow. Sample loading was accomplished through a double-T injector (5) with the aid of a multichannel EOF valving structure (9, 10) consisting of 100 nanochannels on each arm (2 cm long, ~1.5 μm deep). A fused silica capillary (10 mm long, 20 μm i.d. x 90 μm o.d.) from Polymicro Technologies (Phoenix, AZ) was inserted into the LC channel for ESI generation (12).

Mass spectra were acquired with an LTQ ion trap mass spectrometer (Thermo Electron Corp., San Jose, CA). Data dependent MS acquisition conditions and database search parameters were described in chapter 2. One of the SCX fractions (#7) was analyzed with the microchip integrated LC system.

4.3.2 Results and discussion

The choice for an EOF pumping system to run the microfluidic LC was dictated by three reasons: first, the EOF pumps are the only miniaturized pumps that can generate high pressures (hundreds/thousands of bars) [18], second, the manufacturing of the pumps is extremely simple and reliable, and third, the same structure can be effectively utilized for sample loading and valving. If a potential differential is applied between reservoirs (2) and (3), EOF will be generated through the connecting microchannels; if the hydraulic resistance of these pumping channels is sufficiently high, eluent will be pumped from reservoir (2) into the microfluidic network of channels on the chip, even if

the pressure in the chip is high, i.e., 10 bar. The large hydraulic resistance of the pumping microchannels will impede flow leakage back into the reservoir (2). Typical configurations in our designs include microchannels that are ~1-2 μm deep and 5-20 mm long, which are capable of delivering flow rates in the 10-400 nL/min range. A valving structure comprised of similar narrow microchannels, as the ones used for pumping, can be used for injecting and processing the sample in a pressurized environment. As the multiple open channel configuration has a much larger hydraulic resistance than any of the other functional elements on the chip, it can basically act as a valve that is open to material transport through an electrically driven mechanism, but is closed to material transport through a pressure driven mechanism. The same multichannel structure can be used as an EOF pump for eluents, and as an EOF valve for sample introduction into a pressurized microfluidic system.

Scanning electron microscope (SEM) images of cross-sections through the pumping channels are shown in **Figure 2A** and **B**. The pumping/valving channels were placed 25 μm apart and were etched to a depth of ~1.5 μm. SEM images of cross-sections through an empty and packed channel are shown in **Figure 3**. Efficient packing of the LC channel with a slurry of particles can be easily accomplished within a few minutes, and once packed, the side channel used for filling can be closed with an appropriate fitting or plugged. The side channel can, however, be used later for fast eluent rinsing of the LC channels. The stability of the packing within the channel was somewhat better for the RPC18-5 μm particles than for the 10 μm Poros ones. The 10 μm particles were too easily dislodged from their place. With the present design, flow rates through the LC channel packed with 5 μm particles were in the 50-80 nL/min range.

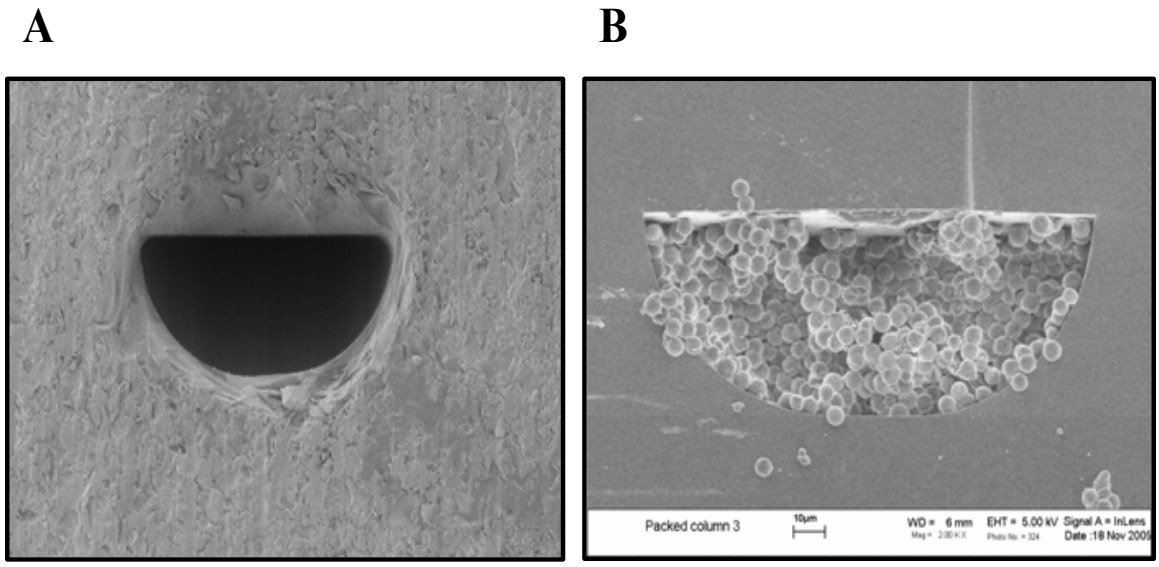


Figure 2. Packed microfluidic LC channel. (**A**) SEM image through an empty microfluidic LC channel; (**B**) SEM image of a cross-section through a packed microfluidic LC channel filled with 5 µm particles.

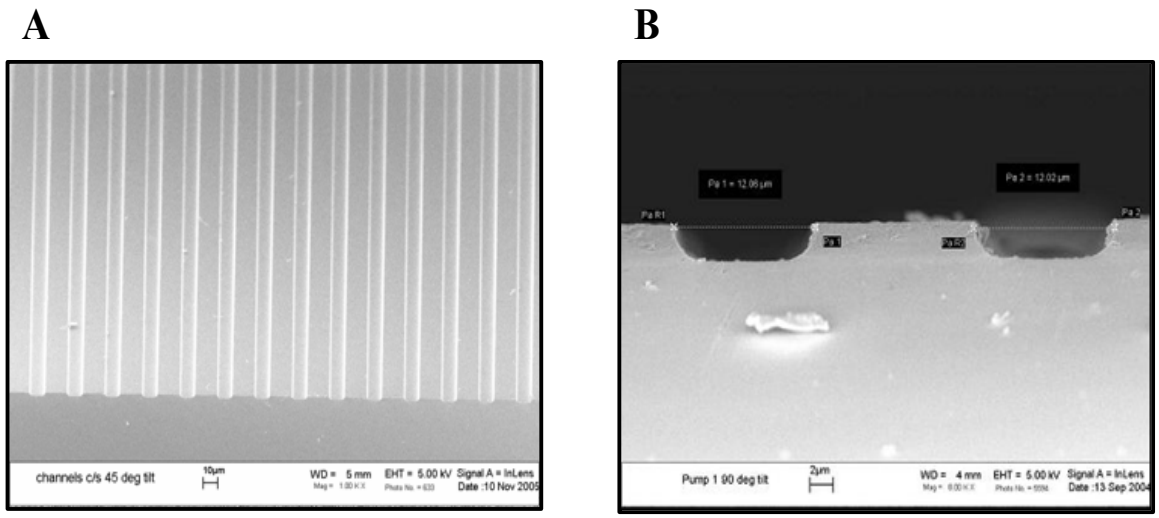


Figure 3. SEM images of pumping/valving channels. (**A**) Top view; (**B**) Cross section.

The sequence of operations necessary to operate the microfluidic LC system is provided in the followings. The microfluidic chip is filled with a low organic content eluent. The sample inlet reservoir (7) is filled with the sample. When a potential differential is applied between the sample inlet (7) and outlet/waste (8) reservoirs, the sample will be loaded through the EOF valve inlet microchannels (9), will be focused at the head of the separation channel, and the depleted sample eluent will be discarded through the EOF valve outlet microchannels (10). While the sample is loaded, there is a very small voltage applied to the EOF pumps to eliminate sample diffusion in the direction of the pumps during the loading process. Once the sample is loaded on the separation channel (6) (see sample plug 5), the voltage on the sample reservoirs is removed. Simultaneously, a potential differential is applied between reservoirs (2) and (3) in order to activate the pump. Due to the fact that EOF is generated in the pumping channels, but backflow through all the pumping and valving microchannels is minimal due to their large hydraulic resistance, most of the flow is directed towards the separation channel. By increasing the potential differential on one of the pumps relative to the other, an eluent gradient can be generated to favor the elution of highly retained components at the head of the separation channel. The voltage necessary for ESI generation is established through the voltage applied to the exit of the pump in reservoirs (3).

4.3.2.1 MCF7 analysis on a chip

Sample loading on the chip was evaluated initially by infusing a 20 μM solution of fluorescent Rhodamine 610 in a solution of CH_3OH/H_2O (5:95 v/v) containing NH_4HCO_3 (15 mM) through the EOF valve. The LC separation column had an enlarged area at the loading point to enable the capture and preconcentration of a large amount of

sample. The dimensions of this on-column preconcentrator were ~ 400 μm x 400 μm. The Rhodamine gradual removal from the preconcentrator was dependent on the composition and flow rate of the eluent. High organic content eluents (>80 % CH_3OH) were able to remove Rhodamine almost instantly.

The SCX fraction that was loaded on the LC chip was used as eluted from the SCX column, without further desalting. The fraction contained a relatively large amount of NaCl (~50-70 mM), and the infusion of this buffer system at 500 V/cm resulted occasionally in the generation of gas bubbles within the EOF valve. While this is an undesired scenario, these bubbles were eventually eliminated once the EOF pumps started pumping. Sample clean-up with a proper desalting cartridge would have been beneficial and would have prevented such an outcome.

