University of Alberta

Development and Applications of Mass Spectrometric Techniques for Comprehensive Proteome Analysis

by

Nan Wang ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Chemistry

Edmonton, Alberta

Fall 2008

Your file Votre référence
ISBN: 978-0-494-46443-4
Our file Notre référence
ISBN: 978-0-494-46443-4

To my dearest parents

Abstract

Proteomics has become a powerful tool in biological and biomedical research. However, current proteome analysis technology only detects a fraction of the entire proteome in a sample. My thesis work is focused on the development and applications of new mass spectrometric techniques to generate as comprehensive proteome coverage as possible, with an ultimate goal of analyzing the entire proteome of cells or tissue samples. The initial work of proteomic analysis of human tear fluids illustrated the analytical challenges of detecting a large number of proteins from a complex sample. Three techniques have been developed to improve the proteome coverage by a shotgun proteome analysis method. Proteome fractionation using sequential protein solubilization and digestion was effective in simplifying a complex sample and enhancing the analysis of membrane proteins. Implementation of an optimized precursor ion extraction (PEI) strategy in producing fragment ion spectra of peptides using liquid chromatography (LC) electrospray ionization (ESI) quadrupole time-of-flight (QTOF) mass spectrometry (MS) was demonstrated to be very useful in improving peptide and protein identification efficiency. An off-line two-dimensional LC separation of peptides based on strong cation exchange fractionation, followed by peptide desalting/quantification using LC UV detection and then injecting a maximal amount of sample to reserved-phase LC QTOF MS was a powerful technique that greatly increased the number of peptides and proteins identified. Finally, these techniques have been combined to analyze the proteomes of MCF-7 cells, zebrafish liver and *E. coli* with 2911, 5710, and 3730 proteins identified, respectively. Sensitive proteome profiling of samples from only thousands of MCF-7 cells was demonstrated. Finally, microwave-assisted acid hydrolysis combined with LC-ESI QTOF MS was useful for mapping protein sequences and analyzing posttranslational modifications of proteins.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor, Professor Liang Li, for granting me the opportunity to study in his research group and for his invaluable guidance, inspiration, encouragement and advice during the course of my research. I have learnt a significant amount from him and I am sure it will continue to benefit my life and career.

I would like to thank the other members of my supervisory committee, Professor Charles A. Lucy and Professor Hicham Fenniri, and the other members of my thesis examining committee, Professor Greg Goss, Professor Norman Dovichi and Professor Robert E. Campbell, for their active participation during my oral examination, their thorough reviews and comments on this thesis, and their valuable advice on my research and career.

My deep gratitude goes to the people with whom I collaborated. I especially thank Professor Greg Goss and his student, Lauren MacKenzie, from the Department of Biological Sciences at the University of Alberta for the collaborative work on the zebrafish liver proteome analysis (Chapters 3 and 6). I also thank Dr. Nan Li for her professional training on sample preparation, LC-MALDI instrument use and many other techniques involved in the tear proteome work (Chapter 2). I am very grateful to Dr. Chuanhui Xie for his contributions to the SCX separation optimization and desalting system setup (Chapter 5). I also thank Professor Joel H. Weiner from the Department of Biochemistry at the University of Alberta for his contributions to the bioinformatic characterization of the proteins identified from *E. coli* and helpful discussions in this exciting area of research (Chapter 7). Many thanks to Andrea De Souza, Fang Wu, Xiaoxia Ye and Mingguo Xu for their great help on the proteomic sample analysis (Chapters 3, 6, 7 and 8, respectively).

I would like to thank many members in Professor Li's research group including Dr. Nan Zhang, Dr. Huizhi Liu, Lidan Tao and Dr. Hongying Zhong for their practical advice and demonstration on using instruments and protocols. My appreciation also extends to other members in the group, in no particular order, Dr. Ying Zhang, Zhihui Wen, Hui Dai, Xinlei Yu, Dr. Chengjie Ji, J. Bryce Young, Andy Lo and Dr. Peng Wang for their helpful discussions and assistance.

Special thanks go to Dr. Randy Whittal and Jing Zheng from the Mass Spectrometry Facility in the Department of Chemistry at the University of Alberta for their professional and patient training on trouble-shooting the QTOF and the MALDI-TOF instruments. I must also thank Dr. Sandra Marcus for growing cells and preparing the membrane fractions.

I would also like to thank the departmental staff in the general and purchasing offices, the mailroom, and the electronics and machine shops for their kind help.

I would like to acknowledge NSERC, the Department of Chemistry and the University of Alberta for providing me with financial support while completing my graduate studies.

Finally, I especially thank my parents, Mr. Jianping Wang and Mrs. Yun Meng for their endless love, understanding, patience and encouragement over the years. I realize how big a sacrifice they have made while I grew up, studied abroad and completed my degree. I will be forever grateful for that.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D-LC	Two dimensional liquid chromatography
ACN	Acetonitrile
BRO	Bacteriorhodopsin
BSA	Bovine serum albumin
CE	Capillary electrophoresis
CHAPS	3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate
CID	Collision-induced dissociation
CMC	Critical micelle concentration
CNBr	Cyanogen bromide
DC	Direct current
DHB	2,5-dihydroxybenzoic acid
DTT	dithiothreitol
E. coli	*Escherichia coli*
EGTA	[Ethylenebis(oxyethylenenitrilo)] tetraacetic acid
ESI	Electrospray ionization
FTICR	Fourier-transform ion cyclotron resonance
FWHM	Full width at half maximum
HCCA	α-cyano-4-hydroxycinnamic acid
HEA	Epithelial Specific Antigen
HEPES	N-2-hydroxyethylpiperazine-N’-2-ethanesulfonic acid
HPLC	High performance liquid chromatography
IDA	Iodoacetamide

IE	Ion exchange
IEF	Isoelectric focusing
m/z	mass to charge
MAAH	Microwave-assisted acid hydrolysis
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectrometry
MW	Molecular weight
NP-40	Nonidet P40-substitute
PBS	Phosphate-buffered saline
PI	Isoelectric point
PMSF	Phenylmethyl sulfonyl fluoride
ppm	part(s) per million
PTM	Posttranslational modification
QTOF/QqTOF	Quadruple time-of-flight
RCF	Relative centrifugal force (× g)
RF	Radio frequency
RP	Reversed-phase
SA	Sinapinic acid
SCX	Strong Cation Exchange
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate- polyacrylamide gel electrophoresis
S/N	Signal to noise ratio
TFA	Trifluoroacetic acid

Tris	Tris (hydroxymethyl) aminomethane
Triton® X-100	t-Octylphenoxypolyethoxyethanol
TOF	Time-of-flight
m	milli- (10^{-3})
μ	micro- (10^{-6})
n	nano- (10^{-9})
p	pico- (10^{-12})

Chapter 1

Introduction to Mass Spectrometry and Proteome Analysis

Proteomics is a relatively new research field and plays an increasingly important role in many research areas including biology and medicine.[1] Unlike conventional ways of studying biological systems where one or a few proteins are interrogated to provide information in understanding their functions or properties, proteomics is a tool to characterize many proteins, ideally the entire proteome (e.g., all proteins expressed in a cell or all proteins present in a sample such as tissue), and relate the properties of these proteins or proteome to one or more phenotypes of the biological system of interest to do functional studies. Depending on the type of biological question to be addressed, proteome characterization or proteome analysis takes several different forms. In some studies, proteome analysis only requires the generation of a list of proteins present in a biological sample.[2-5] In other studies, quantitative proteome analysis data are needed.[6-9] Besides the generation of a qualitative and/or quantitative proteome profile, some applications require information on protein-protein interactions.[10, 11] Finally, characterization of protein modifications or post-translational modifications (PMTs) at the proteome level is needed in many studies, such as research in understanding how a gene deletion can affect protein signal pathway changes.[12-14]

My thesis focuses on only one aspect of the proteome analysis, namely the identification of proteins present in a proteomic sample. Specifically, we wish to identify as many proteins as possible from cells, tissues, or human fluids, with an ultimate goal of identifying all of the proteins in a sample, i.e., the entire proteome. Achieving this goal would form a solid foundation from which new analytical techniques could be further developed for proteome-wide quantitative proteome analysis, protein-protein interaction profiling, and post translational modification (PMT) characterization. In my work, mass spectrometry (MS) is used for proteome analysis. All of the technical developments presented in this thesis centre around the use of tandem mass spectrometry (MS/MS) for peptide or protein identification. There are many excellent reviews on each topic of MS and proteome analysis and thus I do not intend to cover all areas in detail in this chapter. Rather I will focus on the discussion of the most relevant topics to my thesis work. Specifically, protein sample preparation methods related to MS/MS analysis will be described, followed by the discussion of two-dimensional liquid chromatography (2D-LC) separation of peptides. The technology of MS and MS/MS, particularly quadrupole time-of-flight (QTOF) MS, will be introduced. Two database search approaches for protein identification will be described. Finally, the scope of my thesis will be given.

1.1 Overview of Protein Identification Methods

There are two widely used methods for proteome analysis by mass spectrometry. One is to separate the proteins in a proteomic sample and then identify individual proteins by MS. Protein separation is commonly done using gel electrophoresis.[15-18] Liquid chromatography can be used to separate proteins, but the separation efficiency is not as good as the gel-based method.[19] After protein separation, the individual proteins are subjected to chemical or enzymatic treatment to be degraded into peptides followed by MS analysis. This method is sometimes called the gel-based method. The second way of identifying proteins is to use a chemical or enzyme to degrade all of the proteins present in a sample and then use liquid chromatography to separate the peptides, followed by mass spectrometric sequencing of the peptides. This method is sometime referred as the bottom-up or shotgun method. In both methods, the resulting mass spectra are loaded into a search engine where intact peptide masses and/or fragment ion masses are matched against a proteome database for peptide and protein identification. Comparing the two methods, the shotgun method is faster, more sensitive and more amendable to automation.[1, 20]

There is another method called top-down protein identification which is based on the separation of proteins often by liquid chromatography and then MS/MS analysis of individual proteins.[21-26] Protein identification is done either manually by interpreting the MS/MS spectrum of an intact protein ion or automatically by searching the MS/MS spectrum against a proteome database.[27] This method is still under development for analyzing complex proteomic samples.[28, 29] At present, it has a relatively low throughput, compared to the shotgun method. However, this method offers much greater possibility for determining modification sites of a protein,[30] as it can generate many more fragment ions from the intact protein ion covering a larger stretch of amino acid sequence than the shotgun method where only a few peptides of a protein are sequenced.

In my thesis work, the gel-based method was investigated for tear proteome analysis along with the shotgun method (see Chapter 2). The rest of the work was based on the shotgun method alone.

1.2 Protein Sample Preparation

There are many different types of proteomic samples. These include cultured cells from a particular cell line such as MCF-7 breast cancer cells, primary cells such as those isolated from tissue sections, tumor tissue samples, and body fluids such as blood, urine, and tear. In proteome analysis, proteins from a given sample must be extracted and processed carefully, prior to their introduction into a mass spectrometer for analysis. In the shotgun method, proteins must also be degraded into peptides. Quite often proteins from a complex sample are fractionated before they are degraded into peptides. Thus, the whole process of protein sample preparation can be quite involved and, unfortunately, it is often done manually. Besides avoiding human error, great care must be given in each step to avoid sample loss and minimize the use of chemicals that may interfere with the down stream separation and mass spectrometric analysis.

Many chemicals are used to prepare protein samples. Some are chosen to precipitate proteins out of a solution so that the protein pellet can be washed to remove salts, buffers and contaminants. Acetone is commonly used for protein precipitation.[31-33] The solvent is added to an aqueous protein solution and the solubilized proteins are denatured to form precipitates. Acetone can be readily removed and will not interfere with the subsequent sample handling processes.

Complete solubilization of the protein precipitates is important, as chemical or enzymatic degradation of proteins requires the proteins be in solution. Solubilization involves the use of a solvent to disrupt the protein-protein interactions caused by van der waal forces, dipole-dipole, ionic, and hydrogen bonds. To determine which type of solvent is best suited to dissolve a protein, protein hydrophobicity should be considered. Hydrophobicity is often gauged by using the GRAVY (Grand Average of Hydropathy) index.[34] The GRAVY index for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence.[34] In essence, proteins with positive values are considered hydrophobic, and those with negative values are considered hydrophilic. Although interaction and positional effects for adjacent residues are not considered in calculating GRAVY index, it still provides a useful indication of the protein property.

There are many different solvents available to solubilize proteins.[35] In selecting a proper solvent, two factors should be considered. A solvent should be capable of solubilize the proteins completely. Second, it should not interference with downstream sample processing. Hydrophilic proteins are generally easy to dissolve in an aqueous solution. A buffer solution such as NH_4HCO_3 is commonly used as it provides a more basic solution condition (pH=7.8). The acidic amino acids and the terminal carboxylic acid are negative charged at this pH, hence improving the solubility of the proteins. One benefit of using NH_4HCO_3 as a reagent to dissolve proteins is that it is compatible with trypsin digestion (see below), which means no additional step is needed to remove the reagent. However, aqueous solution is less effective in dissolving hydrophobic proteins.[35] The interactions among the hydrophobic domains of proteins can be so strong that charging a few amino acids using a basic condition does not break the inter-molecule interaction. Moreover, some chargeable amino acids within or near the hydrophobic domains may be buried and protonation (in basic condition) or de-protonation (in acidic condition) of these residues may not take place, making the proteins even less likely to be solubilized.

To overcome the hydrophobic interactions between proteins, an organic solvent or an aqueous solution containing a surfactant is often used.[35] Organic solvents such as methanol can dissolve hydrophobic proteins by loosing the hydrophobic interactions and exposing the protein surface to the solvent molecules. However, the selection of organic solvents is limited. The reason is that proteins contain both hydrophobic and hydrophilic domains and a balanced interaction between solvent molecules and proteins must be attained to dissolve proteins. If a non-polar solvent (e.g., hexane) is used, it can interact with the hydrophobic domains of the proteins well, but will strengthen the protein-protein hydrophilic interaction. Thus, the overall solubility of proteins may remain to be the same or even reduced, compared to those in an aqueous solution. Non-polar solvents may be used only to dissolve very hydrophobic proteins such as those with many transmembrane domains (TMDs). There are several programs available in predicting TMD in a protein based on the amino acid sequence.[36] The program TMHMM was found to have the best overall performance for predicting membrane spanning regions.[35-37]

A better approach to solubilize hydrophobic proteins is to use an aqueous solution containing a surfactant or detergent.[38-41] Surfactants can be classified into two categories: non-ionic and ionic surfactants. A non-ionic surfactant is a molecule composed of both hydrophobic and hydrophilic groups, while an ionic surfactant consists of a hydrophobic group and an ionic group. When a surfactant solution is added to a protein pellet, the hydrophobic group interacts with the hydrophobic domains of a protein and, at the same time, the hydrophilic group or ionic group of the surfactant interacts strongly with the aqueous solvent, resulting in solubilization of the proteins. To facilitate the solubilization process, a mechanic means, such as agitation of the solution, is often used to assist the penetration of the surfactant molecules into the hydrophobic interaction domains between proteins.

There are many choices of surfactants.[40] Figure 1.1 shows some common surfactants used to dissolve proteins. Among them, SDS is the strongest solubilizing reagent. At a concentration above a certain value (critical micelle concentration), micelles are formed as illustrated in Figure 1.2. The hydrophobic tails of the micelles can strongly interact with the hydrophobic domains of proteins. After the tails bond to the protein, the SDS molecules will regroup in trying to form micelles, which forces the protein going into the solution with the SDS molecules. Thus, to dissolve a protein, it is important to keep the surfactant concentration sufficiently high to form micelles in the

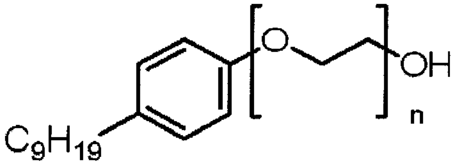
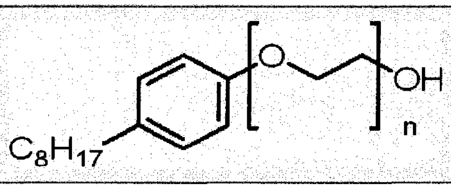
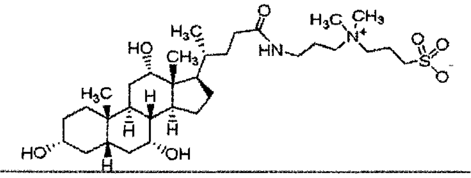
SDS	$CH_3(CH_2)_{10}CH_2O-\underset{\underset{O}{\|}}{\overset{\overset{O}{\|}}{S}}-ONa$	289	8.2 mM
NP-40		680	0.059 mM
Triton X 100		625	~0.23 mM
CHAPS	 O H_3C CH_3 N^+ HN H_3C HO CH_3 H_3C H H H HO H OH O S O O^-	615	4 mM

Figure 1.1 Surfactants commonly used to dissolve proteins.

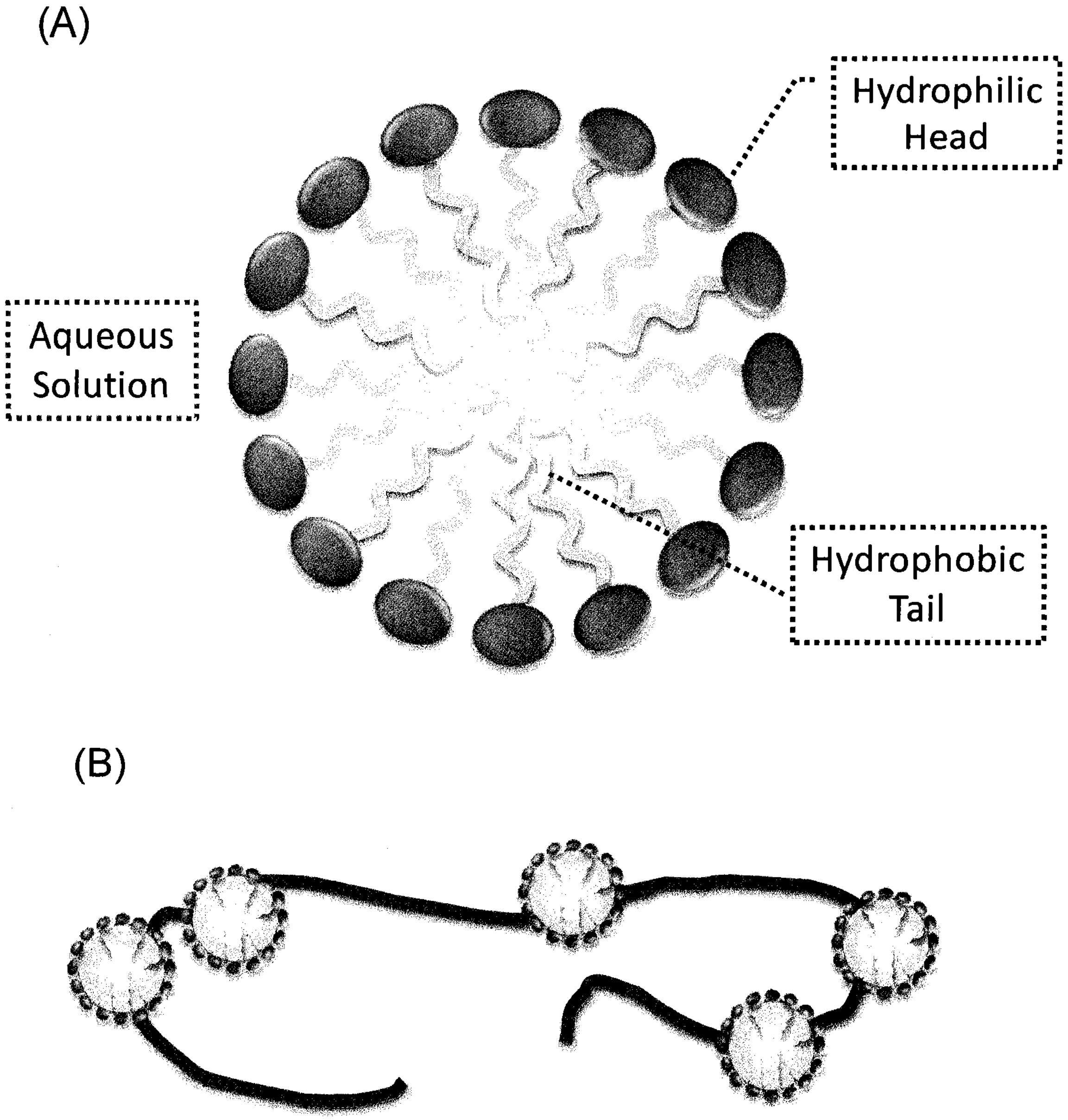


Figure 1.2 (A) SDS micelle and (B) protein and SDS interaction.

solubilizing solution.

While surfactants are good for dissolving proteins, they often interfere with mass spectrometric analysis.[42, 43] The interference is mainly due to their readiness to be ionized, compared to the analytes, during the ionization process. For example, in electrospray ionization (ESI), a commonly used method for generating peptide or protein ions for MS, the ions are formed from a droplet during the ESI process (see Section 1.3). Figure 1.3. shows a charged droplet where the charged molecules reside on the surface of the droplet and these charged species have a chance to be "liberated" into the gas phase through the so called Columbic Repulsion process when the droplet shrinks in size and is no longer able to keep all the charges on the surface.[44-47] In the charged droplet, analytes and other chemical species such as salts and surfactants compete for the limited surface space. A surfactant can easily form ions in the droplet. For example, polyethylene glycol (PEG) derivetives can interact strongly with a sodium ion in the solution (e.g., from sodium chloride impurity present in the sample) to form sodiated PEG. This ion will stay on the surface at the expense of analyte ions. If the concentration of the surfactant in the sample is higher than the analyte, which is often true for trace analysis, all charged species on the droplet surface will be in the form of sodiated PEG. As a consequence, only PEG ions are generated and the mass spectrum registers only these surfactant ions. While PEG is neutral, but readily forms ions with the attachment of alkaline ions often present in a sample, for a cationic surfactant, it is obvious that, in the positive ion detection mode, the pre-formed surfactant ions would dominate the mass spectrum, causing interference with the ionization or detection of the analyte molecules.

Positive ion detection is often used for peptide and protein ionization, as these molecules more readily form positive ions (i.e., protein is more readily protonated than de-protonated in the solution and during the ionization process), resulting in better detection sensitivity. One would think that the use of an anionic surfactant such as SDS might not cause interference, as they are negatively charged and will not compete for the positive charged surface. The problem of using a strong anionic surfactant such as SDS lies in the fact that, after the surfactant molecules are bounded to the proteins, the proteins become negative charged and they will not move the surface. As a result, in an ESI mass

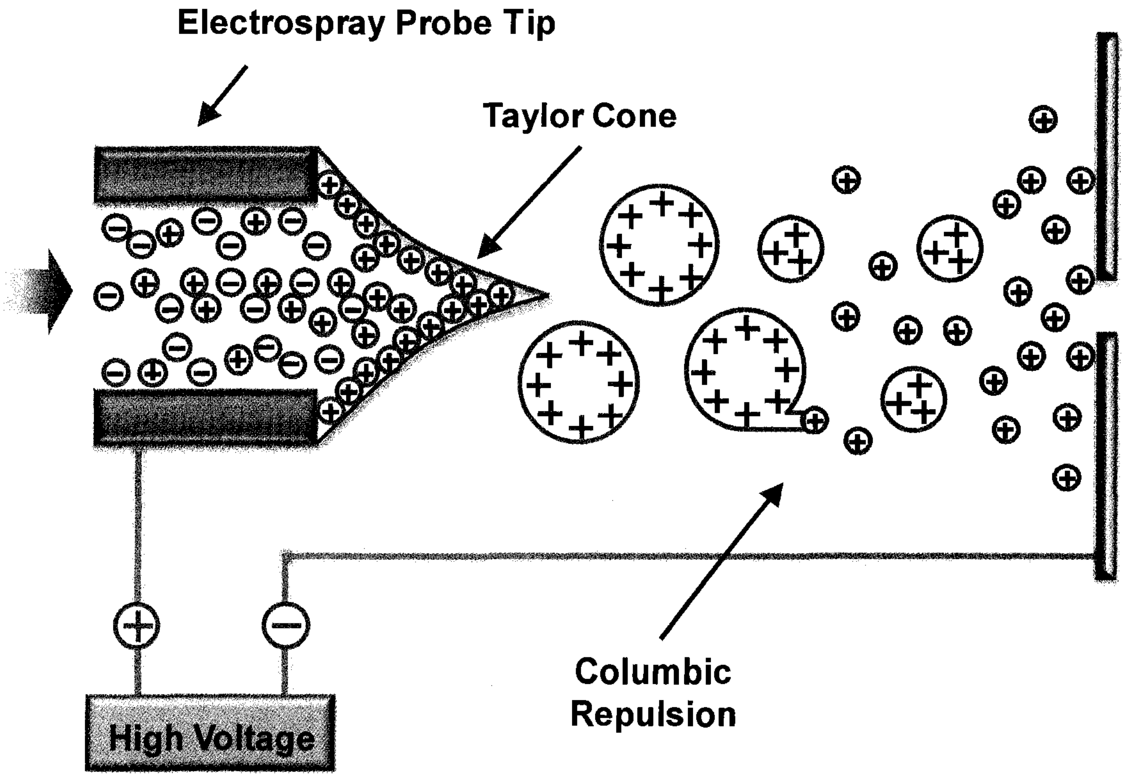


Figure 1.3 Process of electrospray ionization (ESI).

spectrum of a protein sample containing SDS, the dominated peaks are sodium ions plus several SDS sodium adduct ions (e.g., $[SDS+Na]^+$, $[SDS+SDS+Na]^+$, etc.). Of course, if we run the same sample in the negative ion mode, the negative SDS ions dominate the spectrum.

The phenomenon of one type of ion dominating the mass spectrum while reducing the probability of detecting other types of ions in a mixture is called Ion Suppression. In the case of using surfactant as a reagent for dissolving proteins, surfactants can suppress analyte signals in the ionization process, resulting in poor performance. Because of the strong tendency of surfactant molecules to interfere with the detection of analyte molecules in MS, selection of a type of surfactant to be used to dissolve proteins must be carefully considered.[35] In addition, removal of surfactants before injecting the sample into the mass spectrometer for analysis is often required to improve the detection of proteins or peptides.

After proteins are extracted from a sample or solubilized from a protein pellet, they are subjected to chemical or enzymatic degradation to form peptides. This step is needed, as the current mass spectrometric technology can most effectively sequence peptides with molecular weights of less than 3000 Da. There are only a few chemicals having the ability to cleave the amide bonds of the proteins with some specificity to form peptides. For example, CNBr is commonly used to cleave peptide bonds with Met at the C-terminus. However, there are a number of enzymes available for protein digestion. Trypsin is the most commonly used enzyme.[48] This enzyme assists in the hydrolysis of specific peptide bonds to form peptides with an Arg or Lys C-terminui. Compared to other enzymes such as chromotrypsin, trypsin has high specificity. In addition, Arg and Lys are distributed along the protein sequence in spaces that, after trypsin digestion, the peptides generated (also called tryptic peptides) have the molecular weights ranging from 600 to 3000 Da which is ideal for MS sequencing. Finally, the tryptic peptides containing the basic amino acids, Arg or Lys, at the C-terminus can be readily protonated in ESI or matrix-assisted laser desorption ionization (MALDI). Ionization efficiency of tryptic peptides is generally higher than peptides containing no Arg or Lys. For these reasons, trypsin is widely used.

1.3 Liquid Chromatography

After protein digestion, many peptides are generated. In the case of analyzing complex samples, such as a whole extract digest, tens or hundreds of thousands of peptides are expected to be produced. Analyzing a complex mixture of peptides directly by MS is a very challenging task. The problem lies in two fronts. High throughput mass spectrometers do not have the sufficient resolving power to resolve peptide ions of similar masses. In addition, peptides are not ionized with equal efficiency. The ion suppression effect discussed earlier applies here as well, i.e., in a mixture of peptides, only a few peptides with high concentrations and/or high ionization efficiencies will be preferentially ionized and detected. Other peptides in the mixture are suppressed during ionization. Thus, prior to MS detection, efficient separation of the peptides is required in proteome analysis.

There are several techniques available for peptide separation and the most common one used is high performance liquid chromatography (HPLC or LC).[48] LC separation is done based on the differences in how and how strong the peptides interact with a stationary phase packed in a column. There are a number of different stationary phases available and the surface chemistry of the stationary phase determines what separation mechanism is operative. Two LC separation techniques are most commonly used in proteome analysis. The first one is reversed phase (RP) LC. In RP separation, the stationary phase surface contains a non-polar functional group such as C18. Peptide interaction with the stationary phase is mainly through the hydrophobic forces. Peptide elution is done using solvent mixtures with gradual decrease in polarity (e.g., reducing the water content while increasing acetonitrile). The second technique is strong cation ion exchange (SCX) LC. In SCX separation, the LC column is made of stationary phase with surface chemistry containing anions such as $-SO_3^-$ groups. By lowering the peptide solution pH, most peptides in a mixture would be positively charged and thus they can interact with the SCX stationary phase via ionic interaction after injection in the column. Peptide elution is carried out using solvent mixtures with varying cation or salt contents (e.g., increasing the concentration of KCl in the elution solution).

To combine LC with MS, RPLC is used, as it uses a mobile phase compatible with ESI or MALDI. In addition, it offers high separation efficiency compared to other LC techniques such as SCX LC. However, RPLC alone may not be sufficient for separating a complex peptide mixture. To this end, multidimensional separation techniques have been developed.[19, 48] However, increasing the number of dimensions significantly increases the analysis time. At present, most of the shotgun proteome analysis methods are based on the use of two-dimensional (2D) LC where SCX is used as the 1st dimension and RP serves as the 2nd dimension. These two techniques are considered to be orthogonal, i.e., peptide separations are based on two different mechanism or two different properties of the peptides – SCX is for the ionic interaction separation and RP is for the hydrophobic interaction separation. Peptides having similar ionic interaction or retention time from the SCX column separation may have totally different hydrophobic properties, rendering them separable in the RP column.

There are two instrumental configurations for combining SCX with RPLC MS. One is on-line 2D-LC MS and another one is off-line 2D-LC MS (see Figure 1.4). The MudPIT (multidimensional protein identification technology) is one of the well known on-line 2D-LC setups (See Figure 1.4.A).[48] The column used in MudPIT consists of SCX material back-to-back packed with reversed phase material inside a fused silica capillary.[49] The reversed phase LC complements the SCX since it is efficient at removing salts and also compatible with ESI MS. The 2D chromatographic separation takes place in cycles. Each cycle comprises an increase in salt concentration to elute a portion of peptides out of the SCX material, followed by a RP gradient of increasing percentage of an organic solvent (acetonitrile) to progressively elute peptides into the ion source (see Figure 1.3. for the ESI interface), and then to the tandem mass spectrometer. Thus a complex peptide mixture can be separated, prior to sequencing by MS/MS, based on their unique physical properties of charge and hydrophobicity.

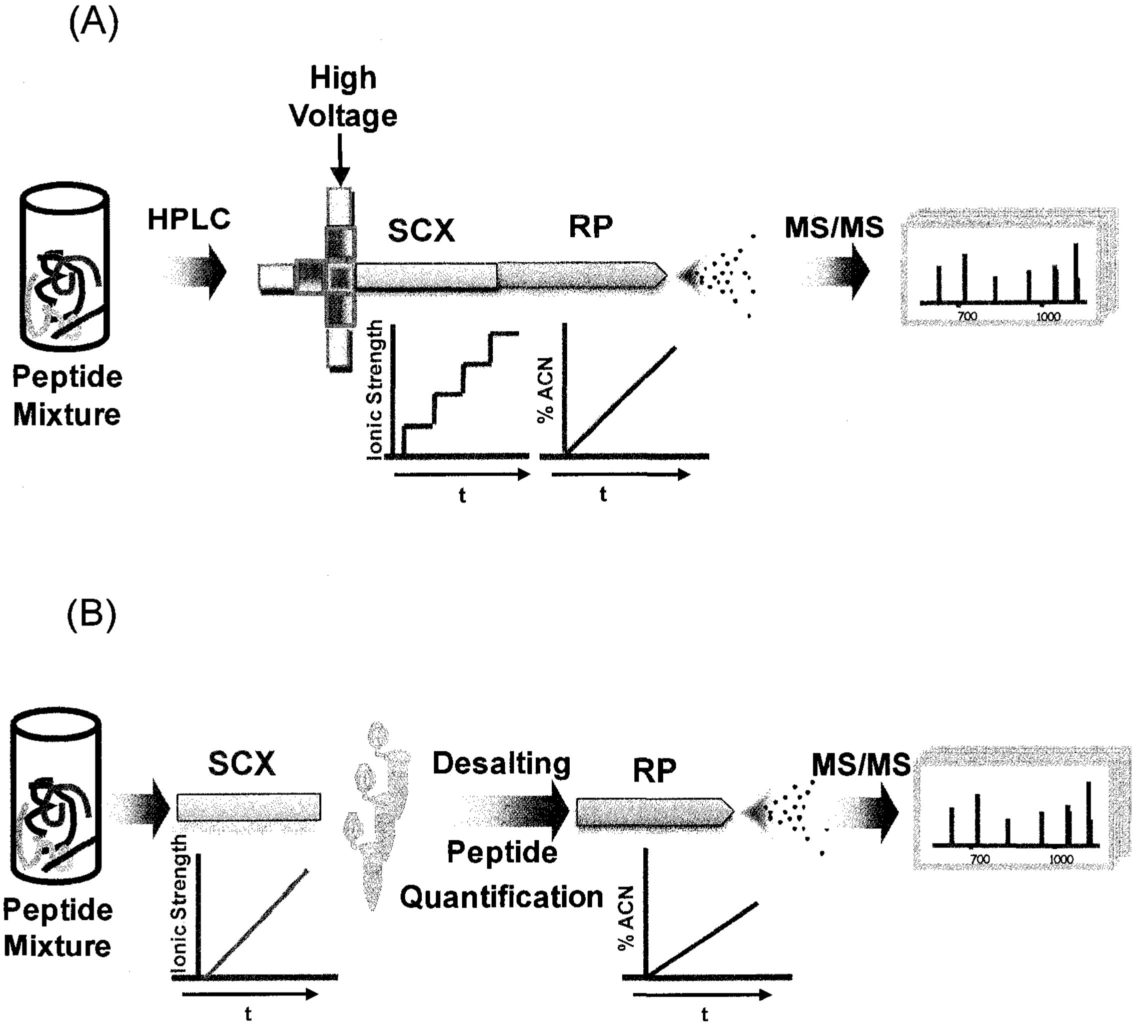


Figure 1.4 (A) On-line 2D-LC MS (MudPIT) and (B) off-line 2D-LC MS.

For the off-line 2D-LC MS (see Figure 1.4.B), instead of packing the SCX and RP materials into one capillary column, the peptide mixtures are first separated on a SCX column which has a relatively large sample loading capacity (e.g., using a column with ID >150 μm, as opposed to 75 μm column used in RPLC). Peptide fractions are automatically collected by a fraction collector commercially available from a HPLC manufacture. Each individual fraction is then loaded to a RPLC MS/MS instrument for sequencing. In Chapter 5 of this thesis, I will describe a new approach of off-line 2D-LC MS where the individual fractions collected from the SCX column are loaded into RPLC with a UV detector and auto-sample collector for desalting and peptide quantification. The amount of the peptides flushing through the RP column can be calculated based on their UV absorbance and a standard calibration curve. The collected peptide fractions are concentrated down to several microliters. An optimal amount of peptides is loaded onto the RPLC separation column and then introduced to the tandem mass spectrometer.

The relative merits of on-line and off-line 2D-LC MS/MS systems will be discussed in Chapter 5. Both methods are well suited for analyzing complex proteome samples such as whole cell lysates or tissues. They are also suitable for purification and analysis of low abundance proteins from highly complex biological matrices. The 2D-LC systems can be used for MALDI MS. In MALDI, an off-line interface such as a heated droplet interface is used to collect the LC fractions onto a MALDI plate in discrete spots.[50-52] During or after fraction collection, matrix is added to the sample spots. The plate is then inserted into a MALDI MS/MS instrument for peptide sequencing.

1.4 Tandem Mass Spectrometry (MS/MS)

In the past decade or so, several new tandem mass spectrometers have been introduced commercially, allowing their wide use in sequencing peptides.[53-56] In my thesis work, I used tandem quadrupole (Q) time-of-flight (TOF) mass spectrometers to generate fragment ion spectra of peptide ions.[55] Figure 1.5 shows the schematic diagrams of MALDI QTOF MS from ABI Sciex and Figure 1.6 shows the schematic diagrams of ESI QTOF MS from Waters.

The tandem mass spectrometer MALDI QqTOF MS (see Figure 1.5) was initially developed in the 1990s' in Standing's lab at the University of Manitoba and subsequently commercialized by ABI Sciex.[55] It can be described in a simple way as a triple quadrupole with the last quadrupole section replaced by a TOF analyzer. The Q refers to a mass-resolving quadrupole, the q refers to a radio-frequency (RF) only quadrupole or hexapole collision cell, and the TOF refers to a time-of-flight mass spectrometer. Ions generated from the MALDI source first pass through Q0, a quadrupole with RF only as ions focusing device in the source region, and enter the quadrupole mass analyzer (Q1). In the MS mode, the quadrupole Q2 is used as an ion focusing device (i.e., RF only and with no collision gas added to the device) which allows all the ions from Q1 to reach the ion modulator after focusing. A pulsed voltage is applied orthogonally to the Q direction, and forces the ions to enter the reflectron TOF and analyzed by a multichannel plate (MCP) detector. In the MS/MS mode, ions of certain m/z are first selected by Q1 (used as mass filter) and introduced to the collision cell (Q2) where collision induced dissociation (CID) occurs with a collision gas (N_2 or Ar). The product or fragment ions generated from CID are analyzed by the reflectron TOF.

The main advantages of this hybrid QqTOF instrument (also called QSTAR as a trade name by ABI Sciex), compared to other commonly used proteome analysis techniques such as quadrupole ion trap MS, are the high mass measurement accuracy (typical mass error < 50 ppm for both peptide and fragment ions) and high resolution (typical resolution of 10,000), resulting in unambiguous determination of charge state and very high specificity in database searches.

To generate MALDI MS and MS/MS spectra from QqTOF, a nitrogen laser emitting 337 nm radiation is operated at 20 Hz and the laser beam is directed via a fiber optical tube to the MALDI sample spot for desorption and ionization. The sample plate is set on a moving x-y stage. After a spectrum is collected from one sample spot (typically 1-100 s depending on the analyte concentration and signal strength), the plate is moved to a different position so that another sample spot is exposed to a laser beam for desorption. The processes of spot movement and spectral collection can be fully automated.

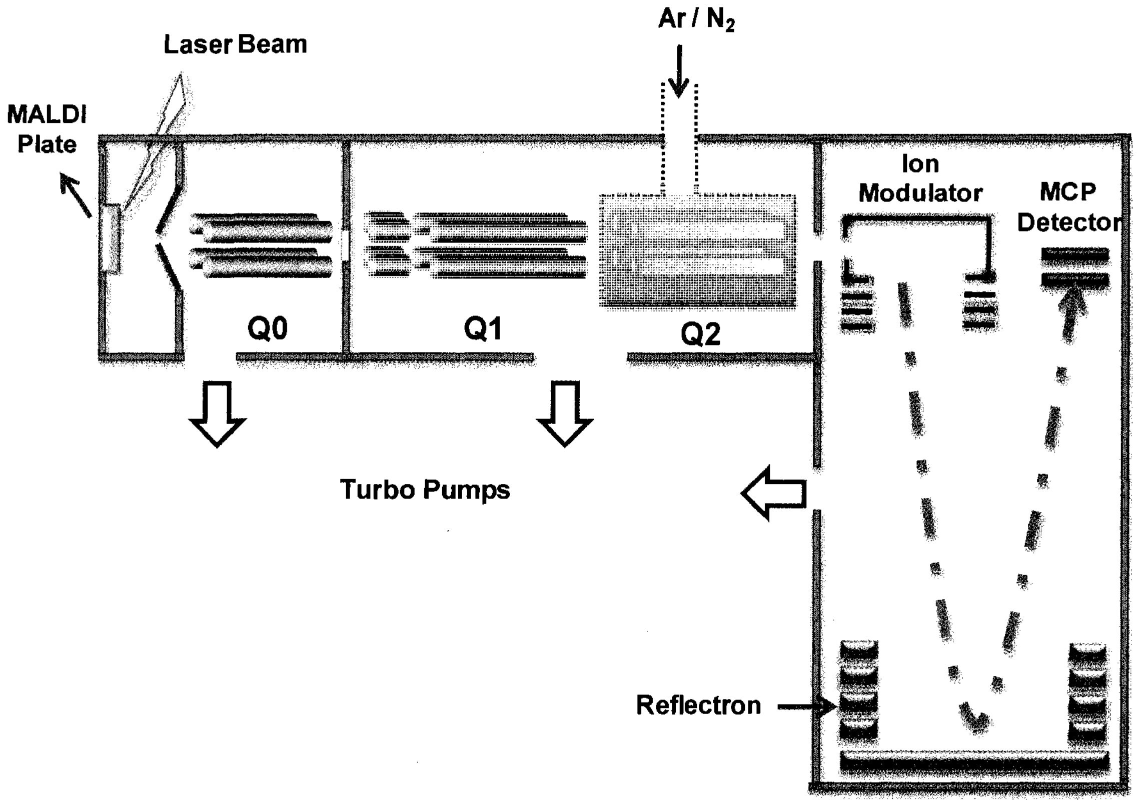


Figure 1.5 Schematic diagram of MALDI QTOF MS from ABI Sciex.

Figure 1.6 shows the schematic diagram of ESI Q-TOF Premier™ mass spectrometer from Waters. It is a hybrid orthogonal acceleration time-of-flight mass spectrometer which enables automated accurate mass measurement of precursor and fragment ions to yield high confidence in structural elucidation and database search results.[57] The Q-TOF Premier combines the high transmission efficiency of ZSpray™ source technology. In ZSpray, the ESI probe is placed orthogonally to the sample cone, instead of facing it straight. This also allows the sample cone to get less dirty than previous designs. The built-in NanoLockSpray™ capability enables routine accurate mass measurement in both MS and MS/MS modes (<30 ppm). The NanoLockSpray interface allows electrospray ionization to be performed in the flow rate range from 5 to 1000 nL/min (i.e., nanoflow).

The electrospray ionization takes place as a result of imparting a high voltage to the eluent as it emerges from the emitter (see Figure 1.3). An aerosol of charged droplets emerges from the emitter and undergoes a reduction in size by solvent evaporation until it has attained a sufficient charge density to allow sample ions to be ejected from the surface of the droplet.[47] For a given sample concentration, the ion currents observed in nanoflow are comparable to those seen in microlitre flow or microflow electrospray (i.e., flow rate of 1 to 50 μL/min). In nanoflow ESI, desolvation of the droplet is relatively easy and the spray cone is not as large as the microflow. Thus, the analyte sampling efficiency in nanospray is higher than in microspray. As a consequence, the sample consumption for nanoflow ESI is significantly less than microflow ESI. This is the main reason nanospray combined with nanoLC is commonly used for proteome analysis.

As in the ABI QSTAR system, the Premier Q-TOF hybrid mass spectrometer provides both quadrupole (MS1) and TOF mass analyzers with an intermediate collision cell for fragmentation if required. This combination allows ions to be selected, individually fragmented and then measured to a high degree of mass accuracy by the reflectron TOF. The quadrupole is available with 4 kDa mass range option and can be operated in three modes:

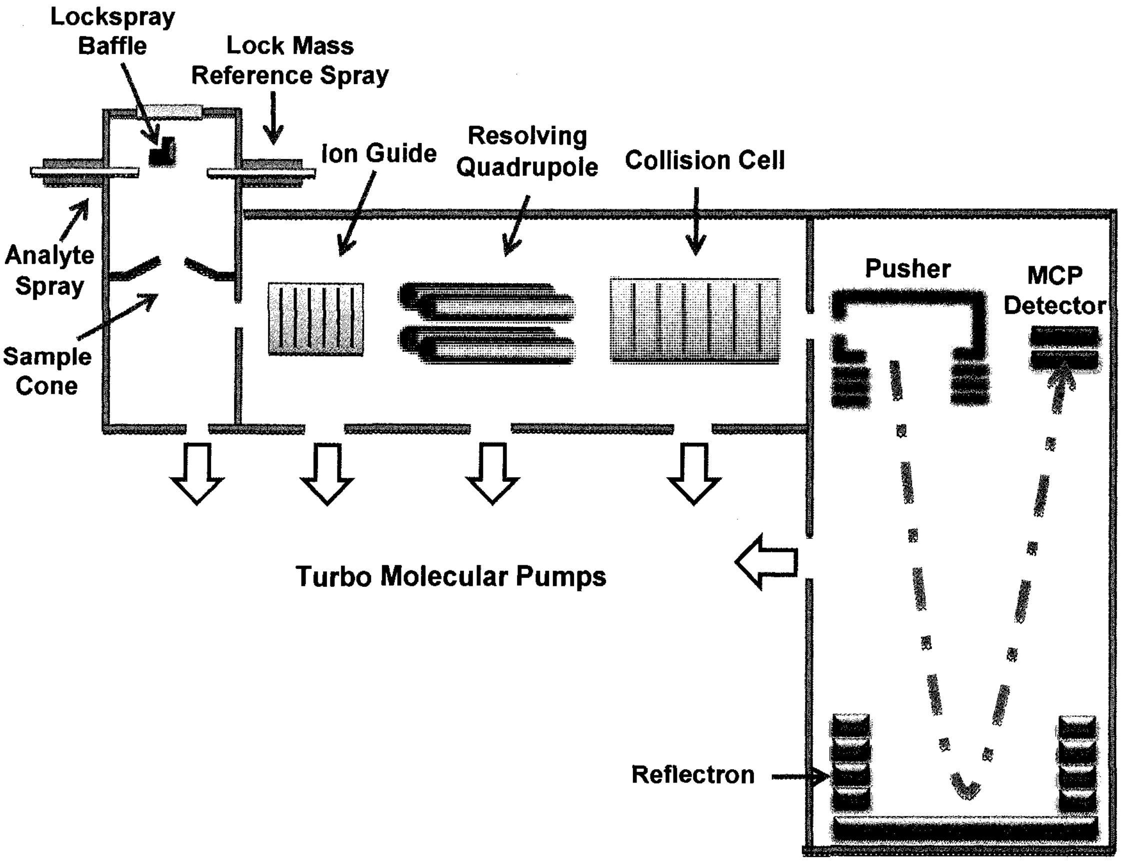


Figure 1.6 Schematic diagram of ESI QTOF MS from Waters.

1). When the quadrupole resolving DC is off (RF only), ions can pass through the quadrupole and be accurately measured by the TOF in what is known as the TOF MS acquisition.

2). When the quadrupole resolving DC is switched on, the quadrupole can either be parked on one specific mass (TOF MS/MS) or can be made to scan through a wide mass range in search of candidate ions for fragmentation (precursor ion scanning).

3). When the instrument is set to automatically switch between TOF MS and TOF MS/MS modes depending on the ions are detected during the TOF MS scan, this is known as data directed analysis (DDA).

The TOF analyzer uses a high voltage pulse to orthogonally accelerate the ions down the flight tube and a reflectron to reflect them back towards the MCP detector (V-optics). A mass spectrum can be generated with a resolution of 10,000.

It should be noted that, in the literature, there are several abbreviations used for describing the quadrupole time-of-flight mass spectrometer (e.g., QTOF, QqTOF, Qq-oaTOF, QTof, etc.). Some of them are trade names. In my thesis, I will generally use QTOF to refer to the quadrupole time-of-flight mass spectrometer.

1.5 Peptide Ion Fragmentation

One of the most popular methods for peptide ion fragmentation in gas phase in a tandem mass spectrometer is collision-induced dissociation (CID).[58] In CID, the molecular or precursor ions are accelerated by an electrical field to high kinetic energy in the vacuum. Then they collide with neutral gas species (e.g., He, N_2 or Ar) in the collision cell of the mass spectrometer. During the collision, part of the kinetic energy is converted into internal energy of the precursor ion. This results in decomposition of the ions and fragmentation of the precursor ion into smaller fragment ions. These fragment ions can then be analyzed by a mass spectrometer. CID can be performed at either high or low collision energies. Low energy CID (10-100 eV) is widely used in most mass spectrometers (triple quadrupole, ion trap, QTOF) for proteome analysis.[58]

For peptide ion fragmentation under CID, the fragment ions are mostly produced by the dissociation of the peptide backbone. The fragmentation pattern of peptide ions in the gas phase at low energy CID is dominated by fragment ions resulted from cleavage of the amide bonds. Several bonds along the peptide backbone can possibly be broken during CID (see Figure 1.7A). The most common ion types are the *b* and the *y* ions, which refer to the fragmentation at the amide bond with charge retention on the N or C terminus, respectively. The nomenclature differentiates fragment ions according to which end of the fragment retains a charge after fragmentation and where the bond breakage occurs.[58] If the charge associated with the peptide ion retains on the amino-terminal side of the amide bond, the fragment ion is named a *b* ion (See Figure 1.7 B). If the charge retains on the carboxyl-terminal of the broken amide bond, this ion is then named a *y* ion (See Figure 1.7 B). In theory, every peptide bond can be fragmented into a *b* ion or a *y* ion. Therefore, subscripts are used to designate the specific amide bond fragmented to generate the observed fragment ions: *b* ions are designated by a subscript that reflects the number of amino acid residues present on the fragment ion counted from the amino-terminus, whereas the subscript of *y* ions indicates the number of amino acids present, counting from the carboxyl-terminus.