A base peak and 2D-chromatogram of the microfluidic LC separation of the MCF7 cellular extract is given in **Figure 4**. Sample volumes loaded on the chip were estimated to be around 1 μL. The efficiency of the separation was in the 45,000-180,000/channel and was dependent on the nature of the peptides. Peak widths at half height were 15-30 s, allowing for a triple play data dependent MS analysis. Peak capacity was estimated to be around 80-100. The micropump that operated this LC system comprised a total 400 pumping channels that delivered eluent flow rate at approximately 60-70 nL/min.

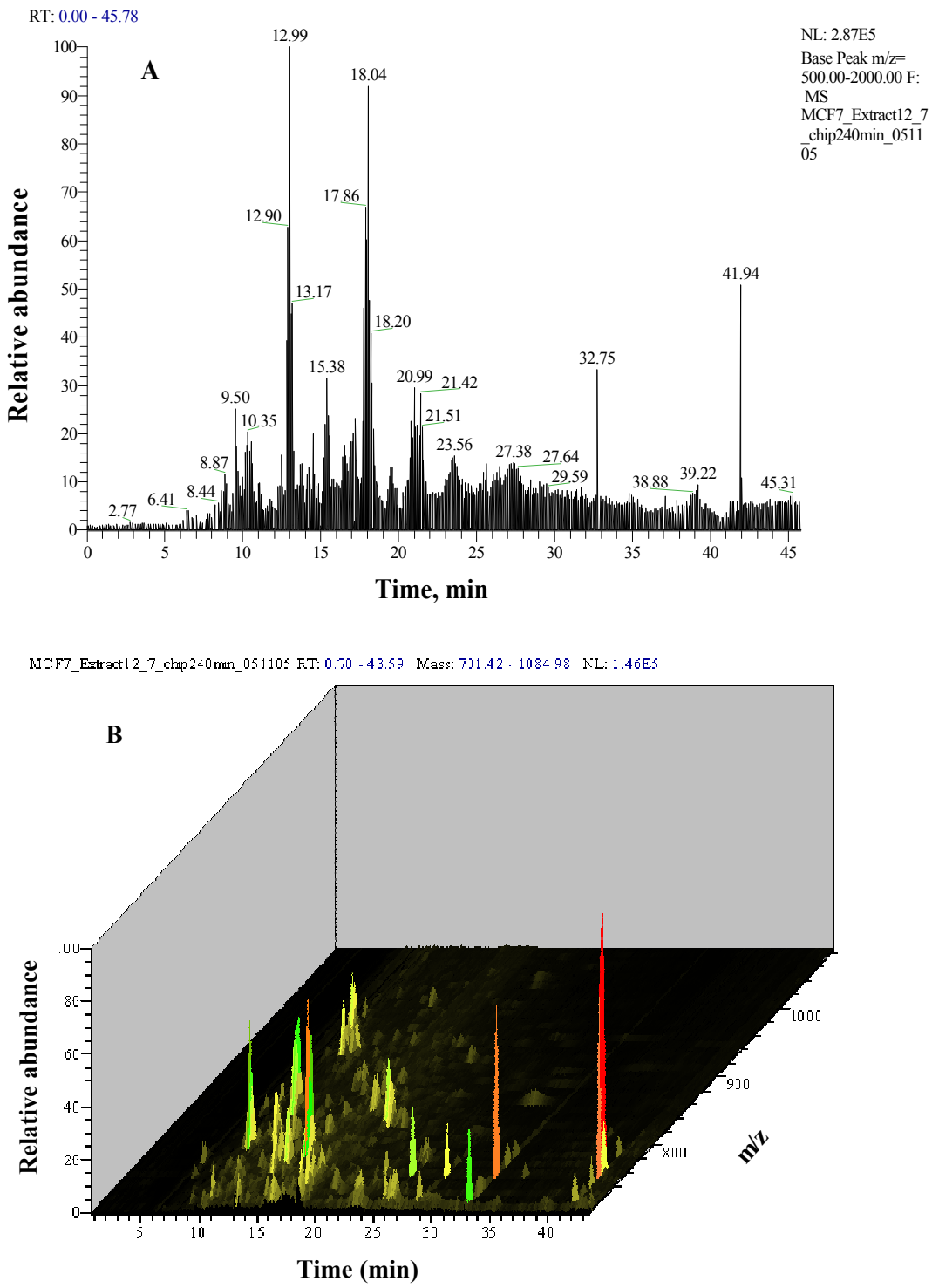


Figure 4. Data dependent microfluidic LC-MS/MS analysis of the MCF7 breast cancer cell line (SCX fraction eluted with~50-70 mM NaCl). (**A**) Base peak chromatogram; (**B**) 2D-view of a relevant m/z region.

The LC separation was performed using isocratic conditions with no gradient provided for the elution of the analytes. The LC eluent was NH_4HCO_3 (15 mM) in H_2O/CH_3OH (40:60), pH~8. The high organic eluent resulted in good peak shapes for the eluted peptides, and surprisingly, a relative uniform distribution of the peptides along the separation time length. This eluent, was not appropriate though to efficiently elute the analytes from Poros packing material. The high pH eluent ensured high EOF in the pumping system while still enabling efficient electrospray ionization in positive ion mode. This eluent is frequently used in our lab as a mobile phase for LC separations and was demonstrated earlier to provide 80-90 % sequence coverage for standard protein digests that were electrosprayed from the chip [19].

Using this microfluidic arrangement, 77 proteins were identified in the SCX fraction using charge dependent cross correlation scores of 1.9, 2.2, and 3.75 as minimum acceptance criteria. Of these, 68 proteins were identified with p-values $p<0.1$, and 39 proteins with $p<0.001$ (**Table 1**). The p-value represents the probability of a random match, as it is calculated by the Sequest software. The total number of proteins identified from the same fraction using micro-HPLC (100 μm x 12 cm columns filled with 5 μm Zorbax SB-C18 reversed phase packing material that were operated at about 170 nL/min with typical eluents containing H_2O/CH_3CN acidified with 0.01 % TFA) was 935 (754 with $p<0.1$ and 573 with $p<0.001$). What concerns the proteins identified with high confidence ($p<0.001$), a ~10 fold drop in the number of identified proteins is observed when switching from the bench-top HPLC to the microfluidic platform. However, repeating the analysis with the bench-top system, and using conditions similar to the chip

(2 cm separation column, basic buffer system, ~1 μL sample injection volumes), the total number of identified proteins was very similar to the results obtained from the chip (see **Table 1**): 91 protein matches (76 with $p<0.1$ and 48 with $p<0.001$). Moreover, there was ~75 % overlap between the proteins identified by 2 unique peptides. A detailed study was conducted to identify the reasons for the drop in the number of proteins identified with the microfluidic platform. Another MCF7 SCX extract was analyzed using various conditions (**Table 2**). The major factor that affected the number of identified proteins, in going from typical HPLC analysis conditions to experimental conditions that mimicked the microfluidic environment, was the sample amount (volume) subjected to analysis. Changing the column length or pH conditions had much smaller effect than decreasing the sample injection volume. Decreasing the volume from 16 to 4 and then to 1 μL, the number of identified proteins with $p<0.001$ decreased from 444 to about 160-180, and to 16, respectively. While going from the acidic buffer conditions to the basic buffer, somehow reduced the number of proteins identified, but the basic buffer condition was ideal for the glass microchip platform. Basic buffer conditions for analysis were chosen to ensure the generation of high EOF in the microfluidic pumping system. Other experiments using the conventional HPLC platform were performed to optimize conditions for the MCF7 analysis and have confirmed this outcome. The analysis of the entire batch of 16 SCX fractions yielded 2,329 protein matches for 8 μL injections, and 4,534 protein matches for 40 μL injections. Data filtering parameters for all these proteins were the same. The overall dimensions of the microchip integrated LC system were 0.5" x 2.5," enabling the integration of 2 LC systems on a 1" x 3" chip, or of 6 LC systems on a 3" x 3" chip substrate.

Table 1 Total number of proteins identified with the microfluidic LC and the bench-top HPLC using columns of different lengths.

Platform	Eluent additive	Injection volume (µL)	Protein matches (total)	Protein matches (p<0.1)	Protein matches (p<0.001)
ChipLC (2 cm)	NH_4HCO_3 (15 mM, pH~8)	~ 1	77	68	39
HPLC (2 cm)	NH_4HCO_3 (15 mM, pH~8)	1	91	76	48
HPLC (12 cm)	TFA (0.01 %)	16	935	754	573

Table 2 Effect of the injection volume and eluent pH on the number of proteins identified in a SCX fraction of the MCF7 protein digest.

LC column length (cm)	Eluent additive	Injection volume (µL)	Protein matches (total)	Protein matches (p<0.1)	Protein matches (p<0.001)
12	TFA (0.01 %)	16	965	591	444
10	TFA (0.01 %)	4	307	228	178
10	NH_4HCO_3 (15 mM, pH~8)	4	164	143	100
2.5	NH_4HCO_3 (15 mM, pH~8)	4	286	229	159
2.5	NH_4HCO_3 (15 mM, pH~8)	1	31	26	16

4.3.2.2 Biomarker detection on chip

The list of identified proteins on the chip was searched for known biomarkers that were identified in the same fraction using conventional HPLC. Five protein matches were found: PCNA, cathepsin D and cytokeratins 8, 18, 19. PCNA was identified from the chip by 1 peptide with p=6.6E-08. The bench-top HPLC-MS experiment yielded 17 unique peptide matches for this protein. MS^2 spectra for the common peptide are shown in **Figure 5**, as acquired from the chip platform and the bench-top HPLC. Major peaks are common for both these spectra. Complete y-ion series and characteristic b-ions are observed. Cathepsin D was identified from the microchip by a peptide with p<2.23E-05. MS^2 spectra for this peptide are given in **Figure 6**. A complete y-ion series is observable in both spectra.

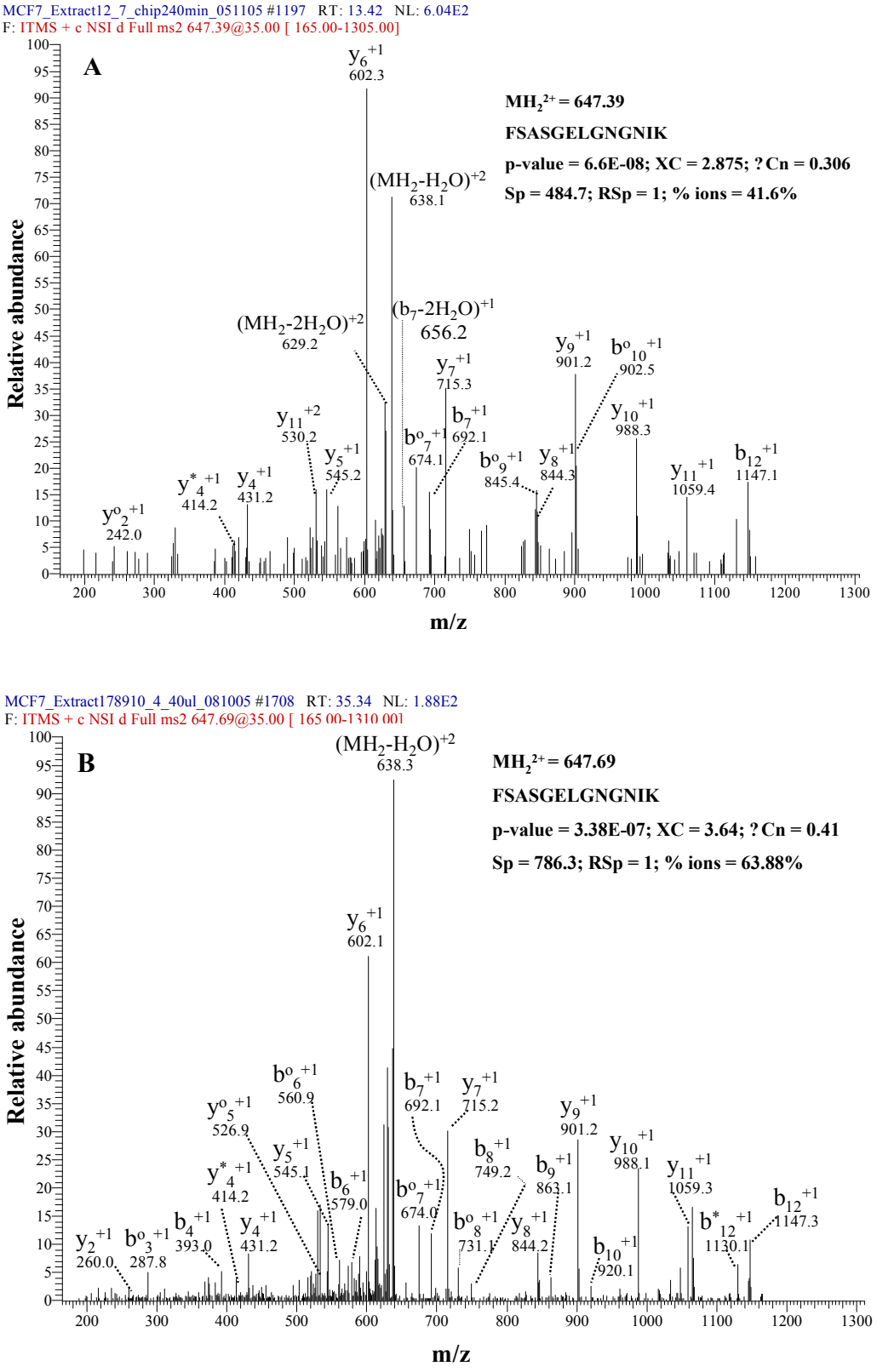


Figure 5. Tandem mass spectra of a "PCNA" peptide generated from: (**A**) microfluidic LC-MS platform, and (**B**) bench-top HPLC-MS system.

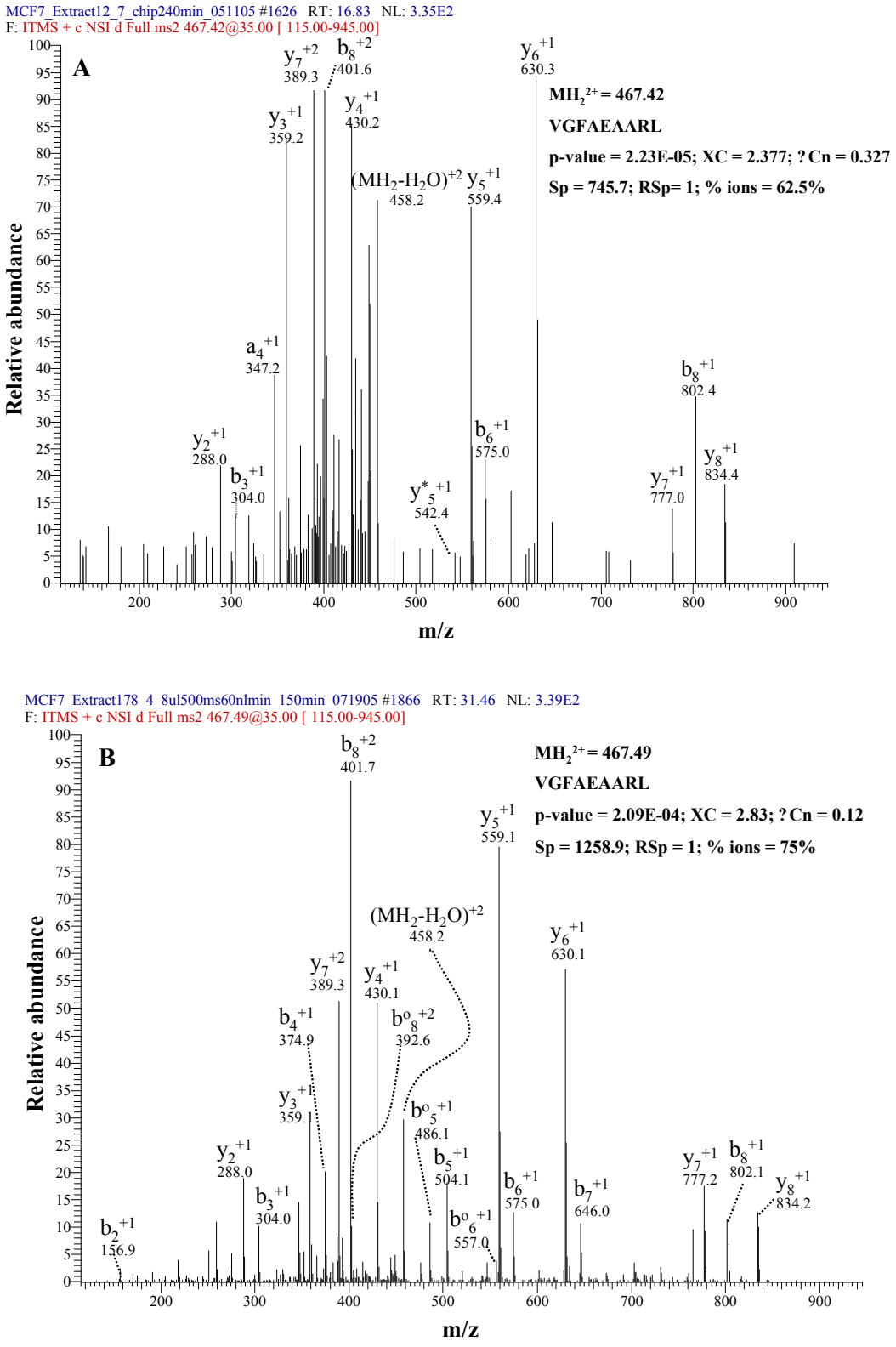


Figure 6. Tandem mass spectra of a "cathepsin D" peptide generated from: (**A**) microfluidic LC-MS platform, and (**B**) bench-top HPLC-MS system.

4.4 Analysis of protein phosphorylation on a chip

Phosphorylation is one of the most common and important posttranslational modifications. Protein phosphorylation is involved in a number of regulatory mechanisms such as cell division, cell growth, cell differentiation, and metabolism. Reversible phosphorylation plays a pivotal role in signal transduction events that involve transmission and amplification of signals from the transmembrane receptors to the nucleus. In eukaryotic cells, approximately one third of all the proteins are phosphorylated at any given time, and over 100,000 potential phosphorylation sites are present in the human proteome [20, 21]. Eukaryotes exhibit phosphorylation on serine, threonine, and tyrosine; however, phosphorylation on serine and threonine residues is more often observed as compared to tyrosine [20]. Approximately 2-5% of the human genome encodes for kinases (~500) which phosphorylate proteins, and phosphatases (~100) that remove a phosphate attached to an aminoacid residue [20].

The detection of phosphoproteins is not an easy task, and continues to be a challenge for several reasons. First, proteins involved in signaling are present in low copy numbers, and hence enrichment becomes a necessary step before analysis. Second, only a small fraction of a protein is phosphorylated at any given time, individual sites being only partially phosphorylated. Third, phosphoproteins can exist in several different phosphorylated forms, and the phosphorylated sites may vary. Fourth, dephosphorylation of phosphoproteins can be caused by phosphatases, if appropriate care is not exercised. Fifth, the dynamic range of most of the analytical techniques used to study phosphorylation is limited. Finally, antibodies work well for phosphoproteins but not for phosphopeptides [20].