1.6 Protein Identification Based on MS and MS/MS

There are generally two ways to identify a protein based on mass spectrometric results, namely peptide mass fingerprinting (PMF) and MS/MS search. The PMF is a method developed in the early 1990s by several groups for protein identification.[59-63] As shown in Figure 1.8, the unknown protein of interest is first digested by a chemical or protease with high sequence specificity, such as CNBr or trypsin, to produce a set of peptides. MALDI or ESI MS is used to determine the accurate masses of the peptides. MALDI TOFMS is usually more favorable for PMF because MALDI produces exclusively singly charged ions for low mass peptides. In addition, MALDI can tolerate buffers and salts in the samples. The mass list is compared *in silico* to the database containing the known protein. During the database searching, the search engine uses computer programs to theoretically cut the known proteins in a database (e.g., the entire set of human proteins predicted from the genome) into peptides according to the

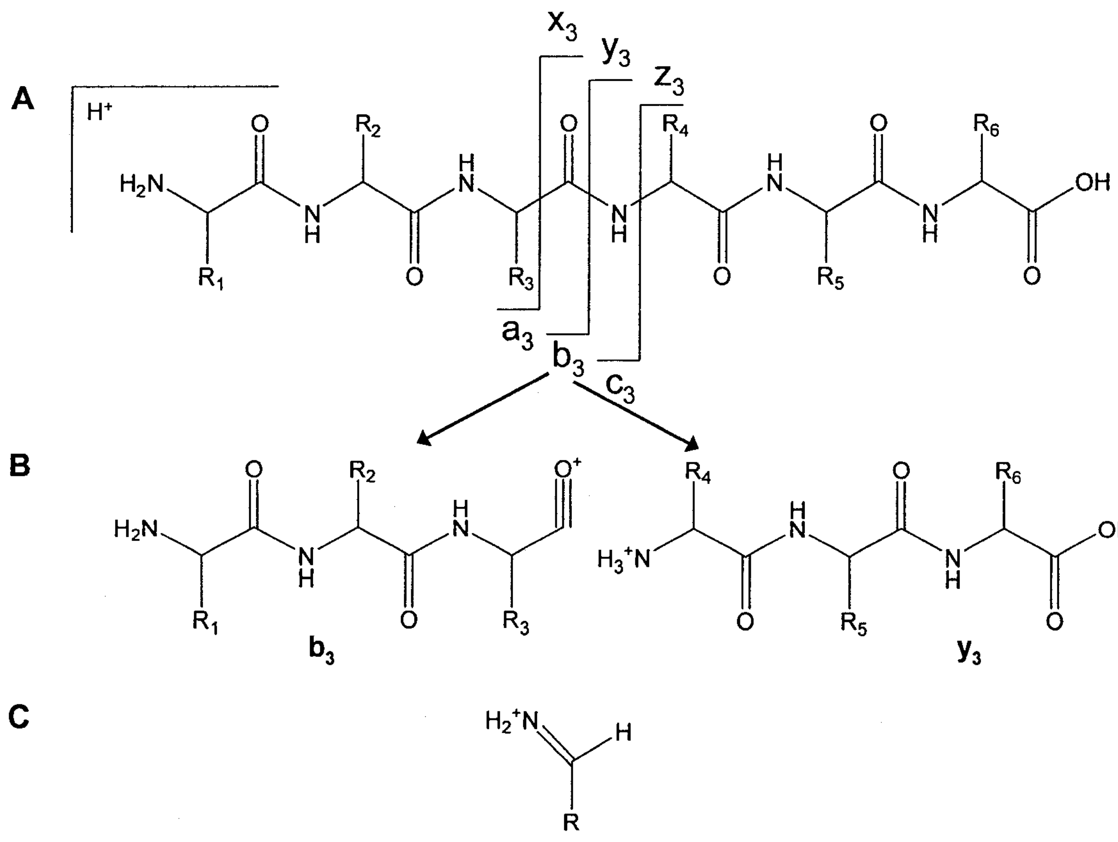


Figure 1.7 (A) CID fragmentation pattern of a peptide ion, (B) an example of b ions and y ions, and (C) an example of immonium ions.

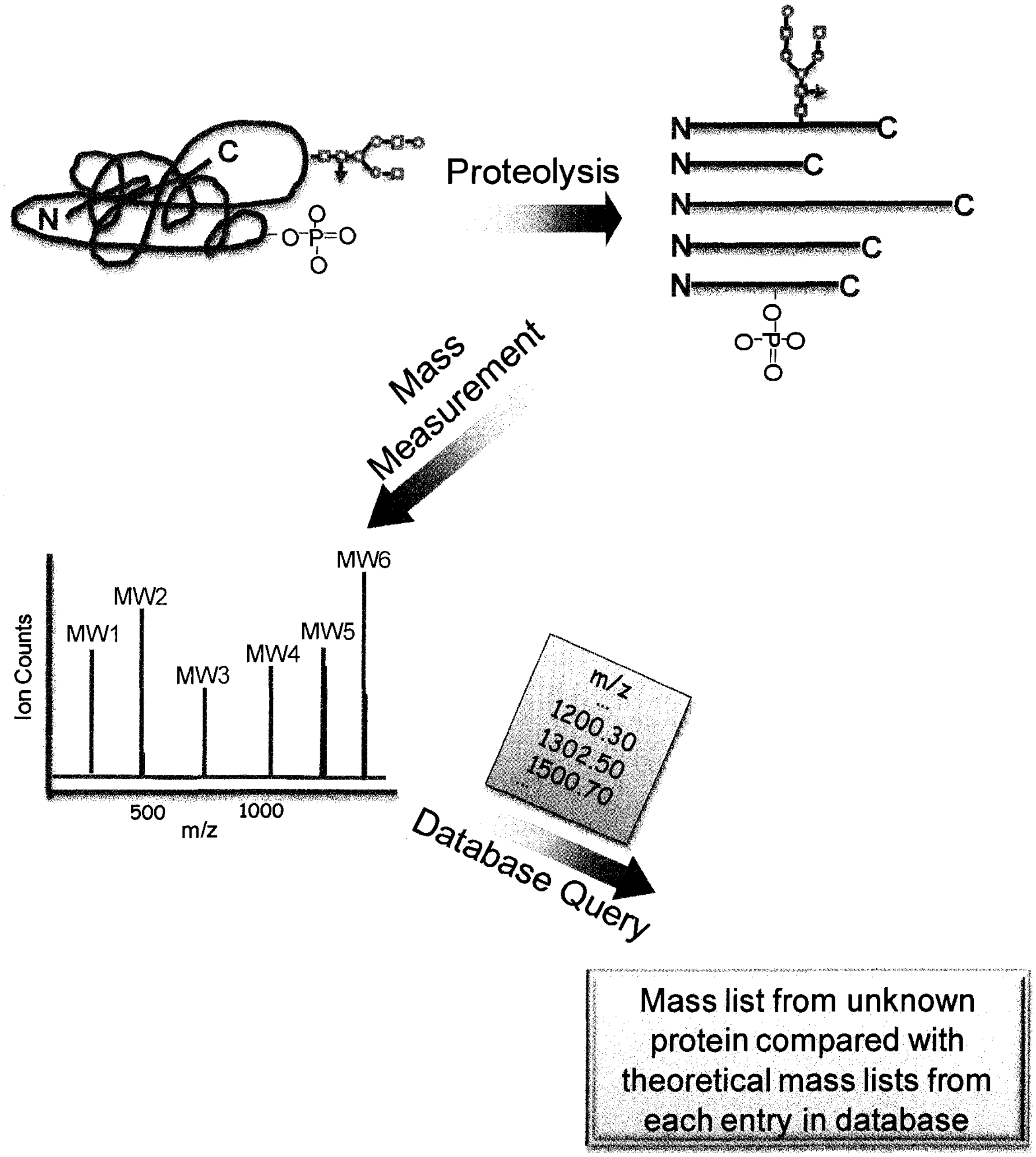
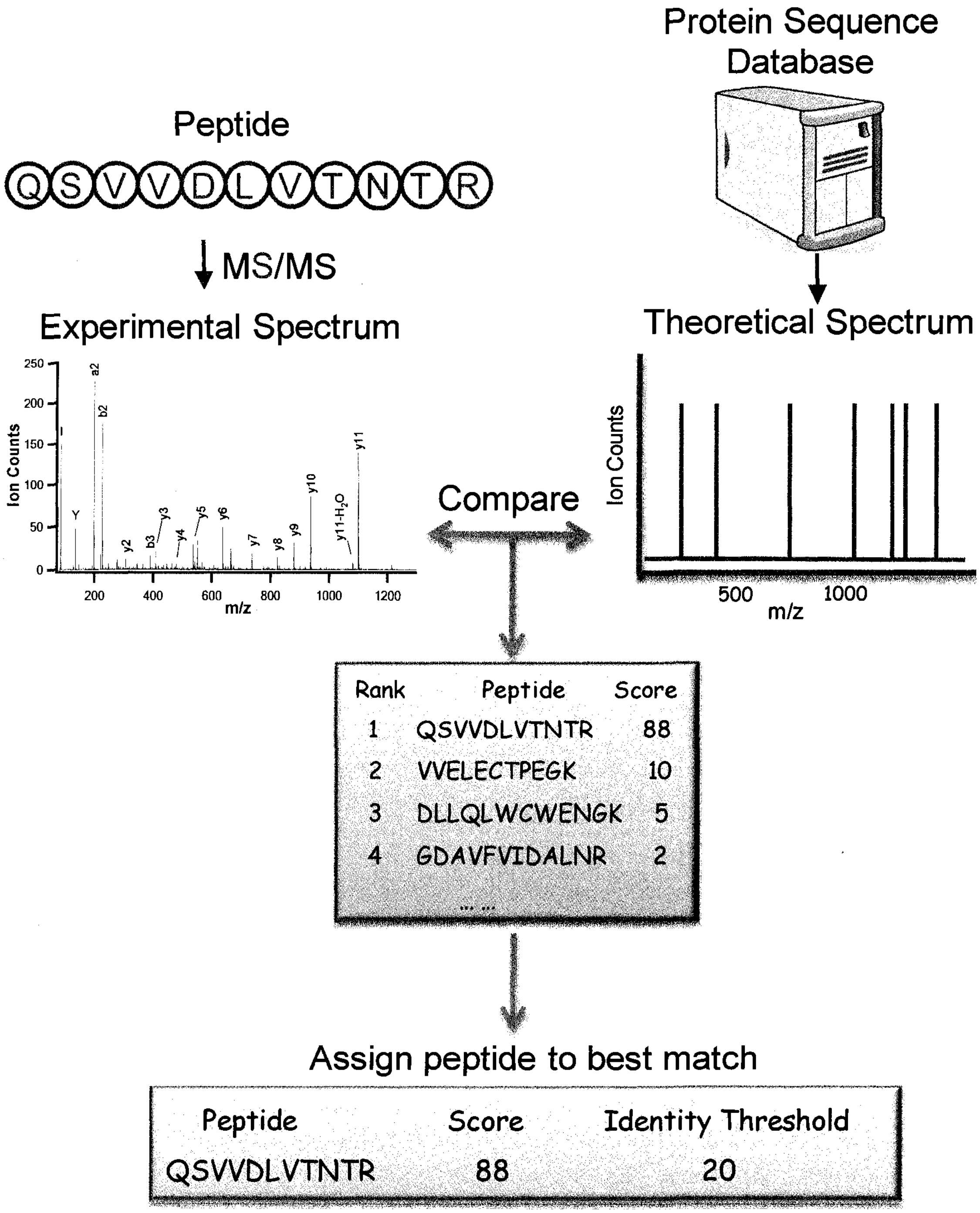


Figure 1.8 Workflow of peptide mass fingerprinting.

proteases used. Then it calculates the absolute masses of the peptides from each protein. Afterwards it compares the peptide masses of the unknown protein (i.e., the mass list entered into the search engine) to the theoretically calculated peptide masses of each protein in the database. The best match is determined statistically. Search engines on the Internet include MASCOT (http://www.matrix-science.com) PeptIdent (http://ca.expasy.org/tools/peptident.html), MOWSE (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse), MS-FIT (http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm) and PeptideSearch (http://www.mann.embl-eidelberg.de/GroupPages/PageLink/peptidesearchpage.html). In all search engines, a few search parameters are used to define the confidence of protein identification, such as the number of masses submitted, the accuracy of peptide mass determination, the number of matched peptides, the sequence coverage and the size of the sequence database.

The advantage of this method is that only the masses of the peptides have to be known, which potentially makes it high throughput. The masses of the peptides can be used as a fingerprint with great specificity. Therefore it is often possible to identify the protein from this information alone. However, one major disadvantage is that peptide mass mapping alone is not sufficient for reliable identification of a protein in many cases, particularly when dealing with protein mixtures or low abundance proteins of which only a few peptides can be detected. In such cases, additional information, such as peptide fragment information is required for a confident identification.

Proteins can be identified by tandem MS analysis of their peptides (see Figure 1.9).[64-67] Since a tandem mass spectrum or MS/MS spectrum contains the amino acid sequence information of the peptide, rather than the peptide masses alone, these searches normally generate more specific and discriminative results than PMF. In generating the MS/MS spectra from a tandem mass spectrometer, the first MS scan assigns each peptide a mass/charge ratio, performs on-the-fly data process and ranks the peptide ions according to their relative intensities (this spectrum is sometimes called survey scan). Then it will pick the four or five most intense peptide ions for MS/MS analysis. It first selects the most intense peptide ion, fragments it in the collision cell during the MS/MS scan, and then collects a MS/MS of the peptide ion which serves as a unique "fingerprint"



Rank	Peptide	Score
1	QSVVDLVTNTR	88
2	VVELECTPEGK	10
3	DLLQLWCWENGK	5
4	GDAVFVIDALNR	2
... ...		

Peptide	Score	Identity Threshold
QSVVDLVTNTR	88	20

Figure 1.9 Workflow of peptide sequencing by tandem MS and database search.

of the peptide. Afterwards, the mass spectrometer selects the second most intense ion, does MS/MS scan, and then analyze the third and the fourth most intense ions. Upon completion of MS/MS spectral collection, the next survey scan is performed, followed by four MS/MS scans. This cycle continues till the LC separation of peptides is completed. Then all the acquired MS/MS spectra from the LC-MS/MS run are processed and imported into a MS/MS database search engine for protein identification.

Many different algorithms for database searching of these un-interpreted MS/MS spectra have been developed. They include SEQUEST (http://fields.scripps.edu/sequest/), MASCOT (http://www.matrixscience.com/) and X!Tandem (http://www.thegpm.org/TANDEM/). During the database search, experimental fragment ion spectra or MS/MS spectra are matched against theoretical fragment ion spectra for all the peptides in the databases that have the same precursor ion mass within the experimental error. Peptides that turn out to be the first hits along with the identification scores equal or higher than the identity threshold defined by the database are generally considered as positive matches. The matched peptides are sequence-linked to their corresponding proteins, resulting in the identification of proteins. The protein identification results can be exported in several formats (e.g., .xls, .xml, .csv) together with their individual peptide sequence, m/z values, molecular weight, peptide score, mass error, etc. to a spreadsheet for further processing or data presentation.

While MS/MS search for protein identification is fully automated, manual interpretation of the MS/MS spectra can sometimes be useful.[68-70] The fragmentation reaction of peptides by low energy CID is believed to be initiated by a mobile proton and directed by a charge-site.[58] In the absence of strongly basic residue (Arg), the migration of a mobile proton to carbonyl oxygen or amide nitrogen initiates the cleavage of various peptide bonds via a cyclic intermediate. Residues that tend to localize mobile protons, for example, His, or donate protons (Asp, Glu) will have a significant effect on peptide fragmentation. In addition, a selective cleavage of the protonated peptide bonds are often observed in the N-terminal side of Pro, C-terminal side of Asp, Glu due to charge localization or donation. Knowledge of the selective cleavage site is important for predicting the fragment pattern of the MS/MS spectra of a peptide, and can increase the protein identification confidence where a manual search is required for the MS/MS spectra that are not in good quality. Some of the rules used for manual interpretation in my thesis work will be highlighted in the Experimental sections of the relevant chapters.

1.7 Scope of the Thesis

In Chapters 2 and 3, I will describe our work related to proteome analysis of human tear fluid and zebrafish liver, respectively. In Chapters 4 to 6, I will focus on the discussion of three new techniques or methods developed for improving protein identification in shotgun proteome analysis. In Chapters 7 and 8, I will demonstrate the applications and analytical performances of the newly developed techniques for zebrafish liver proteome analysis and for comprehensive analysis of the *E. coli* proteome, respectively. In Chapter 9, I will describe the combination of microwave-assisted acid hydrolysis of proteins with the newly developed LC-ESI QTOF MS method for protein sequence analysis. The thesis ends with a conclusion chapter (Chapter 10) where I also briefly comment on future work related to my research.

1.8 Literature Cited

(1) Yates, J.; Osterman, A. *Chemical Reviews* **2007**, *107*, 3363-3366.

(2) Wu, L.; Han, D. *Expert Review of Proteomics* **2006**, *3*, 611-619.

(3) Leitner, A.; Lindner, W. *Proteomics* **2006**, *6*, 5418-5434.

(4) Sprenger, R.; Horrevoets, A. *Proteomics* **2007**, *7*, 2895-2903.

(5) Dworzanski, J.; Snyder, A. *Expert Review of Proteomics* **2005**, *2*, 863-878.

(6) Kito, K.; Ito, T. *Current Genomics* **2008**, *9*, 263-274.

(7) Smith, J.; Figeys, D. *Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire* **2008**, *86*, 137-148.

(8) Rivera-Monroy, Z.; Bonn, G; Guttman, A. *Current Organic Chemistry* **2008**, *12*,

424-440.

(9) Isserlin, R.; Emili, A. *Current Opinion in Molecular Therapeutics* **2008**, *10*, 231-242.

(10) Kocher, T.; Superti-Furga, G. *Nature Methods* **2007**, *4*, 807-815.

(11) Tekirian, T.; Thomas, S.; Yang, A. *Expert Review of Proteomics* **2007**, *4*, 573-583.

(12) Paradela, A.; Albar, J. *Journal of Proteome Research* **2008**, *7*, 1809-1818.

(13) Hoorn, E.; Pisitkun, T.; Yu, M.; Knepper, M. *Proteomics in Nephrology - towards Clinical Applications* **2008**, *160*, 172-185.

(14) Yang, W.; Steen, H.; Freeman, M. *Proteomics* **2008**, *8*, 832-851.

(15) Mathy, G.; Sluse, F. *Biochimica Et Biophysica Acta-Bioenergetics* **2008**, *1777*, 1072-1077.

(16) Matt, P.; Fu, Z.; Fu, Q.; Van Eyk, J. *Physiological Genomics* **2008**, *33*, 12-17.

(17) Ahmed, F. *Expert Review of Proteomics* **2008**, *5*, 469-496.

(18) Carpentier, S.; Panis, B.; Vertommen, A.; Swennen, R.; Sergeant, K.; Renaut, J.; Laukens, K.; Witters, E.; Samyn, B.; Devreese, B. *Mass Spectrometry Reviews* **2008**, *27*, 354-377.

(19) Tomas, R.; Kleparnik, K.; Foret, F. *Journal of Separation Science* **2008**, *31*, 1964-1979.

(20) Wu, C.; Maccoss, M. *Current Opinion in Molecular Therapeutics* **2002**, *4*, 242-250.

(21) Bogdanov, B.; Smith, R. *Mass Spectrometry Reviews* **2005**, *24*, 168-200.

(22) Ge, Y.; Lawhorn, B.; Elnaggar, M.; Strauss, E.; Park, J.; Begley, T.; Mclafferty, F. *Journal of the American Chemical Society* **2002**, *124*, 672-678.

(23) Patrie, S.; Ferguson, J.; Robinson, D.; Whipple, D.; Rother, M.; Metcalf, W.; Kelleher, N. *Molecular & Cellular Proteomics* **2006**, *5*, 14-25.

(24) Zabrouskov, V.; Giacomelli, L.; Van Wijk, K.; Mclafferty, F. *Molecular & Cellular Proteomics* **2003**, *2*, 1253-1260.

(25) Patrie, S.; Charlebois, J.; Whipple, D.; Kelleher, N.; Hendrickson, C.; Quinn, J.; Marshall, A.; Mukhopadhyay, B. *Journal of the American Society for Mass Spectrometry* **2004**, *15*, 1099-1108.

(26) Boyne, M.; Pesavento, J.; Mizzen, C.; Kelleher, N. *Journal of Proteome Research* **2006**, *5*, 248-253.

(27) Frank, A.; Pesavento, J.; Mizzen, C.; Kelleher, N.; Pevzner, P. *Analytical Chemistry* **2008**, *80*, 2499-2505.

(28) Cooper, H.; Hakansson, K.; Marshall, A. *Mass Spectrometry Reviews* **2005**, *24*, 201-222.

(29) Parks, B.; Jiang, L.; Thomas, P.; Wenger, C.; Roth, M.; Boyne, M.; Burke, P.; Kwast, K.; Kelleher, N. *Analytical Chemistry* **2007**, *79*, 7984-7991.

(30) Pesavento, J.; Bullock, C.; Leduc, R.; Mizzen, C.; Kelleher, N. *Journal of Biological Chemistry* **2008**, *283*, 14927-14937.

(31) Chang, M.; Ji, Q.; Zhang, J.; El-Shourbagy, T. *Drug Development Research* **2007**, *68*, 107-133.

(32) Granvogl, B.; Ploscher, M.; Eichacker, L. *Analytical and Bioanalytical Chemistry* **2007**, *389*, 991-1002.

(33) Lill, J.; Ingle, E.; Liu, P.; Pham, V.; Sandoval, W. *Mass Spectrometry Reviews* **2007**, *26*, 657-671.

(34) Kyte, J.; Doolittle, R. *Journal of Molecular Biology* **1982**, *157*, 105-132.

(35) Speers, A.; Wu, C. *Chemical Reviews* **2007**, *107*, 3687-3714.

(36) Punta, M.; Forrest, L.; Bigelow, H.; Kernytsky, A.; Liu, J.; Rost, B. *Methods* **2007**, *41*, 460-474.

(37) Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E. *Journal of Molecular Biology* **2001**, *305*, 567-580.

(38) Kashino, Y. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **2003**, *797*, 191-216.

(39) Carboni, L.; Piubelli, C.; Righetti, P.; Jansson, B.; Domenici, E. *Electrophoresis* **2002**, *23*, 4132-4141.

(40) Jones, M. *International Journal of Pharmaceutics* **1999**, *177*, 137-159.

(41) Hinze, W.; Pramauro, E. *Critical Reviews in Analytical Chemistry* **1993**, *24*, 133-177.

(42) Zhang, N.; Li, L. *Rapid Communications in Mass Spectrometry* **2004**, *18*, 889-896.

(43) Loo, R.; Dales, N.; Andrews, P. *Protein Science* **1994**, *3*, 1975-1983.

(44) Snyder, A. *Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry* **1996**, *619*, 1-20.

(45) Fenn, J. *Annual Review of Physical Chemistry* **1996**, *47*, 1-41.

(46) Kebarle, P. *Journal of Mass Spectrometry* **2000**, *35*, 804-817.

(47) Kebarle, P.; Peschke, M. *Analytica Chimica Acta* **2000**, *406*, 11-35.

(48) Fournier, M.; Gilmore, J.; Martin-Brown, S.; Washburn, M. *Chemical Reviews* **2007**, *107*, 3654-3686.

(49) Wolters, D.; Washburn, M.; Yates, J. *Analytical Chemistry* **2001**, *73*, 5683-5690.

(50) Young, J.B.; Li, L. *Analytical Chemistry* **2007**, *79*, 5927-5934.

(51) Zhang, B.; Mcdonald, C.; Li, L. *Analytical Chemistry* **2004**, *76*, 992-1001.

(52) Young, J.B.; Li, L. *Journal of the American Society for Mass Spectrometry* **2006**, *17*, 325-334.

(53) De Hoffmann, E. *Journal of Mass Spectrometry* **1996**, *31*, 129-137.

(54) Douglas, D.; Frank, A.; Mao, D. *Mass Spectrometry Reviews* **2005**, *24*, 1-29.

(55) Chernushevich, I.; Loboda, A.; Thomson, B. *Journal of Mass Spectrometry* **2001**, *36*, 849-865.

(56) Hu, Q.; Noll, R.; Li, H.; Makarov, A.; Hardman, M.; Cooks, R. *Journal of Mass Spectrometry* **2005**, *40*, 430-443.

(57) Bristow, T.; Constantine, J.; Harrison, M.; Cavoit, F. *Rapid Communications in Mass Spectrometry* **2008**, *22*, 1213-1222.

(58) Wysocki, V.; Resing, K.; Zhang, Q.; Cheng, G. *Methods* **2005**, *35*, 211-222.

(59) Cottrell, J. *Peptide Research* **1994**, *7*, 115-&.

(60) Gevaert, K.; Vandekerckhove, J. *Electrophoresis* **2000**, *21*, 1145-1154.

(61) Johnson, R.; Davis, M.; Taylor, J.; Patterson, S. *Methods* **2005**, *35*, 223-236.

(62) Chen, G.; Pramanik, B. *Expert Review of Proteomics* **2008**, *5*, 435-444.

(63) Lahm, H.; Langen, H. *Electrophoresis* **2000**, *21*, 2105-2114.

(64) Nesvizhskii, A.; Aebersold, R. *Drug Discovery Today* **2004**, *9*, 173-181.

(65) Sadygov, R.; Cociorva, D.; Yates, J. *Nature Methods* **2004**, *1*, 195-202.

(66) Maccross, M. *Current Opinion in Chemical Biology* **2005**, *9*, 88-94.

(67) Delahunty, C.; Yates, J. *Methods* **2005**, *35*, 248-255.

(68) Peng, J.; Elias, J.; Thoreen, C.; Licklider, L.; Gygi, S. *Journal of Proteome Research* **2003**, *2*, 43-50.

(69) Boutilier, K.; Ross, M.; Podtelejnikov, A.; Orsi, C.; Taylor, R.; Taylor, P.; Figeys, D. *Analytica Chimica Acta* **2005**, *534*, 11-20.

(70) Chen, Y.; Kwon, S.; Kim, S.; Zhao, Y. *Journal of Proteome Research* **2005**, *4*, 998-1005.

Chapter 2

Characterization of Human Tear Proteome Using Multiple Proteomic Analysis Techniques*

2.1 Introduction

Tear film is a multi-functional interface between the external environment and the ocular surface. It consists of three layers: the outer lipid layer, the aqueous layer with soluble proteins, and the inner mucin layer, with each secreted by a different set of orbital glands. As a result, tear is a complex, physically heterogeneous fluid composed of many proteins, lipids, carbohydrates and electrolytes.[1] Proteins in the tear film are believed to play a central role in the innate defense of the ocular surface where they protect the external surface from potential pathogens and modulate the wound healing process.[2-5] Physiological factors have been shown to alter the balance of the protein components in the tear film.[6-8] This has triggered an increased interest in studying the protein composition of tear fluids as well as the relationship between protein composition and physiological variations.

Studies on tear proteins have been carried out using a wide range of techniques, including various types of immunological and enzymatic assays of mainly one or a few proteins of interest.[9,10] A number of chromatographic techniques, such as size exclusion high-performance liquid chromatography (HPLC),[11,12] reversed-phase HPLC[13] and ion-exchange HPLC,[13] have been used for tear protein isolation and analysis. Gel electrophoresis techniques, such as sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and two-dimensional (2D) gel electrophoresis, also have been used for tear protein separation.[14-17] The analysis of tear protein patterns by SDS-PAGE has been used as a diagnostic tool for the detection of dry eyes.[18-22] More recently, mass spectrometry (MS) based proteomic methods have been developed and applied for tear protein analyses. For example, using matrix-assisted laser desorption ionization time-

*This work was done under the supervision of Dr. Nan Li. Nan Wang contributed partially to LC ESI/MALDI data acquisition, data acquisition and figures. A form of this chapter was published as: N. Li, N. Wang, J. Zheng, X. M. Liu, O. W. Lever, P. M. Erickson, and L. Li, **2005**, "Characterization of Human Tear Proteome Using Multiple Proteomic Analysis Techniques", *J. Proteome Res.* **4**, 2052-2061.

of-flight (MALDI-TOF) MS, Mulvane and co-workers detected the low molecular weight substances in human tears.[23] MALDI MS was also used to detect proteins adsorbed onto contact lenses.[24,25] Surface-enhanced laser desorption ionization (SELDI)TOF ProteinChip technology was employed to compare the tear profiles before and after a scheduled surgery of an eye,[26] to profile tear proteins from dry eye patients[27] and to yield tear protein patterns for diagnosis of Sjögren's syndrome.[28] Liquid chromatography (LC) electrospray ionization (ESI) MS has been developed to determine the status of the ocular surface by detection of the protein components of tears.[5] In that report, a mass increment of 203 Da in the molecular ion region was observed in protein molecular mass measurements.[5] No MS/MS experiments were performed, but protein glycosylation by GlcNAc or GalNAc (203 Da) was speculated.[5] Fung et al. reported 2D-gel analysis of tear proteins.[29] They identified and characterized different forms of lacrimal-specific proline-rich protein using MALDI MS and LC-ESI MS/MS.[30] A truncated form of this protein was found to be present at a significant level in human tears.[30]

Despite the reported success of these studies, a sensitive and reliable method of protein identification and characterization from a single tear collection is still needed to handle clinical samples in a case study or a population based study. Tear proteome is dynamic and very sensitive to many factors, such as sample collection method,[31-34] age of the person,[35] and physiological states.[6-8] Inter-day variation of protein gel patterns has been reported.[36] Consequently, analytical procedures that require a large volume of tear fluids, obtained by the pooling of samples of different batches or groups, may not be useful for analyzing clinically relevant samples. For example, in a time-course study of proteome changes before and after wearing a contact lens, individual tear samples would be taken at a given time interval and their proteome patterns would be analyzed without sample pooling to achieve good time resolution. In a population-based study, the cost of collecting a large volume of tear fluids from an individual can be a major concern and thus single tear collection is desirable. Another consideration is related to the type of tears collected. Significant differences exist among reflex, open-eye and closed-eye tears in terms of both the protein composition and sample size obtainable from one single collection.[37] Protein concentrations increase in the order: reflex, open-eye, and closed-eye tears, while the single collectable sample volume decreases in that order.[37] On average, about 5 µL closed-eye or open-eye tears can be collected from an individual. Our research goal is to develop proteomic techniques that can generate as extensive as possible information on tear proteins using this low volume of sample.

While small volumes of biological samples (i.e., <5 µL) can be handled by techniques such as MALDI-TOF MS and SELDI MS, these methods do not provide protein identification information, which is critical for functional studies and biomarker discovery. A recent report by Kim and co-workers[38] has demonstrated the feasibility of combining 2D gel electrophoresis and ESI MS to study tear protein expression changes in blepharitis patients using only 2-3 µL of tears. They detected hundreds of protein spots in a 2D-gel image and a dozen of down- or up-regulated proteins in different patients were identified using an in-gel digestion/ESI MS proteomic method. However, no PTM characterization was reported. It is conceivable that different disease states may differ in type or extent of PTM. Thus, detection and characterization of PTMs should be as important, or even more important than protein identification.

In this study, we have examined several proteome analysis techniques and explored the complementary natures of these techniques to identify proteins and characterize PTMs using less than 5 µL of tears. These include SDS-PAGE of tear proteins with in-gel digestion of protein spots followed by peptide mass mapping and MALDI MS/MS sequencing, and in-solution digestion of tear fluids followed by LC-ESI MS/MS and LC-MALDI MS/MS. We show that, using less than 5 µL of reflex tear fluid, a total of 54 proteins could be identified with high confidence. In particular, 44 of these proteins were identified by LC-MALDI MS/MS and a range of PTMs on several proteins were observed and some of them could be characterized by this technique. We envision that LC-MALDI MS/MS, with further development in the areas of protein quantitation and fragmentation pattern recognition for PTM detection, has the potential to become a powerful high-throughput technique for tear proteomics.

2.2 Experimental

2.2.1 Chemicals and Reagents

α-Cyano-4-hydroxycinnamic acid (CHCA), 2,5-dihydroxybenzoic acid (DHB), bovine serum albumin (BSA), bovine trypsin, dithiothreitol (DTT), iodoacetamide, trifluoroacetic acid (TFA), and sodium dodecyl sulfate (SDS) were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). HPLC-grade acetonitrile was from Fisher Scientific Canada (Edmonton, Canada). Water was obtained from a Milli-Q Plus purification system (Millipore, Bedford, MA). CHCA was recrystallized from ethanol (95%) before use. The open-eye tear fluids were collected through a standard capillary collection technique using a 10 μL capillary tube[39] and were stored at -20 °C. The collection time from an individual was less than 1 min. An informed consent was obtained from the volunteer and ethics approval for this work was obtained from the University of Alberta (Arts, Science & Law Research Ethics Board certificate TJ-0605-280).

2.2.2 SDS-PAGE, In-gel Digestion, Peptide Mass Mapping and Sequencing

SDS-PAGE was carried out in a Bio-Rad mini-PROTEAN 3 system using 4%/12% stacking/separating polyacrylamide mini-gels. The SDS sample buffer contained 2% mercaptoethanol (v/v), 1% SDS, 12% glycerol, 50 mM Tris-HCl and a trace amount of bromophenol blue. Prior to electrophoresis, 1 μL of the tear sample was mixed with 9 μL sample buffer and heated at 95 °C for 5 min. Protein bands were detected by silver staining and visible bands were excised for trypsin digestion. The gel pieces were cut into small segments and then washed with water for 20 min and dehydrated in acetonitrile until the gel pieces became white. Reduction and alkylation were then applied in 20 mM DTT and 50 mM iodoacetamide, respectively. After dehydration, gel pieces were covered with 10 ng/L trypsin in 0.1 M NH_4HCO_3 and 2 mM $CaCl_2$ for overnight digestion at 30 °C. Peptides were extracted using 25%, 50%, 75%, and 100% acetonitrile in 0.1% TFA with 20 min of shaking each time. The pooled extracts were Speed-vac (Thermo Savant, Milford, MA) dried to a final volume of approximately 1 μL for all the sample bands.

A two-layer sample preparation method was employed for MALDI analysis of the tryptic digests. Peptide mass mapping using MALDI MS was first carried out and tentative protein identifications were obtained by database search using the MASCOT search engine. MALDI MS/MS was then applied to all the digests either to confirm the identification or identify the protein directly.

2.2.3 In-Solution Digestion, LC-ESI and LC-MALDI MS/MS

A 5-μL portion of tear sample was mixed with 2.5 μL 0.2 μg/μL bovine trypsin and 2 μL 20 mM $CaCl_2$ after reduction and alkylation with 20 mM DTT and 50 mM iodoacetamide, respectively. After overnight trypsin digestion at 30 °C, tear digests were desalted using three C_{18} ZipTips (Millipore, Bedford, MA). The eluates were then dried to evaporate the organic solvent and diluted to a final volume of 20 μL with 0.1% TFA/H_2O solution for the following LC-ESI MS/MS and LC-MALDI MS/MS experiments.

In the LC-ESI MS/MS experiments, 2 μL of the tryptic digest solution (equivalent to 0.5 μL of the starting tear fluid sample) was injected and separated by a 150 μm × 150 mm C_{18} column (particle size 3 μm, pour size 300 Å, Vydac, Hesperia, CA) followed by detection in a Finnigan LCQ Deca ESI ion trap mass spectrometer. A flow-rate of 1 μL/min was used for all separations. Gradient elution was performed with solvent A (0.1%, v/v, aqueous acetic acid) and B (0.1%, v/v, acetic acid in acetonitrile). In the 120-min run, the gradient was as follows: 0 min: 5%B, 5 min: 5%B, 10 min: 10%B, 80 min: 35%B, 110 min: 80%B, 120 min: 5%B. Four MS/MS scans were applied after one MS scan. In the 200-min run, the gradient was as follows: 0 min: 5%B, 5 min: 5%B, 10 min: 10%B, 100 min: 25%B, 160 min: 35%B, 180 min: 80%B, 190 min: 80%B, 200 min: 5%B. Each MS scan was followed by 5 MS/MS scans.

LC-MALDI MS and MS/MS were done through the use of an in-house developed offline heated-droplet LC-MALDI interface.[40] Peptide separation was first performed on an Agilent (Palo Alto, CA) 1100 series capillary HPLC equipped with an auto sampler. Chromatographic analysis was performed using 8 μL tear digest (equivalent to 2 μL of the starting tear fluid) on a 1.0 mm × 150 mm Vydac C_{18} reversed-phase column (Vydac, Hesperia, CA). A flow-rate of 40 µL/min was used for separation. Gradient elution was performed with solvent A (0.1%, v/v, aqueous TFA) and B (0.1%, v/v, TFA in acetonitrile) using the program: 0 min: 0%B, 5 min: 0%B, 15 min: 10%B, 75 min: 25%B, 90 min: 40%B, 100 min: 80%B, 105 min: 0%B, 110 min: 0%B. The UV detector wavelength was set at 210 nm. Fractions from 11 to 110 min were collected onto a 100-well MALDI plate (Applied Biosystems, Foster City, CA). After the plate was cooled to room temperature, 1 µL of 80 mg/mL DHB 50% acetonitrile/water (v/v) matrix solution was added on top of each spot and allowed to air-dry. The sample spots were then ready for both MALDI MS and MALDI MS/MS analysis. The high abundance peptide peaks that had already been identified from the SDS-PAGE MALDI MS and LC-ESI MS/MS experiments were manually excluded from MALDI MS/MS sequencing. On average, approximately 20 MS/MS spectra were obtained from each MALDI spot.

2.2.4 Mass Spectrometers

MALDI MS experiments were carried out on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany) using the reflectron mode of operation. Ionization was performed with a 337-nm pulsed nitrogen laser. A Thermo-Finnigan LCQ Deca ion trap instrument equipped with a Surveyor LC system (San Jose, CA) was used for the LC-ESI MS/MS experiments. MALDI MS/MS experiments were carried out by using an MDS Sciex QSTAR Pulsar QqTOF mass spectrometer equipped with an orthogonal UV-MALDI source (Concord, ON, Canada).

2.2.5 Database Search

Mascot search engine against both SwissProt and NCBInr was used for all database searches in this study. For peptide mass mapping, the peptide mass tolerance was 100 ppm. For the .dta files generated from the LCQ Deca instrument, the precision tolerance for LCQ MS/MS data was 1.4 Da for the parent peptides and 0.8 Da for the fragment ions in database searches. For MALDI MS/MS database searching, the precision tolerance was 0.3 Da for both the parent peptide and the fragment ions.

In general, an automatic database search followed by manual inspection was applied on LC-ESI data for protein identification. In the automatic database search process, trypsin was specified as the proteolytic enzyme, 1 missed cleavage site per peptide was allowed, carbamidomethylation of cysteine was set as the fixed modification and methionine oxidation was set as the variable modification. For protein identification using LC-MALDI data, besides this automatic search, another round of manual database searching on unmatched spectra was also performed where a combination of no specific enzyme type and additional variable modifications was tested. For the high quality MALDI MS/MS spectra that did not yield positive protein identification through database searching, manual inspection was carried out to examine the possibility of protein PTMs. Delta Mass (http://www.abrf.org/index.cfm/dm.home) was used as a reference guide for the study of different types of PTMs.

For spectral presentation, all data were reprocessed using the Igor Pro Software package (WaveMetrics, Lake Oswego, OR).

2.2.6 Manual Inspection for Protein Identification

Manual evaluation was applied on all searched results to minimize falsely identified proteins. For ESI search results, if two or more potential matches were reported for one mass spectrum, only peptide hits with the highest matching score (i.e., No. 1 ranking) for the corresponding spectra were manually inspected. For example, if the MS/MS spectrum of the *m/z* 567.5 (2+) ions yielded 3 possible peptide sequences in the protein identification result with scores of 30, 20, and 10, then only the peptide hit with the score of 30 was manually inspected. If more than one spectrum was assigned to a peptide, then only the spectrum with the highest score was used for manual analysis. The rules used to evaluate the peptide spectral identification were similar to those reported by Chen et al.[41] For peptide candidates of doubly charged ions, at least 5 isotopically resolved y-, b-, or a-ions or associated peaks must match theoretical peptide fragments. Most of the high-intensity fragment ions (top 5) must belong to y-, b-, or a-ions, instead of internal fragments. Major peaks with intensities of higher than 10-20% of the maximum intensity must match theoretical peptide fragments. The percentage threshold was lowered to 5-

10% for peaks with *m/z* values larger than that of the doubly charged parent mass. More random fragmentation was allowed in the low *m/z* region. For peptide candidates of singly or triply charged ions, the match scores were often lower. If the scores were higher than the threshold for significant homology and if most fragment peaks matched the theoretical fragments with high intensity peaks matching the b-, y-, or a-ion series, the identifications were considered positive. In a few cases where the match scores were not higher but very close to the threshold scores, if the match scores were significantly higher than the 2nd highest match score and most fragment peaks matched the theoretical fragment with high intensity peaks matching the b-, y-, or a-ion series, then the identifications were also considered positive. In spectra generated from triply charged parent ions, doubly charged fragment ions were inspected to make sure that basic amino acids such as lysine, arginine, and histidine were in the fragments. In all situations, peaks generated from electronic noise were not considered in the inspection process.

Because of the different fragmentation and detection mechanisms in MALDI MS/MS, somewhat different inspection criteria and processes were used for evaluating peptide identifications. As was done in ESI data evaluation, only peptide hits with the highest score for the corresponding MS/MS spectra were manually inspected. In MALDI MS/MS experiments, most sequenced parent ions are singly charged. The overall peak intensity and quantity were generally not as high and as many as those in ESI spectra. If the match scores were higher than the threshold for significant homology (or the scores were very close to the threshold value and significantly higher than the next possible match) and most of the dominant fragment peaks matched theoretical fragments, and if the fragmentation patterns agreed with the accepted rules for peptide fragmentation in MALDI Q_qTof MS/MS sequencing, then the identifications were considered positive. These rules were based on our accumulated experience over the past three years. They are as follows: (1). y-ions are often of higher intensity than b-ions. However, if the C-terminus of the parent peptide is lysine (K) instead of arginine (R) and the N-terminus is a basic amino acid such as histidine (H), we expect to see higher intensity with b-ion series than y-ions. (2). If the C-terminus of the peptide is lysine (K) instead of arginine (R), y1 may not be visible. (3). Fragmentations on the C-terminal side of aspartic acid (D) and N-terminal side of proline (P) are mostly favored in MALDI QTOF MS/MS. They usually generate the most dominant fragment ions (not only y-ions or b-ions, but also possible internal fragments) in the MS/MS spectrum. (4). Fragmentations on the C-terminal side of glutamic acid (E), asparagine (N), glutamine (Q), alanine (A), glycine (G), valine (V), leucine (L) and isoleucine (I) are preferred to some extent over other amino acids. But the resulting fragment intensity increases are not as significant as that in the D or P cases. (5). Fragments containing basic amino acids, such as arginine (R), lysine (K), or histidine (H) yield relatively higher peak intensity.

2.3 Results and Discussion

We started our tear proteome analysis with SDS-PAGE. Figure 2.1 shows the gel image generated after 10 min silver staining on a 12% separating gel with a sample loading and separation of 1 μL tear fluid. Only a few major bands were observed and the bands from low-abundance proteins were difficult to see, which indicates that proteins were present in tears with a wide concentration range. A total of 10 bands labeled in Figure 2.1 were excised and subjected to trypsin digestion, peptide mass mapping and MS/MS sequencing. A total of 17 proteins were identified and they are listed in Table 2.1. Detailed identification results are shown in Supporting Information 2.1. Proteins detected in this gel-based proteomic method include major tear proteins, such as lactotransferrin (P02788) and lysozyme C (P61626), and serum proteins, such as serum albumin (P02768), immunoglobin λ chain C region (P01842) and immunoglobin κ chain C region (P01834). The serum proteins were probably derived from plasma leakage from the conjuctival blood vessels. Using MALDI MS/MS, proline-rich proteins previously found to be difficult to identify in gel-based experiments using MALDI MS and ESI MS/MS[42] were identified with high confidence. Relative concentrations of these proteins can also be estimated by comparison of the gel-band intensities.

The gel-based approach is straightforward and semiquantitative and this can help gauge the complexity of the sample. Evidence on protein PTMs can also be gathered by judging the molecular weight shift of a given band among different samples. The use of 2D gel electrophoresis should improve the resolution and the success rate of protein

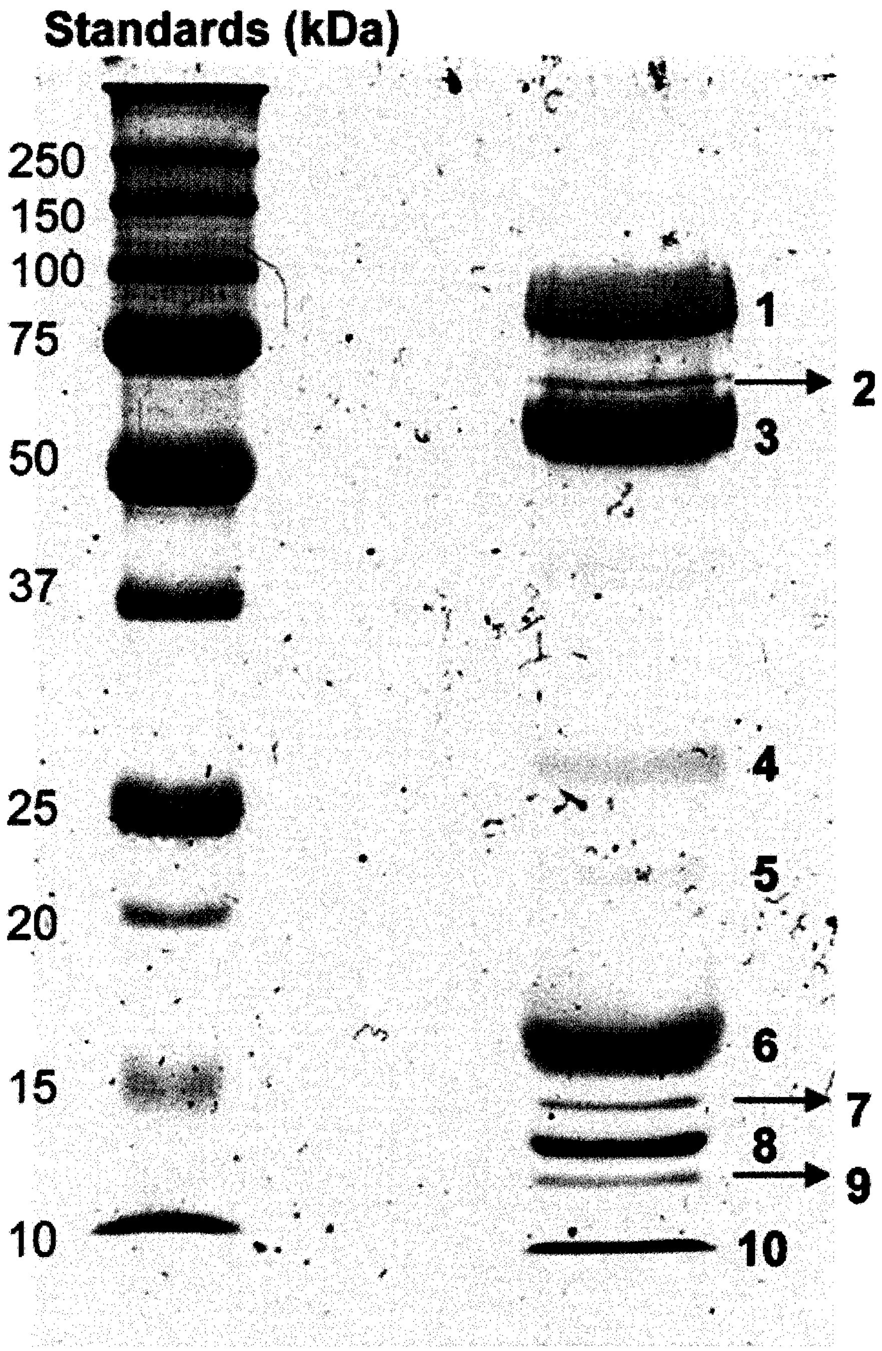


Figure 2.1 Gel image of SDS-PAGE from 1 μL of tear fluid of an individual donor.

identification, as shown by Kim and co-workers.[38] However, the gel-based approach has certain intrinsic limitations. For example, very high molecular weight or low molecular weight proteins may not be separated in gel electrophoresis. In-gel digestion and peptide extraction may be difficult for certain proteins, resulting in difficulty of identifying them. More significantly, automation for high-throughput sample analysis is still difficult. Thus, instead of an in-gel method, which was shown to be effective for a small population based study,[38] our focus is mainly on developing in-solution approaches as a new generation of high-throughput technique for tear proteomics.

We examined both LC-ESI and LC-MALDI in-solution proteomic techniques for tear proteome analysis. A 5-μL portion of tear fluid was digested and desalted by C_{18} Ziptipping. The final product was diluted into a volume of 20 μL with 0.1% TFA/H_2O. LC-ESI MS/MS was first employed as the detection method. A 2-μL portion of the 20 μL digest solution was injected into the LCQ Deca instrument equipped with a 150 μm × 100 mm C_{18} column. A 120-min gradient was applied and Figure 2.2A shows the base peak ion chromatogram. As shown in Figure 2.2A, high quality separation and detection were achieved. From this LC-ESI MS/MS experiment, a total of 2165 MS/MS spectra were obtained. After database searching and careful manual examination of the search results, 28 proteins were identified. They are also listed in Table 2.1 and detailed identification results including peptide sequences and search scores are shown in Supporting Information 2.2. Although more proteins were identified than with the gel-based approach, we were surprised by the low number of proteins identified, considering that both the quantity and quality of the MS/MS spectra were good. It appeared that only the high abundance proteins were identified.

Another LC-ESI MS/MS experiment was performed using 2 μL of the digest solution. In this case, the separation gradient was extended from 120 to 200 min and the 20 most intense peaks found in the first LC-ESI experiment were excluded for MS/MS. In addition, 5 MS/MS scans were performed after each MS scan. After these modifications, 3838 MS/MS spectra were generated. The base peak chromatogram is shown in Figure 2.2B. Only seven new proteins were identified as listed in Table 2.1. These seven proteins seem to include somewhat less abundant ones such as oxygen-

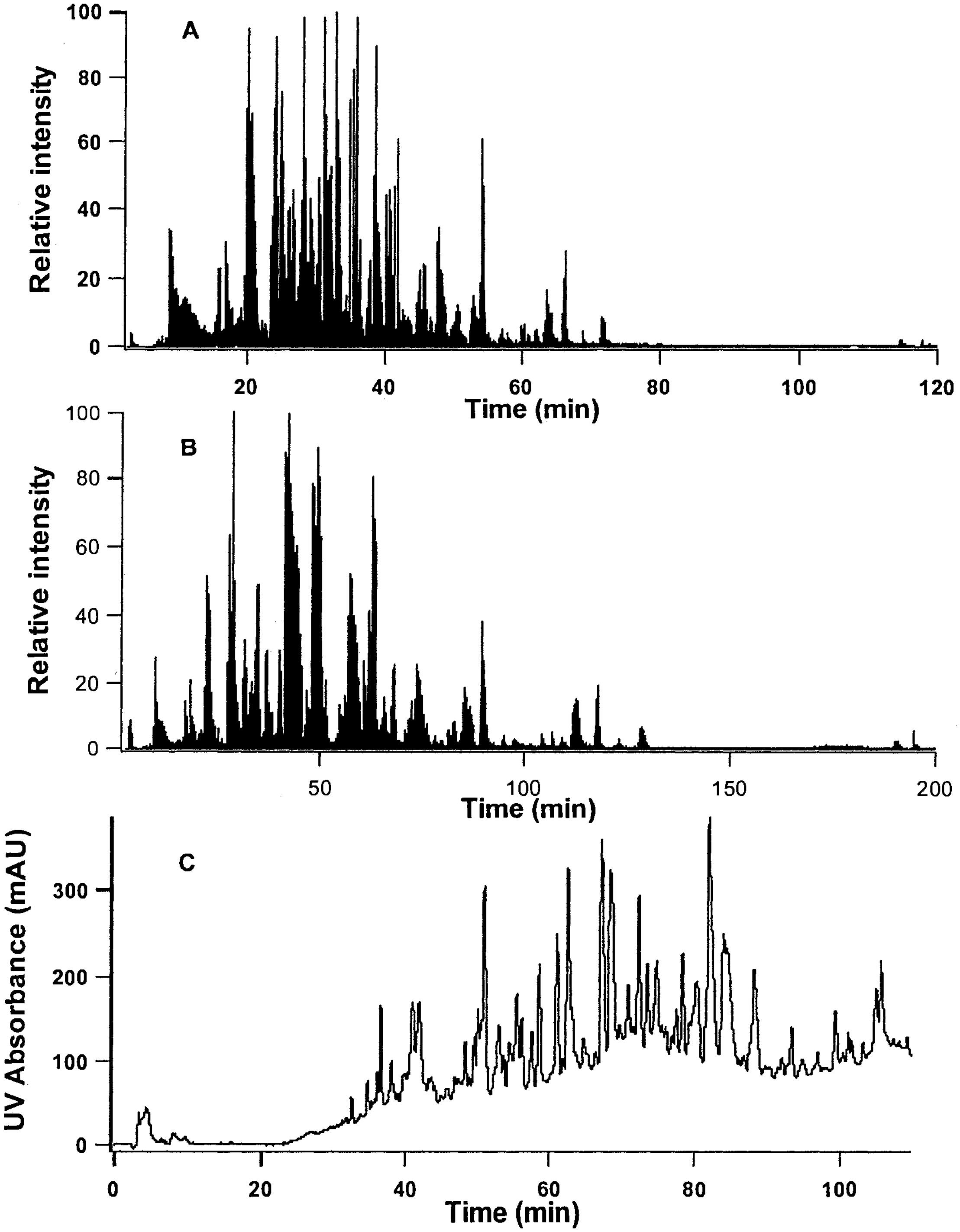


Figure 2.2 Base peak ion chromatograms of tryptic digests of tear fluids from LC-ESI MS/MS using (A) 120 min and (B) 200 min LC gradient. (C) UV chromatogram of the tryptic digest for LC-MALDI MS and MS/MS. The corresponding amount of tear fluids used for (A) or (B) was 0.5 μL and (C) 2 μL.

regulated protein 1 (retinitis pigmentosa RP1 protein), P56715. As indicated in the Swiss-Prot database, this protein is expressed specifically in the retina and has a potential role in the differentiation of photoreceptor cells. Defects in RP1 are the cause of retinitis pigmentosa type 1, a disease characterized by constriction of the visual fields, night blindness, and fundus changes.[43-45] With a molecular mass of about 240 kDa, oxygen-regulated protein 1 would be very difficult to identify by gel-based methods. While these finds are interesting, their biological significance and relevance as potential biomarkers require further studies.