There are many techniques available for the study of phosphorylation, such as ^{32}P radioactive labeling, western blotting with phospho-specific antibodies, Edman sequencing, and MS based approaches [20, 22]. Of these, ^{32}P radiolabeling is most sensitive, but is only applicable to cells in culture [22] and is very labor intensive, with difficulty in obtaining full protein coverage. On the other hand, Edman sequencing is less sensitive, and requires a purified protein and adequate amount of sample for successful microsequencing [20]. Antibody-based methods have limitations related to antibody specificity [22]. Mass spectrometry has long been used for the identification and characterization of modifications associated with an increase or decrease in mass. It is a highly sensitive method which can provide definite localization of modified sites [20]. However, there are difficulties associated with the identification of phosphopeptides with MS, as well. First, ESI for peptide analysis works best in positive ion mode, and this makes the detection of negatively charged phosphopeptides difficult. Second, phospho-serine and phospho-threonine are labile. Third, phosphopeptides generate low intensity peaks in the presence of their non-phosphorylated counterparts. In addition, the presence of isobaric peptides complicates the analysis. Enrichment strategies can be used to improve some of the conditions for low abundant phosphopeptide detection. The enrichment procedures include phosphopeptide recovery by chromatographic methods that use oligo R3 resin, porous graphitic carbon, and metal affinity columns. Also, chemical modification methods that employ ß-elimination (phopho-serine & phospho-threonine) in strongly basic solution, followed by modification with ethenedithiol [23], can be used, as well. Nevertheless, these methods require several chemical alterations and purification steps, and consequently large amounts of sample. Posttranslational

modification analysis of cancer cell line proteomes has been demonstrated [24, 25]. Vasilescu *et al.* has shown the analysis of ubiquitinated proteins by affinity purification followed by LC-MS/MS. 70 ubiquitinated proteins were identified in the MCF7 breast cancer cell line [24]. Phosphoproteome analysis of human colon adenocarcinoma (HT-29) cells, using immobilized metal affinity chromatography (IMAC) followed by LC-MS/MS, resulted in the identification of 213 phosphorylation sites from 116 proteins [25].

Given the importance of protein phosphorylation, the purpose of this work was to evaluate the applicability of the microfluidic LC system for the fast analysis of phosphorylated peptides. As a result of the fact that multiply phosphorylated peptides do not produce an ESI-MS signal in positive ion mode, a strategy was developed that enabled the identification of singly phosphorylated peptides by analyzing a simple protein digest, and the identification of multi-phosphorylated peptides, after treatment with alkaline phosphatase. Treatment with the enzyme resulted in the removal of the phosphate groups from the peptide, consequently rendering them detectable with ESI-MS.

4.4.1 Experimental section

4.4.1.1 Preparation of enzymatic digests

5 mg of a-casein was dissolved in 20 mM ammonium bicarbonate (pH 8.1) resulting into a 50 µM solution of a-casein. Trypsin (20 µg) was then added to 1 mL of a-casein solution in a ratio of substrate:trypsin of 62:1 (w/w). Digestion was performed at

37°C overnight and stopped by the addition of 10 µL acetic acid glacial. The digest was stored at -20°C.

4.4.1.2 Alkaline phosphatase treatment

1 mg of alkaline phosphatase (2,200 units) of calf intestine (Calzyme, San Louis Obispo, CA) was dissolved in 50 mM ammonium bicarbonate, resulting into an enzyme activity of 22 units/µL solution. 1 mL solution of a-casein (2.5 µM) and alkaline phosphatase (0.5 units/ µL) was prepared in 50 mM ammonium bicarbonate. The enzyme and protein were mixed only at the time of analysis for immediate detection of phosphatase activity.

4.4.1.3 Mass spectrometric analysis of phosphorylated peptides

2.5 µM of a-casein digest in $H_2O/CH_3OH/HCOOH$ (78:20:2) was directly infused into the electrospray ionization-ion trap mass spectrometer. The spray voltage was 2.2kV and the capillary temperature was 200°C. The infusion was established using an external syringe pump, at 0.2 µL/min. The top 10 most intense peaks were chosen for fragmentation from the data dependent MS acquisition scans (5 microscans averaged). 1 MS scan was followed by 1 zoom scan and 1 MS^2 on each of these ions. The rest of the conditions for data acquisition were similar to those described in chapter 2. Using similar conditions, the a-casein digest treated with alkaline phosphatase, was also infused to check for the dephosphorylated peptides.

8 µL of a-casein digest (0.25 µM), with and without alkaline phosphatase treatment, were also analyzed using RPLC interfaced to MS. The experimental setup,

RPLC column specifications, solvent composition, and RPLC gradient for peptide elution, were the same as described in chapter 2.

4.4.1.4 Microfluidic chip for the analysis of phosphorylated peptides

The design of the microfluidic chip used for the identification of phosphopeptides was the same as described in the previous section (**Figure 1),** except that for this particular experiment, the sample inlet/outlet channels and reservoirs (7), (8), and (11), were not used. Instead, the sample and the buffer solutions were introduced in the pump reservoirs (2) and (3), and the voltage for EOF generation was applied to reservoir 3. The pumping system was initially filled with 10 mM ammonium bicarbonate buffer in H_2O/CH_3OH (75:25). The flows were visualized for optimization purposes with a Nikon epi-fluorescent microscope. Then, the buffers in the two reservoirs (2) were replaced by a digest of 5 μM a-casein and an acidic buffer solution, $CH_3OH/H_2O/CH_3COOH$ (20:80:1). The chip was placed in front of the mass spectrometer, and the two solutions from reservoirs (2) were infused simultaneously and mixed in a serpentine mixer on the chip, to acquire MS and MS^2 data for the identification of phosphorylated peptides. For dephosphorylation studies, the acidic buffer from reservoir (2) was exchanged with alkaline phosphatase solution (2.2 units/mL). The alkaline phosphatase and the a-casein solutions were next infused through the chip, to identify the new peptides that were generated after dephosphorylation.

4.4.2 Results and discussion

The results from all three experiments conducted from direct infusion, benchtop LC-MS, and microchip-MS, with a-casein digest before and after alkaline phosphatase treatment, were analyzed and compared. However, more emphasis was given on the data acquisition from the microfluidic chip-MS experiment. Database searching was performed with the Turbo Sequest software against the bovine database. The database search parameters were the same as described in chapter 2, with the addition of dynamic modifications at serine, threonine, and tyrosine, for confirmation of the sites of phosphorylation. Bovine a-casein has been extensively used as a model phosphoprotein for MS analysis because it has a large number (10) of phosphorylated residues. The tryptic fragments of a-casein along with their mass, sequence, and phosphorylation information (derived from the Expasy webpage) are summarized in **Table 3**.

A total ion chromatogram collected from the chip, before and after alkaline phosphatase treatment is shown in **Figure 7**. It is worth noting that the intensity of the peaks drops after 22 min. This is due to the addition of alkaline phosphatase in basic solution, and the beginning of the dephosphorylation reaction. In basic solution, the overall intensity of peptide ions is smaller than in acidic solutions, when positive ESI-MS is used for detection. To note also that this was an infusion experiment with data dependent acquisition. The peaks in this TIC are not separated peptides, but unique MS scan events.

Table 3 Theoretical tryptic fragments of a-casein with their mass, position, peptide sequence, and phosphorylation information (generated from the SWISSPROT database).

Tryptic fragment	Position	Mass	$(MH)^{+2}$	$(MH)^{+3}$	#MC	Modifications	Mass	$(MH)^{+2}$	$(MH)^{+3}$	Peptide sequence
T1	16-18	400.2667	200.6334	134.0889	0					RPK
T1+T2	16-22	875.5573	438.2787	292.5191	1					RPKHPIK
T2	19-22	494.3085	247.6543	165.43617	0					HPIK
T2+T3	19-37	2235.2356	1118.118	745.7452	1					HPIKHQGLPQEVLNENLLR
T3	23-37	1759.9449	880.4725	587.31497	0					HQGLPQEVLNENLLR
T3+T4	23-49	3125.657	1563.329	1042.5523	1					HQGLPQEVLNENLLRFFVAPFPEVFGK
T4	38-49	1384.7299	692.865	462.2433	0					FFVAPFPEVFGK
T4+T5	38-51	1641.8675	821.4338	547.95583	1					FFVAPFPEVFGKEK
T5	50-51	276.1554	138.5777	92.718467	0					EK
T5+T6	50-57	946.5204	473.7602	316.17347	1	PHOS: 56	1026.4867	513.74335	342.8289	EKVNELSK
T6	52-57	689.3828	345.1914	230.46093	0	PHOS: 56	769.3491	385.17455	257.11637	VNELSK
T6+T7	52-73	2438.1239	1219.562	813.37463	1	PHOS: 56, 61,63,68	2757.9891	1379.4946	919.99637	VNELSKDIGSESTEDQAMED IK
T7	58-73	1767.7589	884.3795	589.91963	0	PHOS: 61, 63,68	2007.6578	1004.3289	669.88593	DIGSESTEDQAMEDIK
T7+T8	58-94	4069.8223	2035.411	1357.2741	1	PHOS: 61, 63,68,79,81,82 ,83,90	4709.5527	2355.2764	1570.5176	DIGSESTEDQAMEDIKQMEA ESISSSEEIVPNSVEQK
T8	74-94	2321.0813	1161.041	774.36043	0	PHOS: 79, 81,82,83,90	2720.9128	1360.9564	907.6376	QMEAESISSSEEIVPNSVEQ K
T8+T9	74-98	2827.3778	1414.189	943.12593	1	PHOS: 79, 81,82,83,90	3227.2093	1614.1047	1076.4031	QMEAESISSSEEIVPNSVEQ KHIQK
T9	95-98	525.3143	263.1572	175.77143	0					HIQK