However, the identification efficiency of this 200-min LC-ESI MS/MS experiment was still low compared to the analyses of other proteomic samples, such as cell extracts, that produced a similar quality of ion chromatograms to those shown in Figure 2.2B and a comparable number of high quality MS/MS spectra. It was therefore hypothesized that complicated PTMs on tear proteins were present. The presence of extensive PTMs would complicate the fragmentation of the modified peptide ions, resulting in difficulty of protein identification through database searching against the human proteome in which protein modifications are poorly characterized at present. We note that in LC-ESI MS/MS one uniform set of fragmentation conditions are generally applied during the data collection. For peptides with PTMs, optimal conditions can be quite different for different types and degrees of PTMs. In addition, when using ESI, the injected sample was consumed in the experiment and could not be re-examined using different conditions when needed.

The operation of LC-MALDI MS/MS is different from LC-ESI MS/MS. In LC-MALDI experiments where the peptide mixture is fractionated and deposited on a plate, we can select certain peaks of interest for analysis at any given time after sample deposition. This feature allows for fine-tuning of experimental conditions to produce optimal MS/MS results and for examining or re-examining related peptides within a chromatographic run to characterize PTMs of a protein. We thus applied LC-MALDI MS/MS as the final step in our attempt to characterize the tear proteome.

In LC-MALDI, 8 μL tear digest solution (equivalent to 2 μL of the original tear fluid sample) was injected into a 1.0 mm × 100 mm C_{18} column. A 110-min gradient separation was carried out (see Figure 2.2C) and a total of 100 fractions were collected for MALDI analysis. Prior to MS/MS sequencing, the peptides belonging to the abundant proteins were excluded from MALDI MS/MS based on the results obtained from the LC-ESI MS/MS experiments. This experimental arrangement took advantage of the complementary nature of MALDI and ESI and consequently allowed more information to be generated from the MALDI experiment. A great deal of effort was also devoted toward database searching. After the first round of automated search using Mascot, the unmatched spectra were manually searched again. No enzyme specificity was applied in this round of searching, and some additional modifications, such as phosphorylation and methylation, were selected. In the third round, MS/MS spectra showing similar fragmentation patterns were pooled together for spectral interpretation. MS/MS sequencing of a few selected peptide ions that showed possible modifications was re-done with more emphasis on generating database searching fragmentation patterns. This manual process of data collection, database searching, and spectral interpretation is best illustrated with several examples shown below. We recognize that this type of manual operation is by no means high throughput. However, by accumulating our knowledge on PTM analysis, we hope that a certain set of rules may emerge which can guide us to develop semi- or fully-automated data collection and interpretation processes for PTM characterization in the future.

In this work, a total of 15 new proteins were identified from the LC-MALDI MS/MS experiment and they are listed in Table 2.1. Some proteins such as KFLA590 (gi/37183242; precursor MW: 9.04 kDa; p*I*: 4.4) have not been annotated with extensive information. For this protein, the only information given in the database is that it is a secreted transmembrane protein.[46] Some proteins are found to be variants of well-characterized proteins. An example is gi/3402149, a protein indicated in the database as having a contribution of hydrophobic effect to the conformational stability of human lysozyme, which shares an almost identical sequence to that of lysozyme C (P61626) except for a mutation from isoleucine (I) to tyrosine (Y) in amino acid position 59 (aa59).

Our LC-MALDI MS/MS experiment detected a peptide sequence covering this region, thus determining the presence of this mutation.

Many other identified proteins are found to have mutations/variants that are either already annotated in the SwissProt database or not yet reported. Figure 2.3 depicts one of the examples. Despite the seemingly good quality of the spectrum shown in Figures 2.3A, no protein identification was obtained through database searching. In Figure 2.3B, a peptide sequence SVSLQEASSFFR from aa109-120 in proline-rich protein 4 (Q16378) was identified. By comparing Figure 2.3A with 2.3B which shared striking similarities, it was concluded that spectrum 2.3A was a result of sequence *G*SVSLQEASSFFR (aa108-120). This result was also confirmed by an MS/MS spectrum of a peptide ion belonging to sequence aa108-121 (data not shown). Thus, it was concluded that aa108 was mutated from proline (P) to glycine (G) in this proline rich protein 4. Since this mutation/variation has not been reported and annotated in the database, no identification result was obtained from database searching using the MS/MS spectrum shown in Figure 2.3A.

Using the approach of comparing similar MS/MS spectra as described above, many other variants were discovered or confirmed. For example, variation of valine (V) to isoleucine (I) was found at aa128 in Von Ebner's gland protein (tear prealbumin, P31025), and arginine (R) to glutamine (Q) was found at aa120 in proline-rich protein 4 (Q16378). In some cases, multiple mutations were found on the same peptide sequence region which complicated spectral interpretation. Thus, the exact mutation could not be pinpointed in those cases. For example, in Von Ebner's gland protein (P31025), multiple mutations were found in sequences aa19-35 and aa113-130.

Phosphorylation, one of the most important PTMs, was also observed in some of the tear proteins identified. Figure 2.4 shows three MS/MS spectra from three phosphopeptides. A characteristic loss of 98 Da was noticed in all three spectra, suggesting the attachment of one phosphate group in each peptide. By choosing phosphorylation as the variable modification, database searching using the three spectra led to the identification of three overlapping sequences in Von Ebner's gland protein. The phosphorylation site could only be narrowed down to either aa32 (serine) or aa34

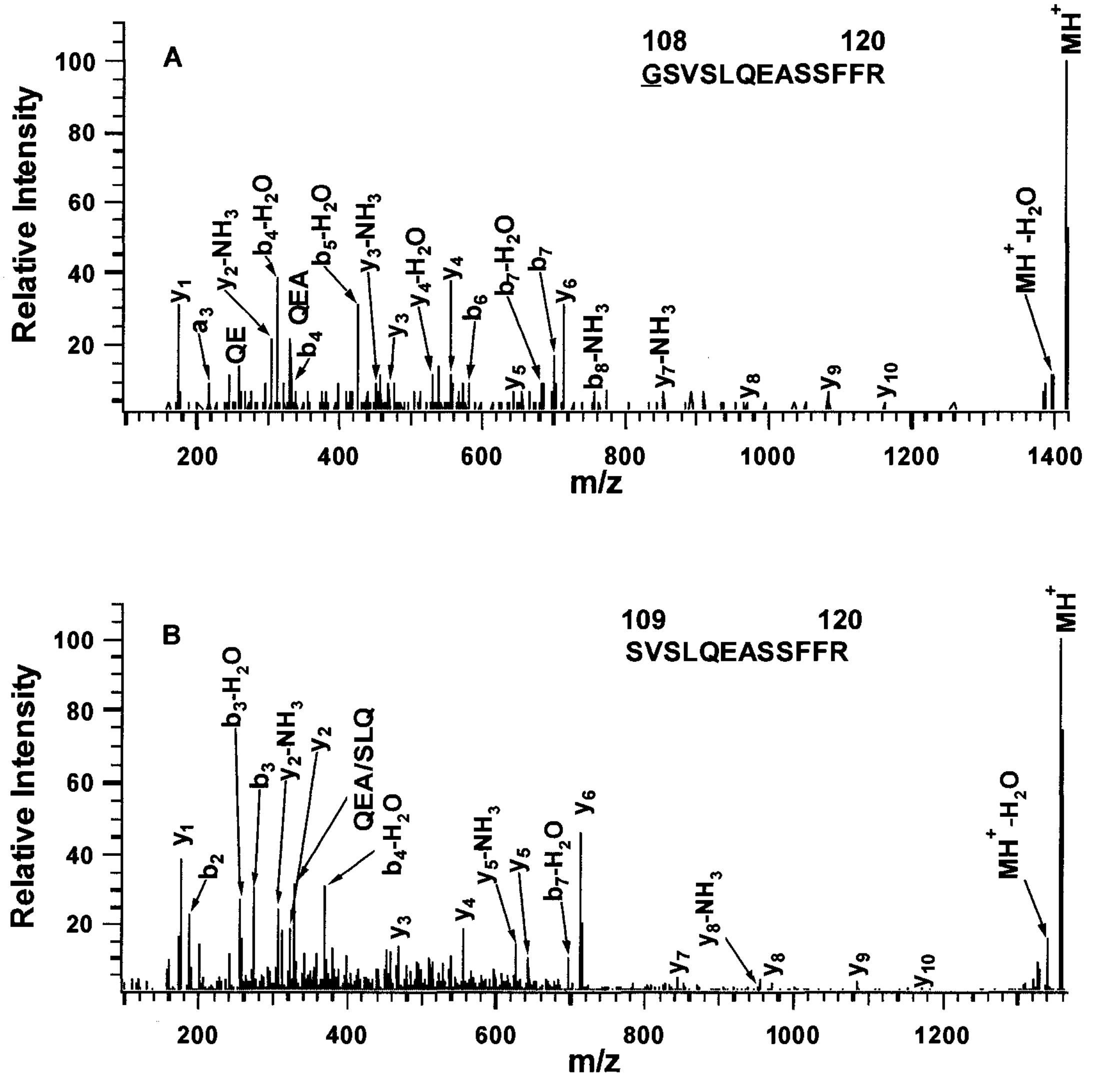


Figure 2.3 MALDI MS/MS spectra of two peptide ions used to determine an amino acid substitution in proline rich protein 4 (Q16378).

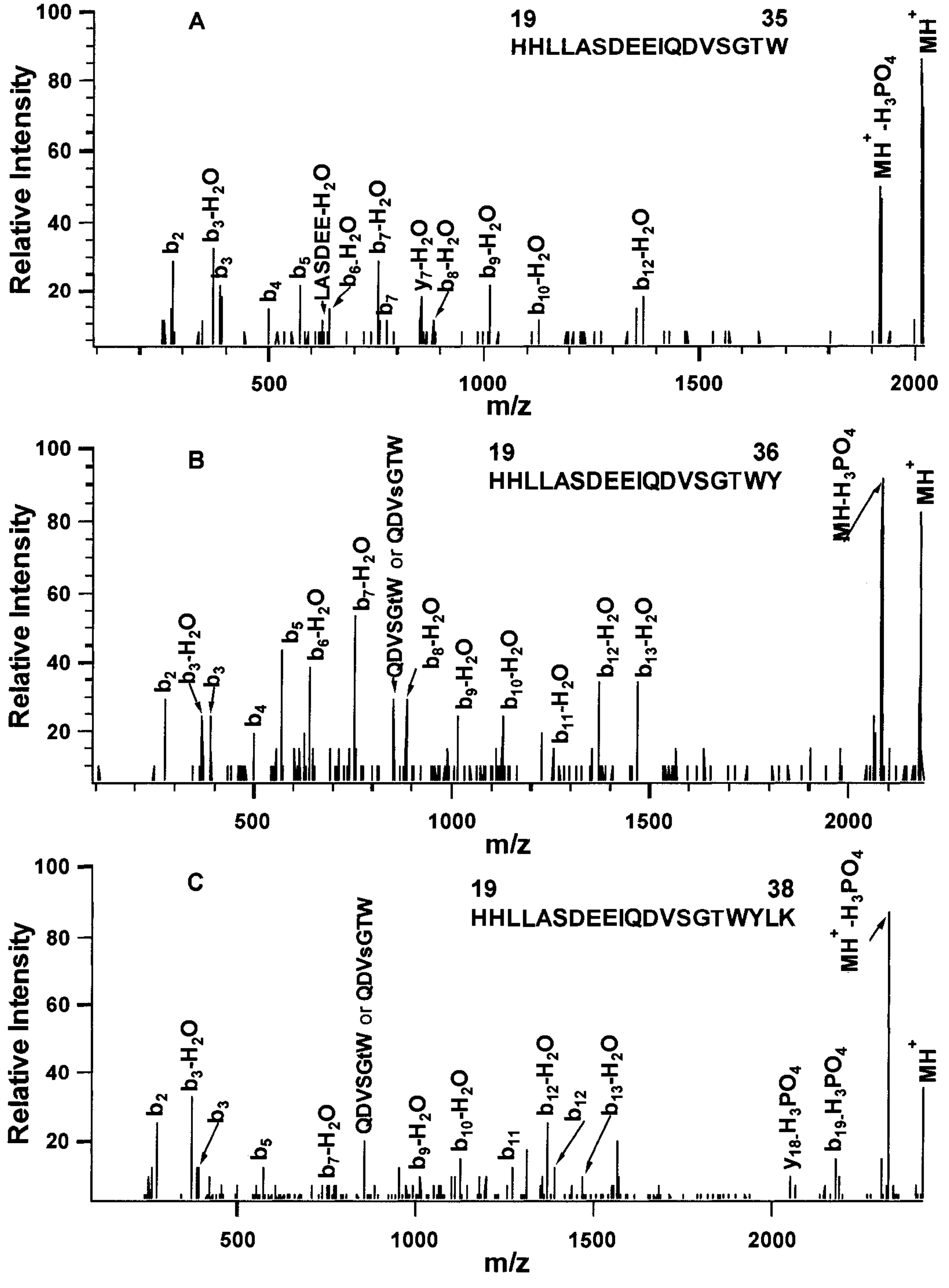


Figure 2.4 MALDI QqTOF MS/MS spectra of peptide ions used to determine phosphorylation on Von Ebner's protein.

(threonine). It is interesting to note that sequences identified from Figure 2.4A and B, although not tryptic peptides, might be the products of chymotryptic activity on the protein (chymotrypsin could be a contaminant in the trypsin reagent).

One other major PTM observed in the LC-MALDI MS/MS experiments was glycosylation. Figure 2.5 shows an example of glycosylation on a peptide which can be used to demonstrate how this form of protein PTM was only discovered in the LC-MALDI MS/MS experiment. In this case, a high quality MALDI MS spectrum on the LC fraction at 58 min was obtained during the initial run of the MALDI sample plate (see Figure 2.5A). As one of the high intensity peaks, the *m/z* 1389.78 peak was selected for MS/MS sequencing and the resulting spectrum is shown in Figure 2.5B. A standard database search based on this MS/MS spectrum did not yield any peptide or protein identification. A manual inspection of the spectrum revealed a mass difference of 203 Da between the parent ion peak (1389.78 Da) and the dominant fragment ion peak (1186.70 Da). This mass appeared to match to the mass of GalNAc or GlcNAc. We then went back to check the MALDI MS spectrum from this fraction and found that a peak at *m/z* 1186.70 was also present, albeit at a very low intensity. An MS/MS experiment was then done on this 1186.70 peak during the actual database search stage of the study. The MS/MS spectrum from this peak (see Figure 2.5C) resulted in the identification of the sequence SILLTEQALAK in extracellular glycoprotein lacritin (Q9GZZ8). We also realized that there was a striking resemblance in the low fragment ion regions of the MS/MS spectra shown in Figure 2.5B and C, suggesting that the peak at *m/z* 1186.70 in the MALDI spectrum was most likely from the fragmentation of the *m/z* 1389.78 peak during ionization and transport to the TOF mass analyzer. Fragmentation of fragile peptide bonds or weak bonds between a modification group and a peptide is not uncommon in MALDI Qq-TOF experiments. Thus, the mass difference of 203 Da seen in Figure 2.5A is believed to be the product of *O*-linked GalNAc modification on the serine (S) residue (aa91). In fact, using the low mass fragment ions shown in Figure 2.5B and using the peptide mass of 1186.70 instead of 1389.78 led to the identification of the same sequence mass in the database search. Glycosylation of residue S in this protein was not annotated in the Swiss-Prot database.

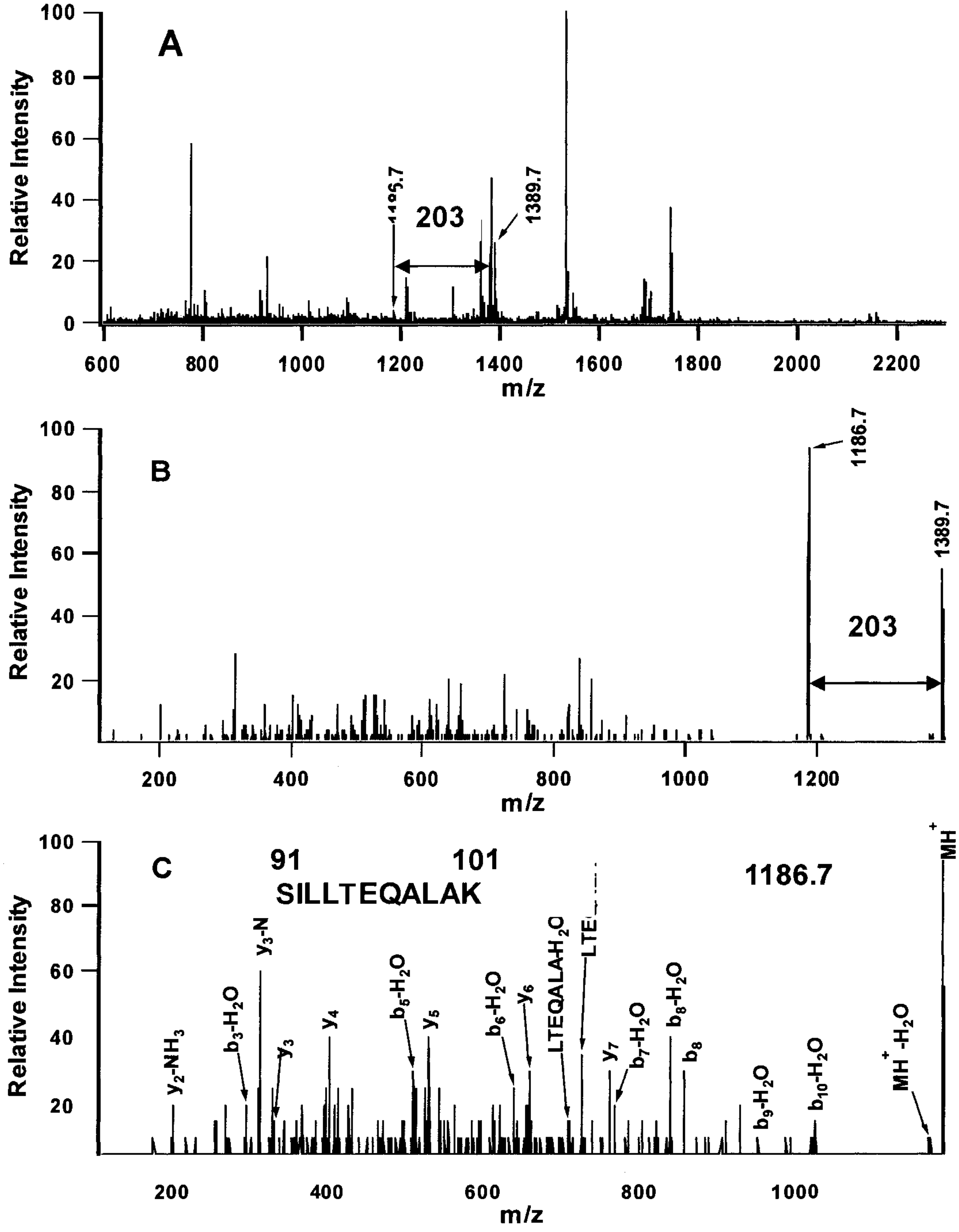


Figure 2.5 (A) MALDI QqTOF MS spectrum and (B, C) MS/MS spectra used to determine glycosylation on extracellular glycoprotein lacritin.

Figure 2.6 shows the MS/MS spectra of a series of peptides depicting a more complex glycosylation pattern. Figure 2.6A is the product ion spectrum of a precursor ion at *m/z* 1494.85 which matches well with a peptide sequence WVPPSPPPPYDSR, leading to the identification of proline-rich protein 1 (Q99935). From the MS/MS spectra of several peptides with masses of greater than 1494.85 (Figure 2.6B-F), it can be concluded that extensive glycosylation on the peptide sequence WVPPSPPPPYDSR (aa44-56) was observed. Glycosylation was on a serine at either aa48 or aa55 through *O*-linked GalNac. The mass difference of 146 corresponds to either deoxyhexoses (Fuc, Rha) or pentosyl, 162 corresponds to hexoses (Fru, Gal, Glc, Man), 291 corresponds to *N*-acetylneuraminic acid (NeuAc, sialic acid), and 365 corresponds to Hex-HexNAc. Similar types of glycosylation were detected in other proline-rich proteins, including proline-rich protein 4 (Q16378) and proline-rich protein 1 (Q99935). Details on glycosylation characterization of the peptides are shown in Supplementary information S2.5. Interestingly, *O*-linked glycosylation on proline-rich proteins has been reported in saliva.[47] In contrast, a recent PTM characterization on proline-rich proteins in tears primarily showed protein truncation.[29] This example clearly illustrates the power of LC-MALDI MS/MS for glycopeptide detection. As shown in Figure 2.6, the technique generates useful fragmentation patterns from glycopeptides. However, detailed characterization, such as determination of glycan structures and exact modification sites, is still a major challenge. Nevertheless, the ability to generate rich glycopeptide fragment ions by MALDI MS/MS does open the possibility of using a fingerprint approach to compare MS/MS spectra of unknowns to a yet to be established glycopeptide MS/MS spectral database (see below).

Many other PTMs were also detected using LC-MALDI MS/MS. These include N-terminal pyroglutamic acid formation from glutamine (Q) or glutamic acid (E) and methylation of C-terminal aspartic (D) or glutamic acids (E) and C-terminus amide formation. Future experiments involving isolation of protein isoforms with various types of modifications will be needed to characterize these modifications fully.

The above work indicates that many different types of protein modifications and point mutations are present in proteins found in tears, making tear protein identification

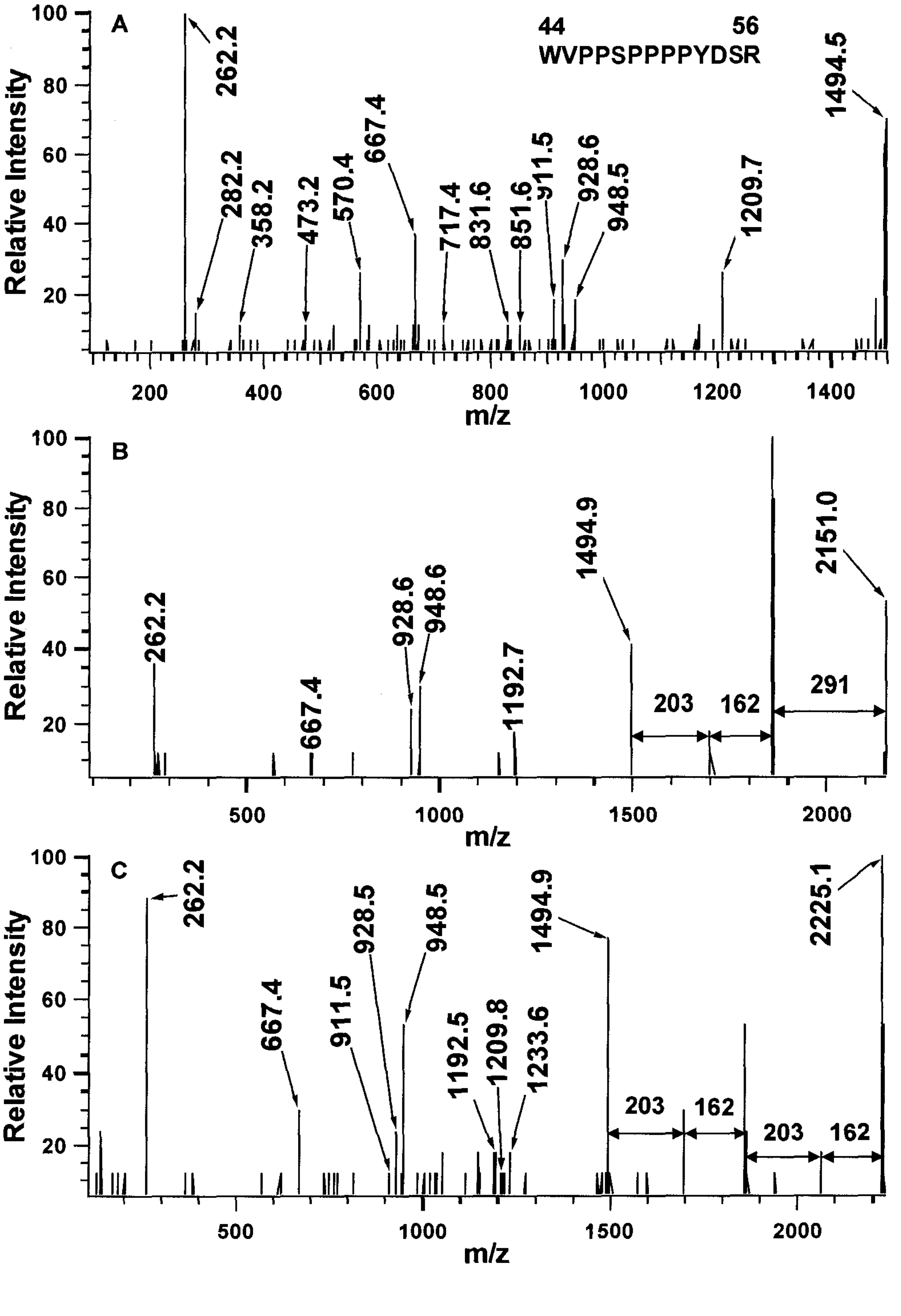


Figure 2.6 MALDI MS/MS spectra used to determine glycosylation on proline rich protein 1 (Q99935).

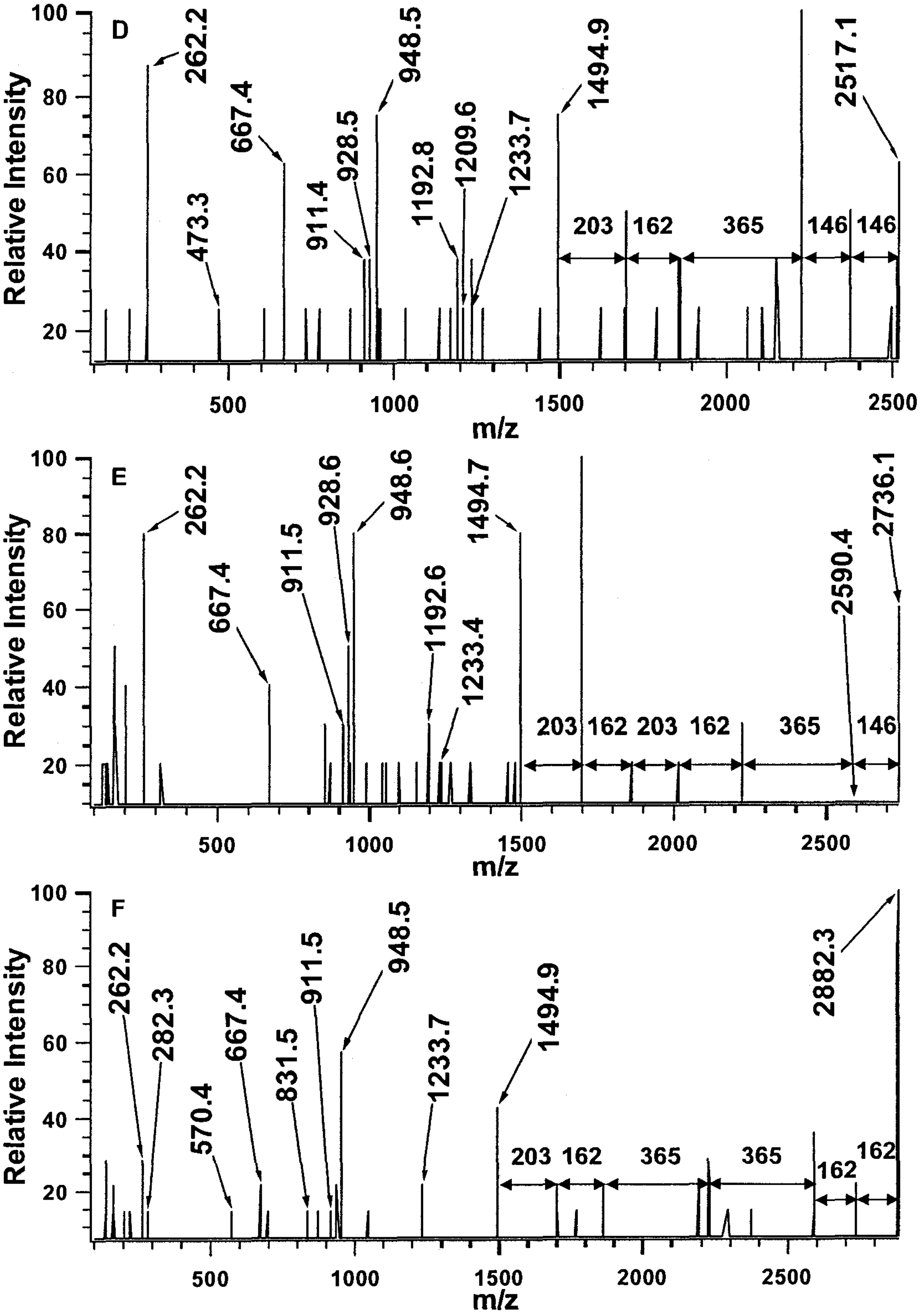


Figure 2.6 MALDI MS/MS spectra used to determine glycosylation on proline rich protein 1 (Q99935).

difficult using conventional database searching methods. LC-MALDI MS/MS facilitates the detection of these modifications, mainly due to the possibility of carrying out result-dependent experiments where a peptide suspected of having modifications can be subjected to re-sequencing using different MS/MS conditions on the same sample.

2.4 Conclusions

We have applied several proteomic techniques to characterize the tear proteome with an ultimate goal of developing a high-throughput technique that can handle a large number of samples in population based studies, such as in the compatibility study of a new contact lens or in biomarker discovery of diseases. In this work, a total of 6 μL reflex tear fluid was taken from a single tear collection with actual usage of about 4 μL for the entire analyses by the SDS-PAGE in-gel digestion MALDI MS method and in-solution methods of LC-ESI and LC-MALDI MS/MS. Fifty four proteins were identified with high confidence and most of these proteins (44/54) could be detected by LC-MALDI MS/MS with the consumption of 2 μL tear. Many of these identified proteins except the high abundance tear proteins detected by the gel-based method were explicitly detected in human tears for the first time. Furthermore, result-dependent experiments could be carried out in LC-MALDI MS/MS that allowed the detection and characterization of peptides with PTMs. Unlike LC-ESI, the MALDI experiments could be carried out under different optimal conditions from the same sample for PTM detection and analysis. With further development in software, it should be possible to automate the process of examining and reexamining the sample spots deposited from LC separation on a MALDI plate using a set of different conditions, either in parallel (i.e., applying a series of varying conditions to a spot before moving onto the next spot) or in sequence (i.e., applying one set of conditions to all spots first and then applying another set of conditions to all spots, and so on). Thus, LC-MALDI is particularly useful for PTM analysis.

The results shown in this study indicate that tear proteome profiling is quite challenging. The number of different proteins identified from a clinically relevant volume of tear sample (i.e., < 5 μL) is small, but these proteins display extensive posttranslational modifications. Thus, the actual number of proteins that can be profiled, including protein isoforms with different degrees of modification, may be quite large using the in-solution LC-MALDI method. In light of the fact that extensive PTMs are present in tear proteins, we need to consider tear proteome profiling as a task of both protein identification and protein modification analysis. In terms of protein quantitation, we have recently demonstrated that differential isotope labeling of N-termini of peptides, based on guanidination of lysines and N-terminal dimethylation, can be used for relative protein quantitation.[48] This labeling chemistry is compatible with modified proteins, since mild chemical reagents are used and the mass tag is attached to the N-terminus of a peptide. Thus, relative quantitation of peptides resulting from modified or unmodified proteins of different tear samples can potentially be done by using this isotope labeling method.

While protein identification can be made by MS/MS database searching, protein modification analysis is not trivial. Ideally, during the tear proteome profiling, the structures and sites of modification groups on a peptide would be defined. Unfortunately, this is nearly impossible to achieve, considering that there are so many different modification possibilities. An alterative to de novo characterization of modified peptides is to use a fingerprint approach to compare the MS/MS spectrum of an unknown peptide to those stored in an MS/MS spectral database. The MS/MS spectral database may be created by using well-characterized tear proteins. To characterize fully a protein with PTMs, this protein along with modified forms needs to be isolated from a relatively large amount of tear fluid. Both top-down and bottom-up proteomic methods can be applied to characterize PTMs, after which the protein sample can be digested and MALDI MS/MS spectra of the resulting peptides, including modified peptides (perhaps under different instrumental conditions to generate optimal fragmentation patterns), can be collected and stored in the spectral database.

In summary, the tear proteome is very complex; it includes many modified proteins, such as phosphoproteins and glycoproteins. Small volumes of samples available for analysis also present a major challenge in tear proteome profiling. However, we envision that LC-MALDI MS/MS, combined with an MS/MS database of peptides produced from well characterized tear proteins, may potentially become a powerful tool for generating tear proteome profiles for comparative proteomics in disease biomarker discovery and functional studies related to eye health. To this end, many practical issues related to sample collection, storage, reproducibility, and interrelation of sample throughput and proteome coverage (e.g., multidimensional LC combined with MALDI would result in better proteome coverage, but require a longer analysis time) will need to be addressed. Experiments designed to address these issues by quantitative LC-MALDI MS and MS/MS are currently underway.

Table 2.1 Summary of the protein identification results.

	Identified Proteins	Accession Number	Gel-MALDI	Soln-ESI-120min	Soln-ESI-200min	In solution LC-MALDI
1	Lactotransferrin	P02788	√	√	√	√
2	Proline-rich protein 1	Q99935	√	√	√	√
3	Ig alpha-1 chain C region	P01876	√	√	√	√
4	Ig alpha-2 chain C region	P01877	√	√	√	√
5	Ig lambda chain C region	P01842	√	√	√	√
6	Ig kappa chain C region	P01834	√	√	√	√
7	Von Ebner's gland protein (Tear prealbumin)	P31025	√	√	√	√
8	Prolactin-inducible protein	P12273	√	√	√	√
9	Cystatin SN precursor (Salivary cystatin SA-1)	P01037	√	√	√	√
10	Lysozyme C precursor (EC 3.2.1.17) (1,4-beta-N-acetylmuramidase C)	P61626	√	√	√	√
11	Cystatin S precursor (Salivary acidic protein-1) (Cystatin SA-III)	P01036	√	√	√	√
12	Mammaglobin B	O75556	√	√	√	√
13	Proline-rich protein 4	Q16378	√	√	×	√
14	Serum albumin	P02768	√	×	√	√
15	Ig heavy chain V-III region BRO /or TEI	P01766 /or P01777	√	×	×	×
16	Nasopharyngeal carcinoma-associated proline rich 4	gi/22208536	√	×	×	√
17	Phospholipase A, membrane-associated	P14555	√	×	×	×
18	Polymeric-immunoglobulin receptor precursor (Poly-Ig receptor) (PIGR)	P01833	×	√	√	√
19	Cystatin SA precursor (Cystatin S5)	P09228	×	√	√	√
20	Extracellular glycoprotein lacritin precursor	Q9GZZ8	×	√	√	√
21	Lipophilin A precursor (Secretoglobin family 1D member 1)	O95968	×	√	√	√
22	Immunoglobulin J chain	P01591	×	√	√	√

23	Cystatin D precursor	P28325	×	√	√	√
24	Cystatin C precursor (Neuroendocrine basic polypeptide)	P01034	×	√	√	√
25	Beta-2-microglobulin precursor (HDCMA22P)	P61769	×	√	√	√
26	Proline-rich protein 5 precursor (Proline-rich protein PBI)	Q99954	×	√	√	√
27	Antileukoproteinase 1 precursor (ALP)	P03973	×	√	√	√
28	Brain-specific angiogenesis inhibitor 3 precursor	O60242	×	√	×	×
29	Aspartyl aminopeptidase (EC 3.4.11.21)	Q9ULA0	×	√	×	×
30	G-rich sequence factor-1 (GRSF-1)	Q12849	×	√	×	×
31	5'-AMP-activated protein kinase, catalytic alpha-2 chain (EC 2.7.1.-) (AMPK alpha-2 chain)	P54646	×	√	×	×
32	Ig heavy chain V-I region SIE	P01761	×	√	×	×
33	Oxygen-regulated protein 1	P56715	×	×	√	√
34	Clusterin precursor (Complement-associated protein SP-40,40)	P10909	×	×	√	√
35	Mesothelin precursor (CAK1 antigen).	Q13421	×	×	√	√
36	Endothelial transcription factor GATA-2.	P23769	×	×	√	√
37	Nuclear RNA export factor 1 (Tip associating protein) (Tip-associated protein) (mRNA expor	Q9UBU9	×	×	√	×
38	Leucine-rich primary response protein 1 (Follicle-stimulating hormone primary response pro	Q92674	×	×	√	×
39	60S ribosomal protein L18a	Q02543	×	×	√	×
40	Leucine-rich repeat transmembrane protein FLRT3 precursor	Q9NZU0	×	×	×	√
41	Chloride intracellular channel protein 2 (XAP121)	O15247	×	×	×	√
42	Proline-rich protein 3 precursor	P02814	×	×	×	√
43	Basic salivary proline-rich protein 4 allele M	P10161	×	×	×	√
44	Ig mu chain C region	P01871	×	×	×	√
45	Serotransferrin precursor	P02787	×	×	×	√
46	deleted in malignant brain tumors 1 isoform a precursor	gi4758170	×	×	×	√
47	KFLA590	gi37183242	×	×	×	√
48	Zinc-alpha-2-glycoprotein precursor	P25331	×	×	×	√
49	hypothetical protein	gi34365249	×	×	×	√
50	similar to common salivary protein 1	gi21687060	×	×	×	√
51	Ig heavy chain variable region	gi2808997	×	×	×	√
52	Phospholipid transfer protein precursor	P55058	×	×	×	√
53	hypothetical protein	gi34365344	×	×	×	√

54	Contribution Of Hydrophobic Effect To The Conformational Stability Of Human Lysozyme	gi3402149	×	×	×	√

2.5 Literature Cited

1. Harding, J. J.; Ed. *Biochemistry of the Eye;* Chapman and Hall: London, 1997.

2. Vesaluoma, M.; Teppo, A. M.; Gronhagen-Riska, C.; Tervo, T. *Curr. Eye Res.* **1997**, *16*, 19-25.

3. Tervo, T.; Vesaluoma, M.; Bernnett, G. L.; Schwall, R.; Helena, M.; Liang, Q.; Wilson, S. E. *Exp. Eye Res.* **1997**, *64*, 501-504.

4. Lembach, M.; Linenber, C.; Sathe, S.; Beaton, A.; Vcakhan, O.; Asbell, P.; Sack, R. *Curr. Eye Res.* **2001**, *22*, 286-294.

5. Zhou, L.; Beuerman, R. W.; Barathi, A.; Tan, D. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 401-412.

6. Claudon, E. V.; Baguet, J. *Adv. Exp. Med. Biol.* **1994**, *350*, 411-416.

7. Sack, R. A.; Sathe, S.; Hockworth, L. A.; Willcox, M. D.; Holden, B. A.; Morris, C. A. *Curr. Eye Res.* **1996**, *15*, 1092-1100.

8. Kijlstra, A.; Polak, B. C.; Luyendijk, L. *Curr. Eye Res.* **1992**, *11*, 123-126.

9. Berta, A. *Enzymology of the Tears*; CRC Press: Boca Raton, FL, 1992.

10. Willcox, M.; Peace, D.; Tan, M.; Demirci, O.; Carney, F. *Adv. Exp. Med. Bio.* **2002**, *506* (Lacrimal Gland, Tear Film, and Dry Eye Syndromes 3), 879-884.

11. Fullard, R. J. *Curr. Eye Res.* **1988**, *7*, 163-179.

12. Grus, F. H.; Augustin, A. J. *Eur. J. Ophthalmol.* **2001**, *11*, 19-24.

13. Baier, G.; Wollensak, G.; Mur, E.; Redl, B.; Stoffler, G. *J. Chromatogr.* **1990**, *525* (2), 319-328.

14. Herber, S.; Grus, F. H.; Sabuncuo, P.; Auguston, A. J. *Electrophoresis* **2001**, *22*, 1838-1844.

15. Grus, F. H.; Sabuncuo, P.; Augustin, A. J. *Electrophoresis* **2001**, *22*, 1845-1850.

16. Herber, S.; Grus, F. H.; Sabuncuo, P.; Auguston, A. J. *Adv. Exp. Med. Bio.* **2002**, *506* (Lacrimal Gland, Tear Film, and Dry Eye Syndromes 3), 623-626.

17. Peach, H. C.; Mann, A. M.; Tighe, B. J. *Adv. Exp. Med. Bio.* **2002**, *506* (Lacrimal Gland, Tear Film, and Dry Eye Syndromes 3), 967-971.

18. Bjerrum, K. B.; Prause, J. V. *Graefe's Arch. Clin. Exp. Ophthalmol.* **1994**, *232*, 402-405.

19. Grus, F. H.; Augustin, A. J. *Electrophoresis* **1999**, *20*, 875-880.

20. Grus, F. H.; Augustin, A. J.; Evangelous, N. G.; Toth-Sagi, K. *Eur. J. Ophthalmol.* **1998**, *8*, 90-97.

21. Grus, F. H.; Dick, B.; Augustin, A. J.; Pfeiffer, N. *Ophthalmologica.* **2001**, *215*, 430-434.

22. Grus, F. H.; Sabuncuo, P.; Herber, S.; Augustin, A. J. *Adv. Exp. Med. Bio.* **2002**, *506* (Lacrimal Gland, Tear Film, and Dry Eye Syndromes 3), 1213-1216.

23. Mulvenna, I.; Stapleton, F.; Hains, P. G.; Cengiz, A.; Tan, M.; Walsh, B.; Holden, B. *Clin. Exp. Ophthalmol.* **2000**, *28*, 205-207.

24. McArthur, S. L.; McLean, K. M.; St. John, H. A. W.; Griesser, H. J. *Biomaterials* **2001**, *22*, 3295-3304.

25. Kingshott, P.; St. John, H. A. W.; Chatolier, R. C.; Griesser, H. J. *J. Biomed. Mater. Res.* **2000**, *49*, 36-42.

26. Zhou, L.; Huang, L. Q.; Beuerman, R. W.; Grigg, M. E.; Li, S. F. Y.; Chew, F. T.; Ang, L.; Stern, M. E.; Tan, D. *J. Proteome Res.* **2004**, *3*, 410-416.

27. Grus, F. H.; Podust, V. N.; Bruns, K.; Lackner, K.; Fu, S.; Dalmasso, E. A.; Wirthlin, A.; Pfeiffer, N. *Invest. Ophthalmol. Vis. Sci.* **2005**, *46*, 863-876.

28. Tomosugi, N.; Kitagawa, K.; Takahashi, N.; Sugai, S.; Ishikawa, I. *J. Proteome Res.* **2005**, *4*, 820-825.

29. Fung, K. Y. C.; Morris, C.; Duncan, M. *Adv. Exp. Med. Bio.* **2002**, *506* (Lacrimal Gland, Tear Film, and Dry Eye Syndromes 3), 601-605.

30. Fung, K. Y. C.; Morris, C.; Sathe, S.; Sack, R.; Duncan, M. W. *Proteomics* **2004**, *4*, 3953-3959.

31. Fullard, R. J.; Snyder, A. C. *Invest Ophthalmol. Vis. Sci.* **1990**, *31*, 119-126.

32. Fullard, R. J.; Tucker, D. L. *Invest Ophthalmol. Vis. Sci.* **1992**, *32*, 2290-2301.

33. Fullard, R. J.; Snyder, A. C. *Adv. Exp. Med. Biol.* **1994**, *350*, 309-314.

34. Baguet, J.; Claudon, E. V.; Sommer, F.; Chevallier, P. *CLAO J.* **1995**, *21*, 114-121.

35. McGill, J. L.; Liakos, G. M.; Goulding, N.; Seal, D. V. *Br. J. Ophthalmol.* **1984**, *68*, 316-320.

36. Ng, V.; Cho, P.; Mak, S.; Lee, A. *Graefe's Arch. Clin. Exp. Ophthalmol.* **2000**, *238*, 892-899.

37. Sitaramamma, T.; Shivaji, S.; Rao, G. N. *Curr. Eye Res.* **1998**, *17*, 1027-1035.

38. Koo, B.-S.; Lee, D.-Y.; Ha, H.-S.; Kim, J.-C.; Kim, C.-W. *J. Proteome Res.* **2005**, *4*, 719-724.

39. Berta, A. *Am. J. Ophthalmol.* **1983**, *96*, 115-116.

40. Zhang, B.; McDonald, C.; Li, L. *Anal. Chem.* **2004**, *76*, 992-1001.

41. Chen, Y.; Kwon, S. W.; Kim, S. C.; Zhao, Y. J. *Proteome Res.* **2005**, *4*, 998-1005.

42. Wilmarth, P. A.; Riviere, M. A.; Rustvold, D. L.; Lauten, J. D.; Madden, T. E.; David, L. L. *J. Proteome Res.* **2004**, *3*, 1017-1023.

43. Sullivan L. S.; Heckenlively, J. R.; Bowne, S. J.; Zuo, J.; Hide, W. A.; Gal, A.; Denton, M.; Inglehearn, C. E.; Blanton, S. H.; Daiger, S. P. *Nat. Genet.* **1999**, *22*, 255-259.

44. Pierce, E. A.; Quinn, T.; Meehan, T.; McGee, T. L.; Berson, E. L.; Dryja, T. P. *Nat. Genet.* **1999**, *22*, 248-254.

45. Guillonneau, X.; Piriey, N. J.; Danciger, M.; Kozak, C. A.; Cideciyan, A. V.; Jacobson, S. G.; Farber, D. B. *Hum. Mol. Genet.* **1999**, *8*, 1541-1546.

46. Clark, H. F. et al. *Genome Res.* **2003**, *13*, 2265-2270.

47. Carpenter, G. H.; Proctor, G. B. *Oral Microbiol. Immunol.* **1999**, *14*, 309-315.

48. Ji, C.; Li, L. *Differential Dimethyl Labeling of N-termini of Peptides for Quantitative Proteome Analysis and Peptide Sequencing*; In Proceedings of the American Society for Mass Spectrometry conference, San Antonio, TX, June 4-9, 2005.

Chapter 3

Proteome Profile of Cytosolic Component of Zebrafish Liver Generated by LC-ESI MS/MS Combined with Trypsin Digestion and Microwave-Assisted Acid Hydrolysis*

3.1 Introduction

Determination of the biological effects that water contaminants have on aquatic organisms is the subject of intense research worldwide. The use of genomic-based biomarkers (toxigenomics), whereby analysis of toxin-induced changes in mRNA expression is used to assess toxicity of a particular compound,[1] have been reported for a number of aquatic organisms, including trout,[2-5] goldfish,[6] daphnia,[7] tilapia,[8-10] and zebrafish.[11,12] Genes whose expression are thought to be indicative of toxin exposure in an aquatic environment include metallothionein as an indicator of metal exposure in rainbow trout,[4] tilapia,[8,9] and molluscs;[13] hsp70 as a measure of cadmium toxicity in zebrafish;[11] and cytochrome P4501A1 as a sign of persistent organic pollutants in tilapia.[10] However, toxigenomics alone cannot give a comprehensive assessment of aquatic toxicity, as there are potential pitfalls associated with this technique. For example, changes in gene expression may be short term, and not actually correlated to any biological change/response.[14,15] Microarray-based toxigenomics can also be problematic, since changes in transcript levels may be due to changes in mRNA stability, rather than transcription itself. Finally, it is well-documented that changes in mRNA transcription do not always correlate with protein expression.[16-18]

Given these potential difficulties associated with the genomic-based methods, the use of proteomic methods to complement toxigenomics is a significant improvement in the detection of new biomarkers. The identification of proteins that specifically respond to various toxicants in a reproducible manner would enable us to determine the function

* A form of this chapter was published as: N. Wang, L. MacKenzie, A. De Souza, H. Zhong, G. Goss and L. Li, **2007**, "Proteome Profile of Cytosolic Component of Zebrafish Liver Generated by LC-ESI MS/MS Combined with Trypsin Digestion and Microwave-Assisted Acid Hydrolysis ", *J. Proteome Res.*, **6**, 263-272. L. MacKenzie contributed to the bioinformatics analysis.

of these proteins and ultimately help in understanding the mechanisms of toxicity for specific compounds.[19] The examination of an organism's proteome not only provides a robust snapshot of its physiology, but also takes into account protein interactions, modifications, and stability.[15,16,20] Previous research has looked at the response of individual protein biomarkers, such as heat shock proteins,[21,22] vittellogenin,[23,24] malate dehydrogenase,[25] and superoxide dismutase,[26] to specific environmental stressors. Recent attempts to uncover the proteomic profiles of various fish species using gel-based proteomic analysis have been moderately successful.[27-29] However, this research has been hampered by the lack of an annotated genome and high-throughput proteome analysis tools.