T9+T10	95-105	1337.6808	669.3404	446.56027	1					HIQKEDVPSER
T10	99-105	831.3843	416.1922	277.79477	0					EDVPSER
T10+T11	99-115	2080.0709	1040.535	694.02363	1					EDVPSERYLGYLEQLLR
T11	106-115	1267.7045	634.3523	423.23483	0					YLGYLEQLLR
T11+T12	106-117	1508.8835	754.9418	503.62783	1					YLGYLEQLLRLK
T12	116-117	260.1968	130.5984	87.398933	0					LK
T12+T13	116-118	388.2918	194.6459	130.09727	1					LKK
T13	118-118	147.1128	74.0564	49.704267	0					K
T13+T14	118-120	438.2711	219.6356	146.75703	1					KYK
T14	119-120	310.1761	155.5881	104.0587	0					YK
T14+T15	119-134	1871.9861	936.4931	624.66203	1	PHOS: 130	1951.9524	976.4762	651.31747	YKVPQLEIVPNSAEER
T15	121-134	1580.8278	790.9139	527.60927	0	PHOS: 130	1660.7941	830.89705	554.2647	VPQLEIVPNSAEER
T15+T16	121-139	2177.1383	1089.069	726.37943	1	PHOS: 130	2257.1046	1129.0523	753.03487	VPQLEIVPNSAEERLHSMK
T16	135-139	615.3283	308.1642	205.7761	0					LHSMK
T16+T17	135-147	1506.7845	753.8923	502.92817	1					LHSMKEGIHAQQK
T17	140-147	910.4741	455.7371	304.15803	0					EGIHAQQK
T17+T18	140-166	3207.5931	1604.297	1069.8644	1					EGIHAQQKEPMIGVNQELAY FYPELFR
T18	148-166	2316.1369	1158.568	772.7123	0					EPMIGVNQELAYFYPELFR
T18+T19	148-208	7013.2919	3507.146	2338.4306	1					EPMIGVNQELAYFYPELFRQFYQL DAYPSGAWYYVPLGTQYTDAPSF SDIPNPIGSENSE K
T19	167-208	4716.1728	2358.586	1572.7243	0					QFYQLDAYPSGAWYYVPLGTQYT DAPSFSDIPNPIGSENS EK
T19+T20	167-214	5445.5248	2723.262	1815.8416	1					QFYQLDAYPSGAWYYVPLGTQYT DAPSFSDIPNPIGSENS EKTTMPLW
T20	209-214	748.3698	374.6849	250.12327	0					TTMPLW

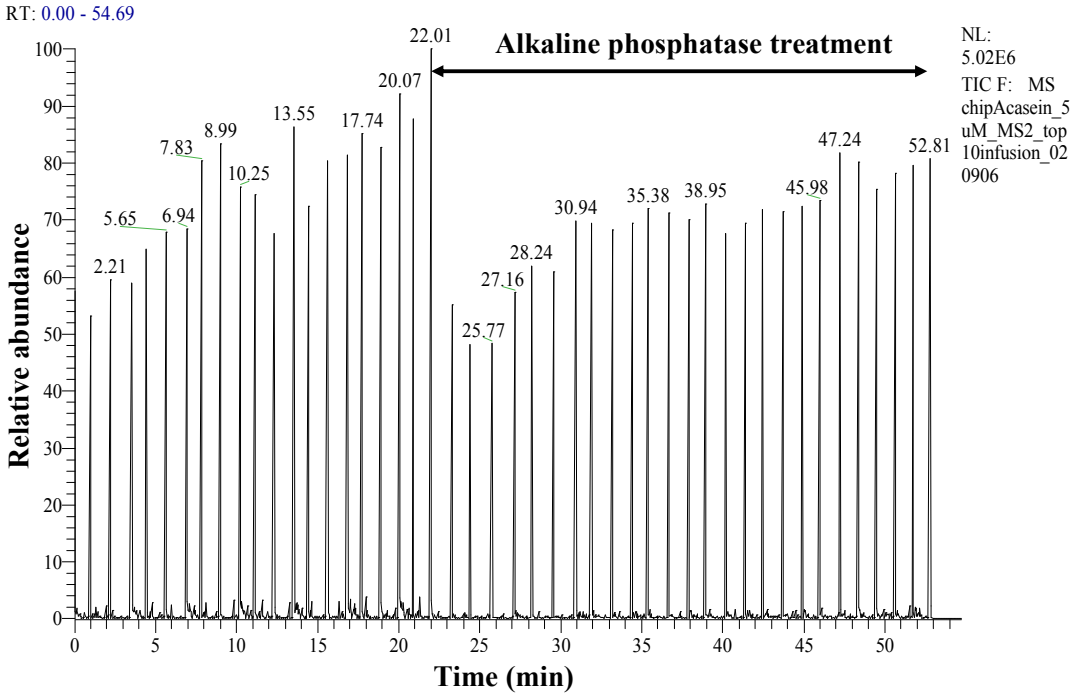


Figure 7. Total ion chromatogram (TIC) of an infusion experiment of the a-casein digest from the microfluidic chip platform.

MS spectra of the a-casein before and after dephosphorylation are shown in **Figures 8A** and **B**. Peptides that belong to a-casein are marked in the spectra. The ions labeled with 'T' are identified as trypsin autolysis products. A total of 9 tryptic fragments before dephosphorylation, and 10 tryptic fragments after dephosphorylation were identified. While 2 additional fragments, that initially were phosphorylated, did show up in the spectrum after dephosphorylation, another fragment has disappeared, probably as a result of electrospraying in the second case from a basic solution.

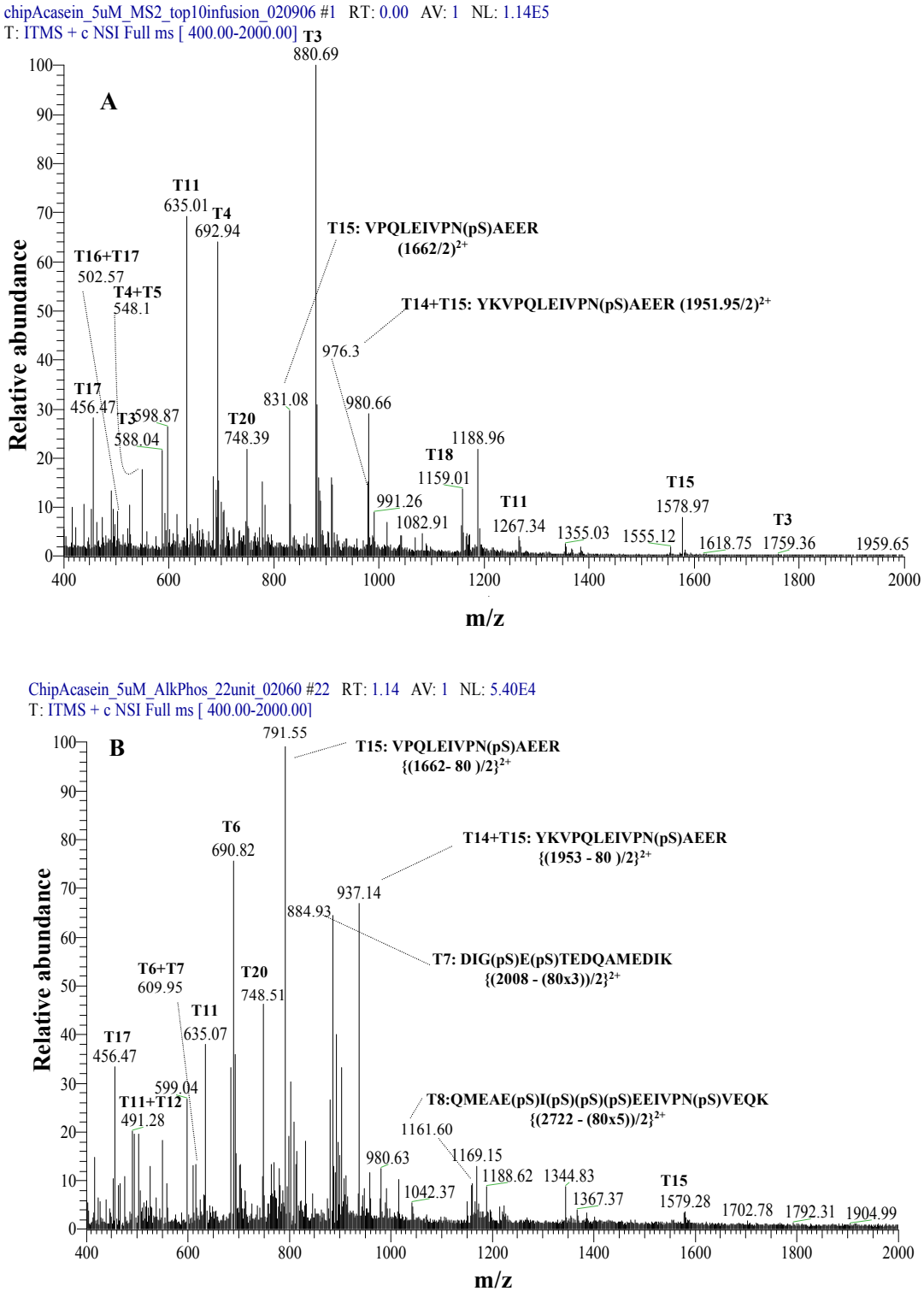


Figure 8. Mass spectra of an a-casein digest from the microchip platform. (**A**) before dephosphorylation; (**B**) after dephosphorylation (T: tryptic fragment).