With this in mind, our long-term aim is to aid in the development of the zebrafish as a toxicological model system. The zebrafish, an established vertebrate model,[30] was chosen because its entire genome has been sequenced and annotated.[31] With the genome sequenced, high-throughput solution-based proteomic analysis can be conducted.[32] We chose to look specifically at the liver, since it is the site of metabolism and detoxification and, therefore, is likely to respond markedly to chemical applications in later quantitative analysis.

The liver proteome is expected to be very complex. Vertebrate cell types may express up to 20 000 proteins, with a predicted concentration dynamic range of up to 5 to 6 orders of magnitude.[33] Thus, the liver proteome should be composed of proteins with a wide range of molecular weight, relative abundance, acidity/basicity, and hydrophobicity. It would be ideal to analyze the entire liver proteome. However, due to current technical challenges associated with high-throughput large-scale proteome analysis, examining a subset of the proteome is a more realistic goal, at least for the foreseeable future. Many of the reported proteome studies on human, mouse, and rat liver samples or cell lines are based on orthogonal electrophoretic separation of the complex protein mixtures (e.g., 2-D PAGE), followed by MALDI MS peptide mass fingerprinting and/or MALDI or ESI MS/MS analysis of proteolytic digests extracted from individual spots. Some of these studies have looked at protein profiling, many focusing on disease-related protein profiling in liver sample and cell lines.[34-36] Other studies have focused on the changes in protein expression in human, mouse, or rat livers under conditions such as partial hepatectomy, surgical stress, and exposure to dietary folate deficiency.[37-39] For example, Martin et al.[27] studied changes in rainbow trout liver proteins during short-term starvation, by looking at changes in the abundance of 780 detected protein spots.

Using 2-D PAGE, Tay et al. recently detect 361 protein spots from samples of developing zebrafish embryos, and 108 of them were analyzed by MS, resulting in the identification of 55 unique proteins.[40] Their work illustrates the power of the proteomic approach for discovering biologically significant proteins relevant to zebrafish embryo development. On the other hand, the relatively small number of identified proteins highlights the difficulty in characterizing the proteome of the zebrafish in complex tissues using gel-based technology. Although the 2-D PAGE MS approach has led to important findings, it discriminates against hydrophobic membrane proteins, very large and small proteins, extremely basic and acidic proteins, and lower abundance proteins. Newer strategies for proteomic analysis involve the use of solution-based analysis, which employs global proteolytic digestion of cell or tissue lysate or subfractions, followed by analysis of the resulting complex peptide mixtures by 1-D or 2-D LC coupled with MS/MS.[41-46] Several studies have shown that the 2-D LC approach is capable of detecting proteins of very wide dynamic range of concentration.[47-50] For example, 564 rat proteins were reported to be identified in a subcellular liver proteome study,[51] while about 2000 unique human liver proteins were identified in three human liver cell lines via isotope-coded affinity tag (ICAT) coupled with MS/MS.[52]

While it is impossible to completely identify all proteins in a complex liver tissue sample using a simple single-step operation with current technology, the development of more advanced solution-based technologies should allow for a more complete coverage of the zebrafish proteome. This, in turn, will aid in the understanding of both the responses to, and the mechanisms of, toxicity upon exposure to specific toxicants. Thus, our first research goal is to develop a methodology for the characterization of the protein complement of the cytosolic component of the zebrafish liver proteome. Proteins in the cytosolic component should be relatively easy to handle, compared to other components, such as the membrane fraction. Cytosolic compartment should also contain a large number of metabolites-an excellent source for metabonomic profiling which should provide complementary information to the proteome results. In this work, we wish to address several questions related to the analysis of cytosolic fraction of a liver tissue sample using a solution-based proteome analysis technique: how do we best apply newly developed sample processing methods, such as microwave-assisted acid hydrolysis of proteins,[53] for the analysis of a subproteome from a complex tissue sample? How many proteins can be identified from the cytosolic component? What types of proteins can we identify from this compartment? Are these proteins relevant to toxicity functional studies?

3.2 Experimental

3.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide, phenylmethylsulfonyl fluoride (PMSF), trifluoroacetic acid (TFA), heparin, sodium bicarbonate, sodium chloride (NaCl), and SDS were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC-grade formic acid, acetone, methanol (MeOH), and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). Tricaine methane sulfonate (TMS) was obtained from Aqualife. Water was obtained from a Milli-Q Plus purification system (Millipore, Bedford, MA). Cellular fractionation was accomplished using a compartmental protein extraction kit (Kit K3013010 from Biochain Institute, Inc., Hayward, CA).

3.2.2 Zebrafish Liver and Sample Preparation.

Zebrafish were taken from the Zebrafish Breeding Facility at the University of Alberta (courtesy of Dr. Andrew Waskiewicz), in the University of Alberta Biosciences Aquatics Facility, where they were kept at 28 °C. Strain A/B zebrafish were used in this study. All animals were treated according to established approved animal care protocols. Briefly, six zebrafish were anesthetized with TMS (0.15 mg/mL), and perfused with heparinized phosphate buffered saline (PBS) (25 IU/mL) via caudal puncture. The livers were excised and placed in a 1.5 mL flat-bottomed Eppendorf (Flex-Tube, Eppendorf) containing lysis buffer (Buffer C + protease inhibitor cocktail from compartmental extraction kit + 2 M added PMSF), and placed on ice. Livers were combined and homogenized using a pestle (30 strokes minimum). This mixture was sonicated in a bath of ice water, using four 10 s sonication bursts, with a 1 min rest between sonications. The protein extraction kit was used to separate the proteins into cytosolic, nuclear, membrane, and cytoskeletal components via differential centrifugation according to the kit's instructions (see Figure 3.1). Components were aliquoted and stored at -80 °C for later analysis. This chapter deals only with the cytosolic component of the liver homogenate. It should be noted that we have not done a comprehensive check of the purity of the cytosolic component (e.g., for contaminating nuclear or mitochondrial proteins) in this analysis. This is because we carried out compartment separation for the purpose of simplifying the liver proteome, not for the purpose of generating a pure subcellular compartment.

3.2.3 Acetone Precipitation and In-Solution Digestion.

Standard reduction of the disulfide bonds and alkylation were carried out on the cytosolic component extracted from the liver sample. A general outline of the procedures applied to the sample can be seen in Figure 3.1. Briefly, 3.5 mg of the cytoplasmic protein component was split equally into two 1.5 mL siliconized vials. These were then each reduced with 28 μL of 900 mM DTT for 1 h at 37 °C. Free thiol groups were blocked by reaction with a double volume of 900 mM iodoacetamide for 1 h at room temperature in the dark. The extracts were acetone-precipitated to remove detergent, unreacted DTT, and iodoacetamide. Acetone, precooled to -80 °C, was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The mixture was kept at -20 °C overnight and centrifuged (14 000 rpm, 10 min, 4 °C). The supernatant was decanted and properly disposed. Acetone was evaporated at room temperature.

Ammonium bicarbonate (50 mM, pH 8.0) was used to redissolve the pellet in each vial. Intermittent vortexing was applied, each of the two vials was centrifuged at 14 000 rpm for 5 min at 4 °C, and these two supernatants were pooled. Trypsin

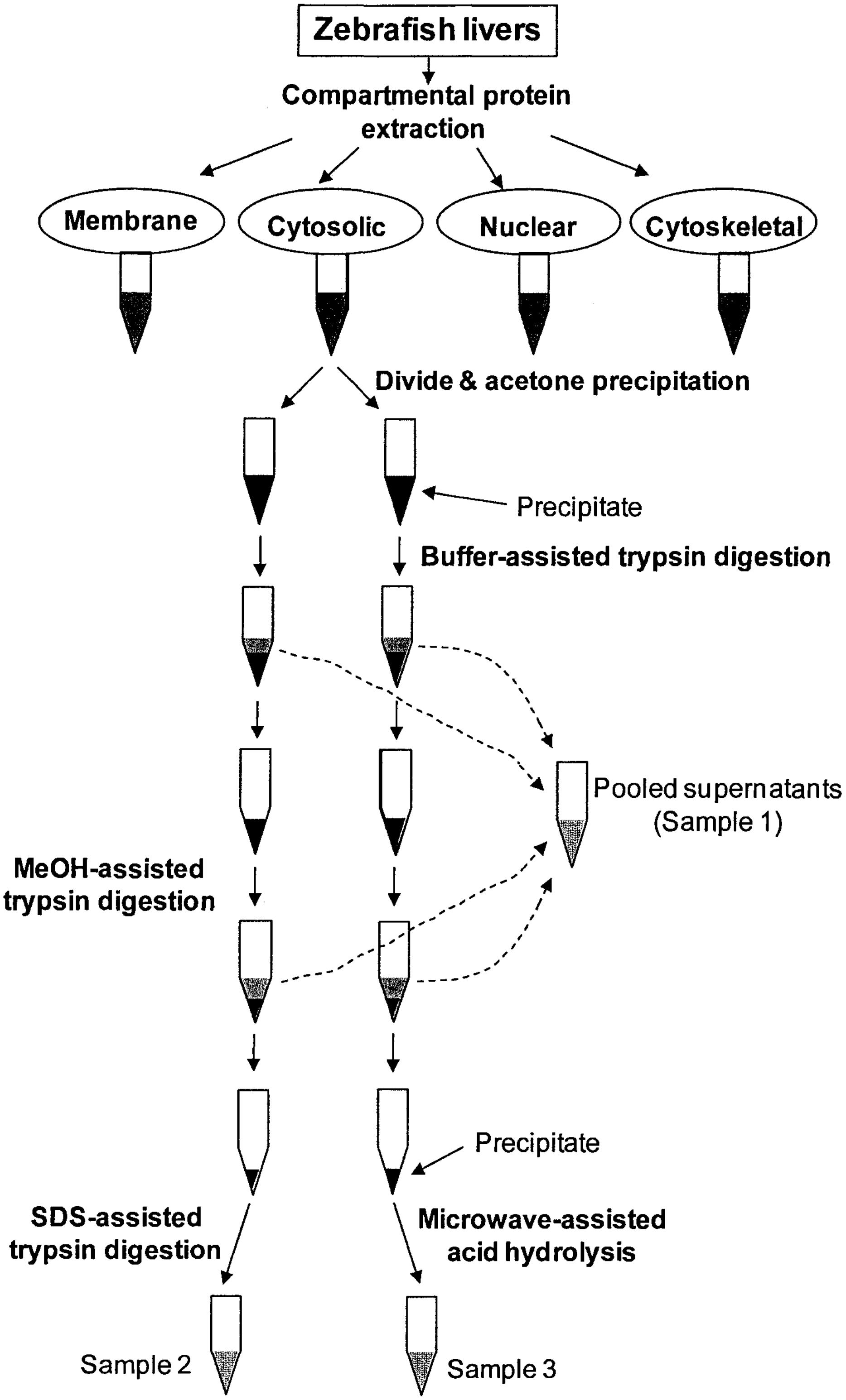

Zebrafish livers
Compartmental protein extraction
Membrane
Cytosolic
Nuclear
Cytoskeletal
Divide & acetone precipitation
Precipitate
Buffer-assisted trypsin digestion
Pooled supernatants (Sample 1)
MeOH-assisted trypsin digestion
Precipitate
SDS-assisted trypsin digestion
Microwave-assisted acid hydrolysis
Sample 2
Sample 3

Figure 3.1 Workflow for sample preparation.

solution was added into the supernatant for an enzyme/protein ratio of 1:45, and digestion was conducted at 37 °C overnight.

To maximize the digestion efficiency of the remaining pellets, the following three additional digestion techniques were applied: (1) methanol-assisted solubilization and subsequent proteolysis,[54-56] (2) SDS-assisted solubilization and subsequent proteolysis,[57,58] and (3) microwave-assisted acid hydrolysis.[53] Each pellet that remained following ammonium bicarbonate treatment was resuspended in 60% MeOH, with sufficient vortexing. Trypsin was added at an enzyme/protein ratio of 1:30, and the solution was incubated at 37 °C for 5 h. The solution was then centrifuged (14 000 rpm, 5 min, 4 °C), and the supernatants from the two parallel vials were pooled. MeOH in the supernatant was evaporated by SpeedVac (Thermo Savant, Milford, MA), and the supernatant was then pooled together with the digests from the buffer trypsin digestion to form Sample 1 (see Figure 3.1). After this MeOH-assisted solubilization and digestion, undissolved pellet in one of the two vials was centrifuged and redissolved in 40 μL of 1% SDS (SDS-assisted solubilization), followed by 20-fold dilution. Trypsin was added to achieve a final enzyme/protein ratio of 1:45. The sample was incubated at 37 °C overnight to complete the digestion (i.e., Sample 2 in Figure 3.1). The protein pellet in the other vial was resuspended in 120 μL of 25% TFA and then microwaved for 10 min.[53] The hydrolysate was dried using a SpeedVac. This vial was then filled with 50% ACN and vortexed, and the ACN was evaporated using a SpeedVac to form Sample 3 (see Figure 3.1). All digestion solutions (i.e., Samples 1-3) were stored at -80 °C until further analysis.

3.2.4 Cation Exchange Chromatography

Peptide mixtures were separated by strong cation exchange (SCX) chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 2.1 × 150 mm Hydrocell SP 1500 column (5 μm, catalog no.: 24-34 SP, BioChrom Labs, Inc., Terre Haute, IN). The buffer solutions used were 20% (v/v) ACN in 0.1% TFA (Buffer A) and 20% (v/v) ACN in 0.1% TFA and 1 M NaCl (Buffer B). Protein digests from each method (i.e., Samples 1-3) were loaded separately onto the SCX column, and peptides were eluted using linear gradients (0-8% Buffer B for 2 min, 8-25% Buffer B for 10 min, 25-50% Buffer B for 2 min) at 0.30 mL/min, with collection of 1 min fractions. In all, 10 fractions were collected based on the chromatography signal recorded at 214 nm. However, the last three fractions were pooled because of their low UV absorbance signals. Therefore, in total, eight fractions per sample were obtained and concentrated to ~20 µL using a SpeedVac.

3.2.5 LC-ESI QTOF MS and MS/MS Analysis.

The peptides in each SCX fraction were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, 2 µL of peptide solution from each SCX fraction was injected onto a 75 µm × 100 mm Atlantis dC18 column (particle size 3 µm, Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were first separated using 120 min gradients (6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min) and electrosprayed into the mass spectrometer (fitted with a nanoLockSpray source) at a flow rate of 300 nL/min. Mass spectra were acquired from *m/z* 300-1600 for 1 s, followed by 4 data-dependent MS/MS scans from *m/z* 50-1900 for 1.5 s each. The collision energy used to perform MS/MS was varied according to the mass and charge state of the eluting peptide. Leucine Enkephalin and (Glu1)-Fibrinopeptide B, a mixed mass calibrant (i.e., lock-mass), was infused at a rate of 250 nL/min, and an MS scan was acquired for 1 s every 1 min throughout the run. An exclusion list was generated based on MASCOT (Matrix Science, London, U.K.) searching results of peptides with a score of 30, which is, on average, 10 points above the identity threshold. A 180 min gradient run, including the exclusion list, was then performed for the same fraction.

3.2.6 Protein Database Search.

Raw search data were lock-mass-corrected, de-isotoped, and converted to peak list files by ProteinLynx Global Server 2.1.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program.

Database searching was restricted to *Danio rerio* (zebrafish) in the NCBI database. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, ±30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, *N*-Acytyl (Protein), oxidation (M), Pyro_glu (N-term Q), Pyro_glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, Mascot threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides with scores lower than the Mascot threshold score for identity were then deleted from the protein list.

The single peptide hits with a matching score lower than 40 (lower than 55 for non-tryptic peptide hits) but above the MASCOT threshold score for identity (averagely 23 for tryptic peptides and 45 for non-tryptic peptides) were manually analyzed. The peptide hit was considered as positive identification if the fragment ions contained more than 5 isotopically resolved y-, b-, or a-ions and the major fragment ion peaks with high intensity (i.e., peak intensity of >30% in a normalized spectrum). Most of the high-intensity fragment ions (i.e., top 5) must also belong to y-, b-, or a-ions, not internal fragment ions. Single peptide hits which failed to meet these criteria were removed from the protein lists. The protein lists were then manually examined. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

3.2.7 Hydropathy Calculation and Annotation of Functions and Localization.

All peptides identified were examined using the ProtParam program available at the EXPASY Web site (http://us.expasy.org/tools/protparam.html), which allows for calculation of the grand average of hydropathy (GRAVY). Proteins exhibiting positive values were considered hydrophobic, and those that exhibit negative values were considered hydrophilic.[59,60]

Once a peptide was predicted to be part of a specific identifiable protein, a protein characterization scheme, utilized in a microsome proteome assay as performed by Garcia et al. (with minor modifications made to the 'Molecular Function' categorization), was used to characterize the zebrafish cytosolic proteome.[61] Using this framework, we categorized protein three ways: according to their (1) Cellular Process, (2) Molecular Function, and (3) Subcellular Localization. To do this, the protein name was searched on either the Zebrafish Information Network database (http://zfin.org); the ExPASy (Expert Protein Analysis System) Proteomics Server (http://ca.expasy.org); Human Protein Reference Database (http://www.hprd.org), or NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene), and its biological process, molecular function, and subcellular localization were determined.

If the peptide was not identified by Mascot, or did not correspond to a known protein name, its 'unique peptide sequence' was run through NCBI BLAST (National Center for Biotechnology Information Basic Local Alignment Search Tool http://www.ncbi.nlm.nih.gov/BLAST/) to see if a likely homology-based match could be found. If a match could be found to *Danio rerio,* or other organisms in the database, then it was characterized using the methods listed previously. If no apparent matches could be found, the peptide was listed as 'unknown' in all three characterization categories.

Proteins listed in the "Cellular Process" scheme included those involved in Cellular Organization and Biogenesis, Immune Response, Metabolism and Energy Pathways, Protein Metabolism, Signal Transduction and Cell Communication, Transport, Other, and Unknown. Those listed in the "Molecular Function" scheme included those involved in Binding (other than GTP-binding), Catalytic Activity, Chaperone Activity, GTP-binding, Kinases and Phosphatases, Motor Activity, Signal Transducer Activity, Structural Molecule Activity, Transcription or Translation regulation, Transporter Activity, Other, and Unknown. Finally, those proteins listed in the "Subcellular Localization" scheme include the following categories: Cytoplasmic, Endoplasmic Reticular, Extracellular, Golgi Apparatus, Membrane, Mitochondrial, Nuclear, Ribosomal, Other, and Unknown.

3.3 Results and Discussion

To generate a comprehensive view of the liver proteome, the homogenized liver sample was initially fractionated into four components, cytosolic, membrane, nuclear, and cytoskeletal, using a compartmental protein extraction kit (see Figure 3.1). Of these, the cytosolic component is of particular interest as it is the soluble fraction expected to contain potential biomarkers of toxicity. As demonstrated below, with the use of a combined protein digestion protocol and 2-D LC QTOF MS/MS, more than one thousand proteins can already be detected in this fraction alone, providing an excellent starting point for future quantitative proteome analysis for biomarker discovery. The analysis of other components, such as the membrane component, expected to contain many hydrophobic proteins at low concentrations, requires further optimization of the current technique. We will report the proteome analysis results of other components in the future.

As Figure 3.1 shows, the cytosolic component was acetone-precipitated from the extraction buffer. After dissolving the protein pellet in a basic buffer and subjecting it to trypsin digestion, there were non-soluble samples remaining in the vials. These remaining samples were subjected to additional levels of digestions. The first was methanol-assisted trypsin digestion,[54,55] and the second was SDS-assisted trypsin digestion.[56-58] Both methods were found to be useful for handling membrane proteins.[54-58] As shown below, these methods also provided additional proteome coverage of cytosolic proteins. Finally, a newly developed microwave-assisted acid hydrolysis (MAAH) method from our lab[53] was applied to the protein pellet presumably composed of denatured proteins not soluble in the basic buffer and methanol. Digests produced from each of the techniques were subjected to strong cation exchange chromatographic separation on a narrow bore SCX column. Figure 3.2A shows a representative UV absorbance chromatogram of the peptide mixtures from one of the three digests. The tryptic digests from the buffer digestion and the methanol method were pooled into one sample, while the digest from the SDS method was analyzed separately as it contained the interfering surfactant SDS. In a separate experiment, it was found that 2-D LC-ESI MS/MS analysis of a BSA digest from the SDS-assisted digestion method detected fewer number of peptides than that from digestion in a buffer solution. It appears that SDS molecules strongly bound to some peptides might not be removed completely during the LC separation, resulting in suppression of the peptide signals in MS/MS. Thus, the SDS digest was analyzed separately from the buffer and methanol digests. The peptide mixture generated by MAAH was analyzed in a separate run, as it would contain non-tryptic peptides.

Figure 3.2A shows a representative UV absorbance SCX-chromatogram of the peptide mixtures from one of the three samples. In all three samples, no sharp chromatographic peaks were observed, indicating that, despite multi-level of solubilization and digestion, each sample was still a complex peptide mixture. In each sample, 8 SCX fractions were generated. For each fraction, good quality reversed-phase LC separation and detection was achieved, as seen in a representative base-peak ion chromatogram shown in Figure 3.2B. In the MS and MS/MS acquisition settings, each MS scan was set for 1 s. Figure 3.2C shows an example of the MS spectra collected during the run. After the 2-D separations, MS spectra acquired during isolated scans looked relatively simple, and ion suppression was greatly reduced. After one MS scan, one MS/MS scan was acquired from each of the four most intense MS peaks. Each MS/MS scan was set for 1.5 s. Figure 3.2D shows an example of the MS/MS spectra obtained. The relatively high-abundance peptides identified during the 120-min gradient run served as exclusion lists in terms of *m/z* values for the extended 180-min run. This strategy enabled additional, less intense peptides to be identified (see Supporting Information 3.1).

Using the QTOF instrument, when the raw data were processed, the masses of the precursor ions and fragment ions were automatically corrected, resulting in high mass accuracy. The peptide mass tolerance and fragment ion mass tolerance were set to ±30 ppm and 0.2 Da, respectively. The scoring algorithm outlined in the Mascot server shows that by narrowing these mass tolerance parameters, the identification threshold can be decreased, improving the resulting scores for the identified peptides. Because the QTOF instrument combines reasonably high sensitivity with high resolution, the MS/MS data acquired were generally of high quality. Figure 3.3 displays three representative MS/MS spectra of peptides sequenced from the three digest samples. Panels A and B of Figure 3.3 show the MS/MS spectra of tryptic peptides from buffer/methanol and SDS digests, respectively. Figure 3.3C shows the MS/MS spectrum of a non-tryptic peptide from the microwave-assisted acid hydrolysate. Even though non-tryptic peptides are not ionized as favorably as tryptic peptides ending with lysine or arginine on the C-terminus in the ESI mode, the quality of the spectrum is still good. Non-tryptic peptides are considered identified if the scores are above the threshold and the measured peptide mass matches with the predicted mass within the experimental error. The spectrum shown in Figure 3.3C had an identification threshold of 44 and a score of 60.

Taken together, proteome analyses from the three digest samples prepared using different solubilization/digestion methods led to the identification of 1204 unique proteins. Among the 1204 proteins identified, 224 (19%) were found in all three samples, while 113 (9%), 420 (35%), and 214 (18%) proteins or related protein groups were uniquely observed in buffer/methanol digest, SDS digest, and MAAH digest, respectively (see Figure 3.4). An important observation is that, after the buffer and methanol digests, the remaining pellet still contained a large number of proteins. In all, 695 unique proteins (58%) could be identified by using a combination of the SDS trypsin digestion and MAAH methods. Not counting the common proteins identified by the three methods (224), only 61 proteins were commonly identified by the SDS and MAAH methods, justifying the need of carrying out two individual digests. These results demonstrate that the sample preparation protocol outlined in Figure 3.1 is effective for generating more comprehensive proteome coverage of the cytosolic fraction of the zebrafish liver than using one digestion method alone.

We have annotated some of the physicochemical and biological characteristics of the obtained proteome data in order to provide a better understanding of the proteome itself. (Supporting Information 3.2) lists all proteins identified from the cytosolic component of the zebrafish liver sample, along with information about subcellular locations and functions of the proteins, GRAVY index of the proteins, peptide sequences, peptide masses, and matching scores.

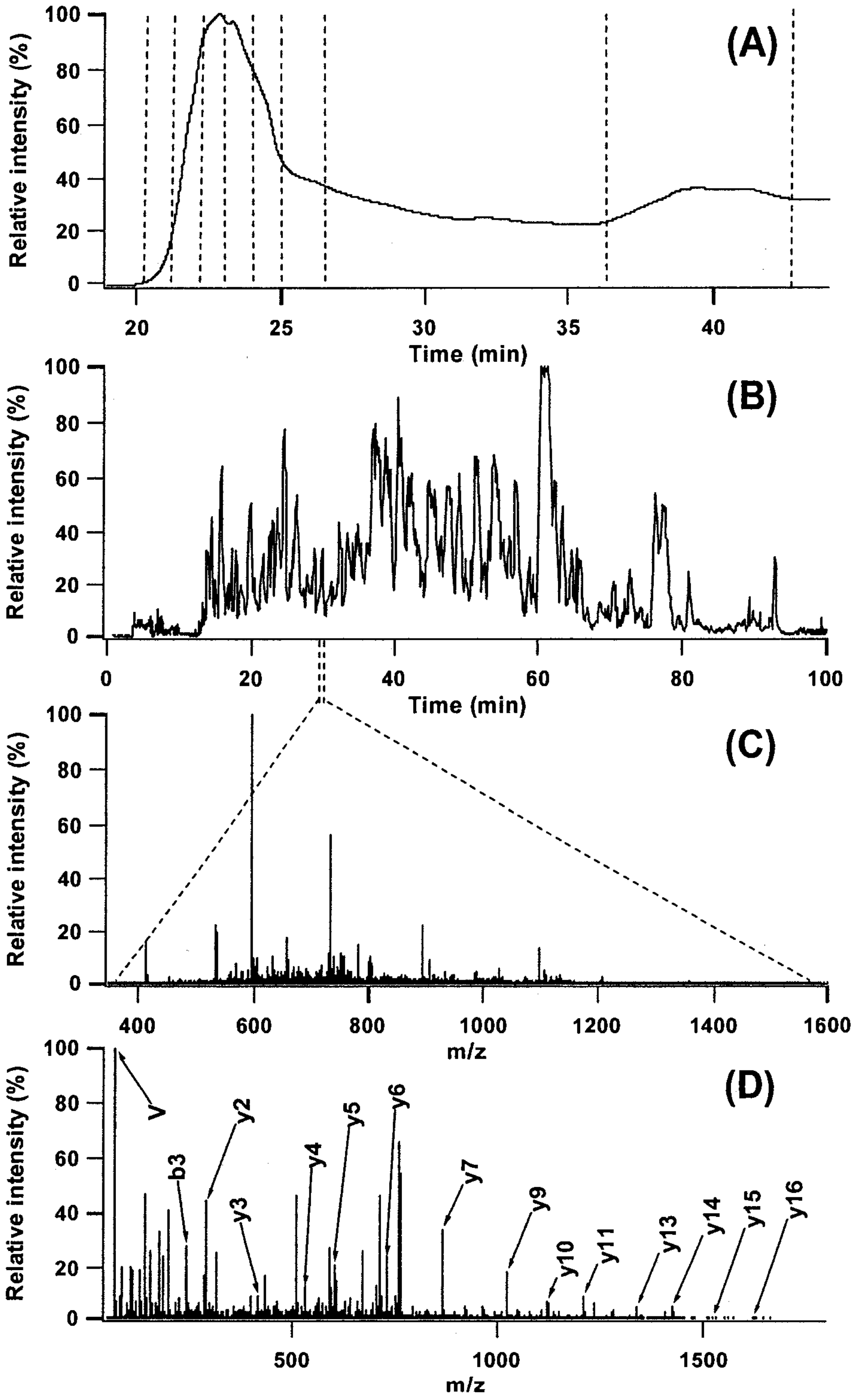


Figure 3.2 2-D LC-MS/MS analysis of the peptide mixture: (A) 214 nm UV chromatogram of SCX separation; (B) ion chromatogram of reversed-phase LC separation of a SCX fraction; (C) MS spectrum of peptides detected in the 1 s time window indicated in panel B; and (D) MS/MS spectrum of a peptide ion selected from panel C.

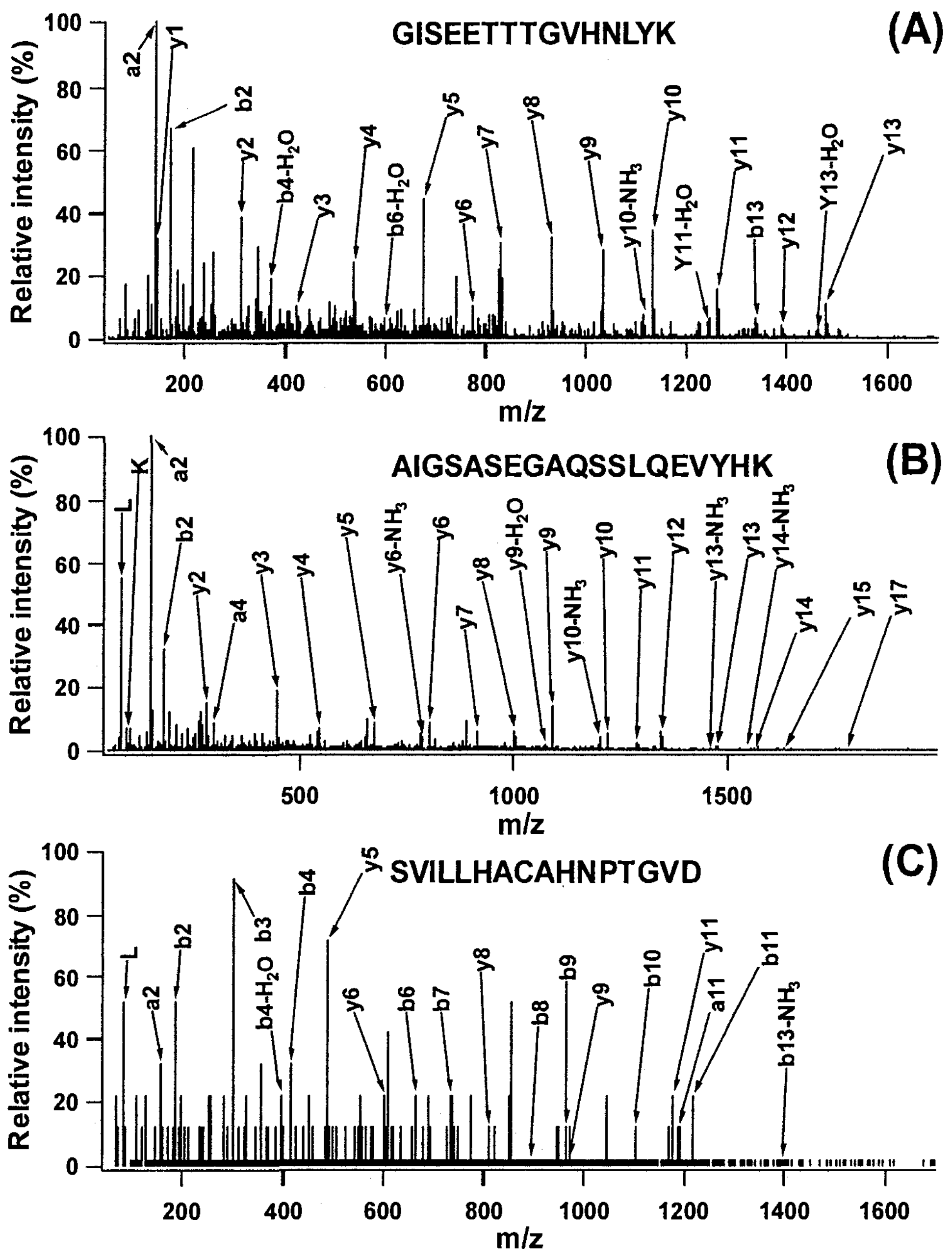


Figure 3.3 ESI-MS/MS spectra of representative peptides corresponding to proteins identified in samples prepared by different digestion methods: (A) from a peptide detected in buffer- and MeOH-assisted tryptic digest belonging to gi|40363541, *S*-adenosylhomocysteine hydrolase [*Danio rerio*]; (B) from a peptide detected in SDS-assisted tryptic digest belonging to gi|45387823, proteasome subunit, alpha type, 5 [*D. rerio*]; and (C) from a peptide detected in microwave digest belonging to gi|47087061, glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2) [*D. rerio*].

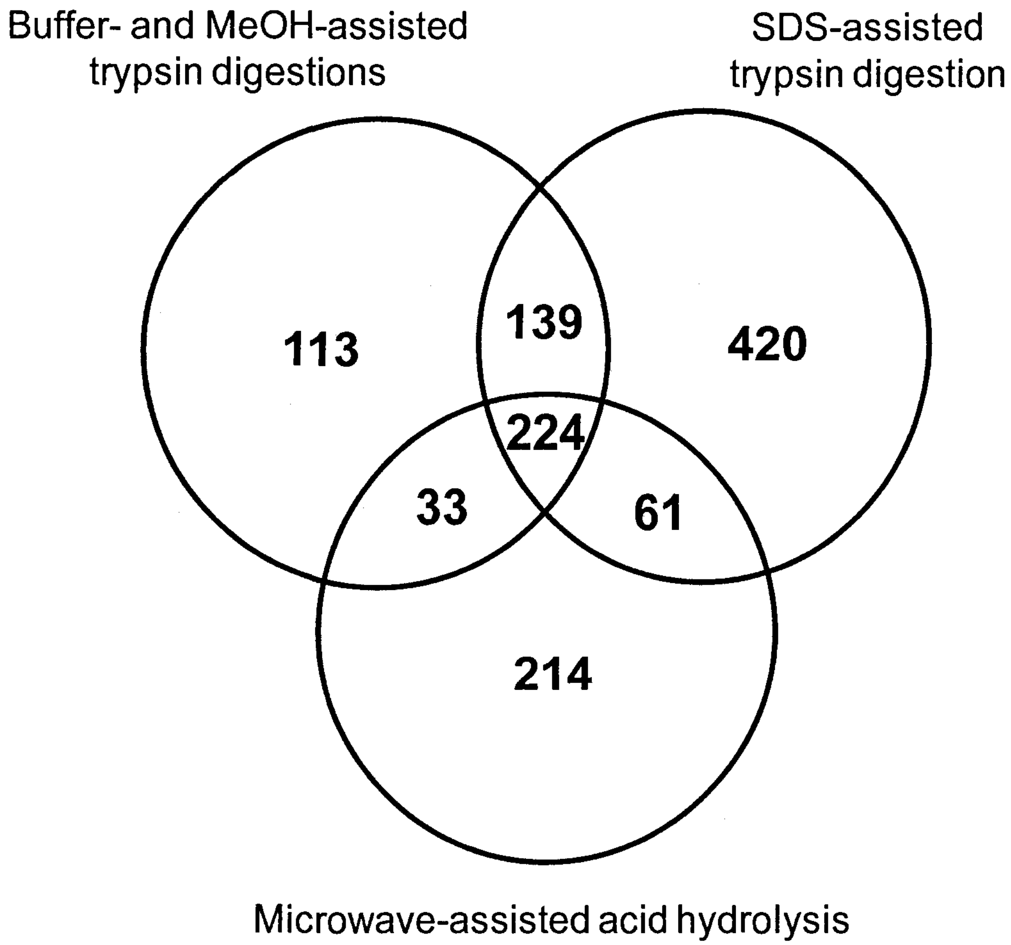


Figure 3.4 Comparison of protein identification results from three digestion techniques. A total of 1204 proteins was identified in the cytosolic component of the zebrafish liver proteome: 509 proteins from the buffer and methanol digests, 843 proteins from the SDS-assisted digest, and 531 proteins from the microwave digest.

GRAVY is a commonly used parameter to gauge the hydropathicity of proteins or peptides. Proteins were categorized into four groups according to their GRAVY indexes: proteins with GRAVY indexes lower than -0.5 were considered hydrophilic; proteins with GRAVY indexes higher than -0.5 but lower than 0 were considered mildly hydrophilic; proteins with GRAVY indexes higher than 0 but lower than 0.5 were considered mildly hydrophobic; and proteins with GRAVY indexes higher than 0.5 were considered hydrophobic. The distribution of GRAVY indexes for the proteins found in each digest is summarized in Figure 3.5. Hydrophilic and mildly hydrophilic proteins were found to take up the majority of proteins found in each digestion (459 out of 509 proteins or 90% in buffer/methanol digest, 732 out of 844 proteins or 87% in SDS digest, and 474 out of 532 or 89% in MAAH digest) (Figure 3.5). This is consistent with the fact that our analysis is specifically from the cytosolic component of the liver sample where we would not expect to find a large amount of hydrophobic proteins.

The SDS and MAAH methods appear to be better at identifying hydrophobic proteins. The SDS and MAAH digestion each contain a slightly higher percentage of mildly hydrophobic and hydrophobic proteins (111 out of 844 proteins or 13% and 57 out of 532 proteins or 11%, respectively) compared to the buffer/methanol digestion (50 out of 509 proteins or 10%) (Figure 3.5). A similar distribution can also be found in Supporting Information 3.3 which shows the distribution of GRAVY indexes of unique proteins identified in each digestion. A much lower percentage of hydrophilic proteins was found in both the SDS digestion (86 out of 420 proteins or 20%) and MAAH digestion (65 out of 214 proteins or 30%) compared to the buffer/methanol digestion (60 out of 113 proteins or 53%) (Supporting Information 3.3). These results indicate that both the SDS digestion and MAAH digestion show better performance in redissolving and digesting proteins with higher hydrophobicity. In addition, many other denatured proteins not soluble in buffer/methanol were digested in the SDS or MAAH method.

Using the Zebrafish Information Network, ExPASy, Human Protein Reference Database, NCBI Entrez, and NCBI BLAST, the proteins were characterized, utilizing a scheme adapted from Garcia et al.[61] Under this framework, proteins were sorted 3 ways,

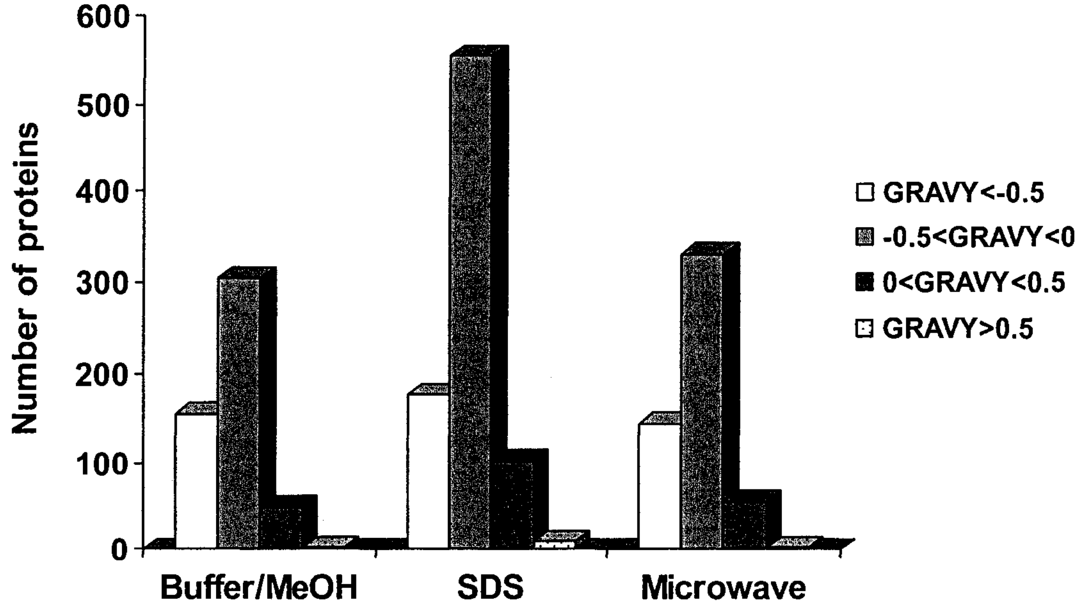


Figure 3.5 Distribution of GRAVY (Grand Average of Hydropathy) indexes of all identified proteins (1204). Proteins were sorted into four groups based on their GRAVY indexes: hydrophilic (GRAVY < -0.5), mildly hydrophilic (-0.5 < GRAVY < 0), mildly hydrophobic (0 < GRAVY < 0.5), and hydrophobic (GRAVY > 0.5).

according to their (1) Cellular Process, (2) Molecular Function, or (3) Subcellular Localization (Figures 3.6-3.8). Subcellular localization is a key functional characteristic of proteins. Proteins can only function optimally in a specific subcellular localization; hence, the determination of subcellular localization of each protein is an important step for large-scale proteomic analysis to provide reliable annotations regarding the biological functions of proteins.

This method of protein characterization yielded a robust snapshot of the liver proteome. Looking at cellular process, we can see that a majority of proteins are involved in some form of metabolism, whether it is general metabolism and energy pathways (41%) or protein metabolism (27%) (Figure 3.6). With regards to molecular function, the largest proportion of proteins performed a catalytic activity (41%), a binding activity (15%), or were structural molecules (13%) (Figure 3.7). The largest proportion of proteins had a subcellular localization in the cytoplasm (36%), and a large portion had an unknown localization (20%) (Figure 3.8). The relatively lower number of proteins identified as truly cytoplasmic (36%) likely is the result of the procedure for isolation. The cytoplasmic fraction is collected first following hypo-osmotic lysis under nondenaturing conditions. Therefore, proteins that have formed stable associations with other organelles (e.g., transmembrane proteins, organellar associated proteins, cytoskeletal elements) would not be found in the cytosolic fraction, even though they have no apparent linkage with these other fractions. Nevertheless, the use of the combined methods of digestion and subsequent MS analyses increases the overall detectability of the cytoplasmic proteins. This is illustrated in Figure 3.9. Out of 433 cytoplasmic proteins, 189 (or 43.6%) proteins were identified in the buffer- and methanol-assisted trypsin digest. Another 177 (or 40.9%) proteins were identified in the SDS-assisted trypsin digest. The use of MAAH allows the identification of an additional 67 (or 15.5%) proteins.

Many known protein biomarkers of toxicity were found, such as epoxide hydrolase, superoxide dismutase, heat shock proteins, vitellogenin, and transaminases. Epoxide hydrolase is a detoxifying enzyme important in drug metabolism,[62] and functions by hydrolyzing toxic epoxides, which include xenobiotics such as styrene

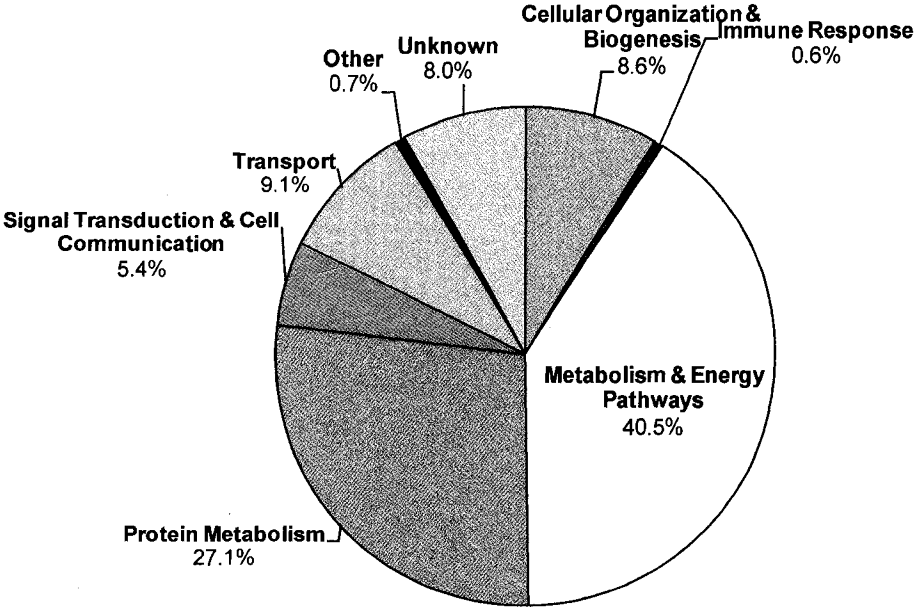


Figure 3.6 Cellular processes of the zebrafish liver proteome, cytosolic component. All 1204 proteins were grouped according to their cellular process, using the Zebrafish Information Network, ExPASy, Human Protein Reference Database, NCBI Entrez Gene, and NCBI BLAST databases.

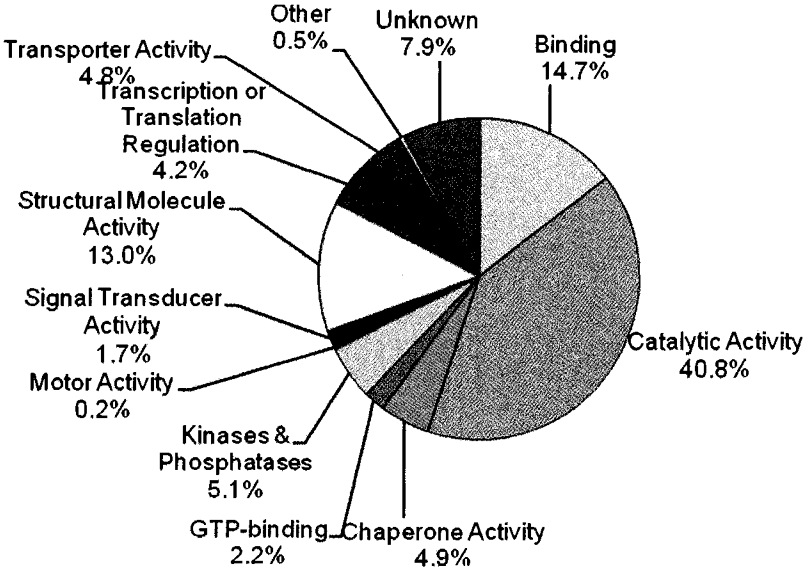


Figure 3.7 Molecular functions of the zebrafish liver proteome, cytosolic component. All 1204 proteins were grouped according to their molecular function, using the Zebrafish Information Network, ExPASy, Human Protein Reference Database, NCBI Entrez Gene, and NCBI BLAST databases.

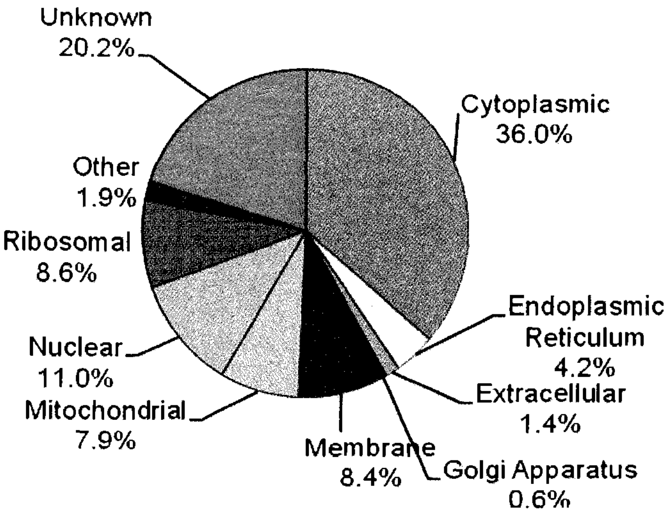


Figure 3.8 Subcellular localizations of the zebrafish liver proteome, cytosolic component. All 1204 proteins were grouped according to their subcellular localization, using the Zebrafish Information Network, ExPASy, Human Protein Reference Database, NCBI Entrez Gene, and NCBI BLAST databases.

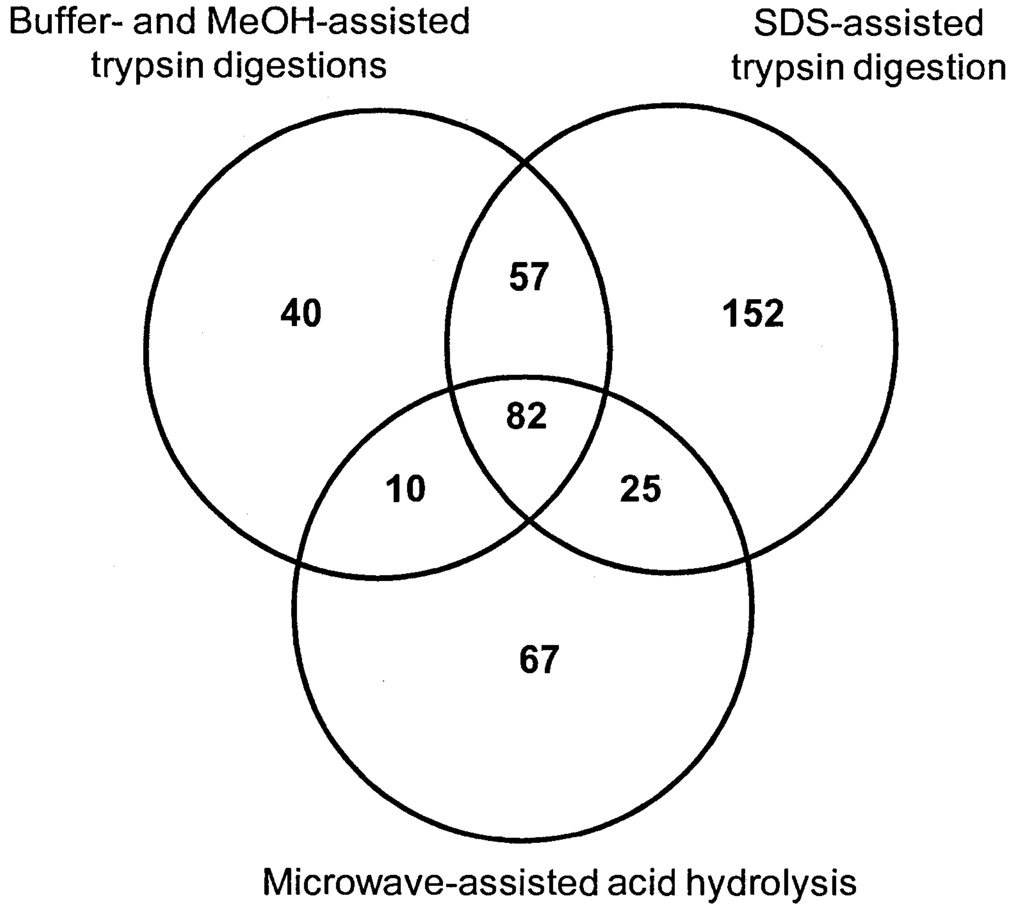


Figure 3.9 Comparison of cytoplasmic protein identification results from three digestion techniques. A total of 433 known cytoplasmic proteins was identified.

oxide.[63] Epoxide hydrolase levels have been found to be decreased in diseased and drug-exposed livers, implicating it as a toxicity biomarker.[62,64] Superoxide dismutase (SOD) is an antioxidant enzyme,[65] and is known to be upregulated in the presence of xenobiotics.[66] Since many xenobiotics induce oxidative stress in an organism, SOD is thought to be a biomarker of xenobiotic exposure.[26] Heat shock proteins are a suite of highly conserved proteins that respond to a variety of cell stresses, including temperature change, heavy metals, and hypoxia.[67] As such, they have a fairly established history as biomarkers.[21,68,69] Vitellogenin is a yolk protein produced by the liver that is normally detected in high amounts only in females.[70] Its upregulation, especially in male organisms, is used as a biomarker to indicate the presence of endocrine disruptors such as nonylphenol and 17 -oestradiol.[23,71] Transaminases are enzymes found primarily in the liver, which catalyze the transfer of amino groups in amino acids.[72] Transaminases, such as alanine aminotransferase, have been used as biomarkers to indicate the presence of organophosphorus pesticides, and their presence is often indicative of liver damage.[73]

The presence of biomarkers such as these in our proteome analysis is invaluable, as they will provide a relative "internal standard" to help in validating our exposures against known responses. If no known markers were detected then one would question the usefulness of the protein profile, and we would not know if the response is similar to previous research. Since a great number of proteins were identified, future work in profiling this list of proteins along with toxicology studies may result in the identification of other markers which may be more specific or sensitive. Toxicants known to induce these particular proteins (above) will be used to validate our exposures and also help to examine the efficacy of each individual toxicant in inducing new sensitive biomarkers of exposure.

3.4 Conclusions

This is the first report on the proteome profile of the cytosolic component of the zebrafish liver. Compartmental protein extraction of liver tissue was used to simplify the proteome at the protein level. The cytosolic fraction was subjected to different levels of protein solubilization and digestion, namely, trypsin digestion in a basic buffer and in methanol, SDS-assisted trypsin digestion, and microwave-assisted acid hydrolysis. The resultant digests were individually analyzed by using 2-D LC-ESI MS/MS. A total of 1204 unique proteins was identified. The number of identified proteins can be considered a good measure of the technical progress of this approach as it greatly increases the coverage of the zebrafish proteome, compared to that of other recently reported techniques.[40,74,75] At this stage, we do not know which of these proteins are unique to the liver. Future analysis of other tissues will allow us to identify proteins which are unique to the liver. As might be expected from the cytosolic component of a liver proteome, the majority of proteins characterized were catalytic, metabolic, or found in the cytoplasm. The demonstration of many protein markers of toxicity found in this particular proteomic analysis bodes well for our future goals of understanding the mechanisms of toxicity and identification of potentially new, more sensitive, and more efficacious biomarkers of exposure. Future work will focus on applying quantitative proteome analysis technique using differential isotope labeling of peptides[76] to compare the proteome changes of cytosolic fractions of livers of zebrafish exposed to different classes of toxicants. In addition, characterizing the remaining nuclear, cytoskeletal, and membrane components of the zebrafish liver proteome will further increase the proteome coverage.

3.5 Literature Cited

1. Choudhuri, S. *Toxicol. Mech. Methods* **2005**, *15*, 1-23.

2. Vetillard, A.; Bailhache, T. *Biol. Reprod.* **2005**, *72*, 119-126.

3. Stephensen, E. K.; Adolfsson-Erici, M.; Celander, M.; Hulander, M.; et al. *Environ. Toxicol. Chem.* **2003**, *22*, 2926-2931.

4. Castaño, A.; Carbonelli, G.; Carballo, M.; Fernandez, C.; et al. *Ecotoxicol. Environ. Saf.* **1998**, *41*, 29-35.

5. Norey, C. G.; Cryer, A.; Kay, J. *Comp. Biochem. Physiol., C* **1990**, *97*, 215-220.

6. Soverchia, L.; Ruggeri, B.; Palermo, F.; Mosconi, G.; et al. *Toxicol. Appl. Pharmacol.* **2005**, *209*, 236-243.

7. Chen, C. Y.; Sillett, K. B.; Folt, C. L.; Whittemore, S. L.; et al. *Hydrobiologia* **1999**, *401*, 229-238.

8. Lam, K. L.; Ko, P. W.; Wong, J. K. Y.; Chan, K. M. *Mar. Environ. Res.* **1998**, *46*, 563-566.

9. Wong, C. K. C.; Yeung, H. Y.; Cheung, R. Y. H.; Yung, K. K. L.; et al. *Arch. Environ. Contam. Toxicol.* **2000**, *38*, 486-493.

10. Wong, C. K. C.; Yeung, H. Y.; Woo, P. S.; Wong, M. H. *Aquat. Toxicol.* **2001**, *54*, 69-80.

11. Blechinger, S. R.; Warren, J. T.; Kuwada, J. Y., Jr.; Krone, P. H. *Environ. Health Perspect.* **2002**, *110*, 1041-1046.

12. Coverdale, L. E.; Lean, D.; Martin, C. C. *Curr. Genomics* **2004**, *5*, 395-407.

13. Roesijadi, G. *Environ. Health Perspect.* **1994**, *102*, 91-95.

14. Marchant, G. E. *Trends Biotechnol.* **2002**, *20*, 329-332.

15. Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837-846.

16. Matsumoto, M.; Hatakeyama, S.; Oyamada, K.; Oda, Y.; et al. *Proteomics* **2005**, *5*, 4145-4151.

17. Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; et al. *Mol. Cell. Biol.* **1999**, *19*, 7357-7368.

18. Gygi, S.; Rochon, Y.; Franza, B. R.; Aebersold, R. *Mol. Cell. Biol.* **1999**, *19*, 1720-1730.

19. Hochstrasser, D. F.; Sanchez, J. C.; Appel, R. D. *Proteomics* **2002**, *2*, 807-812.

20. Fisher, E. H. In *Proteome Research: New Frontiers in Functional Genomics*; Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. Eds.; Springer: Berlin, 1997; preface.

21. Hallare, A. V.; Pagulayan, R.; Lacdan, N.; Kohler, H. R.; et al. *Environ. Monit. Assess.* **2005**, *104*, 171-187.

22. Kohler, H. R.; Bartussek, C.; Eckwert, H.; Farian, K.; et al. *J. Aquat. Ecosyst. Stress Recovery* **2001**, *8*, 261-279.

23. Routledge, E. J.; Sheahan, D.; Desbrow, C.; Brighty, G. C.; et al. *Environ. Sci. Technol.* **1998**, *32*, 1559-1565.

24. Palmer, B. D.; Huth, L. K.; Pieto, D. L.; Selcer, K. W. *Environ. Toxicol. Chem.* **1998**, *17*, 30-36.

25. Amacher, D. E.; Adler, R.; Herath, A.; Townsend, R. R. *Clin. Chem.* **2005**, *51*, 1796-1803.

26. Pedrajas, J. R.; Peinado, J.; Lopez-Barea, J. *Free Radical Res. Commun.* **1993**, *19*, 29-41.

27. Martin, S. A. M.; Cash, P.; Blaney, S.; Houlihan, D. F. *Fish Physiol. Biochem.* **2001**, *24*, 259-270.

28. Chen, T. Y.; Shiau, C. Y.; Wei, C. I.; Hwang, D. F. *J. Agric. Food Chem.* **2004**, *52*, 2236-2241.

29. Berrini, A.; Tepedino, V.; Borromeo, V.; Seochi, C. *Food Chem.* **2006**, *96*, 163-168.

30. Pichler, F. B.; Laurenson, S.; Williams, L. C.; Dod., A.; et al. *Nat. Biotechnol.* **2003**, *21*, 879-883.

31. Jekosch, K. *Methods Cell Biol.* **2004**, *77*, 225-239.

32. Lin, D.; Tabb, D. L.; Yates, J. R. *Biochim. Biophys. Acta* **2003**, *1646*, 1-10.

33. Tyers, M.; Mann, M. *Nature* **2003**, *422*, 193-197.

34. Kim, J.; Kim, S. H.; Lee, S. U.; Ha, G. H.; et al. *Electrophoresis* **2002**, *23*, 4142-4156.

35. Seow, T. K.; Ong, S. E.; Liang, R. C. M. Y.; Ren, E. C.; et al. *Electrophoresis* **2000**, *21*, 1787-1813.

36. Thome-Kromer, B.; Bonk, I.; Klatt, M.; Nebrich, G.; et al. *Proteomics* **2003**, *3*, 1835-1862.

37. Chanson, A.; Sayd, T.; Rock, E.; Chambon, C.; et al. *J. Nutr.* **2005**, *135*, 2524-2529.

38. Strey, C. W.; Winters, M. S.; Markiewski, M. M.; Lambris, J. D. *Proteomics* **2005**, *5*, 318-325.

39. Lopez-Hellin, J.; Gonzalo, R.; Tejeda, M.; Carrascal, M.; et al. *Clin. Sci.* **2005**, *108*, 167-178.

40. Tay, T. L.; Lin, Q.; Seow, T. K.; Tan, K. H.; et al. *Proteomics* **2006**, *6*, 3176-3188.

41. Yan, W.; Lee, H.; Yi, E. C.; Reiss, D.; et al. *GenomeBiology* **2004**, *5*, Art No. R54.

42. Lee, C. L.; Hsiao, H. H.; Lin, C. W.; Wu, S. P.; et al. *Proteomics* **2003**, *3*, 2472-2486.

43. Ou, K.; Seow, T. K.; Liang, R. C. M. Y.; Ong, S. E.; et al. *Electrophoresis* **2001**, *22*, 2804-2811.

44. Welch, K. D.; Wen, B.; Goodlett, D. R.; Yi, E. C.; et al. *Chem. Res. Toxicol.* **2005**, *18*, 924-933.

45. Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Matthews, D. E.; et al. *Anal. Chem.* **2004**, *76*, 4951-4959.

46. Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198-207.

47. Gygi, S. P.; Rist, B.; Griffin, T. J.; Eng, J.; et al. *J. Proteome Res.* **2002**, *1*, 47-54.

48. Julka, S.; Regnier, F. *J. Proteome Res.* **2004**, *3*, 350-363.

49. Ong, S.; Mann, M. *Nat. Chem. Biol.* **2005**, *1*, 252-262.

50. Roe, M. R.; Griffin, T. J. *Proteomics* **2006**, *6*, 4678-4687.

51. Jiang, X. S.; Zhou, H.; Zhang, L.; Sheng, Q. H.; et al. *Mol. Cell. Proteomics* **2004**, *3*, 441-455.

52. Yan, W.; Lee, H.; Deutsch, E. W.; Lazaro, C. A.; et al. *Mol. Cell. Proteomics* **2004**, *3*, 1039-1041.

53. Zhong, H.; Marcus, S. L.; Li, L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 471-481.

54. Simon, L. M.; Kotorman, M.; Garab, G.; Laczko, I. *Biochem. Biophys. Res. Commun.* **2001**, *280*, 1367-1371.

55. Russell, W. K.; Park, Z. Y.; Russell, D. H. *Anal. Chem.* **2001**, *73*, 2682-2685.

56. Blonder, J.; Goshe, M. B.; Moore, R. J.; Pasa-Tolic, L.; Masselon, C. D.; Lipton, M. S.; Smith, R. D. *J. Proteome Res.* **2002**, *1*, 351-360.

57. Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946-951.

58. Zhang, N.; Li, N.; Li, L. *J. Proteome Res.* **2004**, *3*, 719-727.

59. Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 105-132.

60. Phillips, C. I.; Bogyo, M. *Cell. Microbiol.* **2005**, *7*, 1061-1076.

61. Garcia, B. A.; Smalley, D. M.; Cho, H.; Shabanowitz, J.; et al. *J. Proteome Res.* **2005**, *4*, 1516-1521.

62. Guengerich, F. P.; Turvy, C. G. *J. Pharmacol. Exp. Ther.* **1991**, *256*, 1189-1194.

63. Morisseau, C.; Hammock, B. D. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 311-333.

64. Park, J. C.; Han, W. D.; Park, J. R.; Choi, S. H.; et al. *J. Ethnopharmacol.* **2005**, *102*, 313-318.

65. McCord, J. M.; Keele, B. B., Jr.; Fridovich, I. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *68*, 1024-1027.

66. Pedrajas, J. R.; Peinado, J.; Lopez-Barea, *J. Chem.-Biol. Interact.* **1995**, *98*, 267-282.

67. Craig, E. A.; Lindquist, S. *Annu. Rev. Genet.* **1988**, *12*, *631*-677.

68. Wepener, V.; van Vuren, J. H. J.; Chatiza, F. P.; Mbizi, Z.; et al. *Phys. Chem. Earth* **2005**, *30*, 751-761.

69. Dyer, S. D.; Brooks, G. L.; Dickson, K. L.; Sanders, B. M.; et al. *Environ. Toxicol. Chem.* **1993**, *12*, 913-924.

70. Tyler, C. R.; van der Eerden, B.; Jobling, S.; Panter, G.; et al. *J. Comp. Physiol., B* **1996**, *166*, 418-426.

71. Hansen, P. D.; Dizer, H.; Hock, B.; Marx, A.; et al. *Trends. Anal. Chem.* **1998**, *17*, 448-451.

72. Green, D. E.; Leloir, L. F.; Nocito, V. *J. Biol. Chem.* **1945**, *161*, 559-582.

73. de Aguiar, L. H.; Moraes, G.; Avilez, I. M.; Altran, A., E.; et al. *Environ. Res.* **2004**, *95*, 224-230.

74. Shrader, E. A.; Henry, T. R.; Greeley, M. S.; Bradley, B. P.; et al. *Ecotoxicology* **2003**, *12*, 485-488.

75. Bosworth, C. A.; Chou, C. W.; Cole, R. B.; Rees, B. B.; et al. *Proteomics* **2005**, *5*, 1362-1371.

76. Ji, C.; Li, L. *J. Proteome Res.* **2005**, *4*, 734-742.

Chapter 4

Exploring the Precursor Ion Exclusion Feature of LC-ESI Quadrupole Time-of-Flight MS for Improving Protein Identification in Shotgun Proteome Analysis Introduction*

4.1 Introduction

Shotgun or bottom-up proteome analysis is an important technique for generating proteome profiles.[1] This technique is commonly based on the use of liquid chromatography (LC) electrospray ionization (ESI) tandem mass spectrometry (MS/MS) to sequence peptides produced from a digest of a proteome sample. Several tandem MS platforms including ion trap MS and quadrupole time-of-flight (QTOF) MS are currently available with each providing one or several advantageous features including low cost, high speed, sensitivity, specificity, and robustness. Many of these features are intertwined – for example, increasing spectral acquisition speed may result in reduction of detection sensitivity or spectral quality. To increase the proteome coverage or increase the number of proteins identified by the shotgun method, it is vital to optimize the spectral acquisition efficiency which is mainly governed by the speed of spectral acquisition, the quality of the spectra and the frequency of spectral redundancy. For example, on-the-fly or dynamic exclusion of peptides sequenced in previous scans allows efficient use of the instrument time to produce more MS/MS spectra within a LC MS/MS run. Unfortunately, current tandem MS technology is still not adequate to sample (i.e., sequence) all peptides eluted from LC even after extensive pre-fractionation of a complex proteome digest. To mitigate this under-sampling problem, gas phase fractionation has been reported to be useful where multiple runs of a peptide mixture are carried out with each run focusing on detecting peptides in a small m/z window, instead of a single run with a wide m/z window.[2-6] Dividing a wide m/z window into several smaller m/z windows allows more co-eluting peptide ions to be sampled.

Other reported methods of addressing the under-sampling problem include performing replicate runs under the identical conditions[7-15] or with precursor ion exclusion.[16-18] It has been shown that two replicate runs of a complicated proteome digest generally result in the identification of similar number of proteins with about 70 to 80% protein overlaps.[8-15] Thus, by simply running a sample in replicates the number of proteins identified can be increased to some extent, compared to running the sample only once. For example, Liu et al identified a total of about 1375 proteins from three replicate 2D-LC-MS runs of a digest of yeast cell lysate.[9] About 1064 proteins (77.4%) were identified in the first run and 186 additional proteins (13.5%) were identified from the second run and another 125 proteins (9.1%) were identified from the third run. Additional runs beyond the 3rd run could generate more proteins, but the number of unique proteins identified from these runs decreased markedly. In analogy to dynamic exclusion within a LC MS/MS run, exclusion of precursor ions identified from previous run(s) for MS/MS in the new run should, in principle, sample more peptide ions in multiple runs, resulting in identifying more proteins. However, to our knowledge, this strategy has not been adequately developed and certainly is not being widely used.

In an earlier conference report of proteome profiling work of monocytes cell line U937 and their macrophages by using gel-separation of proteins followed by nano-LC ESI MS/MS of protein digests from individual gel-bands, Hui et al identified a total of 1445 proteins from the monocyte protein extract where 1078 (75.0%) proteins were identified in the first run, 226 (15.6%) proteins in the second run using exclusion list of previously acquired precursors in LC MS/MS, and 141 (9.8%) proteins in the third run.[16] Similarly, for the macrophage work, the first run identified 1121 (74.1%) proteins, the second run 273 (18%) and the third run 120 (7.9%). No direct comparison of these data to those generated by running simple replicates without precursor ion exclusion was given. However, comparing their results with those reported in the literature on proteome profiling of similar complexity of protein digests,[8-15] it appears that the number of proteins identified in second or third runs was similar to those identified by using simple replicate runs under the identical running conditions.

Recently, Chen et al reported a method of precursor ion exclusion in offline LC combined with matrix-assisted laser desorption ionization (MALDI) MS/MS.[18] Based on the unique feature of offline LC fractionation onto a MALDI sample plate where the fractionated peptides can be subjected to multiple runs, they first scanned the plate to generate MS spectra which consumed little samples and produced a list of precursor masses of peptides. A subsequent MS/MS run of these peptides resulted in identification of some peptides using database search of the MS/MS spectra. The list of positively identified peptides in this first run then served as the mass exclusion list for the second run, i.e., the precursor masses of the positively identified peptides at a given retention time window were excluded for MS/MS. This process can be repeated for further MS/MS runs. In the analysis of a digest of *E. coli* cell lysate, they demonstrated that the mass exclusion method resulted in a 25% increase in the number of unique peptide identified in the second run, compared to simply pooling MS/MS data from two replicate runs.[18] These encouraging results also illustrate that efficient exclusion of precursor ions is very important for the success of the precursor ion exclusion strategy.

Unlike offline LC-MALDI MS, online LC-ESI MS does not offer the possibility of re-examining the same peptide mixtures at a given retention time which makes it difficult to effectively exclude already sequenced ions. However, in LC-ESI MS, instrument control software is becoming increasingly sophisticated and, in some instruments, it is now possible to enter a list of pre-selected m/z values within a defined retention time window and exclude these ions from carrying out MS/MS scans, in addition to on-the-fly dynamic exclusion.[17] In this work, we report a systematic investigation of the precursor ion exclusion (PIE) strategy in LC-ESI QTOF MS and demonstrate that, with an optimal ion exclusion method, this strategy provides much improved protein identification efficiency compared to running replicates without PIE.

4.2 Experimental

4.2.1 Materials and Reagents.

Dithiolthreitol (DTT), iodoacetamide, trifluoroacetic acid (TFA), sodium bicarbonate and urea were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC-grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). BCA assay kit (Pierce, Rockford, IL).

4.2.2 Sample Preparation.

The MCF7 breast cancer cells (ATCC® number: HTB-22™) were cultured in 15 cm diameter plates at 37 °C in DMEM Gibco media supplemented with 10% fetal bovine serum. The plates were then washed twice with ice-cold 25 mL PBS^{++} buffer (0.68 mM $CaCl_2$, 0.5 mM $MgCl_2$, 1.4 mM KH_2PO_4, 4.3 mM Na_2HPO_4, 2.7 mM KCl, and 137 mM NaCl). The cells were harvested by scraping from the plates into the PBS^{++} buffer and centrifugation at 100 *g* for 8 min at 4 °C. The cell pellet were resuspended in 4 mL phosphate-buffered saline (PBS: 1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM KH_2PO_4, pH 7.4) buffer and passed twice through a mini-cell French Press (Aminco Rochester, NY). This was followed by sonication on ice (4×10 s pulses). The final volume of the lysate was brought up to about 5 mL using PBS buffer. The lysate was then centrifuged at 100 000 *g* for 1 h at 4 °C. BCA assay on an aliquot of the protein solution was performed to determine the protein concentration. Standard reduction of the disulfide bonds and alkylation were carried out on the protein extract. In order to remove the salts and chemicals introduced during the preceding steps, acetone, pre-cooled to -80 °C, was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The mixture was incubated at -20 °C overnight and centrifuged (14 000 rpm for 10 min at 4 °C). The supernatant was decanted and properly disposed. The residual acetone was evaporated at ambient temperature. Ammonium bicarbonate (50 mM, pH 8.0) and 6 M urea was used to redissolve the pellet. Trypsin solution was added into the protein solution for an enzyme/protein ratio of 1:45 after diluting the urea concentration, and digestion was conducted at 37 °C for 48 h. The digestion process was stopped by acidifying the peptide solution. All digestion solutions were stored at -80 °C until further analysis.

Preparation of the yeast sample was done in a similar manner to those reported.[19, 20] Briefly, the yeast strain BY4741 cells (ATCC® 4040002) were grown to mid log phase (O.D. 0.6) overnight in YEPD at 30 °C. The cell culture was centrifugated at 1000g to form a pellet which was then washed twice with 1× PBS. A lysis buffer (0.1 M Na_2CO_3, 310 mM NaF, 3.45 mM $NaVO_3$, 12 mM EDTA, 250 mM NaCl) was added to the pellet and the mixture was frozen in dry ice. The cells were lysed with an assistance of a mortar and pestle. The frozen cell lysate was thawed and placed on ice for 30 min, followed by centrifugation at 14 000 rpm for 10 min at 4 °C. After removing the supernatant, the sample was subjected to trypsin digestion using the procedures as described above.

Before analysis, the protein digest was desalted using an Agilent 1100 HPLC system (Palo Alto, CA) with a 4.6 mm × 50 mm long Polaris C18 column (3 μm, Varian, Palo Alto, CA). After sample loading onto the column, the column was flushed with mobile phase A (0.1% TFA/H_2O) for 5 min to remove the salts. The percentage of mobile phase B (0.1% TFA /ACN) was subsequently increased to 85% to ensure complete elution of the peptides off the column. The collected peptide fractions were then concentrated down to ~5 μL using a SpeedVac and reconstituted in 0.1% formic acid solution.

4.2.3 LC-ESI MS/MS.

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, about 1 μg of the digest solution was injected each time onto a 75 μm × 100 mm Atlantis C18 column (particle size 3 μm, Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120-min gradient (2-6% for 2 min, 6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min) and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 250 nL/min. The column was equilibrated at 2% Solvent B for 20 min before each sample run. A survey MS scan was acquired from m/z 350-1600 for 1 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 1.5 s each. Spectral acquisition switched to MS/MS scan when the intensity change of an individual ion was above 15 counts/s. Charge states of +2 and +3 were chosen for MS/MS, as in our experience of shotgun proteome profiling using this instrument very few peptide ions were detected in +1 or +4 charge state and the MS/MS spectra of the ions of these charge states were generally poor for database searching. For both dynamic and precursor ion mass exclusion, a time window of 150 s and a mass tolerance window of 100 mDa were applied. The collision energy used to perform MS/MS was varied according to the mass and charge state of the eluting peptide ion. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B used as mass calibrants (i.e., lock-mass), was infused at a flow rate of 250 nL/min, and an MS scan was acquired for 1 s every 1 min throughout the run.

The exclusion list of all precursor ions together with their corresponding retention information was obtained directly from the raw LC-ESI data in a previous run and was loaded into the MS acquisition method in the new run. The extended exclusion list (see Results and Discussion) was generated based on the peptides identified from the MASCOT search program (Matrix Science, London, U.K.) (see below). The m/z value, charge state and retention time of each identified peptide were extracted from the database search results and the corresponding raw data. The m/z value of the other charge state (only charge states of +2 and +3 were considered in this case) for each identified precursor ion was calculated. In addition, the +2 and +3 m/z values of all identified peptides consisted of not only the monoisotope value, but also the three additional isotope values. Finally all the m/z values along with their retention time information were loaded into the MS method for the new LC-ESI run. Currently, the process of generating the PIE list and importation of the list to the instrument control software was done manually. The manual operation mainly involved data processing using Excel and the major time consuming step was the cut and paste of the data which took about 30 min to complete for the generation of the selective or complete PIE list. This process should be automatable and an in-house data processing module is being developed which we expect to cut down the processing time to be less than 5 min. Even with manual operation, the total time of data processing including MS/MS data file conversion, MASCOT search, generation of the PIE list was less than 60 min. Since the column washing and equilibrium took about 60 min to complete for optimal column performance in our setup, data processing was done during this period.

4.2.4 Protein Database Search.

Raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files by using ProteinLynx Global Server 2.1.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the Mascot search program. Database searching was restricted to *homo sapiens* (human) in the

SWISSPROT database for the MCF-7 proteome digest and *Saccharomyces cerevisiae* (baker's yeast) for the yeast cell digest. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, (30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, *N*-Acytyl (Protein), oxidation (M), Pyro_glu (N-term Q), Pyro_glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides with scores lower than the Mascot threshold score for identity at the confidence level of 95% were then removed from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

It should be noted that the false positive rate of the protein identification results was not determined. However, in a previous study of shotgun proteome profiling work of *E. coil* membrane fractions using the same instrument and similar database search settings as described above,[21] it was found, through N-terminal amine labeling of peptides and determination of a_1 or a_1-related ions which identify the N-terminal amino acid of each peptide, that the false positive rate of protein identification using this instrument and MASCOT database search program at the confidence level of 95% was less than 1%. While the genome or proteome of MCF7 or yeast cells are clearly different from that of *E coli*, MASCOT search program adjusts the threshold score according to the genome size. It is reasonable to assume that the false positive rate of protein identification is most likely less than 1% for the proteins identified in this work based on peptide matches with scores above the threshold scores at the 95% confidence level.

4.3 Results and Discussion

Because of the complexity of a proteome digest, many peptides coelute in LC even after extensive fractionation of the sample. One common feature of modern tandem MS platforms is to sequence peptides eluted from LC using data dependence acquisition where a survey scan is carried out to generate a MS spectrum of the coeluting peptides, followed by one or more MS/MS scans of the peptide ions according to the ion intensities and the type of detected peptides in the survey scan. Dynamic exclusion is used to exclude the ions whose MS/MS spectra have been taken in previous MS/MS scans within a predefined retention time window (e.g., 150 s). In the precursor ion exclusion strategy, ions whose MS/MS spectra have been acquired or resulted in positive peptide identification from the initial run(s) are excluded for MS/MS sequencing in the subsequent run. As the MS/MS scans are often set to select and sequence the peptide ions according to the decreasing order of their intensities in an MS survey spectrum, one would expect that high abundance peptide ions are sequenced in the first run where the second run would sequence the relatively lower abundance peptide ions. Good reproducibility in chromatographic separation of the peptides routinely achievable with state-of-the-art LC system and column technology can facilitate the selection of peptide ions for exclusion. The MS spectral pattern depends on the composition of the coeluting peptides at a given retention time (strictly over a small time window where a summed spectrum is recorded). If excellent reproducibility in chromatographic separation is achieved, similar sets of peptides will coelute in replicate runs, resulting in similar MS spectral patterns from run to run. These coeluting peptides can potentially be sampled in the order of high to low abundances from replicate runs with precursor ion exclusion.

Figure 4.1 shows three base-peak ion chromatograms obtained from three 2-h replicate runs with PIE from the digest of the MCF cell lysate. Representative MS survey spectra of the coeluting peptides from the three replicate runs are shown in Figure 4.2. The retention times of individual runs are offset by as much as 45 s. However, the spectral patterns are very similar, indicating that at the retention times shown in Figure 4.2 the coeluting peptides had similar compositions. In each spectrum, there are close to 107 m/z values with ion counts above 50 in peak height – a threshold for generating a

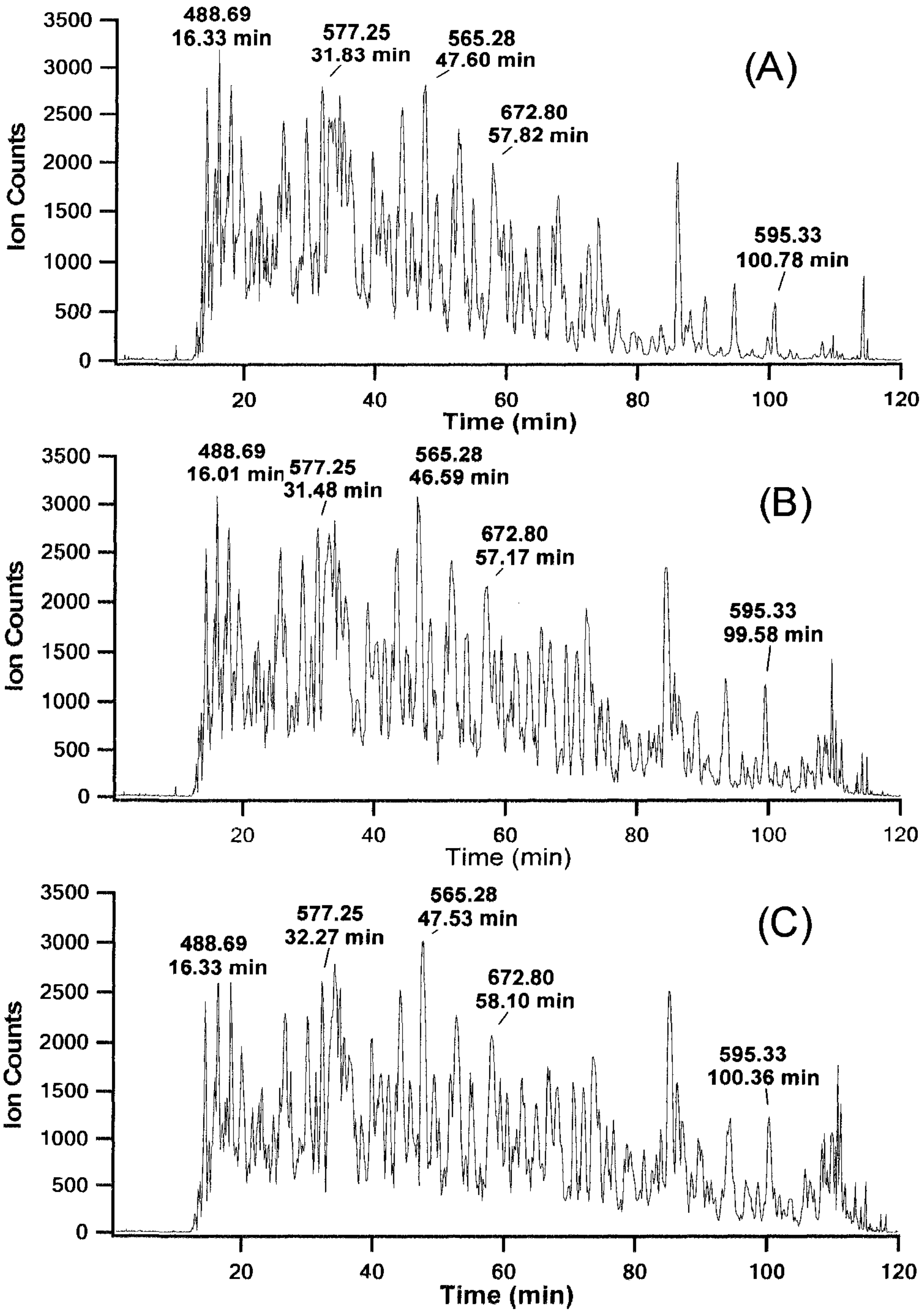


Figure 4.1 Base peak ion chromatograms obtained from three 2-h replicate runs of an MCF7 breast cancer cell lysate digest: (A) the 1st run, (B) the 2nd run with exclusion of precursor ions of the peptides identified from the 1st run, and (C) the 3rd run with exclusion of precursor ions of the peptides identified from the 1st and 2nd runs. The m/z values and retention times of several peaks are labeled.

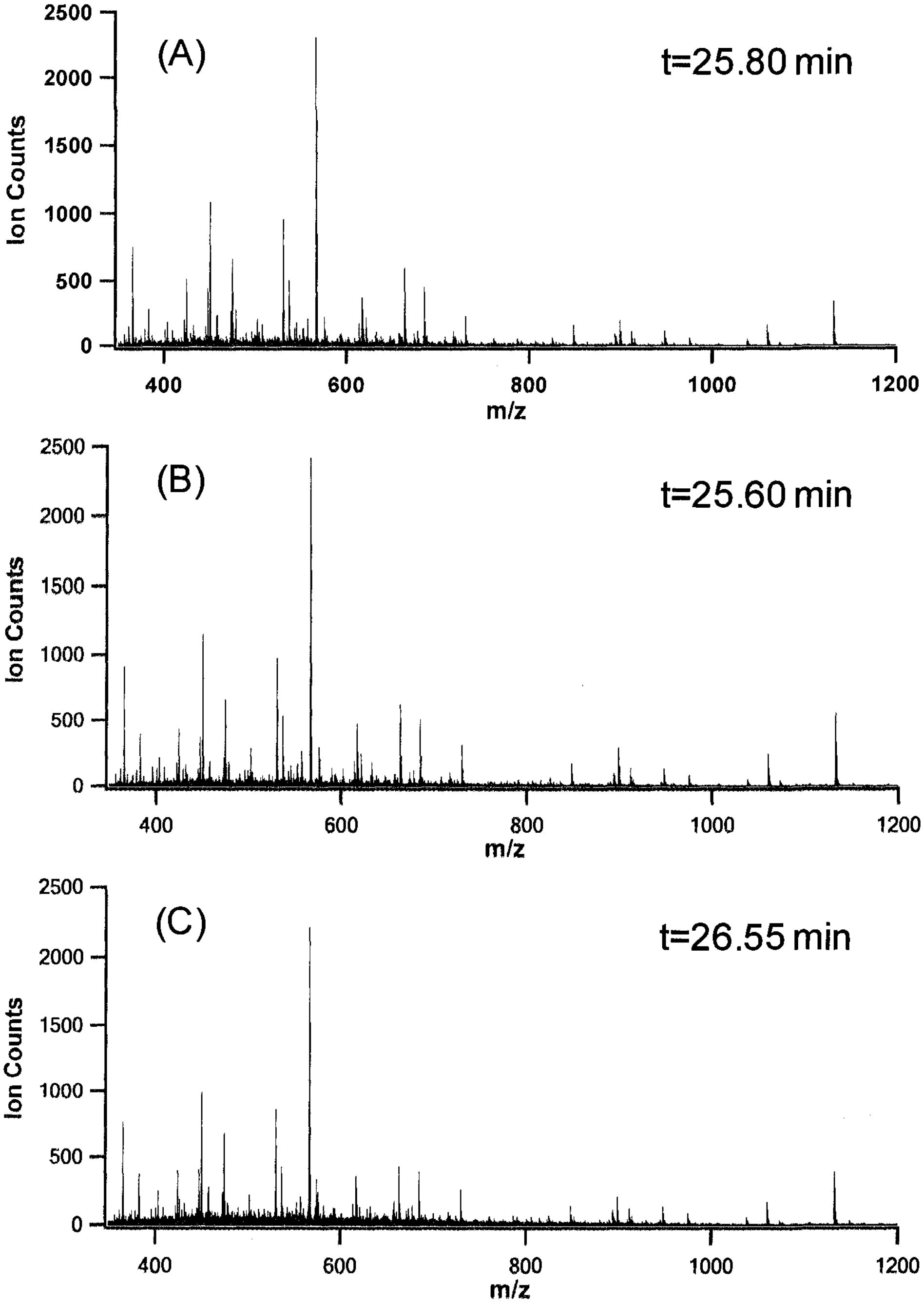


Figure 4.2 Representative MS survey spectra from the corresponding three replicate runs shown in Figure 4.1. The retention time is indicated in each spectrum.

database searchable MS/MS spectrum – below which the successful rate of positive identification is very low (< ~10%) based on our experience in using this instrument for shotgun proteome profiling. Not all 107 m/z values are from different peptides. One peptide may produce more than one dominant peak, because of the co-existence of the peptide ions with different charge states. Taking into account of these co-existence ions in the MS survey spectrum such as the one shown in Figure 4.2A, a total of 93 m/z values are found to belong to different peptide ions. This example shows the complexity of coeluting peptides and performing MS/MS scans on all or most of these ions is clearly a challenging task.

Some of the high abundance peptides elute over a period of a typical chromatographic peak (i.e., 25-35 s at the peak base) and they appear in 4 to 6 neighboring MS survey spectra and these ions can be effectively excluded via dynamic exclusion if they have been subjected to MS/MS sequencing in previous scans. However, the chromatographic peak profiles of the relatively low abundance peptides are severely truncated. They may be detected only in one survey spectrum. Whether these ions will be sequenced by MS/MS depends on their relative intensity ranks among all the ions detected in the survey spectrum. With a limited number of MS/MS experiments available for sequencing, the lower ranks of ions will not be sequenced in the MS/MS runs. In this kind of situation, dynamic exclusion in a replicate run will not increase the chance of sequencing these ions, as it will not change the ranks of the ions detected in the new survey spectrum for MS/MS. Note that, in practice, some variations of replicate runs may change the order of ion intensity ranks, allowing the possibility of sequencing some ions not sequenced in the first run. However, PIE may effectively exclude some of the ions already sequenced in a previous run and thus establish a new order of ion ranks in the new survey spectrum for sequencing, which should allow sampling additional relatively low abundance ions in the new run. An example is shown in Figure 4.3 where the expanded mass spectra over a small m/z region of the survey spectra shown in Figure 4.2 are displayed. As Figure 4.3A shows, some of the higher abundance ions (i.e., peaks labeled with "D") were sequenced in previous scans in this 1st run and, with dynamic exclusion, these ions were not selected for MS/MS sequencing following this MS survey scan. In this case, four ions at m/z 365.72, 424.71, 450.26, and 663.82 were selected for

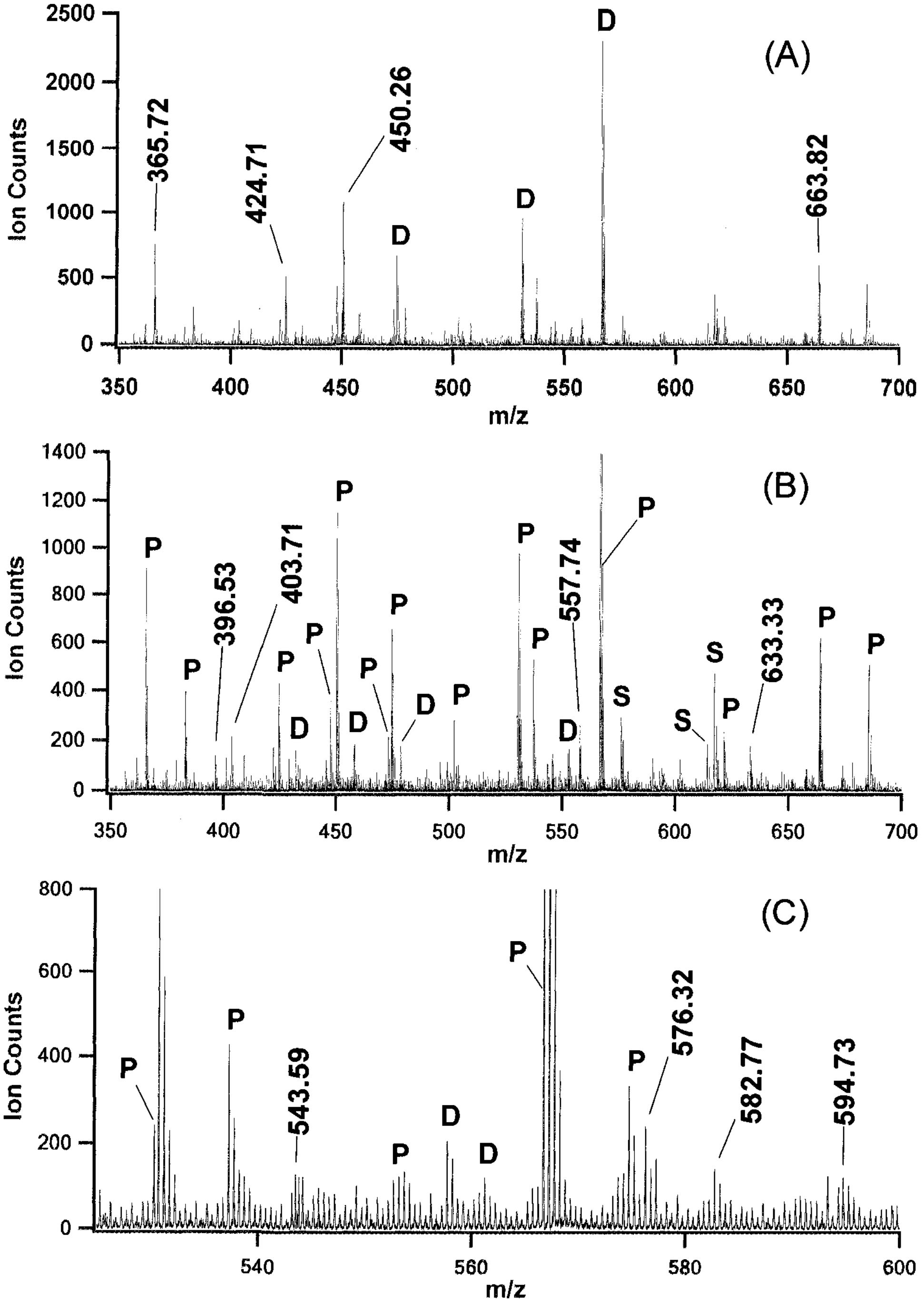


Figure 4.3 Expanded MS spectra from Figure 2. The peaks labeled with m/z values were selected for MS/MS. D = peak with the intensity above the threshold for MS/MS, but already sequenced in the previous scans – it was not selected for MS/MS based on dynamic exclusion. P = peak with m/z matched with one of the ions in the precursor ion list – it was not selected for MS/MS based on PIE. S = singly charged ions which were not programmed for MS/MS.

MS/MS sequencing. Database search from the four MS/MS spectra resulted in the identification of four different peptides. One of the MS/MS spectra (m/z 663.82) is shown in Figure 4.4A along with the peak assignment from the search results.

In the 2nd run, a list of precursor ions detected in the 1st run was excluded (i.e., peaks labeled with "P" in Figure 4.3B) and several additional ions (labeled with "D") were excluded through dynamic exclusion within this run. Four ions (m/z 396.53, 403.71, 557.74 and 633.33) were selected for MS/MS (see Figure 4.3B). One of the MS/MS spectra (m/z 633.33) is shown in Figure 4.4B. As expected, the ion counts of these peptide ions are lower than those selected in the 1st run. Database search of the MS/MS spectra identified one peptide. The identification success rate is decreased, compared to the 1st run, due to the reduction in MS/MS spectral quality from the low abundance ions. In the 3rd run where a list of precursor ions detected in the first two runs was excluded, four ions with even lower abundances were selected for MS/MS (see Figure 4.3C). Only one of the spectra (see Figure 4.4C) resulted in a positive match of a peptide from the database search. Note that three ions sequenced in the 2nd run, but not resulting in peptide identification, were not selected for sequencing in the 3rd run – the peaks at m/z 403.71 and 557.74 were dynamically excluded within this run and the peak at m/z 396.53 was not selected because its intensity fell below those of the other four ions selected for sequencing in the 3rd run.

The above example illustrates that reproducible MS spectral patterns can be generated and lower abundance ions can be sampled with the aid of precursor ion exclusion in replicate runs. However, there are several ways to generate the precursor ion list and implement the ion exclusion in replicate runs. We have examined the effects of four PIE methods on the peptide and protein identification (see Figure 4.5) and compared their data with those obtained by running simple replicates without PIE. The results are summarized in Table 4.1 for the MCF-7 proteome digest.

In the case of running simple replicates, the first 2-h run generated 2110 MS/MS spectra, resulting in the identification of 1153 peptides belonging to 332 proteins. The number ratio of peptide/spectrum is quite high (54.6%), suggesting that good quality MS/MS spectra were acquired from the relatively high abundance peptide ions. The

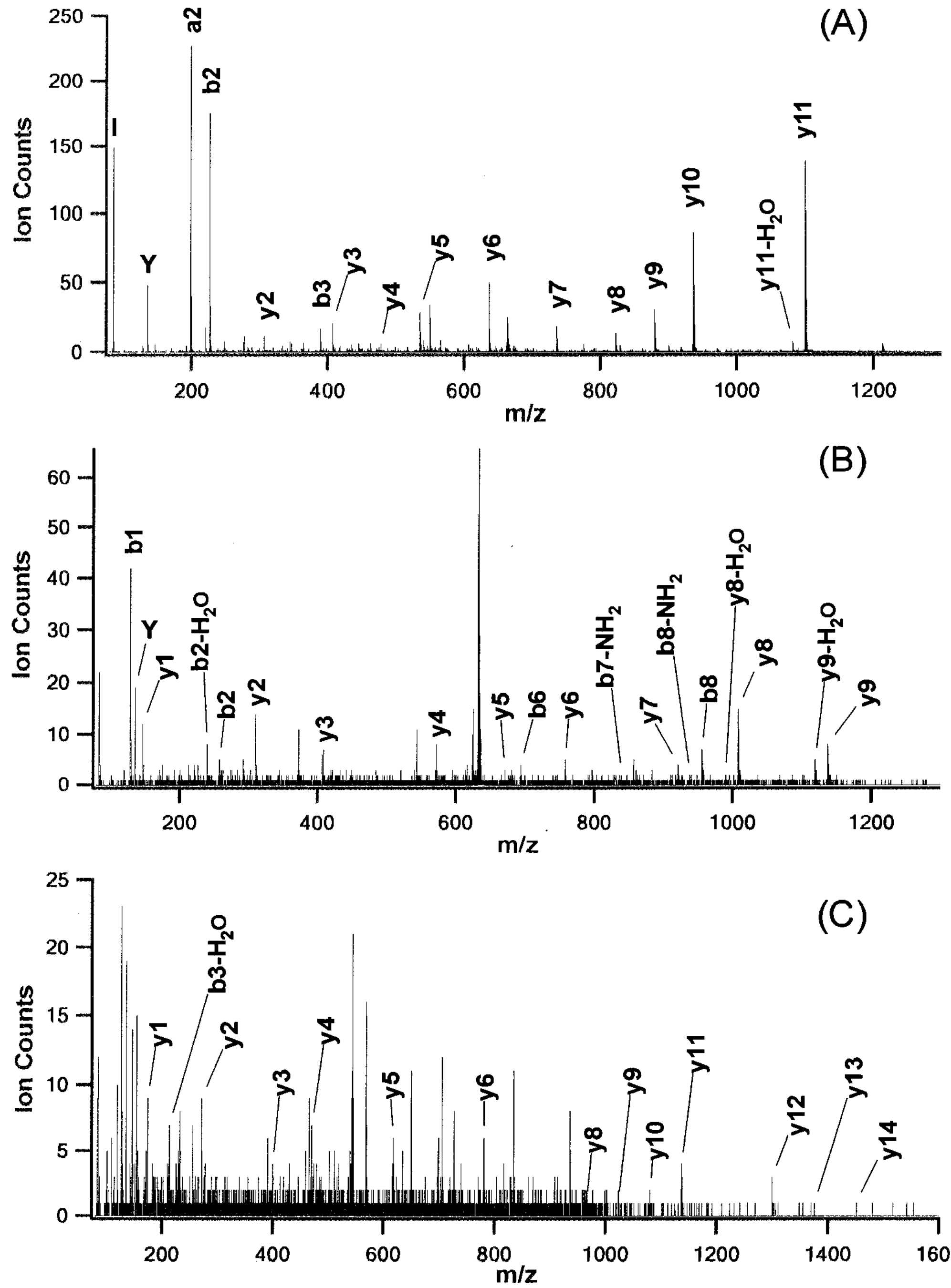
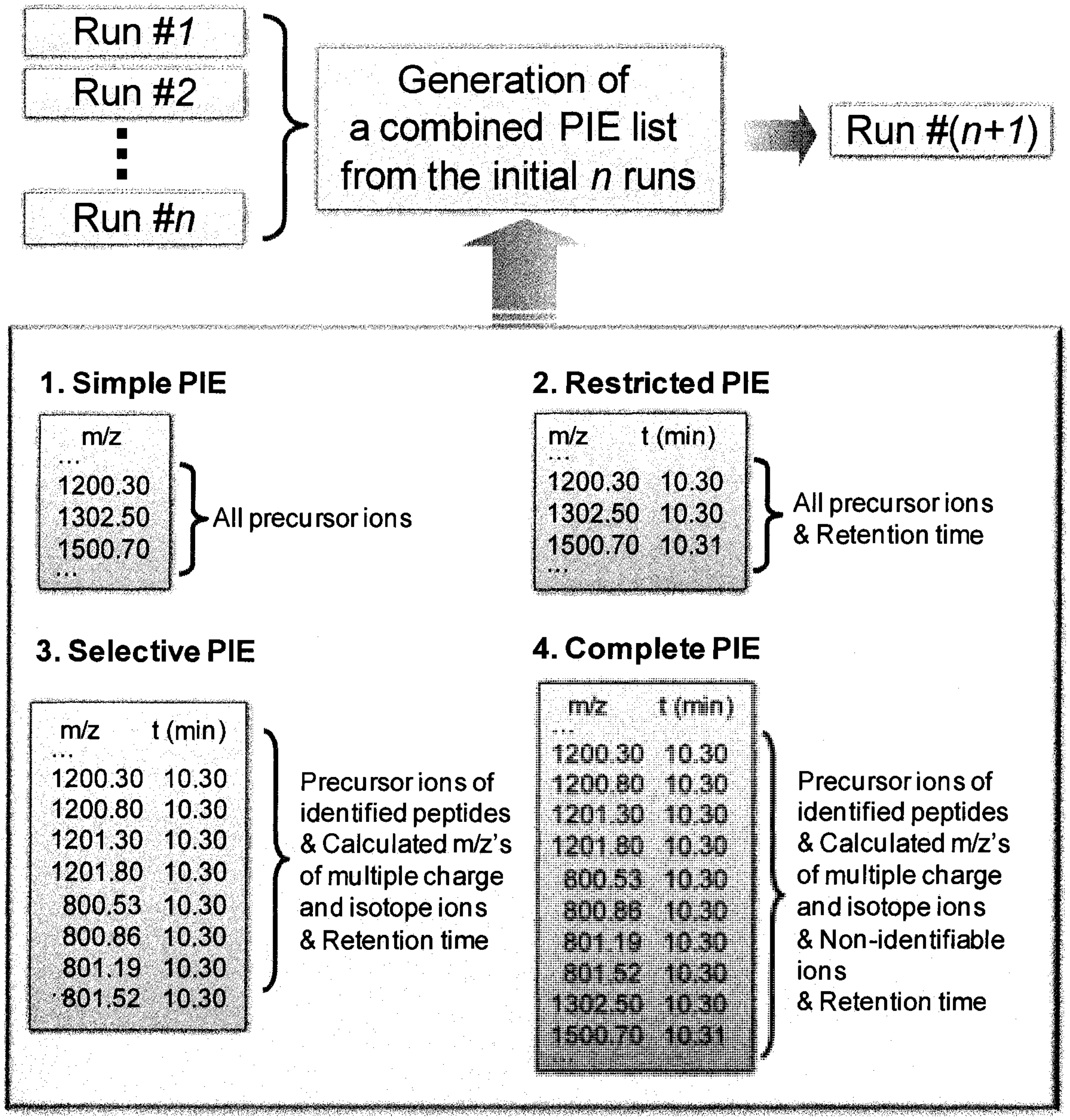


Figure 4.4 MS/MS spectra of the ions at (A) m/z 663.82, (B) m/z 633.33, and (C) m/z 543.59.