Two phosphorylated peptides from Figure 8A, with $(MH_2)^{2+}$ 976.3 and 831.08, were identified as having one phosphorylated site, thus, they were identifiable even before dephosphorylation. For further confirmation, the MS^2 spectra of these peptides indicating the assignment of fragment ions are shown in **Figure 9A** and **B.** A neutral loss of 98 Da, typical to phosphorylated peptides is observable in these spectra. The p-value, Xcorr, ?Cn, Sp, RSp, and %ions for these peptides are given in the spectra. Similar analysis was performed for the dephosphorylated peptides. There were 4 dephosphorylated peptides, with $(MH_2)^{2+}$ 791.55, 937.14, 884.93, and 1161.60, observed in the MS spectrum of the a-casein after alkaline phosphatase treatment. The dephosphorylated peptides lose the phosphate ion, and as a result their mass will be 80 Da smaller (equivalent of 40 Da for a doubly charged peptide). MS^2 spectra for other dephosphorylated peptides are shown in **Figure 10A**, **B**, and **C**.

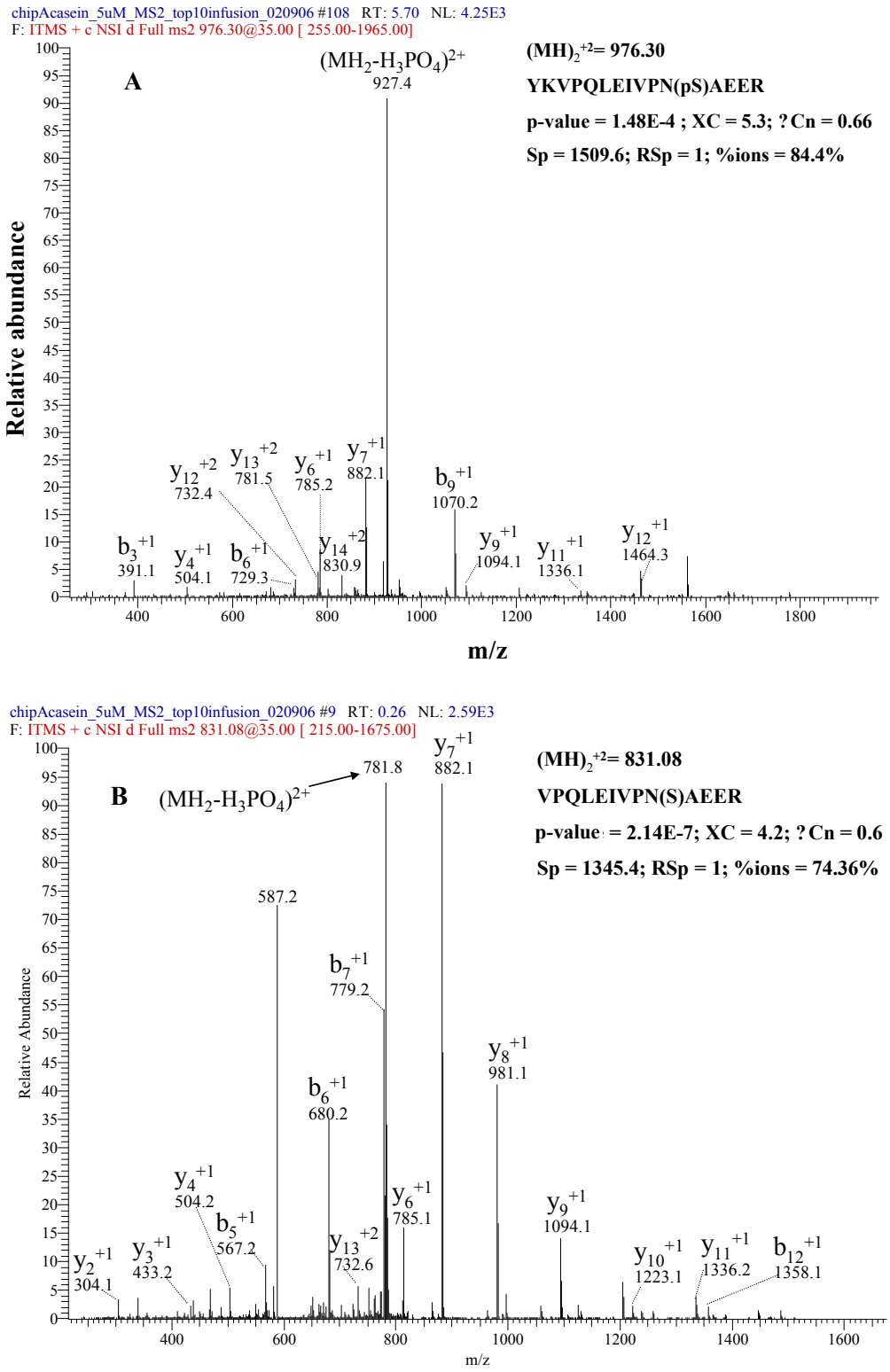
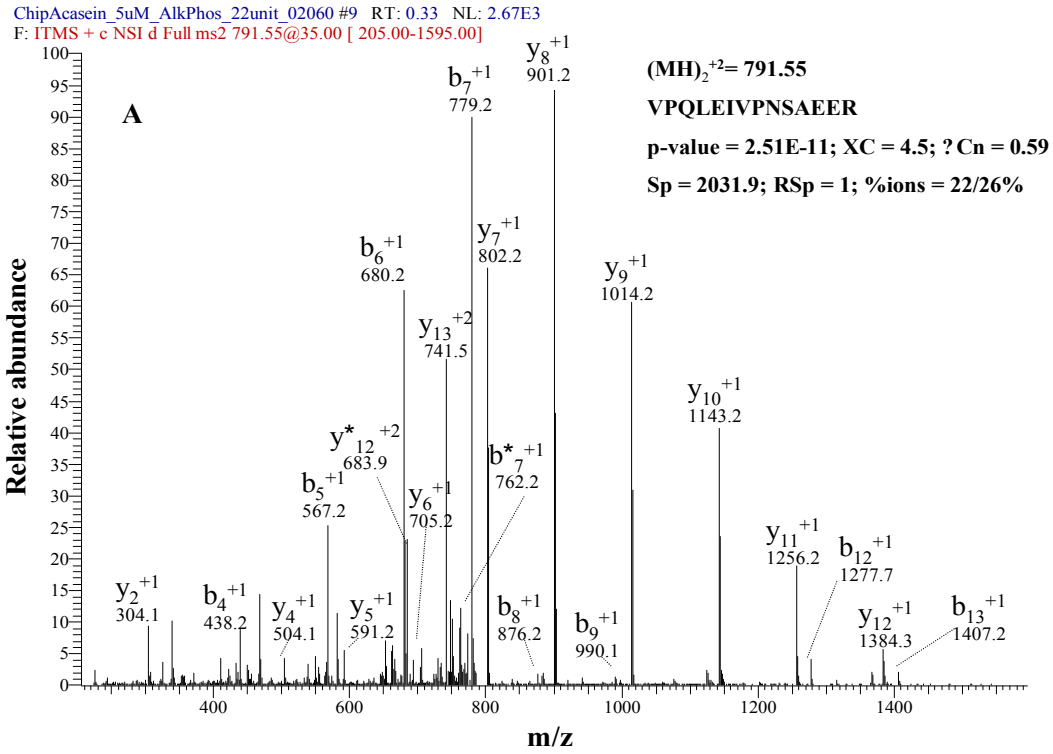
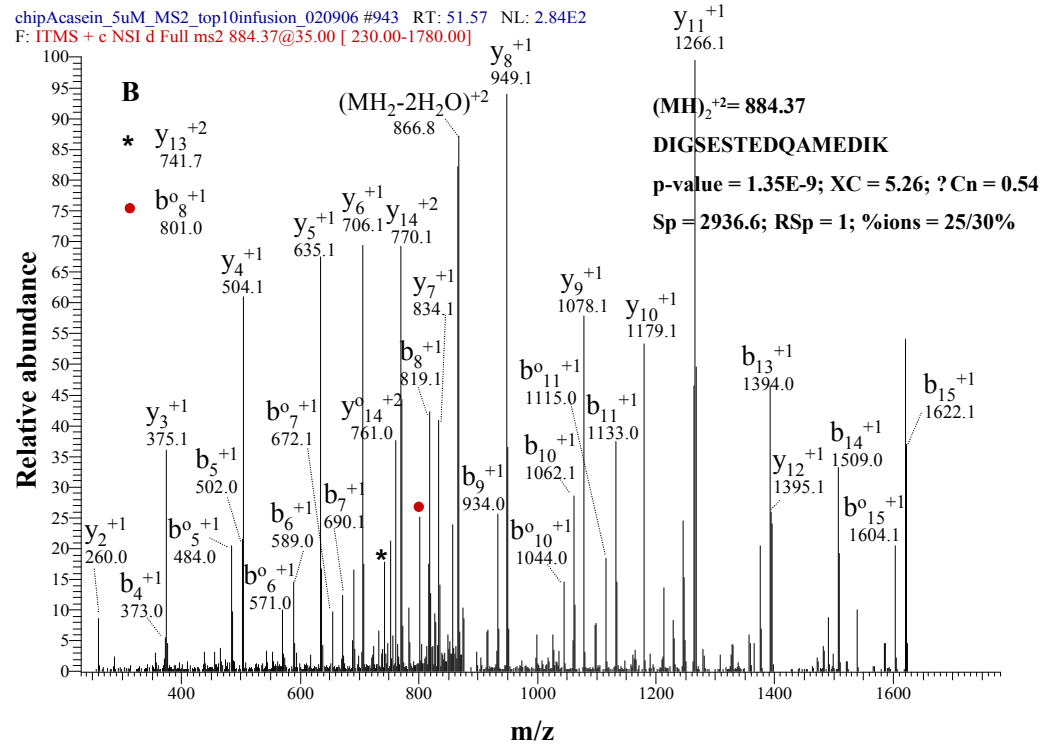


Figure 9. Tandem mass spectra of phosphorylated a-casein peptides generated from the chip. (**A**) $(MH_2)^{2+}$ = 976.3; (**B**) $(MH_2)^{2+}$= 831.08.