Run #1
Run #2
Run #n
Generation of a combined PIE list from the initial n runs
Run #(n+1)
1. Simple PIE
All precursor ions
2. Restricted PIE
All precursor ions & Retention time
3. Selective PIE
Precursor ions of identified peptides & Calculated m/z's of multiple charge and isotope ions & Retention time
4. Complete PIE
Precursor ions of identified peptides & Calculated m/z's of multiple charge and isotope ions & Non-identifiable ions & Retention time

m/z
...
1200.30
1302.50
1500.70
...

m/z	t (min)
...	
1200.30	10.30
1302.50	10.30
1500.70	10.31
...	

m/z	t (min)
...	
1200.30	10.30
1200.80	10.30
1201.30	10.30
1201.80	10.30
800.53	10.30
800.86	10.30
801.19	10.30
·801.52	10.30

m/z	t (min)
...	
1200.30	10.30
1200.80	10.30
1201.30	10.30
1201.80	10.30
800.53	10.30
800.86	10.30
801.19	10.30
801.52	10.30
1302.50	10.30
1500.70	10.31
...	

Figure 4.5 Schematic representations of four precursor ion exclusion (PIE) methods.

second 2-h run generated 2364 MS/MS spectra, leading to identification of 1247 peptides (peptide/spectrum=52.7%) and 336 proteins. Many of them are redundant peptides or proteins and only 346 unique peptides and 89 unique proteins were identified in the 2nd run. Similar results were obtained for the 3rd replicate run except that the unique peptides and proteins identified were 116 and 39 respectively. The trend of diminished return for these replicate runs is consistent with that reported by others.[8-15] Combining the data from the three runs, a total of 1615 unique peptides and 460 unique proteins were identified. The average number of peptides matched with a protein is 3.51. While it is not shown in Column 2 of Table 4.1, an additional run (4th run) only resulted in the identification of 27 more proteins for a total of 487 proteins. It is clear that additional runs will not result in a significant increase in the number of proteins identified.

To implement the precursor ion exclusion strategy, the simplest way is to compile a list of precursor ions from the first run and enter them into the mass exclusion window in the QTOF instrumental control software for the second run (see Figure 4.5). Column 3 in Table 4.1 shows the data produced using this simple PIE method. A list of precursor ions was generated from the first run as that in Column 2. A total of 2148 m/z values were entered into the mass exclusion window for the 2nd run. This number is slightly higher (1.8%) than the number of MS/MS spectra generated in the first run (i.e., 2110). This difference was introduced in the data processing using the MassLynx software where a small number (38 out of 2148) of the MS/MS spectra were discarded due to the selection of extra isotope peaks of some precursor ions. In our experiment, to gauge the reproducibility of the results, two replicate 2nd runs were carried out under the same experimental and PIE conditions and the data obtained are shown in Column 3 of Table 4.1 as 2nd (i) and 2nd (ii). From the results of the replicate runs, it appears that good run-to-run reproducibility was achieved. In the 2nd (i) run, 2253 MS/MS spectra were recorded, resulting in the identification of 803 peptides. Among them, 673 peptides were unique to this run and were not identified in the 1st run. The number of proteins identified was 395 and, among them, 199 proteins were unique to this run. Combined with the 332 proteins identified in the 1st run, the total number of proteins identified was 531. Similarly, for the 2nd (ii) run, 515 proteins were identified from the combined two runs with PIE. Note that the peptide/protein ratio is 803/395 or 2.03 in 2nd (i) and 2.08 in

2nd (ii). In the 1st run, the ratio is 1153/332 or 3.47 peptides/protein. It is clear that, in the 2nd run, on average, fewer peptides matched with a protein, which is consistent with the notion depicted in Figure 4.3 that the 2nd run with PIE sampled the relatively lower abundance peptides.

In a separate experiment, we examined the effect of increasing the MS/MS scan time on the number of peptides identified. We initially thought that, after excluding the high abundance ions, the remaining relatively low abundance ions gave reduced quality of MS/MS spectra and thus the quality of MS/MS spectra might be increased by increasing the MS/MS scan time, albeit at the reduction of the number of MS/MS spectra acquired. It was found that, when the MS/MS scan time was doubled (i.e., from 1.5 to 3 s), the number of unique peptides identified was actually reduced. Thus, we abandoned the idea of increasing MS/MS scan time in the 2nd run.

Another way to implement the PIE strategy is to first compile a list of precursor ions with their corresponding chromatographic retention time information from the first run (see Figure 4.5). The m/z values of these precursor ions along with their retention times are entered into the mass exclusion program which is included in the QTOF control software. A predefined retention time window is used to exclude only the precursor ions detected within this window in the 2nd run. The advantage of this restricted precursor ion exclusion method, compared to simply excluding all the precursor ions at any given retention time as described above, is that different peptides with the same or similar m/z values within a mass tolerance window, but eluted at significantly different retention times, will not be falsely excluded. This becomes increasingly important as the number of peptide ions to be excluded increases. After considering the extent of possible chromatographic retention time shifts from run to run, we chose a time window of ±150 s for exclusion which is conservative to ensure any chromatographic shifts in replicate runs will not cause problems. The results obtained using this restricted PIE method are shown in Column 4 of Table 4.1. Again, two repeated 2nd runs were carried out based on the precursor ion information generated in the 1st run in Column 2.

As Column 4 of Table 4.1 shows, in the 2nd (i) run, 1085 peptides were identified and 863 of them were unique to this run. They were assigned to 405 different proteins and 200 of them were unique to this run. Combined with the proteins identified in the 1st run, a total of 532 proteins were identified from the two runs combined. In the 2nd (ii) run, 190 unique proteins were identified, bringing to a total of 522 proteins in the two runs. The average number of proteins identified (527) by using restricted PIE is similar to that obtained by using simple PIE (523). However, the peptide/protein ratio is 1085/405 or 2.68 in 2nd (i) with restricted PIE and 1068/400 or 2.67 in 2nd (ii). These ratios are greater than the average ratio of 2.06 peptides/protein obtained in the 2nd run with simple PIE. Thus, the overall sequence coverage or the protein identification confidence level is improved in the 2nd run with restricted PIE.

A third PIE method investigated is to exclude only the ions positively identified from the 1st run (see Figure 4.5). After the 1st run, MS/MS spectra are subjected to database search for peptide identification. There are several approaches of building an ion exclusion list for the 2nd run based on the search results. The first approach is to simply enlist the m/z values of the detected ions whose MS/MS spectra resulted in positive peptide identification along with their retention time information. Another approach is to calculate the m/z values from the masses of the identified peptides according to their charge states (i.e., +2 and +3). This list is more comprehensive than that of the first approach and takes into account of the coexistence of multiple charge ions of some identified peptides in survey MS spectra. For example, a peptide may be identified based on the MS/MS spectrum of the +2 ion. If the +3 ion of the peptide is also present, but not sequenced in the 1st run, this ion will not be excluded in the first approach, but will be excluded if the expanded list is used. The third approach is to further expand the ion exclusion list by adding the m/z values of the peptide isotope peaks (e.g., the monoisotope ion plus $^{13}C_1$, $^{13}C_2$ and $^{13}C_3$ isotope ions). In theory, adding the isotope peaks to the list is unnecessary, because the ion exclusion program in the QTOF instrument excludes the monoisotope ion along with their other isotope ions from MS/MS sequencing if the monoisotope m/z value matches with the one in the exclusion list. However, we found that the isotope ions of the same peptide were sometimes sequenced in the 2nd and 3rd runs. This error might be due to difficulty of de-isotoping the peptide ions, particularly when the ion intensities were low where the isotope envelope of the peptide ions did not match well with the theoretical profile. By including the isotope m/z values in the exclusion list, exclusion of these isotope ions of identified peptides are ensured. The downside of this approach is that we may falsely exclude other peptides having the same m/z values (within an m/z tolerance window) as those of the isotope ions of the peptide intended to the excluded. But, by narrowing the retention time window for ion exclusion, this coincidence of event can be minimized. It should also be noted that if peak overlap does occur, the resulting MS/MS spectrum would contain the fragment ions of different types of peptides and, hence, not likely lead to positive peptide identification anyway. We have compared these three approaches for selective PIE and found the third approach to be the most effective. Thus, the selective PIE method discussed below refers to the use of an expanded list of m/z values taking into account of multiple charge states and isotope peaks.

The results from this selective PIE method with two repeat 2nd runs are shown in Column 5 in Table 4.1. In the 2nd (i) run, 11664 m/z values along with their retention time information were entered for exclusion. Despite these many exclusions, 2365 MS/MS spectra were still collected, which is similar to the number collected in the restricted PIE method (see Column 4 of Table 4.1). The number of peptides identified in this run was 959, and 864 of them were unique to this run. Interestingly, there are still 95 common peptides (10%) identified in the 1st and 2nd runs. With ±100 mDa mass tolerance window for exclusion, some ions with a small mass shift to outside this window were not excluded in the 2nd run. In a separate study (data not shown), it was found that increasing the mass tolerance window, e.g., using ±200 mDa, instead of 100 mDa, ran into the problem of overly excluding the ions, resulting in a reduction in the number of peptides identified.

From the 2nd (i) run with selective PIE, 432 proteins were identified and among them 225 proteins were unique to this run. Thus, a total of 557 proteins were identified from the combined two runs with selective PIE. Similar results were obtained in the 2nd (ii) run. The average total number of proteins identified was 553, compared to 527 using restricted PIE and 523 using simple PIE. The number ratio of peptides and proteins identified in 2nd (i) with selective PIE is 2.22 peptides/protein and 2.21 in 2nd (ii). Compared to the ratio obtained with restricted PIE, it appears that selective PIE identified more unique proteins, but the peptide/protein ratio is reduced from 2.68 to 2.22.

However, the peptide/protein ratio is still better than the average coverage (2.06) generated in the 2nd run with simple PIE.

The final PIE method examined is to combine the positive features of the restricted and selective methods to produce a more complete list of precursor ions for exclusion (see Figure 4.5). In the selective PIE method, all the sequenced ions lead to positive peptide identification are included in the exclusion list; but many sequenced ions not generating positive peptide identities are not included. Since the intensities of these non-identifiable ions do not change in replicate runs, their MS/MS spectra generated in the 2nd run will be similar to those from the 1st run and database search of these spectra will not result in peptide identification as in the 1st run. Thus, in the complete PIE method, the m/z values of these non-identifiable ions found in the initial run(s) will be included for exclusion, in addition to those listed in the selective PIE method. The m/z values of the non-identifiable ions can be readily determined by subtracting the m/z values of the peptide ions lead to positive peptide identification from the restricted PIE list. This truncated list is then added to the selective PIE list to produce a complete PIE list.

The 6th column in Table 4.1 shows the results generated from two replicate runs using the complete PIE method. The number of MS/MS spectra collected is similar to that of the selective method, despite an increase in the number of m/z values excluded in the 2nd run. In the 2nd (i) run, a total of 975 peptides were identified, and 943 of them were unique to this run. Only 32 common peptides were identified in the 1st and 2nd (i) runs. Similarly, 31 common peptides were identified in the 1st and 2nd (ii) runs. These results indicate that complete PIE is more effective in excluding already sequenced peptide ions, compared to the other three PIE methods. However, in this case, the increase in the number of unique peptides identified does not translate into an increase in the number of unique proteins identified. The unique proteins identified from the 2nd (i) and (ii) runs are 206 and 207, respectively, bringing in a total of 548 and 549 proteins from two runs. These numbers are slightly lower than 557 and 549 proteins generated from the two runs using selective PIE. However, the number ratio of peptides and proteins identified in 2nd (i) with complete PIE is 2.34 peptides/protein and 2.30 in 2nd (ii), which are higher than those of the corresponding runs from the selective PIE method (i.e., 2.22 and 2.21).

Data generated from the third run of the same digest are shown in Columns 7-10 in Table 4.1. For each PIE method, the 3rd run was done in replicates and using the ion exclusion list generated from the 1st run and the corresponding 2nd (i) run. With simple PIE (see Column 7), the number of MS/MS spectra collected was reduced to about half of the number in the 1st or 2nd run. In contrast, the restricted PIE method still generated the number of MS/MS spectra close to that from the 1st or 2nd run. With selective PIE, the number of MS/MS spectra collected was similar for the 1st, 2nd or 3rd run. For complete PIE, the number was slightly lower than the 2nd run, but higher than the 1st run. These results indicate the importance of using retention time information to reduce false exclusions, particularly when the number of m/z values to be excluded is very large. With simple PIE, 198 proteins were identified from the 3rd (i) run with the peptide/protein ratio of 1.50. However, only 54 proteins were unique to this run. The total number of different proteins identified from the three runs with simple PIE is 585. With 3rd (ii), the total number is 559.

Column 8 in Table 4.1 shows that, with 2063 MS/MS spectra collected in the 3rd (i) run with restricted PIE, 676 peptides were identified. The number ratio of peptides and spectra is considerably smaller than those in the 1st and 2nd runs, indicating that the quality of the MS/MS spectra deteriorates as a result of sampling lower abundance peptide ions. From the 676 peptides, 384 proteins were identified with the peptide/protein ratio of 1.76 and 130 of them were unique to this run. Thus, the total number of proteins identified from the three runs with restricted PIE is 652. With 3rd (ii), the total number is 661. It is clear that both the number of proteins identified and the average peptide/protein ratio using this method of ion exclusion are significantly greater than those obtained by using simple PIE.

From the 3rd run with selective PIE, a total of 18117 m/z values were enlisted for exclusion. As it is shown in Column 9 of Table 4.1, 657 peptides were identified and matched with 385 proteins in the 3rd (i) run. The peptide/protein ratio is 1.71 which is similar to that obtained with restricted PIE. With 3rd (ii), 633 peptides and 379 proteins were identified with an average peptide/protein ratio of 1.67.

The last column of Table 4.1 shows the 3rd run results obtained by using the complete PIE method. A total of 23858 m/z values along with their retention time information were used for exclusion. The number of MS/MS spectra collected was in between the numbers obtained from the restricted and selective PIE methods. Similar to the results found in the 2nd run, the number of unique peptides identified in complete PIE is greater than that obtained with selective PIE. However, in the 3rd run, the number of unique proteins identified is significantly greater than that of selective PIE (i.e., 175 and 180 proteins from the 3rd (i) and (ii) runs with complete PIE respectively vs. 133 and 115 proteins from the 3rd (i) and (ii) runs with selective PIE). The peptide/protein ratio is 1.55 for 3rd (i) and 1.47 for 3rd (ii).

To summarize the combined results obtained from the three replicate runs with different methods of precursor ion exclusion, the average total number of proteins identified from the three runs with complete PIE is 726, representing the highest number of unique proteins identified among the four PIE methods examined. The complete PIE method identified 6.6%, 10.5%, and 26.9% more proteins than the selective, restricted and simple PIE methods, respectively. In all cases, the number of proteins identified is much greater than the 460 proteins found in the 3 replicate runs without PIE. For example, using complete PIE, the protein number increases from 460 to 726, representing a 58% increase. At the peptide level, an average of 2055 unique peptides was identified from the combined three runs with simple PIE with the average peptide/protein ratio of 3.59, compared to 2461 unique peptides identified in the restricted PIE runs (3.75 peptides/protein), 2506 unique peptides identified by using selective PIE (3.68 peptides/protein) and 2659 unique peptides identified by the complete PIE method (3.66 peptides/protein). In the case of three replicate runs without PIE, a total of 1615 unique peptides was identified (3.51 peptides/protein). Thus, the sequence coverage is slightly improved when PIE is used. It can be concluded that the major benefit from the use of PIE is the significant gain in the number of proteins identified. Finally, it is worth commenting on the number of proteins commonly detected in replicate runs. In the replicate runs without PIE, the protein overlap between the first two runs is about 74%. In the case of the 2nd run with simple PIE, the protein overlap between two replicate runs is about 68% (i.e., 199 vs. 183 with 130 common proteins). In the 3rd run, the overlap is

48% (54 vs. 44 with 23 common proteins). In this case, the total number of proteins identified from the two replicate 2nd runs is 252, which is similar to that from the combined 2nd (i) and the 3rd (i) runs (i.e., 253). Thus, the 3rd run with simple PIE did not increase the overall number of proteins identified, compared to the 3rd run without PIE (i.e., the replicate 2nd run). However, in the 2nd run with restricted PIE, the protein overlap is about 72% (200 vs. 190 with 140 common proteins). In the 3rd run, the protein overlap is 55% (130 vs. 129 with 71 common proteins). In this case, the total number of proteins identified from the two replicate 2nd runs is 250. In contrast, the number of proteins identified from the combined 2nd (i) and the 3rd (i) runs is 330. Likewise, in the 2nd run with selective PIE, the protein overlap is about 70% (225 vs. 217 with 153 common proteins). In the 3rd run, the protein overlap is 59% (133 vs. 115 with 73 common proteins). The total number of proteins identified from the two replicate 2nd runs is 289, compared to 358 proteins identified from the combined 2nd (i) and 3rd (i) runs. In the case of complete PIE, the protein overlap is 62% (206 vs. 207 with 128 common proteins) in the 2nd replicate runs. The protein overlap is 63% (175 vs. 180 with 111 common proteins in the 3rd replicate runs. The total number of proteins identified from the two replicate 2nd runs is 285, compared to 381 proteins identified from the combined 2nd (i) and 3rd (i) runs. It is clear that restricted, selective or complete PIE is still more effective in increasing the number of unique proteins identified in the 3rd run, compared to running replicates with simple PIE or without PIE.

To gauge the general applicability of the PIE strategy for improving protein identification and examine the effects of proteome sample complexity on the extent of improvement, we applied the four PIE methods to identify proteins from a tryptic digest of a whole cell extract of yeast cells. The peptide composition of the yeast cell lysate digest is expected to be less complicated than that of the MCF-7 cell lysate digest, as the yeast genome size is considerably smaller than the human genome. Tables 4.2 and 4.3 show the results of 5 replicate LC-ESI MS/MS runs of the yeast digest without PIE and with PIE, respectively. For each PIE run, the sample was run in triplicates. Figure 4.6 shows the plots of the total number of proteins identified from each method vs. the number of runs.

As Table 4.2 shows, in the runs without PIE, an average of 248 proteins were identified in one run which is less than that from the MCF-7 digest (i.e., an average of 331 proteins per run as shown in Column 2 of Table 4.1). The total number of proteins identified from the first three runs combined is 315, compared to 460 proteins identified from three replicate runs in MCF-7. As the data in Column 2 of Table 4.1 and those in Table 4.2 show, the number of MS/MS spectra acquired in a run is very similar for the two samples. However, the number of peptides identified from these spectra is considerably less for yeast (e.g., 692 peptides from the 1st run of yeast vs. 1153 peptides from the 1st run of MCF-7). In addition, the peptide/protein ratio for yeast is significantly smaller (e.g., 2.78 for the 1st run of the yeast sample vs. 3.47 for MCF-7). Thus, the overall peptide and protein identification efficiency is lower for yeast. This can be attributed to the difference in proteome complexity; the yeast proteome is less complex than MCF-7. Within a cell, different proteins are present in a wide range of concentrations. Comparing the protein concentration distributions of the yeast and MCF-7 cell extracts, one would expect that there be a smaller number of proteins present in each concentration range in the yeast cell extract. Likewise, the number of peptides at any given concentration range in the yeast digest should be smaller than that of the MCF-7 digest. Note that the same amount of the cell extract digest was loaded onto the nano-LC ESI MS/MS system for sequencing. Since the LC-MS/MS system can only ionize, sequence, and identify the peptides with concentrations above certain threshold, a larger fraction of peptides in the yeast sample injected into the instrument, compared to the MCF-7 digest, would be expected to have their concentrations below the identification threshold and were consequently not identified.

As shown in Table 4.3, using the simple PIE method, an average of 57 unique proteins was identified in the 2nd run, which is more than the 44 unique proteins identified in the 2nd run without PIE. However, in the 3rd run with simple PIE, only 18 unique proteins were identified, compared to 22 in the 3rd run without PIE. In the 4th and 5th runs with simple PIE, the number of unique proteins identified is even smaller than the corresponding runs without PIE. This is due to the increase in the number of false exclusions as more ions are added to the exclusion list in the subsequent replicate runs (e.g., only an average of 182 MS/MS spectra were acquired in the 4th run with the exclusion of 3641 m/z values). These results indicate that, except for the 2nd run, simple PIE is not an effective method for increasing the number of proteins identified in the replicate runs for this less complicated proteome digest sample. This finding may explain why an earlier attempt of using simple PIE for the analysis of the digests of protein mixtures separated by gel electrophoresis of cell extracts did not result in the identification of more proteins than those obtained by using simple replicate runs without PIE.[16] This finding is also consistent with the 3rd run results from the MCF-7 digest. As indicated earlier, in the 3rd run of the MCF-7 digest, simple PIE increased the overall number of proteins identified to an extent similar to that from the 3rd run without PIE (i.e., the replicate 2nd run). The peptide ions sampled in the 3rd run appear to bear similarity to those of the yeast digest in terms of peptide concentration range and composition.

However, as Table 4.3 and Figure 4. 6 show, the other three PIE methods performed much better than simple PIE and identified many more proteins than the simple replicate runs without PIE. In the 2nd run of the restricted PIE method, an average of 123 unique proteins were identified, bringing in a total of 366 proteins identified from two runs, which is already more than the total number of proteins identified from 5 replicate runs without PIE (i.e., 353). The total number of proteins identified from two runs is 344 in selective PIE and 385 in complete PIE. Among these three PIE methods, a smaller number of proteins were identified in selective PIE. This finding is consistent in all replicate runs except the last run where the number of unique proteins identified is similar for all three methods. A reduced number of proteins identified is mainly due to the decrease in the number of peptides identified – a larger fraction of the MS/MS spectra did not result in peptide identification as the non-identifiable ions were not excluded in selective PIE. Comparing the results of the MCF-7 and yeast digests, it appears that, if the overall peptide identification efficiency from the MS/MS spectra is lower, as in yeast, exclusion of non-identifiable ions become more important in order to allow the subsequent run to sample different ions to increase the chance of identifying more peptides.

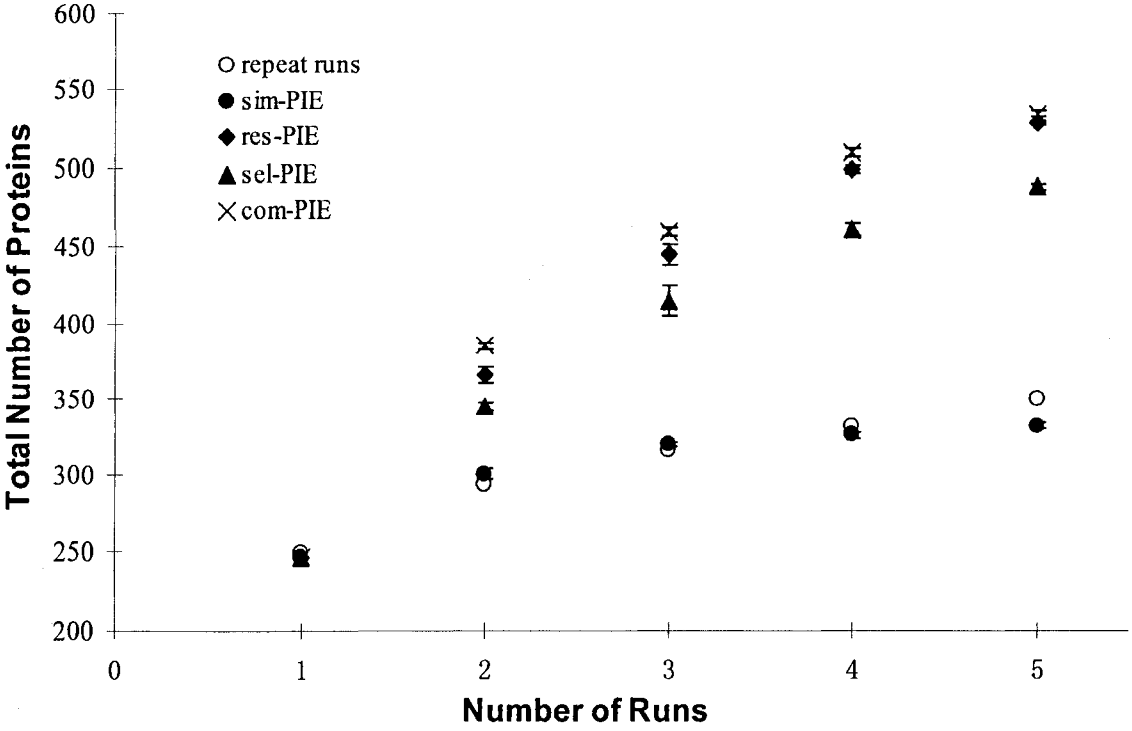


Figure 4.6 Comparison of the numbers of proteins identified from replicate runs with and without PIE for yeast cell lysate digests (see Table 4.3 for more details).

In the case of the 2nd run with complete PIE, the number of peptides identified is less than that with restricted PIE (423 vs. 500), but the number of unique peptides identified is slightly more than that in restricted PIE (381 vs. 374). This result indicates the effectiveness of complete PIE to exclude the already identified peptides including the ions of different charge states and isotope peaks. However, in the 3rd run, the numbers of peptides and unique peptides identified are similar for complete and restricted PIE. This is likely due to an increased number of false exclusions in complete PIE when low abundance peptide ions were sampled for MS/MS. For the 4th and 5th runs, the number of peptides identified in complete PIE becomes less than that with restricted PIE. In both cases, the 5th run did not result in a substantial increase in the number of unique proteins identified. In fact, as Figure 4.6 shows, there is a large increase in the total number of proteins identified from the 2nd run, but the increase is much less pronounced going from 2nd to the 3rd and subsequent runs. Thus, the unique-protein identification efficiency (number-of-unique-proteins/time) decreases significantly in the 3rd and subsequent runs, but is still much higher than the simple replicate runs without PIE.

For yeast, the codon adaptation index (CAI) may be used to gauge the relative protein abundance expressed in a cell.[19, 22-24] This provides an opportunity to examine the PIE method to see if it can indeed probe low abundance proteins as the replicate runs progress. CAI calculation was done using a web-based resource[24] and the CAI values generated from the program (CAI Calculator 2) were generally in agreement with those published by others.[19] Proteins from genes with CAI of less than 0.2 are considered to be expressed in relatively low abundances. In the 1st run, the average percentage of proteins identified with CAI < 0.2 in all the identified proteins in a run is 19±2% (5 repeat runs or n=5) (see Supporting Information 4.1). In the 2th run with complete PIE, the average fraction of relative low abundance proteins with CAI < 0.2 in the uniquely identified proteins in the run is 38±3% (n=3) which is significantly higher than that of the 1st run. In the subsequent runs, 3rd, 4th and 5th, the fraction is 47±4%, 59±2%, and 47±5%, respectively. The fraction leveled off at the 4th run; but, overall, a greater fraction of low abundance proteins were sampled when the replicate runs with complete PIE were carried out. Another way of gauging the relative abundance differences probed by the sequential replicate runs is to examine the abundances of precursor ions selected for MS/MS. The median of precursor ion counts in peak areas in the 1st run with 2206 ions selected for MS/MS is 3774. The average median of precursor ion counts from the complete PIE method is 1453±138 counts in the 2nd run (2028 ions), 1073±58 counts in the 3rd run (1748 ions), 846±71 counts in the 4th run (1290 ions) and 868±46 counts in the 5th run (1037 ions). Similar to the CAI data, the counts leveled off at the 4th run. The overall trend is that lower abundance ions were sampled as the number of replicate runs increased. In contrast, the average median of precursor ion counts is 3656±301 for the 5 replicate runs without PIE. The above CAI data as well as the precursor ion abundance results indicate that the complete PIE method can increase the chance of sequencing low abundance peptides, resulting in the identification of low abundance proteins.

To summarize the yeast results, a total of 533 proteins were identified from five replicate runs using the complete PIE method, representing the largest number of proteins identified among the four PIE methods. Compared to 353 proteins identified from the 5 replicate runs without PIE, 180 additional proteins (51%) were identified. The first three runs without PIE identified a total of 315 proteins, while 459 proteins were identified from three runs with complete PIE, representing an increase of 46%. For the MCF-7 cell digest, as described earlier, an increase of 58% (726 with complete PIE vs. 460 without PIE in three runs) was found. It appears that a greater improvement in protein identification is achieved for running replicates with complete PIE for a more complicated digest than a less complicated digest. Nevertheless, in both cases, the complete PIE method offers a significant improvement in protein identification efficiency over the simple replicate runs without PIE.

It should be noted that, while the data shown in this work were generated in the Waters QTOF instrument, this PIE strategy should be applicable to any QTOF instruments as long as control software allowing for inclusion of a mass exclusion list for MS/MS is available. In principle, this strategy should also be applicable to other types of mass analyzers. However, the extent of improvement may differ, depending on the performance of a LC/MS system, particularly on chromatographic reproducibility and mass resolution/accuracy. For example, in a low resolution ion trap mass spectrometer, one would expect an increased chance of false ion exclusion with PIE, compared to a QTOF instrument. However, if the retention time window for ion exclusion is set to be very narrow, which requires excellent chromatographic retention time reproducibility, the extent of false ion exclusion would be reduced due to a smaller number of ions to be excluded within a very small retention time window.

4.4 Conclusions

We have demonstrated that precursor ion exclusion in replicate runs of LC-ESI MS/MS is an effective shotgun proteome analysis strategy. It significantly increases the number of proteins identified from a cell lysate digest, compared to replicate runs without PIE. Four PIE methods were investigated and their effects on peptide and protein detectability in a quadrupole time-of-flight mass spectrometer were studied. Ion exclusion can be implemented by compiling a list of m/z values of precursor ions that have been subjected to MS/MS in the previous run and entering them in the mass exclusion program in the new run. This simple PIE method was found to be not as effective as the other three methods, namely, restricted, selective and complete PIE methods. In the restricted PIE method, the m/z values of the precursor ions along with their retention times are entered into the mass exclusion program. This method reduces the number of m/z values to be excluded at a given retention time and thus reduces the possibility of false ion exclusions associated with the simple PIE method. The selective PIE method involves the use of peptide mass information from the initial database search results to determine the m/z values of all doubly and triply charged ions of peptides plus additional isotope peaks. The complete PIE method combines the positive features of the restricted and selective PIE methods by including the non-identifiable peptide ions in the exclusion list of selective PIE. It is demonstrated that this expanded ion exclusion allows the identification of a greater number of proteins compared to other three methods. The sequence coverage in terms of the number of identified peptides per protein was slightly increased when an ion exclusion method was used, compared to running replicates without PIE. The performance of the PIE methods was examined and compared in the analysis of the MCF-7 and yeast whole cell lysate digests. It is found that, for the more complex proteome digest of the MCF-7 cells, all four PIE methods provide a significant improvement in protein identification efficiency over running replicates without PIE. For the less complex proteome digest of the yeast cells, the restricted, selective and complete PIE methods provide a significant improvement while the simple PIE method does not.

4.5 Supporting Information

Supporting information includes MASCOT search results including protein names, sequences of matched peptides, MASCOT scores, mass errors, and MASCOT threshold scores obtained from individual LC-ESI QTOF MS/MS runs. Codon adaptation index (CAI) data for the proteins identified from the yeast cell lysate digest. This material is available in the attached disk.

Table 4.1 Summary of the results obtained from replicate 2-h runs of the MCF-7 cell digest with and without precursor ion exclusion[1].

Number	Run 1/2/3 [2]	Run 2 (i)|(ii) (sim-PIE)[3]	Run 2 (i)|(ii) (res-PIE) [4]	Run 2 (i)|(ii) (sel-PIE) [5]	Run 2 (i)|(ii) (com-PIE) [6]
MS/MS spectra	2110/2364/2382	2253|2234	2389|2383	2365|2358	2360|2333
m/z values excluded	-	2148	2148	11664	12423
Peptides identified	1153/1247/1247	803|784	1085|1068	959|920	975|953
Unique peptides	1153/346/116	673|660	863|837	864|814	943|922
Proteins identified	332/336/326	395|376	405|400	432|417	416|414
Unique proteins	332/89/39	199|183	200|190	225|217	206|207
Total proteins	460 (from 3 runs)	531|515 (from 2 runs)	532|522 (from 2 runs)	557|549 (from 2 runs)	548|549 (from 2 runs)

Number	Run 3 (i)|(ii) (sim-PIE)	Run 3 (i)|(ii) (res-PIE)	Run 3 (i)|(ii) (sel-PIE)	Run 3 (i)|(ii) (com-PIE)
MS/MS spectra	1207|1040	2063|2020	2404|2405	2275|2269
m/z values excluded	4049	4567	18117	23858
Peptides identified	297|241	676|632	657|633	642|619
Unique peptides	229|241	477|438	530|497	585|561
Proteins identified	198|166	384|358	385|379	413|420
Unique proteins	54|44	130|129	133|115	175|180
Total proteins	585|559 (from 3 runs)	652|661 (from 3 runs)	690|672 (from 3 runs)	723|728 (from 3 runs)

(1) See supplementary materials for protein names, sequences matched peptides, MASCOT scores, mass errors, and MASCOT threshold scores.

(2) Replicate runs without precursor ion exclusion.

(3) 2nd run using simple precursor ion exclusion (sim-PIE) without retention time information.

(4) 2nd run using restricted precursor ion exclusion (res-PIE) with retention time information.

(5) 2nd run using selective precursor ion exclusion (sel-PIE) with expanded m/z list and retention time information.

(6) 2nd run using complete precursor ion exclusion (com-PIE) with retention time information. Note that, for replicate runs with com-PIE, the first run dataset was different from the 1st run in (2); the 1st run dataset for com-PIE is provided in the supplementary materials.

Table 4.2 Summary of the results obtained from five replicate 2-h runs of the yeast cell digest without PIE[1].

Number	Run 1	Run 2	Run 3	Run 4	Run 5
MS/MS spectra	2227	2197	2232	2220	2206
Peptides identified	692	693	672	674	666
Unique peptides	692	128	49	40	29
Total peptides	692	820	869	909	938
Proteins identified	249	254	246	240	243
Unique proteins	249	44	22	16	19
Total proteins	249	293	315	331	350

(1) See supplementary materials for protein names, sequences matched peptides, MASCOT scores, mass errors, and MASCOT threshold scores.

Table 4.3 Summary of the results obtained from five replicate 2-h runs of the yeast cell digest with the four PIE methods(1, 2).

Number	Run 1	Run 2 (sim-PIE)	Run 2 (res-PIE)	Run 2 (sel-PIE)	Run 2 (com-PIE)
MS/MS spectra	2206	1244±23	2112±10	2146±4	2028±10
m/z values excluded		2247	2247	5821	8315
Peptides identified	666	230±6	500±10	372±12	423±6
Unique peptides	666	184±7	374±13	291±9	381±8
Total peptides	666	850±7	1040±13	957±9	1047±8
Proteins identified	243	151±5	262±4	218±5	258±1
Unique proteins	243	57±3	123±5	101±2	142±2
Total proteins	243	300± 3 (2 runs)	366±5 (2 runs)	344±2 (2 runs)	385±2 (2 runs)

Number	Run 3 (sim-PIE)	Run 3 (res-PIE)	Run 3 (sel-PIE)	Run 3 (com-PIE)
MS/MS spectra	413±23	1833±29	2069±13	1748±34
m/z values excluded	3254	4384	9663	12072
Peptides identified	69±3	322±3	238±16	287±21
Unique peptides	39±2	203±9	179±14	200±6
Total peptides	892±2	1247±9	1141±14	1235±6
Proteins identified	62±2	217±11	180±9	203±6
Unique proteins	18±1	77±7	74±10	76±3
Total proteins	319±1 (3 runs)	445±7 (3 runs)	415±10 (3 runs)	459±3 (3 runs)

Number	Run 4 (sim-PIE)	Run 4 (res-PIE)	Run 4 (sel-PIE)	Run 4 (com-PIE)
MS/MS spectra	182±2	1541±44	1998±12	1290±43
m/z values excluded	3641	6269	12248	18835
Peptides identified	30±4	244±19	148±4	157±5
Unique peptides	17±4	132±7	103±5	118±2
Total peptides	906±4	1381±7	1263±5	1362±2
Proteins identified	29±3	183±10	116±4	133±4
Unique proteins	7±2	61±2	32±4	52±2
Total proteins	325±2 (4 runs)	498±2 (4 runs)	461±4 (4 runs)	509±2 (4 runs)

Number	Run 5 (sim-PIE)	Run 5 (res-PIE)	Run 5 (sel-PIE)	Run 5 (com-PIE)
MS/MS spectra	130±14	1224±30	1913±11	1037±13
m/z values excluded	3818	7829	13672	21610
Peptides identified	19±2	157±9	123±8	127±8
Unique peptides	14±0	79±2	73±5	60±3
Total peptides	916±0	1464±2	1337±5	1425±3
Proteins identified	19±2	124±4	100±6	99±3
Unique proteins	5±2	31±2	29±2	24±2
Total proteins	332±2 (5 runs)	527±2 (5 runs)	487±2 (5 runs)	533±2 (5 runs)

(1) See supplementary materials for the results of individual runs and, for each run, the protein names, sequences matched peptides, MASCOT scores, mass errors, and MASCOT threshold scores.

(2) Except Run 1, the number shown represents the average value of three repeated runs in each setting associated with a standard derivation.

4.6 Literature Cited

(1) Fournier, M. L.; Gilmore, J. M.; Martin-Brown, S. A.; Washburn, M. P. *Chemical Reviews (Washington, DC, United States)* **2007**, *107*, 3654-3686.

(2) Mintz, P. J.; Patterson, S. D.; Neuwald, A. F.; Spahr, C. S.; Spector, D. L. *EMBO Journal* **1999**, *18*, 4308-4320.

(3) Patterson, S. D.; Spahr, C. S.; Daugas, E.; Susin, S. A.; Irinopoulou, T.; Koehler, C.; Kroemer, G. *Cell Death and Differentiation* **2000**, *7*, 137-144.

(4) Spahr, C. S.; Davis, M. T.; McGinley, M. D.; Robinson, J. H.; Bures, E. J.; Beierle, J.; Mort, J.; Courchesne, P. L.; Chen, K.; Wahl, R. C.; Yu, W.; Luethy, R.; Patterson, S. D. *Proteomics* **2001**, *1*, 93-107.

(5) Davis, M. T.; Spahr, C. S.; McGinley, M. D.; Robinson, J. H.; Bures, E. J.; Beierle, J.; Mort, J.; Yu, W.; Luethy, R.; Patterson, S. D. *Proteomics* **2001**, *1*, 108-117.

(6) Yi, E. C.; Marelli, M.; Lee, H.; Purvine, S. O.; Aebersold, R.; Aitchison, J. D.; Goodlett, D. R. *Electrophoresis* **2002**, *23*, 3205-3216.

(7) Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K.-K.; Zhao, R.; Smith, R. D. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 11049-11054.

(8) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. *Nature Biotechnology* **2003**, *21*, 532-538.

(9) Liu, H.; Sadygov, R. G.; Yates, J. R., III *Analytical Chemistry* **2004**, *76*, 4193-4201.

(10) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. *Nature Methods* **2005**, *2*, 667-675.

(11) Chong, P. K.; Wright, P. C. *Journal of Proteome Research* **2005**, *4*, 1789-1798.

(12) Durr, E.; Yu, J.; Krasinska, K. M.; Carver, L. A.; Yates, J. R.; Testa, J. E.; Oh, P.; Schnitzer, J. E. *Nature Biotechnology* **2004**, *22*, 985-992.

(13) Cagney, G.; Park, S.; Chung, C.; Tong, B.; O'Dushlaine, C.; Shields, D. C.; Emili, A. *Journal of Proteome Research* **2005**, *4*, 1757-1767.

(14) Venne, K.; Bonneil, E.; Eng, K.; Thibault, P. *Analytical Chemistry* **2005**, *77*, 2176-2186.

(15) Sarvaiya Hetal, A.; Yoon Jung, H.; Lazar Iulia, M. *Rapid communications in mass spectrometry: RCM* **2006**, *20*, 3039-3055.

(16) Hui, J. P. M.; Tessier, S.; Butler, H.; Jonathan, B.; Kearney, P.; Carrier, A.; Thibault, P. *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics*, Montreal, Quebec, Canada 2003.

(17) Wang, N.; Zheng, J.; Whittal, R.; Li, L., *Proceedings of the 54th ASMS Conference on Mass Spectrometry and Allied Topics*, Seattle, WA, U.S.A. 2006.

(18) Chen, H.S.; Rejtar, T.; Andreev, V.; Moskovets, E.; Karger, B. L. *Analytical Chemistry* **2005**, *77*, 7816-7825.

(19) Washburn, M.; Wolters, D.; Yates, J. *Nature Biotechnology* **2001**, *19*, 242-247.

(20) Washburn, M.; Ulaszek, R.; Deciu, C.; Schieltz, D.; Yates, J. *Analytical Chemistry* **2002**, *74*, 1650-1657.

(21) Ji, C.; Lo, A.; Marcus, S.; Li, L. *Journal of Proteome Research* **2006**, *5*, 2567-2576.

(22) Sharp, P.; Li, W. *Nucleic Acids Research* **1987**, *15*, 1281-1295.

(23) Coghlan, A.; Wolfe, K. *Yeast* **2000**, *16*, 1131-1145.

(24) Wu, G.; Culley, D.; Zhang, W. *Microbiology-SGM* **2005**, *151*, 2175-2187.

Chapter 5

Off-line Two-dimensional Liquid Chromatography with Maximized Sample Loading to Reversed-Phase LC-ESI Tandem Mass Spectrometry for Shotgun Proteome Analysis*

5.1 Introduction

Liquid chromatography (LC) combined with tandem mass spectrometry (MS/MS) has become a widely used technique for shotgun proteome analysis (1). Reversed-phase (RP) LC is a preferred mode of separation for LC MS/MS due to its high separation power and compatibility of the mobile phases with ionization techniques such as electrospray ionization (ESI). Because of the complexity of peptide samples generated from a proteome digest, additional peptide separation is often required prior to RPLC MS/MS. The use of multi-dimensional LC based on different separation mechanisms can increase the detection concentration dynamic range and reduce ion suppression in MS analysis, resulting in a greater number of peptides and proteins identified (1). However, to determine the actual number of LC dimensions to be used for peptide separation in proteome analysis, the overall analysis time must also be considered. At present, two-dimensional (2D) LC with RPLC as the 2nd dimension of separation is widely used in combination with MS/MS for shotgun analysis of a proteome of modest complexity, such as whole cell extracts or organelles, providing useful proteome coverage in a reasonable analysis time. Because of analysis time constraints, it is critical to develop and optimize an analysis strategy for generating the highest proteome coverage in the shortest analysis time. In this work we report an efficient technique based on the use of off-line 2D LC-ESI MS/MS for shotgun proteome analysis.

In 2D-LC MS/MS, strong-cation exchange (SCX) LC is commonly used to provide orthogonal peptide separation, although various other modes of separation have also been

*A form of this chapter will be submitted for publication as: N. Wang, C. Xie, B. J. Young and Liang Li, 2008, "Off-line Two-dimensional Liquid Chromatography with Maximized Sample Loading to Reversed-Phase LC-ESI Tandem Mass Spectrometry for Shotgun Proteome Analysis". Dr. C. Xie and N. Wang contributed equally to this work.

combined with RPLC for shotgun proteome analysis (1). Combining SCX with RPLC can be done either on-line or off-line. On-line 2D-LC MS/MS has the advantages of full automation and minimum sample loss (2, 3). On the other hand, a major attribute of the off-line approach lies in the possibility of optimizing analytical performance in the two modes of separation independently (4). For example, SCX separation of peptides using a mobile phase detrimental to RPLC and/or ESI can be used in the off-line approach (5, 6). While independent optimization of SCX and RPLC has been shown to be beneficial in improving the peptide separation efficiency (1), another important feature that has not been fully explored in off-line 2D-LC ESI MS/MS is related to RPLC sample loading.

The importance of RPLC sample loading for optimizing peptide identification by ESI MS/MS is well recognized and some controls of sample loading to RPLC MS/MS in 2D-LC MS/MS have been attempted (7). For example, peptide amounts in SCX fractions were estimated based on UV absorption signal intensities in the SCX chromatogram and portions of a few selected individual fractions were individually injected to RPLC MS/MS (7). Judging from the results obtained from the initial MS/MS run, a subsequent run was carried out after adjustment of the injection volume or concentrating/diluting the SCX fraction. While this approach offers some control of the sample amount to be injected to RPLC MS/MS, it requires initial test runs which consume samples and time as well as potentially overloading the column, if the initial sample concentration is too high. Column overloading can be a serious problem which requires lengthy column cleaning and equilibration, prolonging the total analysis time. In addition, because the test run provides only an estimate of the peptide concentration in the SCX fractions, the sample loading to RPLC is still not fully optimized.

To optimize the sample loading with a goal of maximizing the performance of RPLC MS/MS, a technique to accurately measure the concentration of a small amount of peptides in a SCX fraction containing high amounts of salt, buffer or organic modifier is needed. To this end, we have developed a fully automated system based on RPLC using a step solvent gradient and UV detection to measure the peptide concentration. This system also purifies the peptide mixtures by getting rid of salt, buffer and other chemicals present in SCX fractions. In this paper, the setup and performance of the system are first described. The results of our investigation on the effects of sample loading on the detectability of peptides and proteins by RPLC-ESI MS/MS are then discussed. A strategy of off-line 2D-LC with maximized sample loading to RP-LC MS/MS and its application for proteome profiling of breast cancer MCF-7 cells are presented. In addition, the biological significances of the proteome profile generated are discussed within the context of reported breast cancer biomarkers.

5.2 Experimental

5.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide, trifluoroacetic acid (TFA), sodium bicarbonate and urea were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). The BCA assay kit was from Pierce (Rockford, IL).

5.2.2 Cell culture and protein sample preparation

The MCF7 breast cancer cells (ATCC® number: HTB-22™) were cultured in 15 cm diameter plates at 37 °C in DMEM Gibco medium supplemented with 10% fetal bovine serum. The plates were then washed twice with ice-cold 25 mL PBS^{++} buffer (0.68 mM $CaCl_2$, 0.5 mM $MgCl_2$, 1.4 mM KH_2PO_4, 4.3 mM Na_2HPO_4, 2.7 mM KCl, and 137 mM NaCl). The cells were harvested by scraping them from the plates into the PBS^{++} buffer and centrifugation at 100 *g* for 8 min at 4 °C. The cell pellet was resuspended in 4 mL PBS buffer (1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM KH_2PO_4, pH 7.4) and passed twice through a mini-cell French Press (Aminco Rochester, NY). This was followed by sonication on ice (4×10 s pulses). The final volume of the lysate was brought up to about 5 mL using PBS buffer. The lysate was then centrifuged at 100 000 *g* for 1 h at 4 °C.

Protein concentration was determined by BCA assay using a BCA assay kit. The protein solutions was made into aliquots and stored at -80 °C. An acetone-precipitation step was introduced to remove the salts from the extracted proteins. Acetone was precooled to -80 °C and added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The mixture was then incubated at -20 °C for 60 minutes and centrifuged at 14 000 rpm for 10 min. The supernatant was decanted and properly disposed, the residual acetone was evaporated at ambient temperature and the pellet was stored at -20 °C for further use.

A solution containing 100 mM ammonium bicarbonate and 6 M urea was used to redissolve the pellet in the vial. The proteins were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide, followed by dilution to reduce urea concentration for subsequent trypsin digestion. Trypsin digestion of BSA and the 4-protein mixture was performed using the same standard procedure.

5.2.3 Desalting and quantification

The desalting and quantification setup consisted of an Agilent 1100 HPLC system (Palo Alto, CA) with a UV detector. The desalting of the tryptic peptides was performed on a 4.6 mm × 5 cm Polaris C18 A column with a particle size of 3 μm diameter and 300 Å pore. (Varian, CA). After loading of an appropriate amount of peptide sample, the column was flushed with mobile phase A (0.1% TFA in water), thus salts and other interferences, such as DTT and IDA, were effectively removed. Subsequently, the concentration of mobile phase B (0.1% TFA in acetonitrile) was step-wise increased to 85% to ensure the complete elution of the peptide fractions from the column. For separation of the BSA digest, the mobile phase B concentration was set at 2.5% for 5.0 min, and then gradually increased to 35% in 30 min to elute the peptide components evenly, followed by an increase to 85% in 5 min and held for 10 min. The peptides were fractionated between 4.55 and 5.05 min.

5.2.4 Cation exchange chromatography

In this work, highly hydrophilic polysulfoethyl A column (2.1 mm i.d. x 250 mm with particle size of 5 μm diameter and 300 Å pore) from PolyLC was used for the strong-cation exchange separation of the tryptic peptides of the proteins of MCF-7 cells.

Gradient elution was performed with mobile phases A (10 mM KH_2PO_4, pH 2.76) and B (10 mM KH_2PO_4, pH 2.76, 500 mM KCl), The gradient profile was as follows: 0 min: 0% B, 5 min: 0% B, 100 min: 20% B, 150 min: 60% B, 160 min: 100% B, 200 min: 100% B. The fractions were collected every 10 min for the first 40 min, every 4 min between 40 and 100 min and every 10 min thereafter. Therefore, a total of 28 fractions were collected and directly desalted and quantified by the desalting setup described above.

5.2.5 LC-ESI MS/MS

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, the desalted and quantified digests were concentrated using a SpeedVac (Thermo Savant, Milford, MA) to ~5 μL and reconstituted to a specific concentration using 0.1% formic acid. Then the intended amount of digest solution was injected onto a 75 μm × 100 mm Atlantis dC18 column (Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120 min gradient (2-6% Solvent B for 2 min, 6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min) after column equilibration at 2% Solvent B for 20 min and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 1 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 1.5 s each. Charge states of +2 and +3 were chosen for MS/MS, as in our experience of shotgun proteome profiling using this instrument very few peptide ions were detected in +1 or +4 charge state and the MS/MS spectra of the ions of these charge states were generally poor for database searching. For both dynamic and precursor ion mass exclusion, a mass tolerance window of 80 mDa was applied. The collision energy used to perform MS/MS was varied according to the mass and charge state of the eluting peptide ion. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants (i.e., lock-mass), was infused at a flow rate of 300 nL/min, and a 1 s MS scan was acquired every 1 min throughout the run.

For the SCX fractionated samples, each fraction was analyzed twice on the LC-MS system with a precursor ion mass exclusion list involved in each second run. The exclusion list (see Results) was generated based on the peptides identified in the first run from the MASCOT search program (Matrix Science, London, U.K.) (8). The m/z values of each identified peptide with a MASCOT score 10 points equal to or higher than the identification threshold were extracted from the database search results and loaded into the MS method for the new LC-ESI run.

5.2.6 Data analysis

Raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files by using ProteinLynx Global Server 2.1.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program (version 2.2.1). Database searching was restricted to *Homo sapiens* (human) in the SWISSPROT database (October 4, 2007) and 17317 entries were searched. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, oxidation (M), pyro-Glu (N-term Q) and pyro-Glu (N-term E) [for phosphopeptide identification, variable modifications were phosphor(ST) and phosphor(Y)]. The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides including phosphopeptides with scores lower than the MASCOT threshold score for identity at a confidence level of 95% were then removed from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list. Specifically, the final unique protein or peptide list was generated by merging all the protein or peptide lists from individual runs according to the following roles: only unique proteins (under unique gene names) and peptides with the highest scores were kept; each peptide was only associated to one unique protein; only the first hit within each identified protein group was kept in the list as a representative protein. Redundant peptides with lower identification scores were removed. And redundant proteins with either lower scores or lower number of peptides were also removed.

To gauge the false positive peptide matching rate in our analysis, we applied the target-decoy search strategy by searching the MS/MS spectra against the forward and reversed human proteome sequences (9). Briefly, the matched spectra from the MASCOT database search using the forward or correct proteome sequence were re-searched against the reversed proteome sequence as decoy. The decoy peptide matches with scores above the threshold scores at the 95% confidence level were then compared to those in the forward sequence search. If the score of a MS/MS spectrum matched with a decoy peptide was higher than that of the same spectrum matched with a normal peptide, a false positive match was registered. The false positive matching rate was calculated by using the equation $2\times n(\mathrm{rev})/[n(\mathrm{forward})+n(\mathrm{rev})]$, where n(rev) and n(forward) are the number of matches from the reserved (decoy) and forward (correct) sequence, respectively (9).

The Gene Ontology (GO) terms of cellular component, molecular function and biological process for the identified proteins were extracted from the ExPASy (Expert Protein Analysis System) Proteomics Server according to their Swiss-Prot IDs (http://ca.expasy.org). This information was referenced from the QuickGO GO browser (http://www.ebi.ac.uk/ego). Thus, proteins without any GO terms assigned in the Swiss-Prot database were not involved in the GO term classification process.

5.3 Results

5.3.1 Peptide Quantification and Desalting

There are a number of techniques useful for measuring peptide concentrations. UV spectrometry is one of the simplest techniques and provides micromolar sensitivity when the absorption wavelengths corresponding to the peptide backbones (i.e., the carbonyl groups) absorption are probed at about 214 nm. However, for SCX fractions, each fraction contains a small amount of peptides with high concentrations of salts and their counter anions that absorb or reflex at similar UV wavelengths as those of peptides. Desalting of the SCX fractions followed by UV spectrometric measurement may be applied to quantify the peptide amount in each fraction. Since desalting can be carried out effectively by using HPLC with a RP column, it makes more sense to combine the desalting step with the UV measurement in an HPLC system equipped with a UV detector. The entire process of desalting and peptide quantification can thus be fully automated. However, in developing such a system, a number of technical issues need to be worked out and are discussed below.

The main purpose of using HPLC in the LC-UV peptide quantification system is to get rid of the salts. Separation of peptides is neither needed nor desirable as it increases the analysis time. Thus, a step solvent gradient was used to desalt with one solvent and elute peptides with another. In our experiment, 2.5% acetonitrile with 0.1% TFA was used to pre-equilibrate the Polaris C18 A column as well as to flush out salts and other low-retention substances. A solvent containing 85% acetonitrile with 0.1% TFA was introduced for the rapid desorption of all the peptides simultaneously, producing a sharp peptide elution peak (<0.50 min in peak width). The area of the peptide peak was used to provide quantitative information. To quantify the peptide amount, a calibration curve from a standard peptide mixture is required.

Figure 5.1 shows the overlaid elution profiles of different concentrations of BSA digests from the desalting and elution LC-UV runs. The peaks shown between 1 and 3 min are from the impurities in the digests and the peaks with retention time around 4.7 min are from the BSA tryptic peptides. As Figure 5.1 shows, the elution patterns of the BSA digests are very similar. As the concentration of the BSA digest increases, the peptide peak area increases accordingly. When a blank run was performed, there was still a peak appearing at the same retention time as the BSA digests in the chromatogram. This peak, named as the system peak, comes from the rapid solvent change during the step gradient due to differences in absorption coefficients and reflective indexes of the two changing solvents. However, the system peak is very reproducible. First, it has a reproducible migration time from multiple runs (e.g., the variation is less than 0.02% in five replicate runs). Secondly, the peak area does not change from run to run (e.g., the variation is <0.78% in five replicate runs). It should be noted that, since the solvents contain 0.1% TFA and TFA absorbs at 214 nm, any variation of TFA concentration in the two solvents will have an influence on the peak area of the system peak. Thus caution should be exerted to avoid any changes in TFA concentration between the mobile phases to ensure the constancy of the system peak.

Figure 5.2(A) shows the calibration curve of peak area difference versus the amount of sample injection for the BSA digest. Peak area difference was calculated from the peak area measured at a given sample injection minus the system peak measured in a blank run. Very good correlation (R^2=0.999) by linear regression between 0.25 and 15 μg was achieved. The intercept of the curve has a small negative value, indicating that the system peak area in the sample runs is slightly smaller than that from a blank run. The presence of the analyte at the elution time corresponding to the system peak appears to alter the magnitude of the changes from the solvent absorption and refractive index, compared to those of a blank run. The calibration curve has a linear range from 0.25 to 15 μg, whilst the peak area increases nonlinearly for injections from 15 to 30 μg. Nonlinear responses were due to saturation of the UV absorbance detection. Note that the system peak is quite large and thus the linear response range is reduced, compared to absorbance measurement in conventional HPLC-UV experiments where there is no or small baseline absorbance. To quantify less than 0.25 μg of peptides (i.e., to extend the quantification dynamic range), a smaller RP column may be used. However, switching to a smaller column to extend the lower limit is not required. As is illustrated below, the optimal amount of injection to capillary RPLC-ESI MS/MS is in the microgram range. Thus the linear calibration range shown in Figure 5.2(A) is adequate for quantification of peptides in individual fractions collected from a typical SCX run.

Because the peptide composition of an SCX fraction of a proteome digest is quite

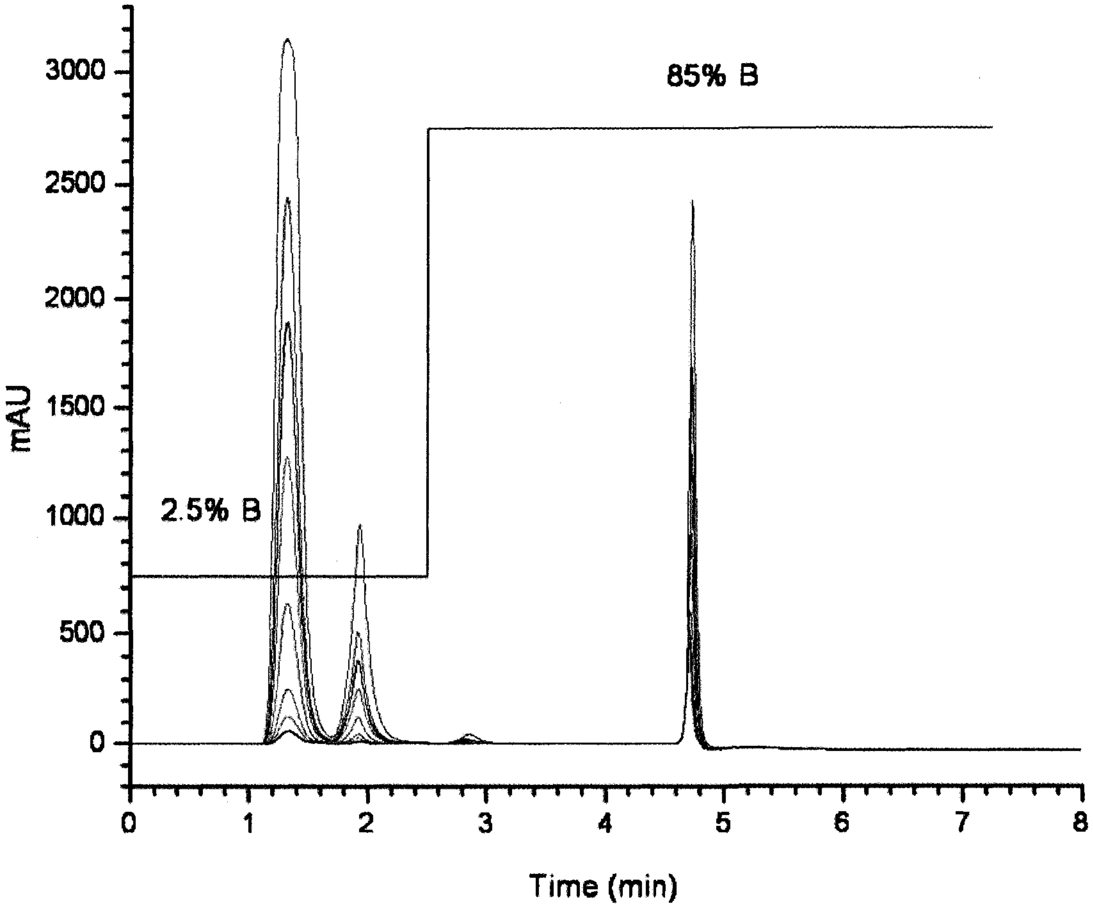


Figure 5.1 Overlay of the elution profiles of 0.125, 0.25, 0.5, 1.0, 2.5, 5.0, 7.5, 10 and 20 μg BSA tryptic digests. Chromatographic conditions: Polaris C18 A column, 3 μm, 50× 4.6 mm I.D.; mobile phase, 0.1% TFA in water (A) and acetonitrile (B); flow rate, 1 mL/min; detection wavelength, 214 nm; gradient: 2.5% B in 0-2.50 min and 85% B for 2.5-17.50 min.

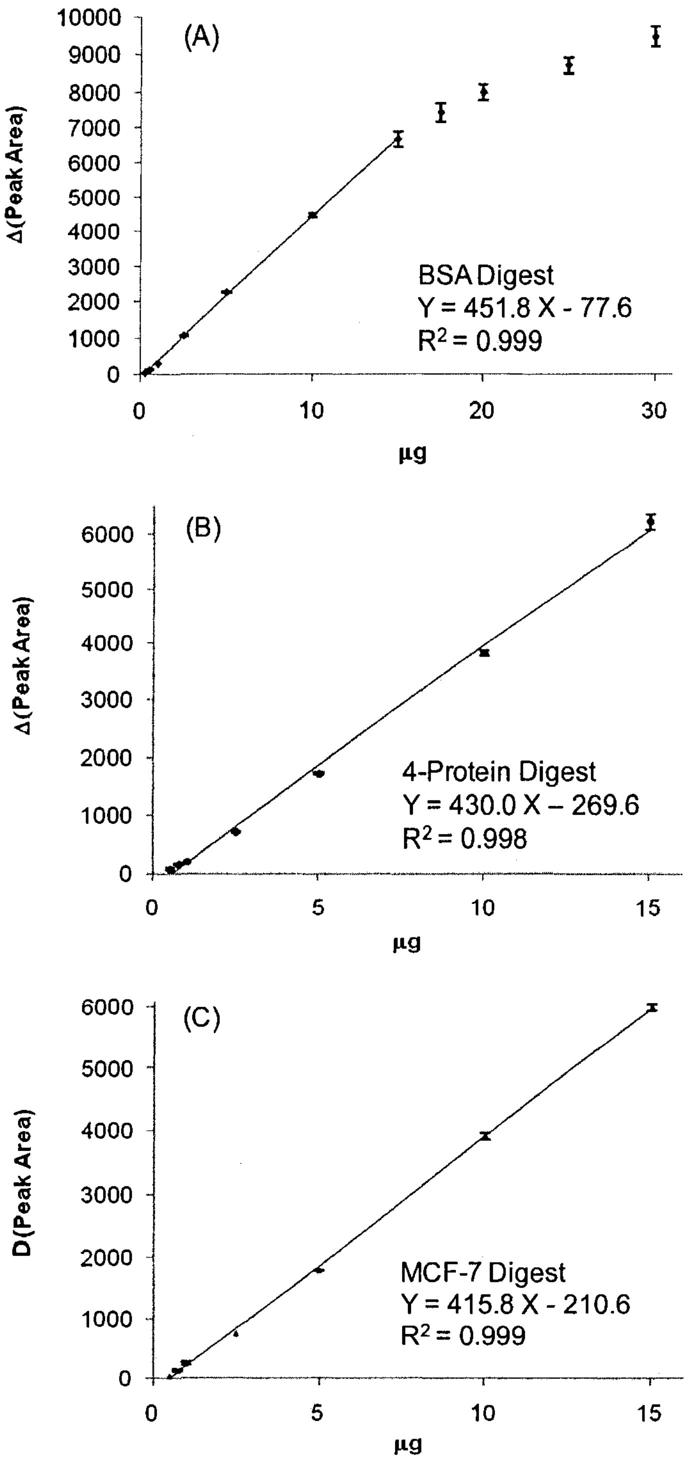


Figure 5.2 Calibration curves constructed from RPLC-UV at 214 nm: (A) BSA tryptic digest, (B) tryptic digest of equimoles of cytochrome C, myoglobin, lysozyme and β-casein, and (C) tryptic digest of MCF-7 whole cell protein extract. Δ(Peak Area) was calculated from the peak area of a sample injection minus the system peak area of a blank injection.

different from that of the BSA digest, the use of BSA digest to establish a calibration curve for quantification of unknown peptide mixtures may introduce some errors. Quantification is based on the UV absorbance at 214 nm mainly arising from the absorption of carbonyl groups along the peptide chains as well as other moieties such as aromatic amino acids. Variations of UV absorption from different peptides are expected. Thus an unknown peptide mixture may have somewhat different absorptivity from that of the BSA digest. To partially gauge the extent of this difference in average molar absorptivity of different peptide mixtures, a mixture of equimolar cytochrome c, β-casein, myoglobin and lysozyme was digested by trypsin and the resulting peptides were quantified by the LC-UV system using conditions identical to those used for the analysis of the BSA digest. The calibration curve for the 4-protein digest is shown in Figure 5.2(B) and the discrepancy of the slopes of the two calibration curves was found to be about 4.8%. Therefore, if the BSA digest is used for calibration, the peptide amount of the 4-protein digest determined from the BSA digest calibration curve is up to 4.8% lower than the expected value. Based on multiple evidences including the absence of any protein signals monitored by LC, gel electrophoresis, MALDI or ESI MS analysis of the digest samples, we believe that the digestion of BSA or 4-protein mixture was complete. Thus, the difference in the calibration slopes reflects the difference of average absorptivity between the BSA digest and the 4-protein digest.

An even more complicated peptide mixture, whole cell protein digest of MCF-7, was examined. The calibration curve of this digest is shown in Figure 5.2(C). Since the amount of the peptides in the digest was unknown, a stock solution of the digest was first injected into the LC-UV system to generate an elution peak from which the peptide concentration of the solution was determined based on the use of the 4-protein digest calibration curve. Various concentrations of the digest were then prepared and injected into the desalting and quantification system to construct the calibration curve shown in Figure 5.2(C). This curve has a very similar slope to that of the 4-protein digest, indicating that the average absorptivity approaches a constant value as the complexity of the mixture increases beyond the 4-protein digest. Peptides from an individual SCX fraction should have compositional complexity somewhere between the 4-protein digest and the whole cell extract digest. Thus, the 4-protein digest is a better choice than the BSA digest as a calibration standard.

The use of a proper step gradient for peptide elution is critical for efficient elution of all peptides from the column at the highest possible speed to increase sample throughput, which is important for processing many SCX fractions collected in a typical SCX run. To gauge if any peptides in a peptide mixture were lost during the desalting and peptide amount measurement process, the UV chromatograms from the separation of the BSA tryptic digest before and after the desalting experiment were obtained (see Figure 5.3). As shown in Figure 5.3, the desalting procedure effectively removes the salts and other interferences which elute at a retention time of less than 4 min and the chromatographic patterns for the peptide peaks are well preserved. Thus, RPLC-UV using step gradient provides a rapid and effective means of desalting and concurrent measurement of peptide quantities of SCX fractions with good accuracy and without affecting peptide composition.

5.3.2 Maximizing Sample Loading to RPLC MS/MS

The use of RPLC-UV for desalting and peptide quantification allows us to precisely control the amount of peptide mixtures to be injected to RPLC-ESI MS/MS. To investigate the effects of sample loading on the detectability of peptides and proteins, a breast cancer cell line, MCF-7, was used to generate a tryptic digest from a whole cell protein extract. This cell line was chosen because of its biological significance (see below) as well as its representation of a complicated proteome often encountered in many biological studies. After trypsin digestion, the peptide sample containing high concentrations of salts and buffers was subjected to desalting and quantification using RPLC-UV. Varying amounts of peptides were injected into the RPLC ESI-QTOF system and two replicate runs for each amount were carried out. The sample loading capacity for the 75 μm × 100 mm Atlantis dC18 column with a short trap column (180 μm × 20 mm) is normally around 2 μg according to the manufacturer's information. Thus, the amounts of the MCF7 digests being tested ranged from 250 ng to 2 μg.

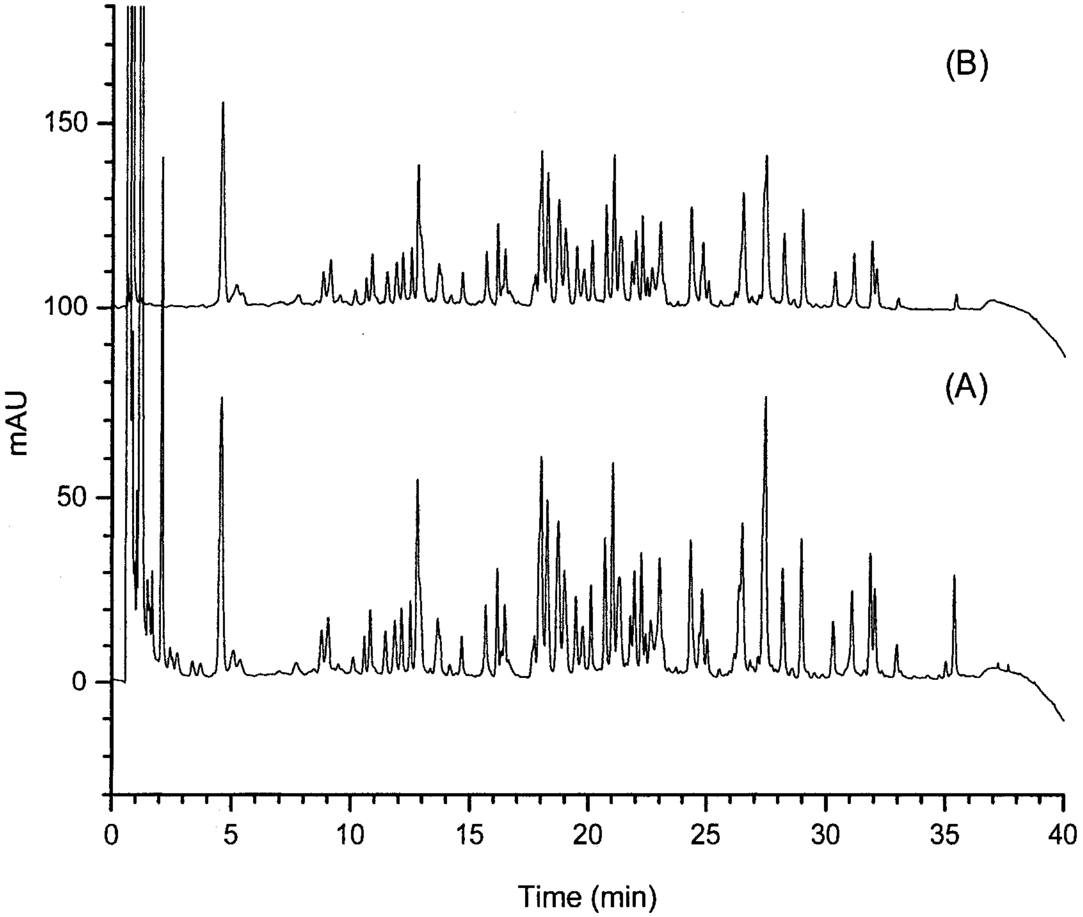


Figure 5.3 Chromatograms generated from the separation of the BSA tryptic digests (A) before and (B) after desalting. Gradient: 2.5%-2.5%-35%-85%-85%-2.5% B in 0-5-35-40-50-50.01 min. Other conditions are the same as in Figure 5.1.

To investigate the relation between the peptide/protein identification yields and the sample loading amount, the numbers of peptides/proteins identified and their identity scores were plotted against their corresponding injection amount (see Figure 5.4). Figure 5.4A clearly shows an important trend, i.e., the number of peptides identified and their identification scores increased as the sample loading increased. There was a significant increase (65%) in the number of identified peptides when the sample loading increased from 250 ng to 500 ng, and another 29% increase when 1 μg of sample was loaded. A smaller increase (6%) was observed when 2 μg of sample was loaded. Figure 5.4A also shows that the number of high-score peptides (e.g., scores of above 100) increased dramatically as the sample loading increased. However, the trend was not as pronounced when 2 μg of sample was loaded. It is worth noting that good run-to-run reproducibility from the replicates (<1% variation in number of peptides) was obtained as shown in Figure 5.4A. But, with a sample loading of 250 ng, a variation of 5% was observed.

At the protein level, as shown in Figure 5.4B, both the number of unique proteins and the identification score of proteins increased significantly (55%) when the sample loading increased from 250 ng to 500 ng. When 1 μg of sample was loaded, a relatively smaller increase (8%) in number of proteins was observed, whereas there was still a remarkable increase in the percentage of proteins with high identification scores (e.g., score of >500). The average protein score for 1 μg sample loading was 192, which was 40 points and 59 points higher than that of 500 ng and 250 ng sample loadings, respectively. The same tread in false positive peptide matching rate was observed. The false positive rate was found to be 0.97%, 0.58% and 0.16% for the search results obtained with 250 ng, 500 ng and 1 μg of sample loading, respectively. In contrast, the average number of unique proteins dropped 6% when 2 μg sample was analyzed, although the average score per protein increased to 231. This can be explained by considering the ion suppression effect in ESI MS/MS. As the sample loading amount increases to a point close to column saturation, the ion suppression effect becomes more severe, so that peptides from the low abundance proteins have less chance to be ionized. This is evident in Figure 5.5 where the distribution of peptide hits per protein for each sample amount is shown. The percentage of "single hit" proteins likely from the low abundance proteins decreased and the "multiple-hit" proteins (i.e., >11 peptides) increased as the sample loading increased from 1 to 2 µg.

Another more serious problem associated with 2 µg sample loading is related to sample carry-over from run to run. With 2 µg loading, sample carry-over on the nano-LC column was found to be very severe and great effort and time were needed to wash the column to ensure its high quality performance for the next run. For example, in contrast to a 1 hour total washing and equilibrium time commonly used after 1 µg sample injection, a total of at least 3 hours (i.e., three cycles) were required to clean and equilibrate the column after 2 µg sample injection before the next run. Thus, the use of 2 µg sample loading consumes more sample with no benefit of increasing the number of unique proteins identified while causing a severe sample carryover problem in a typical nano-LC column used for shotgun proteome work. Our cumulated experience in running the RPLC-ESI MS/MS instrument with various proteomic digests, including many membrane protein samples, indicates that sample carryover becomes a major problem when over ~1.5 µg of sample is loaded to the column. Thus, we generally avoid a sample loading of greater than 1.5 µg for any given sample.

Comparing the peptide and protein results obtained from the 500 ng and 1 µg sample loading, as shown in Figure 5.4B, the number of proteins identified from the 1 µg sample is, on average from the two replicates, slightly higher than that from the 500 ng sample loading. However, the number of peptides identified from the 1 µg sample loading is significantly higher than that of 500 ng sample loading as illustrated in Figure 5.4A. Thus, a sample loading of 1 µg is preferred over the 500 ng sample loading in terms of the confidence of proteins identified judged by the peptide/protein ratio, as also evident from the protein distribution shown in Figure 5.5.

From the above discussion and the results shown in Figures 5.4 and 5.5, it can be concluded that 1 µg sample injection not only allowed the identification of a maximum number of unique proteins and a higher level of protein identification confidence, but also did not cause any column saturation and sample carryover problems on the LC column.

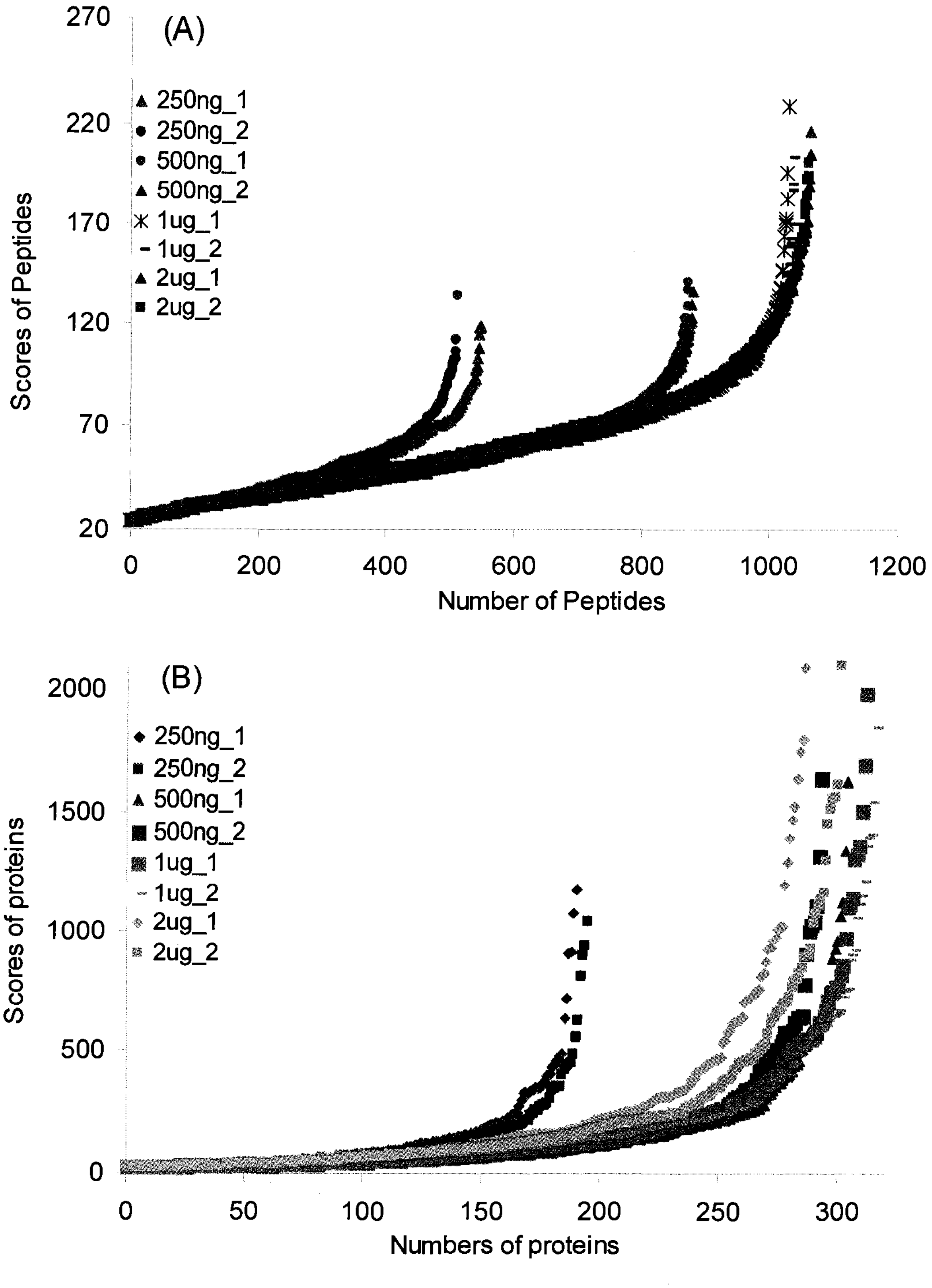


Figure 5.4 Distributions of MASCOT matching scores of (A) peptides and (B) proteins from RPLC MS/MS runs of the MCF-7 cell digests with different amounts of sample loading. Replicates were performed for each amount.

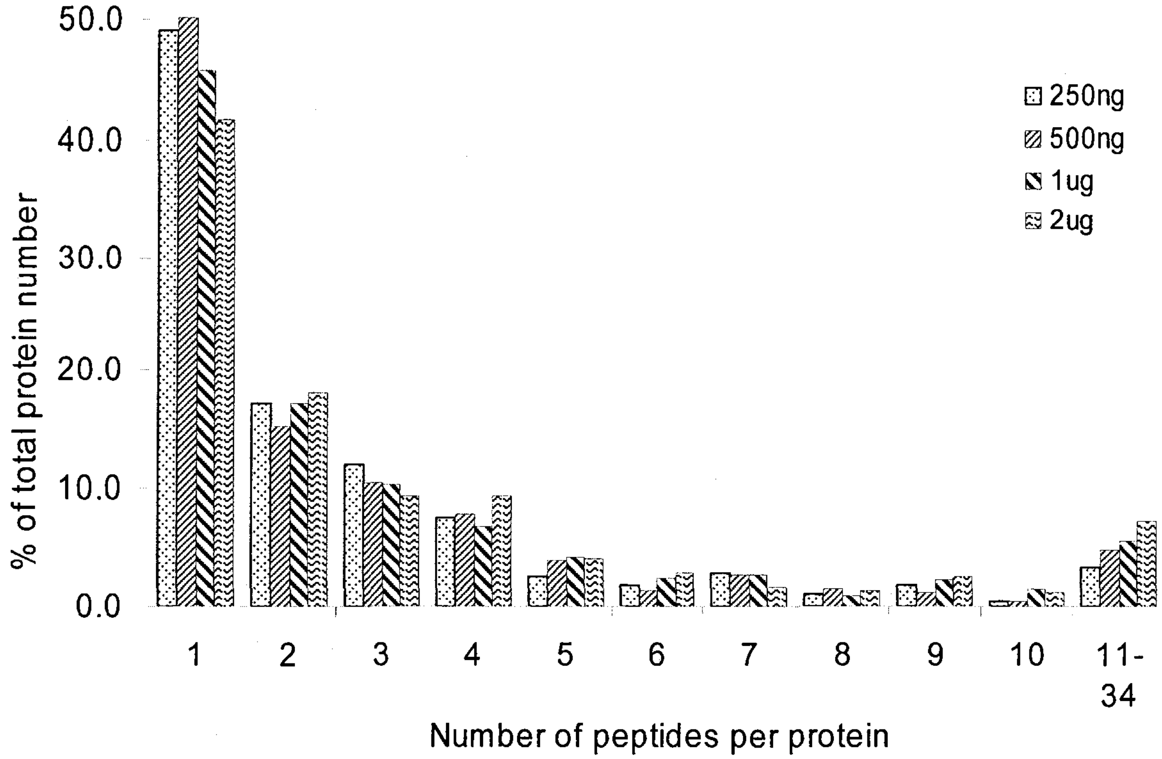


Figure 5.5 Distributions of the percentage of protein numbers as a function of the number of peptides matched to a protein. The dataset were the same as the ones used for plotting Figure 5.4.

While this optimal sample loading value was determined for the instrument and experimental setup described in this work, for different systems it can be determined accordingly if there are any parameters of the particular LC-MS system being changed, such as column inner diameter, column length, etc.

5.3.3 2D-LC MS/MS of MCF-7 Cell Extracts

Direct analysis of the MCF-7 tryptic digest by one-dimensional RPLC MS/MS resulted in the identification of a little over 300 proteins, as shown in Figure 5.4. To increase the proteome coverage, SCX was used to fractionate the proteome digest, followed by desalting and peptide quantification using RPLC-UV. Figure 5.6A shows the UV chromatogram of the SCX separation along with the fractionation time windows of the digest. A total of 28 fractions were automatically collected into a fraction collector – the intense peak from 40-100 min was fractionated more frequently (4 min per fraction vs. 10 min per fraction) as a larger amount of peptides was expected to elute within this retention window from a complex proteome digest. While it is certainly possible to collect samples at equal intervals, this work illustrates the flexibility in the off-line SCX separation and fractionation process. The 28 fractions were loaded into the RPLC-UV autosampler and the fractions were individually desalted, quantified and collected into sample vials. The peptide amount in each fraction was calculated based on the calibration curve of the 4-protein digest and shown as a bar diagram in Figure 5.6B. The amount of each fraction was greater than 1 μg and, as a result, no sample pooling from adjacent fractions was required for maximal sample loading to RPLC MS/MS. In fact, all fractions except the last contained more than 2 μg, rendering the possibility of running replicates for each fraction. Note that the minimum volume of residual sample required to be present in the sample vial is about 0.5 μL. With 5 μL sample injection, a minimum total volume of 5.5 μL sample is needed, which means that, to inject 1 μg of sample into the column, a minimum of 1.1 μg of sample is required. For two injections with each at 1 μg, a total of 2.1 μg of sample is needed.

After drying the desalted samples, each sample was redissolved to a proper volume by adding 0.1% fomic acid to make a peptide concentration of 0.2 μg/μL. For nano-LC

ESI-MS/MS analysis, 1 μg of each fraction was loaded to the column. Each fraction was run twice using the same LC elution gradient and acquisition method, but with a precursor ion exclusion list included in the second run. The exclusion list consisted of all the precursor ions identified in the first run with their identification scores of 10 points equal to or above the identity threshold in MASCOT database search (8).

The bar diagram in Figure 5.6C shows the distribution of unique proteins identified from the two runs for each SCX fraction. From this diagram, it can be seen that 22 out of 28 fractions had at least 300 proteins identified in the first LC MS/MS individual runs. Interestingly, despite a maximized sample amount being injected for each SCX fraction, changes in the numbers of proteins identified were still observed. As Figure 5.6C shows, six fractions, three from the front part of the SCX gradient and three from the end part of the gradient, showed much smaller numbers of proteins identified. Note that the MS/MS scan numbers for the first three fractions (1751 for the 1st fraction, 1844 for the 2nd fraction, 1895 for the 3rd fraction, while 1570 for the 3rd last fraction, 1092 for the 2nd last fraction, and 1199 for the last fraction) were approaching the level of the other 22 fractions (average scan number ~2126). The total ion chromatograms of these six fractions show less intense chromatographic peaks compared to other fractions. Moreover, for the first two fractions, each sample was run again by doubling the amount of sample injected into the column (i.e., 2 μg sample loading). However, similar results were obtained (data not shown). These results indicate that the peptides in these six fractions were not ionized as efficiently as others. It turns out that, in the first three fractions, a total of 25 unique phosphopeptides were identified when Phospho (ST) and Phospho (Y) were selected as the variable modifications for the MASCOT search parameters (see Supporting Information 5.1). Searching the MS/MS data matched with phosphopeptides against the reverse proteome sequence database generated no match. In the Supporting Information 5.1, the assignments of phosphorylation sites by MASCOT were manually checked and confirmed. No phosphopeptides could be identified from other fractions, which is consistent with reports by others, i.e., phosphopeptides having smaller positive net charges, compared to non-phosphopeptides, could be present in a larger portion in the first few SCX fractions than the rest of fractions (10). However, phosphopeptides are generally not as efficiently ionized as other non-phosphopeptides.

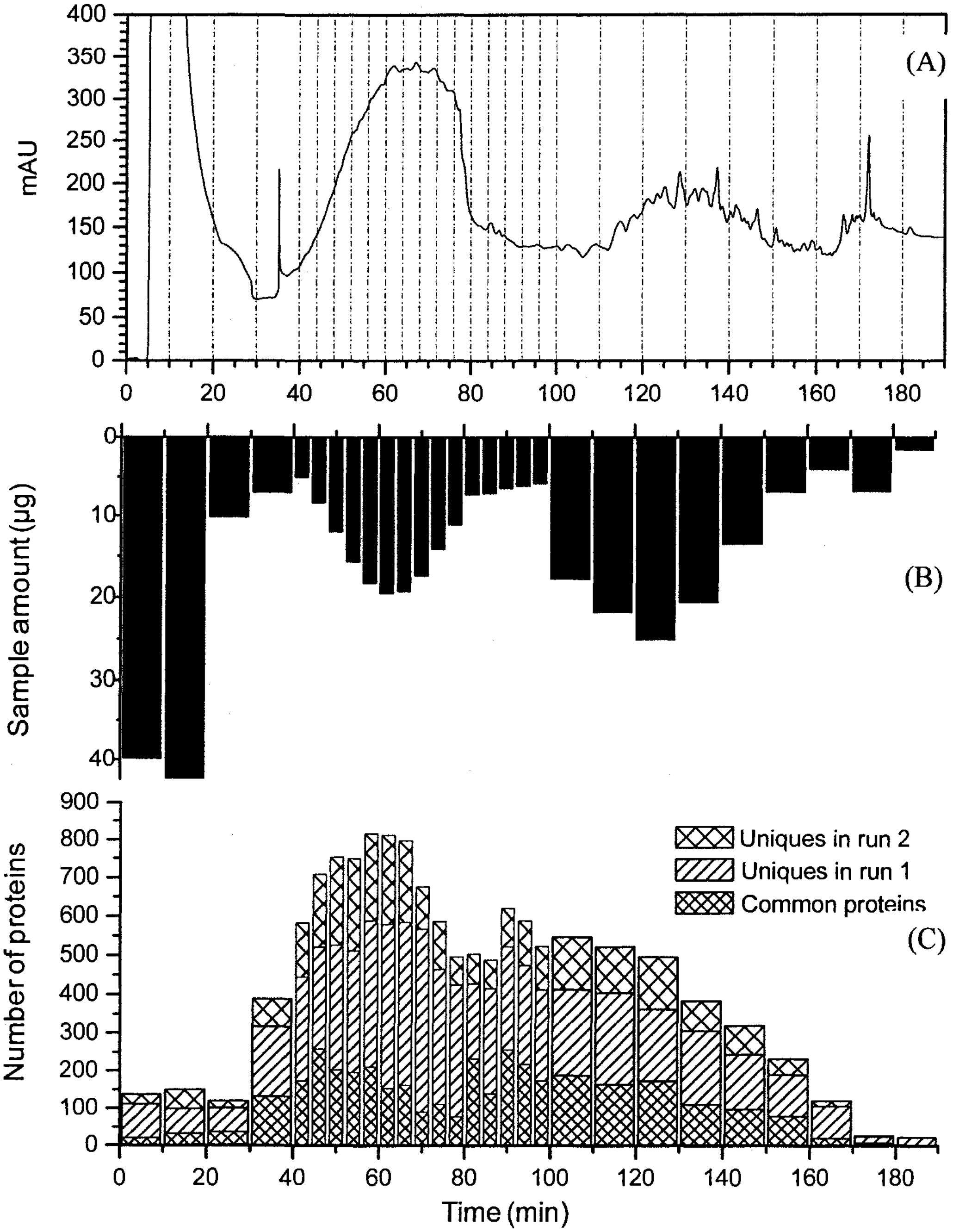


Figure 5.6 (A) SCX separation and fractionation of tryptic peptides of MCF-7, (B) quantification results of the fractions shown in bar-graph, and (C) numbers of unique proteins identified for the 28 SCX fractions. SCX chromatographic conditions: PolySulfoethyl A column, 5 μm, 250 x 2.1 mm I.D.; mobile phase, 10 mM KH_2PO_4 (A) and 10 mM KH_2PO_4, 500 mM KCl (B) with the pH 2.75; flow rate, 0.2 mL/min; detection wavelength, 214 nm; gradient: 0%-0%-20%-60%-100%-100% B in 0-5-100-150-160-190 min; Fractions were collected every 10 min in the first 40 min, every 4 min between 40 and 100 min and every 10 min afterwards.

It appears that the protein number variation shown in Figure 5.6C can be attributed to the difference in peptide complexity in different SCX fractions. The peptide complexity varies according to the type, number and concentration range of the peptides in a sample. For two mixtures with the same quantity, more peptides may be identified from a peptide mixture containing a large number of peptides with similar properties and similar concentrations or ion intensities, compared to a mixture containing a small number of peptides with similar properties but a wide range of concentrations or ion signal intensities (e.g., a few high abundance peptide ions dominate the mass spectra). The exact factors governing the number of proteins identified in each fraction are difficult to ascertain. However, in the future, as the number of proteome samples with varying complexity to be run using this strategy of off-line 2D-LC MS/MS with maximal sample loading increases, we may be able to correlate certain peptide properties, such as retention time in SCX, peptide solution charge state in a given pH and ionization efficiency of peptides, with the peptide detectability. With a better understanding of the relation between the SCX separation and the number of proteins identified, we might be able to minimize the variation of proteins identified from fraction to fraction by applying an ideally balanced salt gradient in which the peptides could be eluted or collected to give similar complexity.

After the first runs of the 28 SCX fractions by RPLC MS/MS, a total of 2362 proteins were identified. For the entire dataset, forward sequence search by MASCOT resulted in 23,552 matches and reversed sequence search found 22 matches. Thus, the false positive rate of peptide matching was estimated to be 2×22/(23552+22) or 0.19%. With a total sample consumption of 28 μg for the analysis, the average protein amount used for identification was about 11.9 ng, assuming 100% conversions of proteins into peptides. Except the last fraction which contained less than 2 μg sample, the other 27 fractions were run again with precursor ion exclusion. A total of 549 additional proteins were identified. In this 2nd run, forward sequence search by MASCOT resulted in 10,920 matches and reversed sequence search found 33 matches. Thus, the false positive rate of peptide matching was estimated to be 2×33/(10920+33) or 0.60%. The increase of false positive rate from 0.19% in the 1st run to 0.60% reflects the decreased quality of peptide matching going from the 1st run to the 2nd run. With the exclusion of high abundance peptide ions identified in the 1st run, one would expect that relatively lower abundance peptide ions were sequenced in the 2nd run. In total, 55 two-hour-LC MS/MS runs of the 28 SCX peptide fractions lead to the identification of 12417 non-redundant peptides assigned to 2911 unique proteins (see Supporting Information 5.2 for the entire list of proteins identified). The average peptides/protein ratio is 4.3. Among these peptides, 89% of them were identified with a MASCOT search confidence level of higher than 99%. The identification confidence level for the rest of the peptides was equal to or higher than 95%. A total of 1123 proteins were identified with a single-peptide match, among which 809 proteins had an identification confidence level of greater than 99%. The overall false positive peptide matching rate from the two runs combined is estimated to be 0.32% by using the target-decoy sequence search strategy.

5.4 Discussion

Several factors need to be considered in deciding whether an off-line or on-line 2D-LC MS/MS strategy is used for shotgun proteome analysis. To offset the inconvenience and potential sample loss associated with the off-line strategy, the off-line method must provide a substantial advantage over other aspects to remain competitive. In this work, we illustrated the importance of sample loading to RPLC MS/MS for increasing the identification efficiency of peptides and proteins. In an on-line method using SCX as the 1st dimension of separation, the sample loading to RPLC cannot be readily optimized. Because of the wide variation of peptide concentrations during the SCX separation, under- or over-loading of peptides eluted from SCX to RPLC MS/MS is unavoidable. If peptide loading is less than the optimal amount, peptide and protein identification efficiency suffers. If RPLC column is overloaded, subsequent runs of peptides eluted from SCX will be seriously compromised. The use of off-line 2D-LC MS/MS overcomes this dilemma.

In off-line 2D-LC MS/MS, the first dimension of separation can be fully optimized using optimal gradient conditions. Peptide fractionation can also be optimized according to the peak elution profile. For example, more frequent fractionation may be done in an elution region where most peptides are eluted whilst other regions may be collected less frequently. Alternatively, fractionation can be carried out at a constant time interval and, after quantification of the individual fractions, the adjacent fractions with each containing less than the optimal amount for RPLC MS/MS can be pooled to produce a sample with a sufficient amount for one or more injections. In any case, the amount of sample injected into RPLC MS/MS can be maximized to achieve the highest peptide and protein identification efficiency. This is critical in the overall shotgun proteome analysis workflow, as the MS/MS instrument must be efficiently utilized to identify the maximum number of peptides and proteins in a given period. Any under- or over-sampling will run the risk of missing identification of many peptides present in a sample.

In addition to the sample amount, the quality of the peptide sample injected to RPLC MS/MS can also have an effect on separation and detection. Desalting of fractions from SCX separation is crucial prior to RPLC-ESI MS/MS, as the salts can affect sample loading and ESI performance. Currently commercial products, such as ZipTip C18 pipette tips packed with a bed of chromatographic medium, or home-made tips with bead packing in pipette tips are designed for this purpose (5, 11). The protocol of using bead-packed pipettes to purify peptides includes four steps: equilibrating the tip, binding of the sample, washing out the salts and other impurities, and eluting the peptides. The whole procedure is generally carried out manually. Using the RPLC-UV system for desalting the peptide samples, as described in this work, offers several advantages, compared to Ziptip or other packed tips. The HPLC column has a much higher loading capacity; a 5.0 × 4.6 mm ID column has a sample loading capacity of up to 1 mg. In addition, the desalting experiment can be performed fully automatically, while quantitative information about each sample can be generated at the same time.

In the RPLC-UV desalting and peptide quantification system, a step gradient can be used to speed up the process so that each cycle generally takes about 6-7 min. A system peak from the rapid switch of the solvents is observed in the chromatogram, but does not affect the quantitative results. The area of the peptide peak is used for quantification with a calibration curve established by using a 4-protein digest standard. BSA digest may be used for calibration; but it gives a systematic error of about 5%. The linear calibration range from 0.5 to 15 μg is sufficient for most shotgun proteome analysis work. In our lab, multiple SCX fractions are routinely desalted and quantified in a fully automated and unattended manner using RPLC-UV. This system can also be used to desalt and quantify a relatively larger amount of peptides such as a whole cell extract digest before SCX fractionation. If the amount of peptides to be quantified is beyond this linear range, sample dilution is required. Alternatively, for a concentrated solution, one can inject a portion of the sample for desalting/quantification and then inject the rest of the sample for desalting. Since RPLC-UV is a non-invasive technique, if a sample of unknown concentration generates an elution peak with its area beyond the linear range for quantification, one has the option of re-injecting the desalted sample after sample volume adjustment.

After examining the effects on sample loading to nano-RPLC ESI-MS/MS on peptide and protein detectability, we determined the optimal sample loading to our system to be about 1 μg. The optimal sample loading should be dependent on various experimental parameters including column size and column length. Considering the great variations of the peptide and protein identification efficiencies with different amounts of sample loading, we suggest that for a given instrumental setup it is useful to determine the optimal sample loading using a complex peptide mixture such as a whole cell digest after desalting and peptide quantification.

Based on the results generated from standard digests and MCF-7 whole cell extract digests, we proposed and applied the off-line 2D-LC MS/MS strategy with maximal sample loading to RPLC MS/MS to generate a proteome profile of MCF-7 cells. SCX separation and fractionation produced 28 fractions from the proteome digest. After RPLC-UV desalting and quantification of the 28 fractions, each fraction was injected into nano-RPLC ESI-MS/MS with a sample loading of 1 μg. An additional MS/MS run was performed on 27 of the 28 fractions that contained more than 2 μg of peptides. From the 55 2 h runs, a total of 16,924 proteins were identified with an average of 308 proteins per run. However, many proteins were identified more than once. In all, a total of 2911 unique proteins were identified. It is clear that additional strategies need to be developed to maximize the performance of RPLC MS/MS to increase the number of unique proteins identified. For example, we will optimize the precursor ion exclusion (PIE) method to

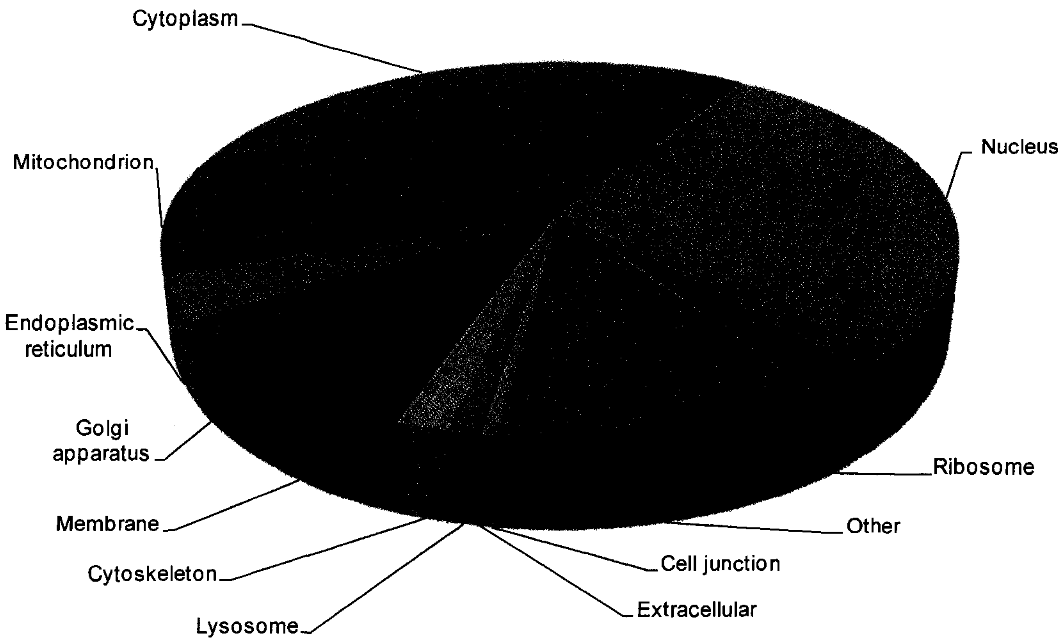


Figure 5.7 Distribution of the cellular components of the proteins identified from the MCF-7 cell protein extracts.