ChipAcasein_5uM_AlkPhos_22unit_02060 #9 RT: 0.33 NL: 2.67E3
F: ITMS + c NSI d Full ms2 791.55@35.00 [205.00-1595.00]
A
(MH)2+2= 791.55
VPQLEIVPNSAEER
p-value = 2.51E-11; XC = 4.5; ?Cn = 0.59
Sp = 2031.9; RSp = 1; %ions = 22/26%
Relative abundance
m/z


chipAcasein_5uM_MS2_top10infusion_020906 #943 RT: 51.57 NL: 2.84E2
F: ITMS + c NSI d Full ms2 884.37@35.00 [230.00-1780.00]
B
(MH)2+2= 884.37
DIGSESTEDQAMEDIK
p-value = 1.35E-9; XC = 5.26; ?Cn = 0.54
Sp = 2936.6; RSp = 1; %ions = 25/30%
Relative abundance
m/z

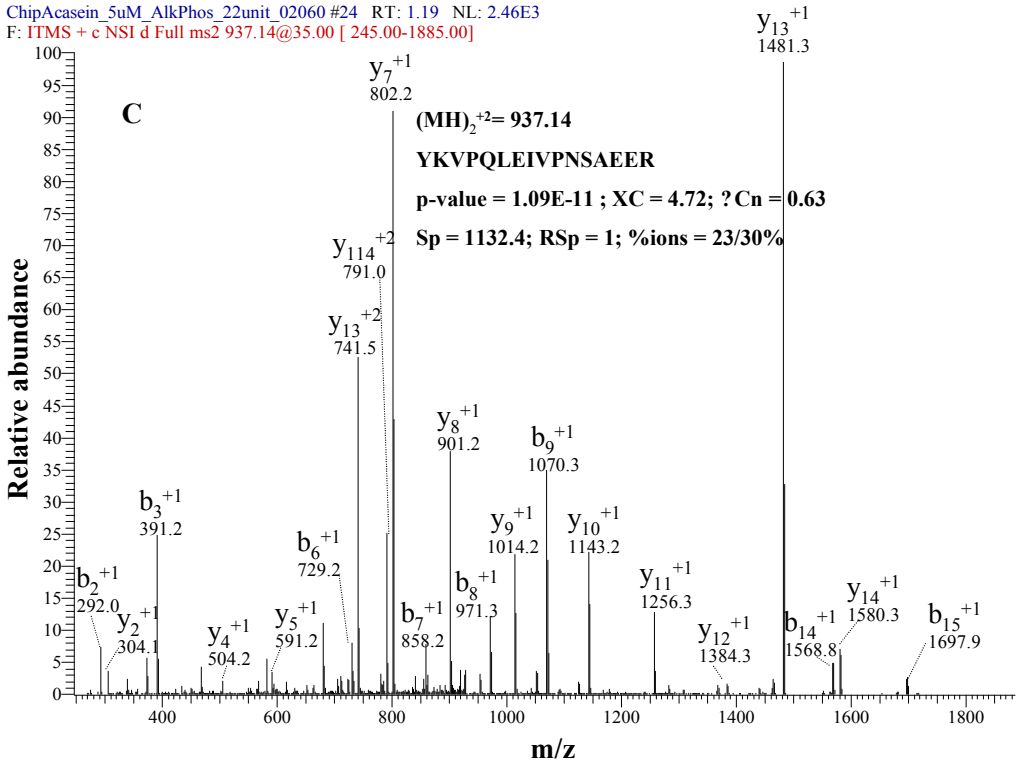


Figure 10. Tandem mass spectra of dephosphorylated a-casein peptides generated from the chip. (**A**) $(MH_2)^{2+}$= 791.55; (**B**) $(MH_2)^{2+}$= 884.37; (**C**) $(MH_2)^{2+}$= 937.14.

For confirming the results from the chip, a bench-top LC-MS/MS experiment was conducted, that enabled the injection of a large sample amount and the generation of intense ion spectra. Base peak chromatograms from the LC-MS/MS analysis of a-casein, before and after dephosphorylation, indicating the identified tryptic fragments, are shown in Figures **11A** and **B**. The results confirm the presence of the non-phosphorylated counterparts of the peptides that were initially phosphorylated. As compared with the chip experiment the benchtop LC-MS/MS experiment identified additional multi-phosphorylated peptide sequences. Improvements in chip design, that will lead to prolonged interaction times between the phosphorylated peptides and alkaline phosphatase (for example, longer infusion channels), will enable a more efficient dephosphorylation process, and will produce more intense and easily detectable ion signals. These experiments demonstrate, however, the applicability of these chips for the identification of phosphorylated peptides and phosphorylation sites, in analysis times as short as 10-15 min.

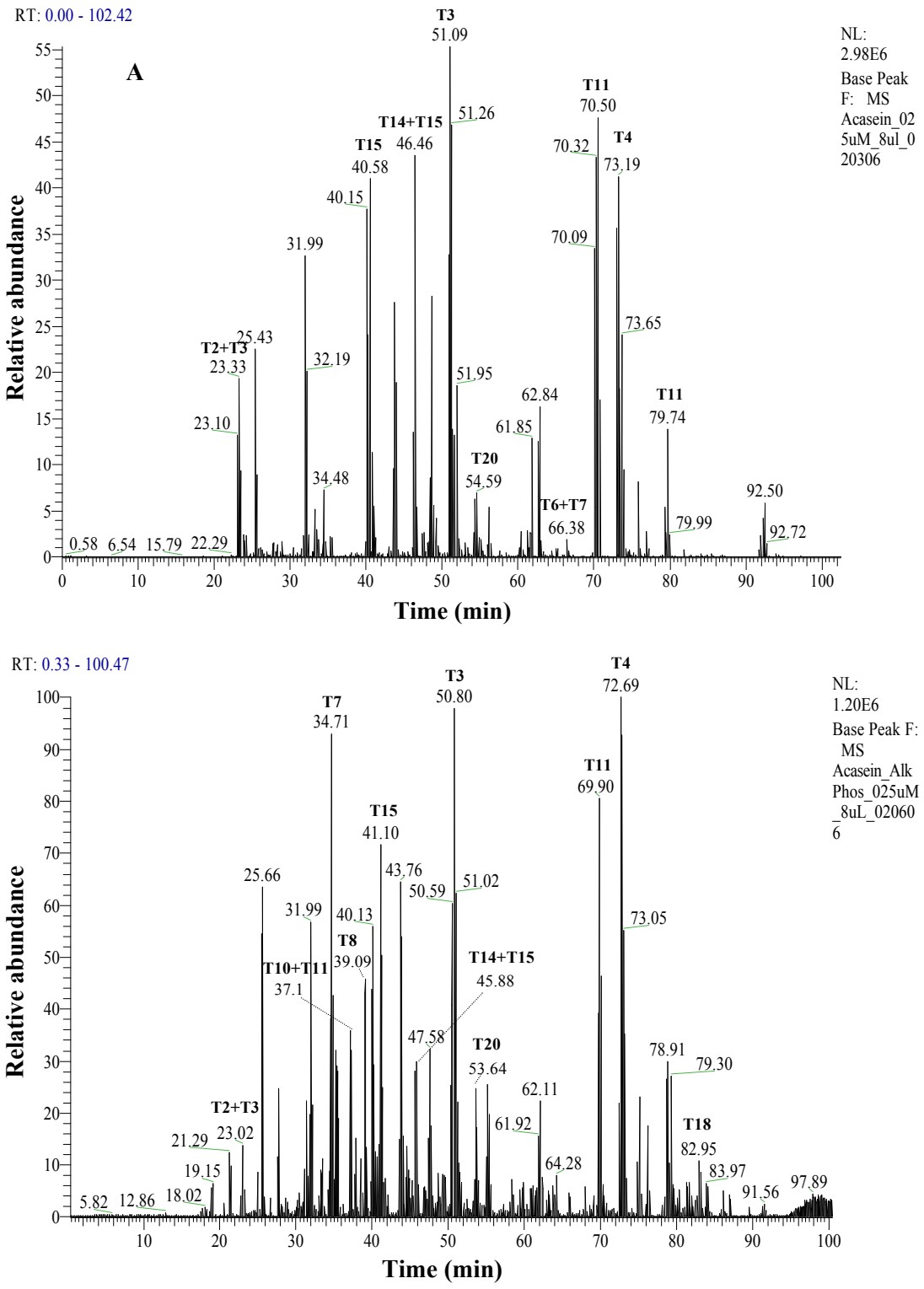


Figure 11. Base peak chromatograms of the a-casein digest generated with bench-top LC-MS/MS. (**A**) Before dephosphorylation; (**B**) After dephosphorylation.

4.5 References

1. Lazar, I. M., Ramsey, R. S., Jacobson, S. C., Foote, R. S., and Ramsey, J. M. (2000) Novel microfabricated device for electrokinetically induced pressure flow and electrospray ionization mass spectrometry. *J. Chromatogr. A.* **892**, 195-201

2. Figeys, D., Aebersold, R. (1998) Nanoflow solvent gradient delivery from a microfabricated device for protein identifications by electrospray ionization mass spectrometry. *Anal. Chem.* **70**, 3721-3727

3. Lazar, I. M., Sundberg, S. A., Ramsey, R. S., and Ramsey, J. M. (1999) Subattomole-Sensitivity Microchip Nanoelectrospray Source with Time-of-Flight Mass Spectrometry Detection. *Anal. Chem.* **71**, 3627-3631

4. Oleschuk, R. D., Harrison, D. J. (2000) Analytical microdevices for mass spectrometry. *Trends Anal. Chem.* **19(6)**, 379-387

5. Harrison, D. J., Glavina, P. G., Manz, A. (1993) Towards miniaturized electrophoresis and chemical-analysis systems on silicon-an alternative to chemical sensors. *Sens Actuators B* **10**, 107-116

6. Harrison, D. J., Manz, A., Fan, Z. H., Ludi, H., Widmer, H. M. (1992) Capillary electrophoresis and sample injection systems integrated on a planar glass chip. *Anal. Chem.* **64**, 1926-1932

7. Jacobson, S. C., Hergenroder, R., Kounty, L. B., and Ramsey, J. M. (1994) High-speed separations on a microchip. *Anal. Chem.* **66**, 1114-1118

8. Jacobson, S. C., Hergenroder, R., Kounty, L. B., and Ramsey, J. M. (1994) Open-channel electrochromatography on a microchip. Anal. Chem. **66**, 2369-2373

9. Jacobson, S. C., Kounty, L. B., Hergenroder, R., Moore, A. W., and Ramsey, J. M. (1994) Microchip capillary electrophoresis with an integrated postcolumn reactor. Anal. Chem. **66**, 3472-3476

10. Lazar, I. M., Sarvaiya, H., Trisiripisal, P., and Yoon, J. H. (2005) Microfluidic LC system for the analysis of proteomic constituents in cancerous cell lines. 53rd Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, USA, June 5-9

11. Figeys, D., Aebersold, R. (1999) Microfabricated modules for sample handling, sample concentration and flow mixing: application to protein analysis by tandem mass spectrometry. J. Biomech. Eng. **121(1)**, 7-12

12. Weinberger, R. (1993) Practical capillary electrophoresis. Boston: Academic Press.

13. Lazar, I. M., Grym, J., and Foret, F. (2005) Microfabricated devices: a new sample introduction approach to mass spectrometry. *Mass Spectrom. Reviews.* **00**, 1-21

14. Ziaie, B., Baldi, A., Lei, M., Gu, Y., Siegel, R. A. (2004) Hard and soft micromachining for BioMEMS: review of techniques and examples of applications in microfluidics and drug delivery. *Adv. Drug Del. Reviews.* **56**, 145-172

15. Jacobson, S. C.; Hergenroder, R.; Koutny, L. B.; Warmack, R.J.; Ramsey, J.M. (1994) Effects of injection schemes and column geometry on the performance of microchip electrophoresis devices. *Anal. Chem*. **66**, 1107-1113

16. Marc Madou, (1997) Fundamentals of Microfabrication, CRC Press, page 405.

17. Lazar, I.M.; Karger, B.L. (2002) Multiple open-channel electroosmotic pumping system for microfluidic sample handling *Anal. Chem*., **74(24)**, 6259-6268

18. Paul, P. H.; Arnold, D. W.; Rakestraw, D. J. (1998) *Proceedings of the Micro Total Analysis Systems Workshop*, Banff, Canada, Oct. 13-16

19. Lazar, I. M.; Ramsey, R. S.; Ramsey J. M. (2001) On-chip proteolytic digestion and analysis using "wrong-way-round" electrospray time-of-flight mass spectrometry *Anal. Chem*. **73**, 1733-1739

20. Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotech.* **20(6)**, 261-268

21. Zhang, H., Zha, X., Tan, Y., Hornbeck, P., Mastrangelo, A., Alessi, D., Polakiewicz, R., and Comb, M. (2002) Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J. Biol. Chem.* **277**, 39379-39387

22. Guerrera, I. C., Atkinson, J. P., Kleiner, O., Soskic, V., and Zimmermann, J. G. (2005) Enrichment of phosphoproteins for proteomic analysis using immobilized Fe(III)-affinity adsorption chromatography. *J. Proteome Res.* **4**, 1545-1553

23. Oda, Y., (2001) Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. Nat. Biotechnol. **19**, 379-382

24. Vasilescu, J., Smith, J. C., Ethier, M., and Figeys, D. (2005) Proteomic analysis of ubiquitinated proteins from human MCF-7 breast cancer cells by immunoaffinity purification and mass spectrometry. *J. Proteome Res.* **4**, 2192-2200

25. Kim, J. E., Tannenbaum, S. R., and White, F. M. (2005) Global phosphoproteome of HT-29 human colon adenocarcinoma cells. *J. Proteome Res.* **4**, 1339-1346

Chapter 5: Conclusions and Future Prospects

5.1 Conclusions

The use of proteomic technologies with mass spectrometry detection has proven to be a promising strategy to analyze biological samples in search for potential cancer biomarkers. The present research was aimed at the development of a 2D LC-MS platform for the characterization of the MCF7 breast cancer cell proteome. A sequence of optimization strategies were performed in order to develop a protocol that enabled the reliable and sensitive detection of a large number of proteins. In addition, a series of established and potential biomarkers, as reported in the literature (TP53RK, cathepsin D, E-cadherin, Ki-67, PCNA, PSA, CA125, 14-3-3 sigma, cytokeratins-18, 19, calreticulin, and heat shock proteins-60, 90), were also identified.

While a precise comparison with data reported in the literature is not possible, due to broad variations in the experimental protocols and content/size of the utilized databases, we believe that this study represents the most comprehensive characterization of a breast cancer related sample. Similar research efforts have typically resulted in the identification of up to 300-500 proteins [1, 2]. Jacobs has reported the identification of 1,700 proteins in human mammary epithelial cells by using a non-redundant database with 76,402 FASTA entries [3]. Data were filtered only with Xcorr cutoff values of 1.9, 2.2, 3.75 and ? Cn>0.1. An additional filtering parameter, the LC normalized elution time of a peptide, was also used to increase the confidence of protein identifications, and this parameter reduced the protein IDs to 1,574. A total of 228 tryptic digest fractions were analyzed from alkylated and non-alkylated proteins, and a large number of MS^2 spectra,

i.e., 700,000, were generated in his study. Alternatively, Tomlinson has reported the identification of 1,966 unique proteins in the KATO III human gastric carcinoma cell line, using manual data interpretation for the validation of results [4]. The protocol involved, however, the analysis of as many as 1,354 peptide subfractions. We are reporting the identification of 1,895 proteins that were selected conservatively with two sets of filters and $p<0.001$; an additional 472 proteins (total 2,367) that passed commonly used selection criteria need more intense scrutiny, possible manual validation, especially if the scope of the analysis is the identification of novel biomarkers. Furthermore, these numbers would be much increased if partial tryptic peptides would have been allowed in the search. These results were generated by analyzing only 16 peptide SCX fractions and 54,843 MS^2 spectra. Over 100 potential cancer markers were identified, of which, ~25 are accepted as established biomarkers.

The development of microfluidic bioanalytical devices has gained much interest over the last few years. The research described in chapter 4 demonstrates the capability of a microchip to perform analytical separations and detect biomarkers and posttranslationally modified peptides. The identification of a total of 77 proteins, 39 proteins with $p<0.001$, was possible with these chips. In addition, the LC microdevice enabled the identification of 5 cancer biomarkers (PCNA, cathepsin D, and cytokeratins 8, 18, 19), and was also applicable for the analysis of phosphopeptides. The key advantages of this device include its miniaturized format, capability to perform rapid analysis, disposability, and contamination free analysis of small quantities of sample. This demonstrates the applicability of these microfluidic chips for proteomic investigations and biomarker screening.

5.2 Future prospects

Confident identification of many cancer specific proteins that can be of further interest to the biomedical community was accomplished in this research. It is very important to be capable of identifying initially a large list of proteins, before planning for a detailed analysis of their expression levels and function. The data collected in this work will be a good resource that will allow for the further development of differential expression analysis protocols, and the identification of novel biomarkers that will enable us to differentiate between healthy and diseased states.

Moreover, the data generated in this study will be used to create a database that integrates information regarding the identity, expression level and function of cancer specific proteins. The database will be made publicly available to serve the broader scientific community. The development of novel microfluidic devices with ESI and MALDI MS detection will be also pursued. The focus will be on microfluidic designs that are appropriate for large scale population screening.

5.1 References

1. Celis, J. E., Gromov, P., Cabezon, T., Moreira, J. M. A., Ambartsumian, N., Sandelin, K., Rank, F. and Gromova, I. (2004) Proteomic characterization of the interstitial fluid perfusing the breast tumor microenvironment: a novel resource for biomarker and therapeutic target discovery. *Mol. Cell. Proteomics*. **3(4)**, 327-344

2. Xiang, R., Shi, Y., Dillon, D. A., Negin, B., Horvath, C., and Wilkins, J. A. (2004) 2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J. Proteome Res.* **3**, 1278-1283

3. Jacobs, J. M., Mottaz, H. M., Yu, L. R., Anderson, D. J., Moore, R. J., Chen, W. N. U., Auberry, K. J., Strittmatter, E. F., Monroe, M. E., Thrall, B. D., Camp, D. G., and Smith, R. D. (2003) Multidimensional proteome analysis of human mammary epithelial cells. *J. Proteome Res.* **3**, 68-75

4. Tomlinson, A. J., Hincapie, M., Morris, G. E., and Chicz, R. M. (2002) Global proteome analysis of a human gastric carcinoma. *Electrophoresis*. **23**, 3233-3240

Vita

Hetal Sarvaiya was born on 1st June, 1979 in Surat, Gujarat, India. She received her Bachelor of Engineering in Electrical Engineering in June 2000 from Gujarat University, Ahmedabad, India. She served as a teaching assistant and continued as a lecturer in the Department of Electrical Engineering after her degree completion. In August 2004, she began her study for the Master of Science in Biomedical Engineering at Virginia Tech. After graduation, Hetal plans to work on research and development of new techniques and methodologies for biomedical applications using mass spectrometry.