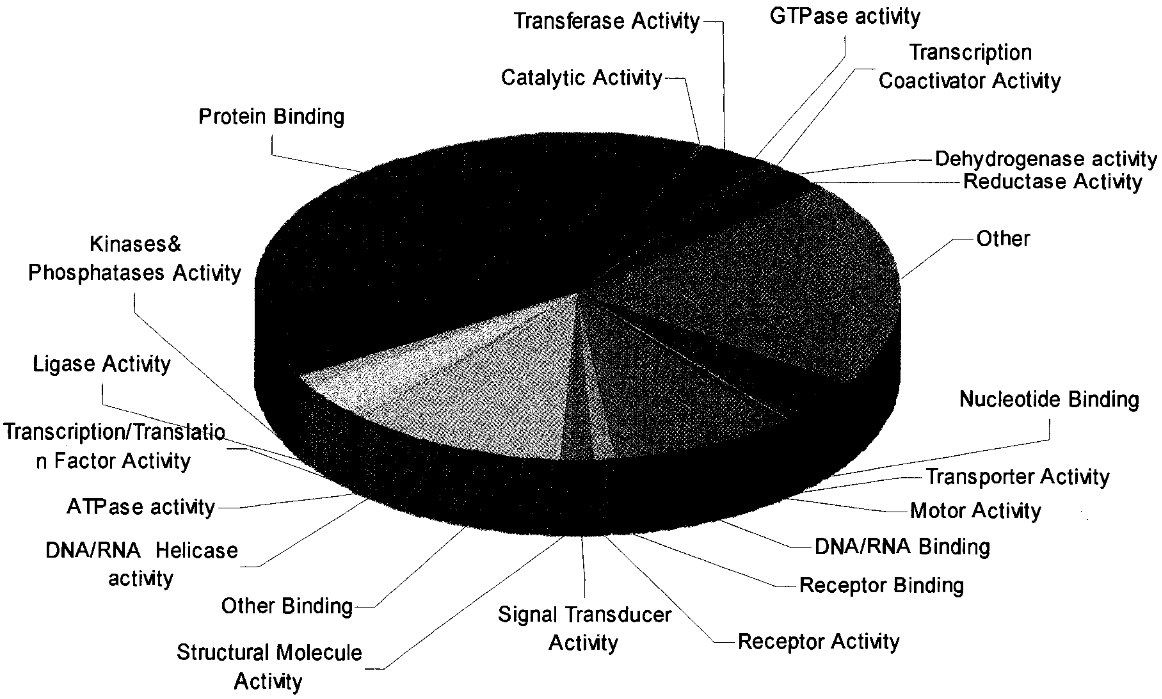


Figure 5.8 Distribution of the molecular functions of the proteins identified from the MCF-7 cell protein extracts.

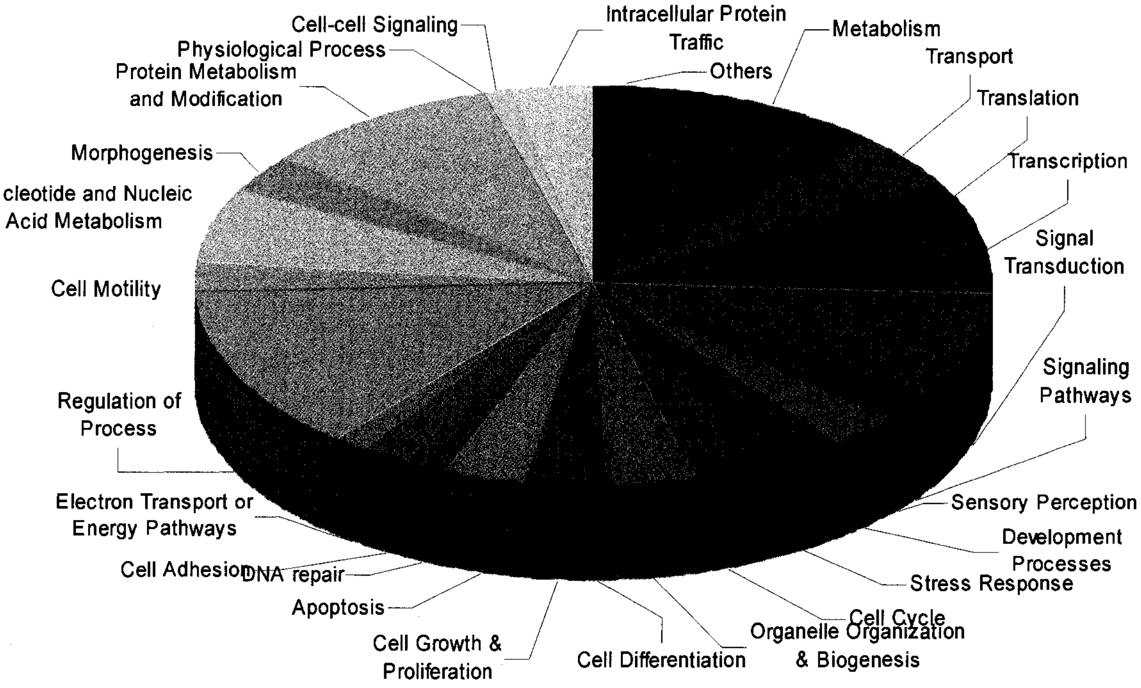


Figure 5.9 Distribution of the biological processes of the proteins identified from the MCF-7 cell protein extracts.

enhance the exclusion power to facilitate the identification of low abundance peptides. With off-line 2D-LC, it should be possible to start with running one SCX fraction in nano-RPLC MS/MS, followed by running the next fraction with the exclusion of precursor ions of the peptides already identified in the previous fraction. This rolling exclusion process should improve the detectability of low abundance peptides that are only present in one SCX fraction, not in multiple fractions.

Using the present off-line 2D-LC MS/MS method, the proteome profile of MCF-7 cells generated already represents one of the most comprehensive profiles reported in the literature (12-17). The identified 2911 proteins can be categorized in three ways according to their cellular component (Figure 5.7), molecular function (Figure 5.8) and biological process (Figure 5.9) after the extraction of GO term information from ExPASy. Since some of the proteins have not been assigned to any GO terms based on the information supplied in the Swiss-Prot database, only the proteins of which their GO terms were validated in the Swiss-Prot database are classified. The pie diagrams for the cellular component, molecular function and biological process classifications cover only 85%, 71% and 60% of the total proteins, respectively.

As shown in Figure 5.7, the largest proportions of proteins had a subcellular localization in the cytoplasm (32%) and nucleus (30%). Proteins located in cell membrane and mitochondria also take up relatively high percentages (9% and 8% respectively) (Figure 5.7). A lower percentage of cell membrane portion was observed which might be due to insufficient protein solubilization in the protein purification step for the whole cell lysate. A large population of proteins were found to be located at several different organelle membranes, such as nucleus, mitochondria, endoplasmic reticulum, and Golgi apparatus (Supporting Information 5.2). For future work to generate more comprehensive profile of the MCF-7 cells, a sequential protein solubilization and digestion method reported previously (18), combined with the off-line 2D-LC MS/MS strategy could be performed.

As demonstrated in Figure 5.8, binding activities, particularly protein binding and DNA/RNA binding activities, take up the largest proportions (35% and 7% respectively) in terms of protein molecular functions. Many of the proteins perform kinase and phosphatase activities, transcription or translation factor activities. With regard to biological processes, the majority of proteins involve regulation of various processes (14%), metabolism (9%) and signal transduction (8%) (Figure 5.9). Other biological processes which play important roles in the onset and development of cancer include 95 proteins involved in apoptosis, defectiveness of which has been implicated in a variety of diseases including cancer; 66 in cell growth and proliferation; 53 in cell adhesion and 42 in DNA repair.

One of the major motivations for proteomic technology development is to facilitate the comprehensive proteome profiling analysis and search for proteins that may serve as biomarkers or therapeutic targets for early disease diagnosis, prognosis and treatment. While cancer still remains one of the major public health challenges, the information to be obtained from improved proteomic technologies may provide insights into the molecular complexity of the disease progress, and thus significantly accelerate cancer research progress and increase rates of survival. A common hindrance for proteomic biomarker discovery work is its limited power of detecting low-abundance cancer markers. Off-line 2D-LC MS/MS strategy offers the potential to detect relatively low abundance proteins including potential biomarkers.

Among the 2911 unique proteins identified from the MCF 7 cells in this work, 260 of them have been reported as potential biomarkers which are normally differentially expressed between normal and malignant cells and tissues (19-21). These proteins are listed in Supplemental Table S5.3 along with their molecular weight, peptide sequences identified, peptide MASCOT scores and identity thresholds, identification confidence level, and their GO terms. 93% (242 unique proteins) of the potential biomarkers were identified in our work with a confidence level of higher than 99%, except 18 proteins identified with a confidence level of greater than 95%. Some of these proteins, including epidermal growth factor receptor (EGFR) were reported to be of low abundance in normal cells. 13 of the biomarkers detected have a citation rate of higher than 500 according to a candidate cancer biomarker list recently reported by Anderson *et al.* in 2006 (19).

These potential breast cancer biomarkers possess different molecular functions and are involved in various biological processes. A few cell cycle associated biomarkers, for example, antigen Ki-67, cyclin-dependent kinase inhibitor p27 (p27kip1) and G1/S-specific cyclin D1 were identified with high confidence. The antigen Ki-67, identified with 18 unique peptides, is a protein strictly restricted with cell proliferation. It can be used in immunostaining as an excellent marker to determine the growth fraction of a given cell population (22). Carcinoma of the breast is one of the best studied examples in the context of correlating the fraction of Ki-67-positive tumor cells with the clinical course of cancer (23). The cell cycle regulator protein p27kip1 possesses the function of stopping or slowing down the cell division cycle. Several studies implicated that abnormally low levels of this protein in tumor cells were associated with poor prognosis of breast cancer, especially among women with hormone-receptor-positive tumors which depend on estrogen and progesterone to grow (24, 25).

Cell adhesion proteins are typically transmembrane receptors located at cell surfaces and bind to another cell, surface or extracellular matrix. The expression of a few of the identified cell adhesion molecules, such as CD44 and E-cadherin, has also been extensively studied in breast cancer research. The overexpression of the membrane glycoprotein CD44 has been reported to be associated with an antagonistic outcome in several breast cancer studies (26-28). The loss of E-cadherin expression in epithelial tissues is consistently observed in breast cancer, which decreases the strength of cellular adhesion and may allow cancer cells to cross the basement membrane and invade surrounding tissues (29, 30). Biomarker 14-3-3 protein sigma was identified with 9 unique peptides and a protein score of 703. It is encoded from a p53-regulated gene which undergoes frequent epigenetic silencing in various cancers, including carcinoma of the breast (31, 32). Two apoptosis regulator proteins, caspase 3 and caspase 6 were also listed. The failure of caspases expresssion is one of the main contributions to breast tumor development (33). Heat shock proteins HSP 27, 60, 70 and 90 were all identified in our approach with multiple peptide hits of high confidence level. The production of high levels of several HSPs has been found in breast cancer cells, which may both augment the aggressiveness of these tumors and make them more resistant to treatment (34).

Lastly, several growth factors and receptors such as insulin-like growth factor I and II receptors (IGF-IR, IGF-IIR), were identified with >99% confidence level. The insulin-like growth factors (IGFs) play an important role in regulating cell proliferation, differentiation, apoptosis, and transformation. IGF-IR was overexpressed in certain cancers and its overexpression is associated with aggressive tumors (35-37). Epidermal growth factor receptor (EGFR), was also listed although the identity confidence level lower at >95%. The overexpression of this membrane tyrosine kinase has been connected with adverse prognosis in breast cancer (38). EGFR is also a marker that has been used clinically (19).

5.5 Conclusions

RPLC-UV with step solvent gradient has been developed as a means of quantifying peptides as well as getting rid of salts and other impurities in a peptide sample. This method allows us to optimize the amount of sample injected into RPLC MS/MS for peptide sequencing in shotgun proteome analysis, which was found to be critical for maximizing the number of peptides and proteins identified in a 2-h run. An off-line SCX/RP 2D-LC MS/MS strategy with maximized sample injection into RPLC MS/MS was proposed and demonstrated to be useful for generating the proteome profile of MCF-7 cells. This strategy is also universal and has already been applied to many other proteome profiling applications that will be reported in the future.

5.6 References Cited

1. Fournier, M., Gilmore, J., Martin-Brown, S., and Washburn, M. (2007) Multidimensional separations-based shotgun proteomics. *Chem. Rev.* 107, 3654-3686.

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

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

4. Takahashi, N., Takahashi, Y., and Putnam, F. W. (1983) Two-dimensional high-performance liquid chromatography and chemical modification in the strategy of sequence analysis Complete amino acid sequence of the lambda light chain of human immunoglobulin D. *J. Chromatogr.* 266, 511-522.

5. Ballif, B., Villen, J., Beausoleil, S., Schwartz, D., and Gygi, S. (2004) Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell Proteomics* 3, 1093-1101.

6. DeSouza, L., Diehl, G., Rodrigues, M., Guo, J., Romaschin, A., Colgan, T., and Siu, K. (2005) Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and clCAT with multidimensional liquid chromatography and tandem mass spectrometry. *J. Proteome Res.* 4, 377-386.

7. Peng, J., Elias, J., Thoreen, C., Licklider, L., and Gygi, S. (2003) 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.

8. Wang, N., and Li, L. (2008) Exploring the Precursor Ion Exclusion Feature of LC-ESI Quadrupole Time-of-Flight MS for Improving Protein Identification in Shotgun Proteome Analysis. *Anal. Chem.* in press.

9. Elias, J., and Gygi, S. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207-214.

10. Beausoleil, S., Jedrychowski, M., Schwartz, D., Elias, J., Villen, J., Li, J., Cohn, M., Cantley, L., and Gygi, S. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12130-12135.

11. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* 75, 663-670.

12. 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.

13. Brown, K. J., and Fenselau, C. (2004) Investigation of Doxorubicin Resistance in MCF-7 Breast Cancer Cells Using Shot-Gun Comparative Proteomics with Proteolytic 18O Labeling. *J. Proteome Res.* 3, 455-462.

14. 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.

15. An, Y. M., Fu, Z. M., Gutierrez, P., and Fenselau, C. (2005) Solution isoelectric focusing for peptide analysis: Comparative investigation of an insoluble nuclear protein fraction. *J. Proteome Res.* 4, 2126-2132.

16. Hardouin, J., Canelle, L., Vlieghe, C., Lasserre, J.-P., Caron, M., and Joubert-Caron, R. (2006) Proteomic analysis of the MCF7 breast cancer cell line. *Cancer Genomics & Proteomics* 3, 355-368.

17. Hou, W., Ethier, M., Smith, J. C., Sheng, Y., and Figeys, D. (2007) Multiplexed Proteomic Reactor for the Processing of Proteomic Samples. *Anal. Chem.* 79, 39-44.

18. Wang, N., MacKenzie, L., De Souza, A. G., Zhong, H. Y., Goss, G., and Li, L. (2007) Proteome profile of cytosolic component of zebrafish liver generated by LC-ESI MS/MS combined with trypsin digestion and microwave-assisted acid hydrolysis. *J. Proteome Res.* 6, 263-272.

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

20. Ross, J. S., Linette, G. P., Stec, J., Clork, E., Ayers, M., Leschly, N., Symmons, W. F., Nortobagyi, G. N., and Pusztai, L. (2003) Breast cancer biomarkers and molecular medicine. *Expert Rev. Mol. Diagn.* 3, 573-585.

21. 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, 169-188.

22. Gasparini, G., Pozza, F., Meli, S., Reitano, M., Santini, G., and Bevilacqua, P. (1991) Breast-Cancer Cell-Kinetics - Immunocytochemical Determination of Growth Fractions by Monoclonal-Antibody Ki-67 and Correlation with Flow Cytometric S-Phase and with Some Features of Tumor Aggressiveness. *Anticancer. Res.* 11, 2015-2021.

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

24. Barbareschi, M., van Tinteren, H., Mauri, F. A., Veronese, S., Peterse, H., Maisonneuve, P., Caffo, O., Scaioli, M., Doglioni, C., Galligioni, E., Dalla Palma, P., and Michalides, R. (2000) P27(Kip1) expression in breast carcinomas: An immunohistochemical study on 512 patients with long-term follow-up. *Int. J. Cancer* 89, 236-241.

25. Porter, P. L., Barlow, W. E., Yeh, I.-T., Lin, M. G., Yuan, X. P., and Elizabeth Donato, G. W. S., Charles L. Shapiro, James N. Ingle, Charles M. Haskell, Kathy S. Albain, James M. Roberts, Robert B. Livingston, and Daniel F. Hayes (2006) p27Kip1 and Cyclin E Expression and Breast Cancer Survival After Treatment With Adjuvant Chemotherapy. *J. Natl. Cancer Inst.* 98, 1723-1731.

26. Guriec, N., Gairard, B., Marcellin, L., Wilk, A., Calderoli, H., Renaud, R., Bergerat, J. P., and Oberling, F. (1997) CD44 isoforms with exon v6 and metastasis of primary N0M0 breast carcinomas. *Breast Cancer Res. Treat.* 44, 261-268.

27. Schumacher, U., Horny, H. P., Horst, H. A., Herrlich, P., and Kaiserling, E. (1996) A CD44 variant exon 6 epitope as a prognostic indicator in breast cancer. *Eur. J. Surg. Oncol.* 22, 259-261.

28. Morris, S. F., O'Hanlon, D. M., McLaughlin, R., McHale, T., Connolly, G. E., and Given, H. F. (2001) The prognostic significance of CD44s and CD44v6 expression in stage two breast carcinoma: an immunohistochemical study. *Eur. J. Surg. Oncol.* 27, 527-531.

29. Charpin, C., Garcia, S., Bonnier, P., Martini, F., Andrac, L., Choux, R., Lavaut, M. N., and Allasia, C. (1998) Reduced E-cadherin immunohistochemical expression in node-negative breast carcinomas correlates with 10-year survival. *Am. J. Clin. Pathol.* 109, 431-438.

30. Yoshida, R., Kimura, N., Harada, Y., and Ohuchi, N. (2001) The loss of E-cadherin, alpha- and beta-catenin expression is associated with metastasis and poor prognosis in invasive breast cancer. *Int. J. Oncol.* 18, 513-520.

31. Mhawech, P. (2005) 14-3-3 proteins - an update. *Cell Res.* 15, 228-236.

32. Lodygin, D., and Hermeking, H. (2005) The role of epigenetic inactivation of 14-3-3 sigma in human cancer. *Cell Res.* 15, 237-246.

33. Vakkala, M., Paakko, P., and Soini, Y. (1999) Expression of caspases 3, 6 and 8 is increased in parallel with apoptosis and histological aggressiveness of the breast lesion. *Br. J. Cancer* 81, 592-599.

34. Fuqua, S. A. W., Oesterreich, S., Hilsenbeck, S. G., Hoff, D. D., Eckardt, J., and Osborne, C. K. (1994) Heat shock proteins and drug resistance. *Breast Cancer Res. Treat.* 32, 67-71.

35. Bonneterre, J., Peyrat, J. P., Beuscart, R., and Demaille, A. (1990) Prognostic-Significance of Insulin-Like Growth Factor-I Receptors in Human Breast-Cancer. *Cancer Res.* 50, 6931-6935.

36. Oh, Y. (1998) IGF-independent regulation of breast cancer growth by IGF binding proteins. *Breast Cancer Res. Treat.* 47, 283-293.

37. Yu, H., and Rohan, T. (2000) Role of the insulin-like growth factor family in cancer development and progression. *J. Natl. Cancer Inst.* 92, 1472-1489.

38. Castellani, R., Visscher, D. W., Wykes, S., Sarkar, F. H., and Crissman, J. D. (1994) Interaction of Transforming Growth-Factor-Alpha and Epidermal Growth-Factor Receptor in Breast-Carcinoma - an Immunohistologic Study. *Cancer* 73, 344-349.

Chapter 6

Comprehensive Proteome Profile of Zebrafish Liver Membrane Fraction Generated by Off-line Two-dimensional Liquid Chromatography QTOF Mass Spectrometry*

6.1 Introduction

Zebrafish is widely used as a model system for studying many biological processes including toxicology studies of organic pollutants.[1-4] The ultimate goal of our research is to develop a proteomic approach to investigate how specific toxicants affect the biology of zebrafish to understand the toxicity mechanisms, and to use the fish proteome as a potential indictor or biomarker for gauging the level of exposure to toxicants. These studies will be carried out by quantitative proteome comparison of zebrafish grown under various aquatic conditions. To reveal subtle changes in the zebrafish proteome, ideally the entire proteome should be examined. Unfortunately current proteome analysis techniques cannot cover the entire proteome and only partial lists of proteins have been reported for zebrafish.[5-12] We wish to develop improved sample handling and mass spectrometric techniques to profile as many proteins as possible. In Chapter 3, a method of protein sub-fractionation combined with off-line two-dimensional (2D) liquid chromatography (LC) quadrupole time-of-flight mass spectrometry (QTOF MS) was described to analyze the cytosolic component of the zebrafish liver. With the recent development in QTOF MS, namely the introduction of the optimal precursor ion extraction (PIE) strategy as described in Chapter 4 and the maximal LC-MS sample loading technique as described in Chapter 5, we expected that a greater number of proteins would be identified from the zebrafish. In this chapter, we focus on the identification of a comprehensive list of proteins from the liver membrane component.

Analyzing the proteome of the membrane component is much more challenging, compared to other components, as it contains many hydrophobic proteins at low

*Fang Wu partially contributed to this work in sample preparation and data acquisition.

concentrations. Hydrophobic proteins are difficult to analyze.[13] They can be easily lost to the sample container walls by adsorption during sample workup. They are difficult to digest by enzymes, because the cleavage sites of a protein may not be accessible by an enzyme in a protein solution. The peptides generated may be too hydrophobic for separation and ionization, resulting in low efficiency of detection. Proteomic technology development for analyzing membrane proteome has remained a very active field of research.[13] Several methods have been reported to analyze membrane proteomes of various biological sources with varying degrees of success.[13] Shotgun proteome analysis combined with surfactant-assisted protein solubilization and digestion has been demonstrated to be particularly powerful in analyzing membrane proteome.[14, 15] Our lab has been involved in developing the surfactant-assisted shotgun proteome method using the strong surfactant, SDS, as a solubilizing reagent.[15, 16] SDS is perhaps the strongest surfactant which is useful in solubilizing very hydrophobic proteins such as integral membrane proteins. However, SDS can also cause severe interference with LC separation and MS detection. Thus, prior to LC-MS analysis of a proteome digest containing SDS, removal of SDS is required. This can be accomplished by using a strong-cation exchange column in a LC system.[15]

In this work, we report the proteome analysis of membrane component of the zebrafish liver by using a shotgun method involving protein fractionation/digestion and off-line 2D-LC ESI QTOF MS analysis of the digests.

6.2 Experimental

6.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide, phenylmethylsulfonyl fluoride (PMSF), trifluoroacetic acid (TFA), heparin, sodium bicarbonate, sodium chloride (NaCl), LC-MS grade formic acid and SDS were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, LC-MS grade water, acetone, methanol (MeOH), and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). Tricaine methane sulfonate (TMS) was obtained from Aqualife. The BCA assay kit was from Pierce, Rockford, IL. Cellular fractionation was accomplished using a compartmental protein extraction kit (Kit K3013010 from Biochain Institute, Inc. Hayward, CA).

6.2.2 Zebrafish Liver Membrane Protein Extraction.

The zebrafish were treated following the same protocol as described in Chapter 3. Briefly, 6 strain A/B zebrafish were anesthetized with TMS (0.15 mg/mL), and perfused with heparinized phosphate buffered saline (PBS) (25 IU/mL) via caudal puncture. The livers were excised and placed in a 1.5 mL flat-bottomed Eppendorf (Flex-Tube, Eppendorf) containing lysis buffer (Buffer C + protease inhibitor cocktail from compartmental extraction kit + 2 M added PMSF), and placed on ice. Livers were combined and homogenized using a pestle (30 strokes minimum). This mixture was sonicated in a bath of ice water, using four 10 s sonication bursts, with a 1 min rest between sonications. The compartmental protein extraction kit was used to isolate the membrane proteins from cytosolic, nuclear, and cytoskeletal components via differential centrifugation according to the kit's instructions (see Figure 6.1). The concentration of the membrane component was measured by BCA protein assay. Components were aliquoted and stored at -80 °C for later analysis.

6.2.3 Acetone Precipitation and In-Solution Digestion.

The membrane fraction was first subjected to standard reduction of the disulfide bonds and alkylation. Briefly, 1.5 mg membrane protein component was reduced with 10 μL of 450 mM DTT for 1 h at 37 °C. Free thiol groups were blocked by reaction with a double volume of 450 mM iodoacetamide for 1 h at room temperature in the dark. The extracts were then acetone-precipitated to remove detergent, unreacted DTT, and iodoacetamide. Acetone, precooled to -80 °C, was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The mixture was

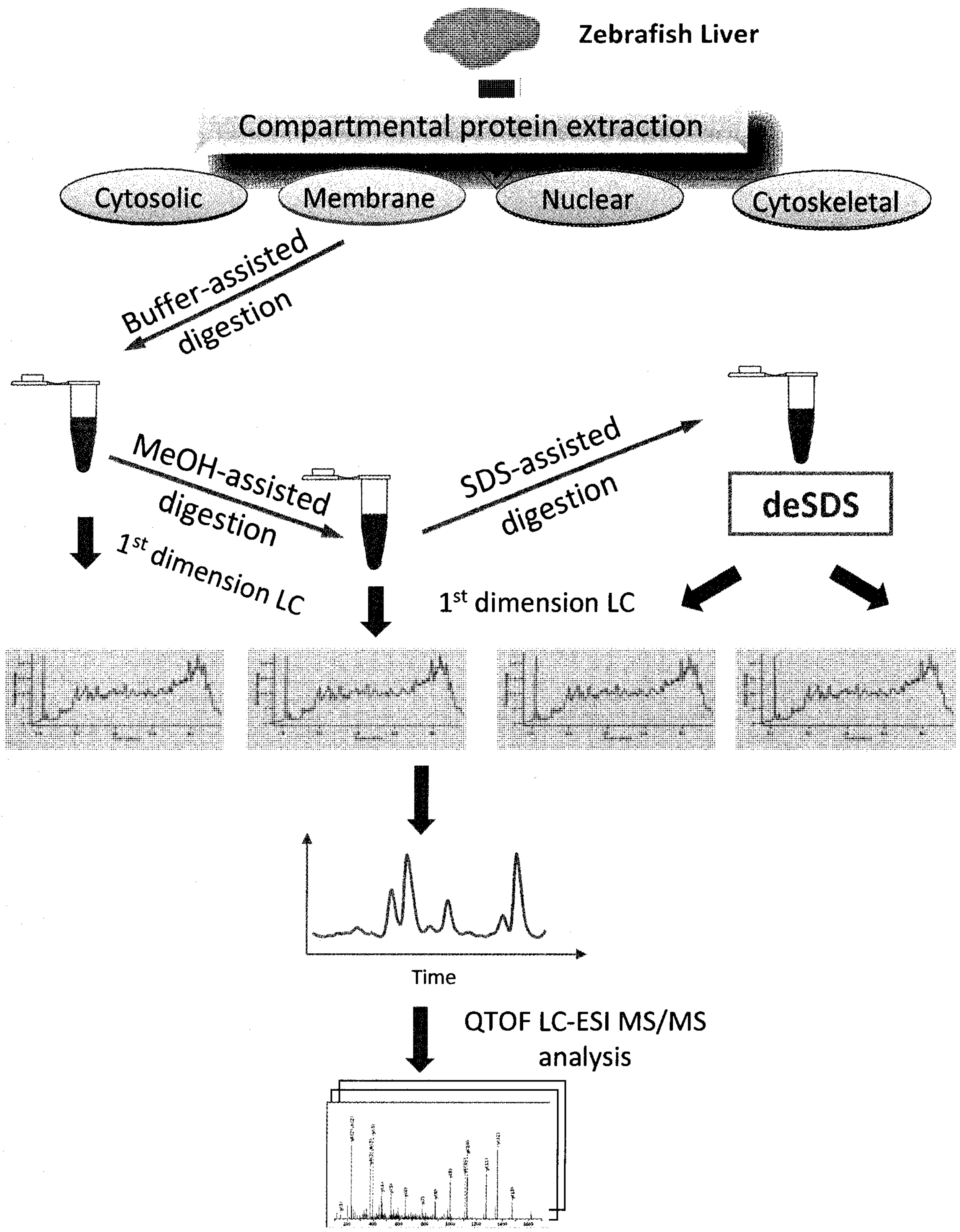


Figure 6.1 Experimental workflow.

kept at -20 °C overnight and centrifuged (14 000 rpm, 10 min, 4 °C). The supernatant was decanted and properly disposed. Acetone was evaporated at room temperature.

The membrane protein pellet was then subjected to the sequential solubilization and digestion protocol described in Chapter 3 (see Figure 6.1). Firstly, ammonium bicarbonate (50 mM, pH 8.0) was used to solubilize the membrane protein pellet with intermittent vortexing applied. The vial was then centrifuged at 14 000 rpm for 5 min at 4 °C. Trypsin solution was added into the supernatant for an enzyme/protein ratio of 1:45, and digestion was conducted at 37 °C overnight.

The pellet remained following ammonium bicarbonate treatment was resuspended in 60% MeOH, with sufficient vortexing. Trypsin was added at an enzyme/protein ratio of 1:30 and the solution was incubated at 37 °C for overnight. The solution was then centrifuged (14 000 rpm, 5 min, 4 °C) and the supernatant was transferred to a different vial. MeOH in the supernatant was evaporated by SpeedVac (Thermo Savant, Milford, MA). After this MeOH-assisted solubilization and digestion, the undissolved pellet was redissolved in 40 μL of 1% SDS (SDS-assisted solubilization), followed by 20-fold dilution. Trypsin was added to achieve a final enzyme/protein ratio of 1:40. The sample was incubated at 37 °C for 2 overnights with 10% (of amount added for the 1st overnight) more fresh trypsin added before the 2nd overnight digestion. A very small amount of protein pellet was still left after these treatments and it was stored at -80 °C for further microwave-assisted acid hydrolysis analysis.

6.2.4 Cation Exchange Chromatography

Peptide mixtures were separated by strong cation exchange (SCX) chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 2.1 × 250 mm highly hydrophilic polysulfoethyl A column (particle size of 5 μm diameter and 300 Å pore, PolyLC Inc., Columbia, MD). Two independent runs of SCX were carried on the sample. In both runs, the gradient elution was performed with mobile phases A (10 mM KH_2PO_4, pH 2.76) and B (10 mM KH_2PO_4, pH 2.76, 500 mM KCl). However, the gradient profiles were different. In the 1st SCX run, the peptides were eluted using linear gradients (0 min: 0% B, 5 min: 0% B, 6 min: 6% B, 29 min: 40% B, 34 min: 60% B, 38 min: 100%

B, 43 min: 100% B.) at 0.20 mL/min, with collection of 1 min fractions. In the 2nd run, the peptides were eluted using a slightly different linear gradient (0 min: 0% B, 7 min: 0% B, 8 min: 6% B, 36 min: 28% B, 44 min: 40% B, 49 min: 60% B, 53 min: 100% B, 58 min: 100% B). In each run, a total of 26 fractions were collected based on the chromatography UV absorption signals recorded at 214 nm.

6.2.5 Peptide Desalting and Quantification by RPLC.

Desalting and quantification were carried out in an Agilent 1100 HPLC system (Palo Alto, CA) using a newly developed method described in Chapter 5. Desalting of tryptic peptides was performed on a 4.6 mm × 5 cm Polaris C18 A column with a particle size of 3 μm and 300 Å pore (Varian, USA). After loading of the peptide sample, the column was flushed with mobile phase A (0.1% TFA in water) and the salts were effectively removed. Subsequently, the concentration of mobile phase B (0.1% TFA in ACN) in the mobile phase was step-wise increased to 85% to ensure complete elution of the peptide fractions from the column. During the peptide elution process, a chromatographic peak was produced and based on the peak area the amount of peptides was determined. Four-protein digests of various amounts were used as standards for the generation of a linear calibration between the peak area and the injected peptide amount. The calibration curve was generated as $y=430.04x-269.56$, where y refers to the peak area of the peptide sample, and x refers to the peptide amount analyzed.

6.2.6 LC-ESI QTOF MS and MS/MS Analysis.

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, the desalted and quantified digests were concentrated using a SpeedVac (Thermo Savant, Milford, MA) to ~5 μL and reconstituted to a specific concentration using 0.1% formic acid. Then 1-μg of digests was injected onto a 75 μm × 100 mm Atlantis dC18 column with a particle size of 3 μm (Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120 min gradient (2-6% Solvent B for 2 min, 6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min) and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 0.8 s each. For both dynamic and precursor ion mass exclusion, a mass tolerance window of 80 mDa was applied. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants was infused at a flow rate of 300 nL/min, and a 1 s MS scan was acquired every 1 min throughout the run.

Peptide precursor ion exclusion (PIE) strategy was applied to exclude relatively high-abundance peptides identified from the adjacent two SCX fractions to enable additional and less abundance peptides to be analyzed and identified. An exclusion list was generated based on MASCOT (Matrix Science, London, U.K.) searching results of peptides with a score 10 points equal to or higher than the identification threshold.

6.2.7 Protein Database Search.

Raw search data were lock-mass-corrected, and converted to peak list files by ProteinLynx Global Server 2.2.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program. Database searching was restricted to *Danio rerio* (zebrafish) in the NCBI database. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, ±30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, *N*-Acytyl (Protein), oxidation (M), Pyro_glu (N-term Q), Pyro_glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides with scores lower than the MASCOT threshold score for identity were then deleted from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

6.2.8 Transmembrane Domain Prediction.

The transmembrane domains for all the proteins identified in all four fractions were predicted using the TMHMM server 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0).

6.3 Results and Discussion

Figure 6.1 shows the workflow for the analysis of the membrane component of the zebrafish liver. After acetone precipitation of the proteins, the NH_4HCO_3 buffer was used to dissolve the proteins. The un-dissolved protein pellet was subjected to methanol-assisted solubilization. It was observed that a large fraction of the pellet was not dissolved. The strong surfactant, SDS (1%), was then added to the pellet for further dissolution. A small amount of pellet still remained. As Figure 6.1 shows, each protein fraction from the above described sequential solubilization process was digested by trypsin except the final pellet not dissolvable even in SDS. This pellet was subjected to microwave-assisted acid hydrolysis using 25% TFA. We found that most proteins (about 2/3 of the total pellet) were dissolved in the SDS solution for this membrane component. It is interesting to note that, in the case of cytosolic component, the largest fraction of the protein pellet was dissolved in the buffer solution. This observation makes sense as the membrane component contains more membrane proteins than the cytosolic component, hence less likely being solubilized in the buffer solution.

Because a large amount of proteins were dissolved in SDS, this protein fraction was expected to have a more complex protein composition than the other protein fractions. As a consequence, the digest from the SDS fraction should contain a much greater number of peptides than the other fractions. Thus, in this work, the SDS digest was subjected to two independent SCX separations. The difference between the two SCX runs was on the gradient conditions used. The rational behind this experimental design was that different gradients might separate the peptides with different retention properties and, thus, the peptide compositions of the SCX fractions collected in the 1st SCX run would be somewhat different from the corresponding fractions from the 2nd SCX run. This changes the concentration ranks of peptides in the SCX fractions, compared to merely running replicates of SCX under the identical conditions. In the reversed phase

(RP) LC-ESI MS/MS analysis of the individual SCX fractions, the peptide composition in the SCX fraction affects the identification results in many ways. Peptide ions are sequenced by MS/MS according to their ion intensity ranks within a MS scan. One scenario is that a peptide in a SCX fraction in the 1st SCX run may not be sequenced because of its low rank in a MS scan at the retention time where the peptide is co-eluted with other peptides. However, in a SCX fraction from the 2nd SCX run, this peptide, compared to other co-eluting peptides, may be present in a relatively higher concentration or less suppressed during the ionization, resulting in a higher rank in the MS scan. As a consequence, this peptide will be sequenced by MS/MS. Thus, this dual-SCX strategy may increase the number of proteins identified from the SDS digest.

Figure 6.2 shows the SCX UV chromatograms obtained under two different gradient conditions as described in the Experimental section. Because of the complexity of the sample, the chromatographic peaks are not well resolved. Judging from the differences in some of the fine features of the two chromatograms, it appears that these two chromatograms are somewhat different. A total of 40 fractions were collected from the first SCX run and 48 fractions from the second SCX run. Each SCX fraction was then desalted and the peptide concentration was determined by using the RPLC-UV desalting and quantitation system as described in Chapter 6. The low abundance fractions were pooled. In the end, a total of 52 fractions were produced for LC MS/MS from both the 1st SCX run and the 2nd SCX.

Figure 6.3 shows two representative MS spectra where a common peptide was identified from the SCX fractions collected from two different SCX runs. The MS scan spectrum in Figure 6.3A was from the 1st SCX run and the one shown in Figure 6.3B was from the 2nd SCX run. It is clear that the two spectra are different in spectral patterns, indicating that the peptide compositions are different from the two samples. This example illustrates that the dual-SCX fractionation experiments can indeed generate peptide mixtures which are different in peptide compositions in corresponding SCX fractions.

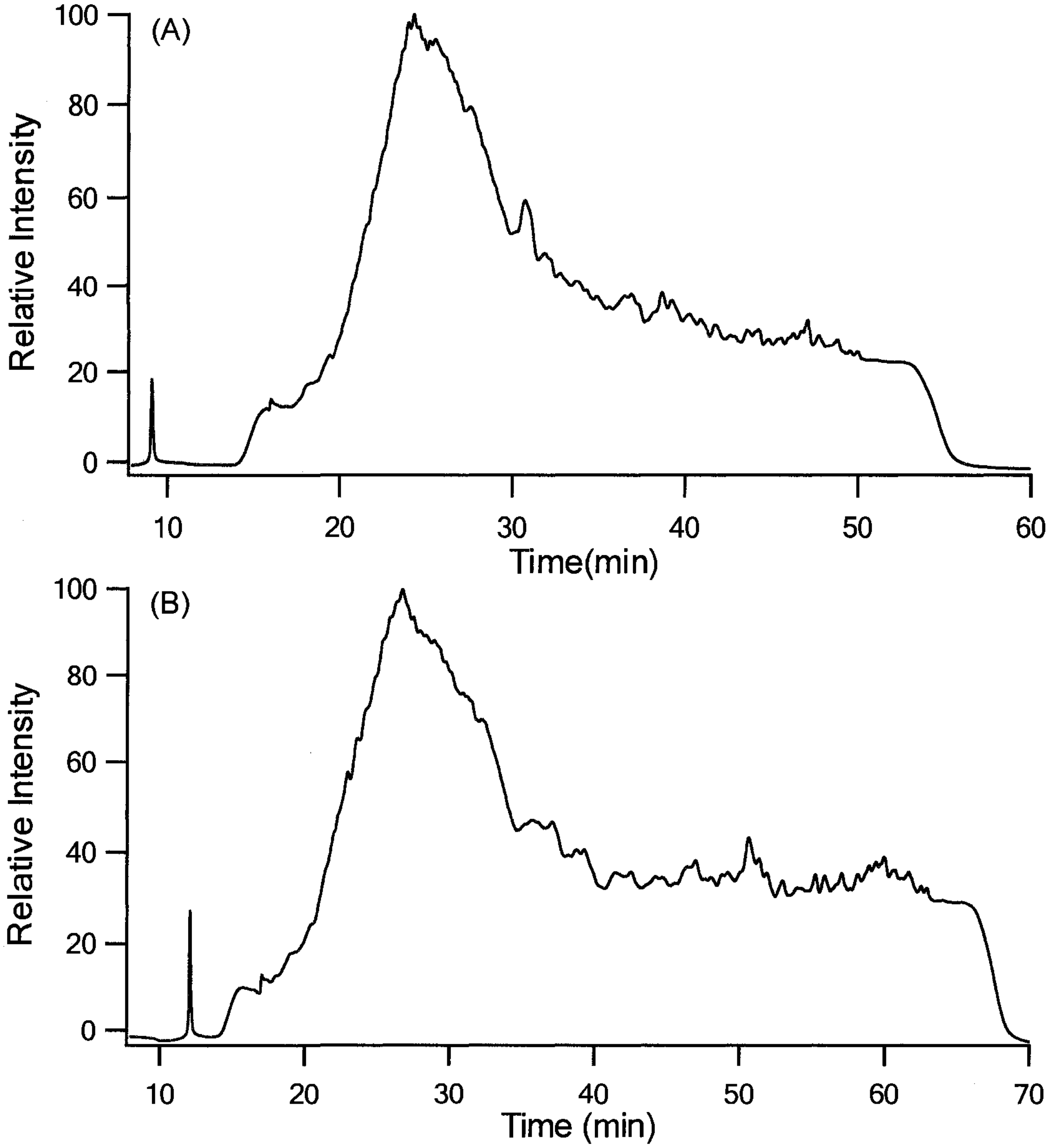


Figure 6.2 SCX UV chromatograms obtained under two different gradient conditions.

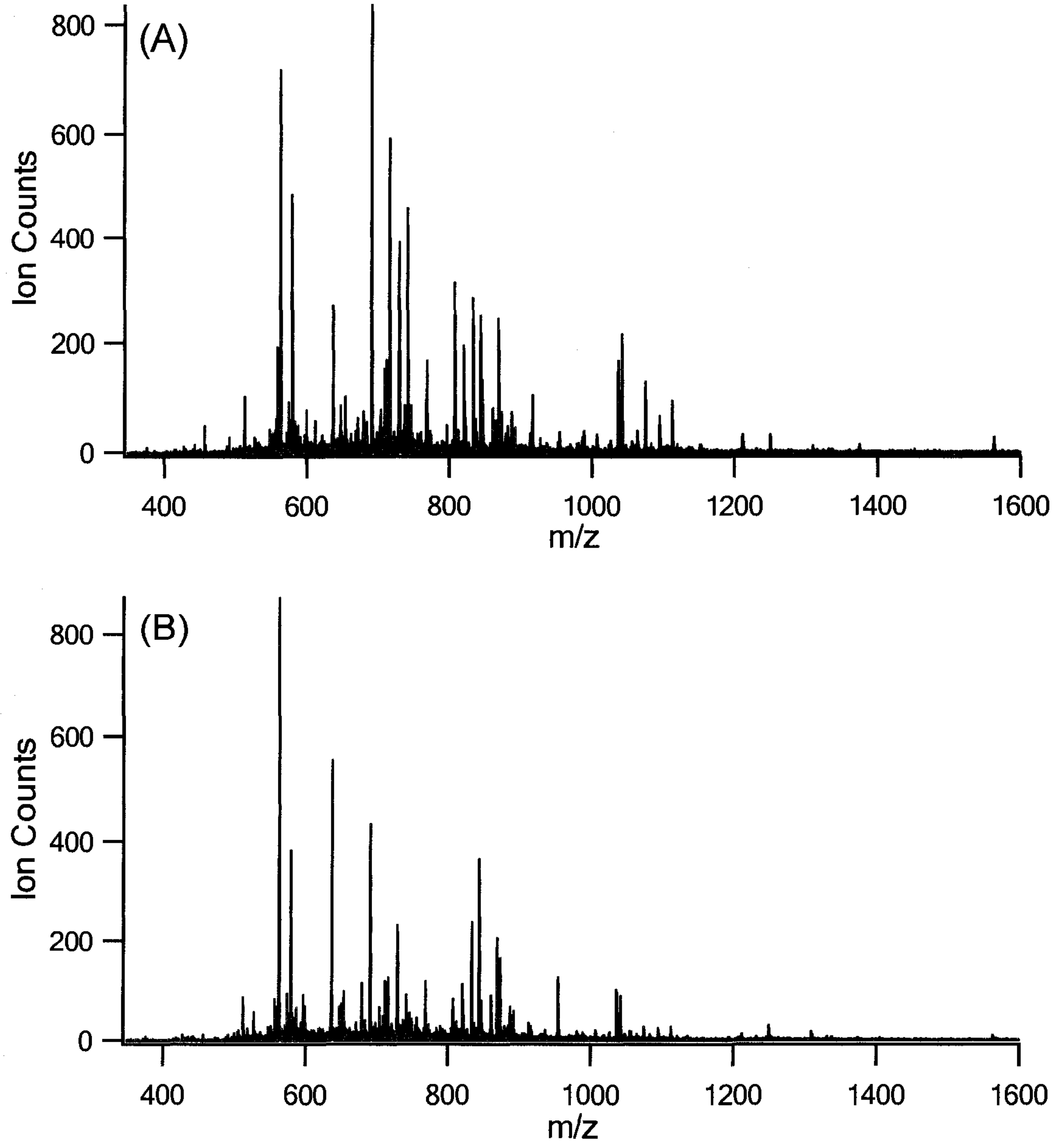


Figure 6.3 Two MS scan spectra from two SCX samples of (A) the 1st run and (B) the 2nd run where a common peptide was identified.

Figures 6.4A and 6.4B summarize the results generated by LC MS/MS analysis of the SCX fractions from the SDS digest. In the 1st SCX run, a total of 11060 different peptides including 3753 unique peptides to this run were identified, compared to a total of 11205 peptides including 3898 unique peptides identified from the 2nd SCX run. Among them, 7307 peptides were in common, bringing in a total of 14958 peptides identified from the two runs. At the protein level, 3386 unique proteins or protein groups were identified in the 1st SCX run, compared to 3528 proteins from the 2nd SCX run. There were 2438 proteins identified commonly in both SCX samples (see Figure 4B), bringing in a combined total of 4476 proteins identified from the two samples. These results indicate that the proteome coverage from the two SCX samples prepared by using different gradient conditions is complementary to each other and the combined results extend the proteome coverage.

It should be noted that, while we did not carry out the direct comparison of this dual-SCX method to simple replicate runs using identical SCX conditions, we would expect that the replicate runs would not identify as many unique proteins as that from the dual-SCX approach. This can be inferred from the work described in Chapter 6 on the MCF7 proteome analysis using replicate runs with precursor ion exclusion (PIE). In that case, a total of 2362 proteins were identified from the 1st SCX run. With PIE of the peptide identified from the 1st SCX run, the replicate run of the SCX sample by LC MS/MS resulted in the identification of an additional 549 proteins. This represents an increase of 23.2%. For the current zebrafish work, the 1st SCX run identified 3386 unique proteins and the 2nd SCX run identified an additional 1090 unique proteins, representing an increase of 32.2%. If we switch the order of the two SCX dataset (i.e., considering the 2nd SCX run as the first run), a total of 3528 was identified in the 1st run and 948 additional proteins were identified from the 2nd run, representing an increase of 26.9%, which is still greater than 23.2% increase gained from replicate runs with PIE. This comparison work also indicates that, for the future work, we need to further optimize the SCX separation conditions so that the two SCX runs will generate significant differences in peptide composition in the SCX fractions collected. Alternatively, an orthogonal separation method such as the use of pH gradient chromatography may be used to fractionate the peptides prior to SCX chromatography.

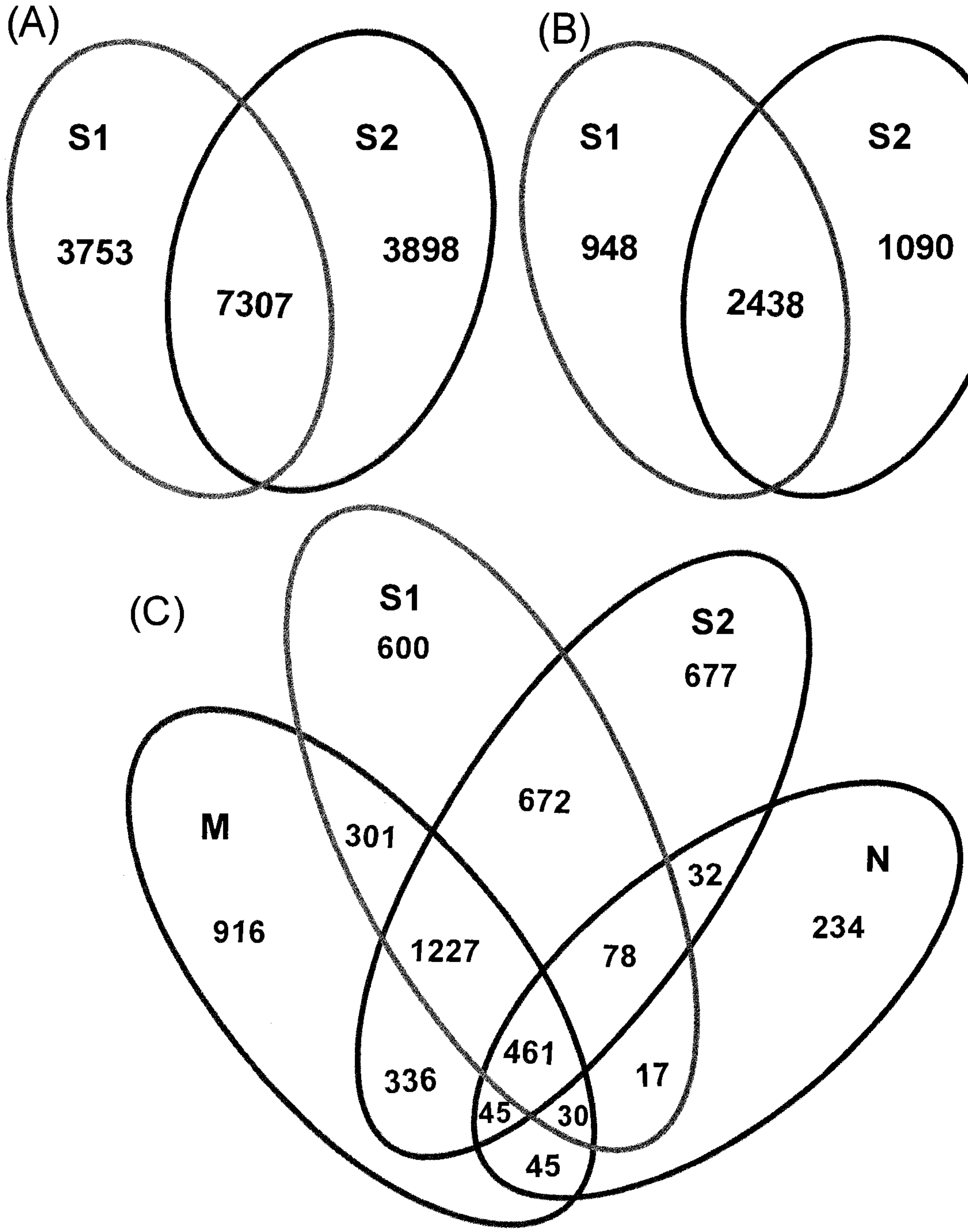


Figure 6.4 (A) Distribution of the numbers of peptides identified in the two SCX samples of the SDS fraction. Distributions of the numbers of proteins identified (B) in the four protein fractions and (C) in the two SCX samples of the SDS fraction.

Indeed, a recent work in our laboratory of using three-dimensional LC separations (pH gradient, SCX and RP) has demonstrated that 3D-LC can significantly improve the peptide identification efficiency (e.g., about 7900 proteins were identified from MCF-7 whole lysate digests, compared to about 3000 proteins identified from 2D-LC ESI QTOF) [Xie, Wang, and Li, unpublished 2008].

Figure 6.4C compares the numbers of proteins identified from different samples including buffer fraction (N), methanol fraction (M), SDS 1st SCX run (S1), and SDS 2nd SCX run (S2). Combining all the identification results from the three protein samples, 5671 unique proteins were identified from the membrane component of the zebrafish liver. Only a small number of unique proteins were identified from the buffer fraction (234 out of 5671 or 4.1%). This work suggests that, in dealing with the membrane component of the zebrafish, the buffer protein fraction did not significantly expand the proteome coverage. For future work, analysis of this fraction is perhaps not needed in order to save analysis time. By contrast, 916 unique proteins (16.2%) were identified from the methanol fraction. There were many common proteins identified between the methanol fraction and the SDS fraction (i.e., 2445 proteins). Interestingly, 4476 out of 5671 or 78.9% of the proteins from the membrane component of the zebrafish liver were identified from the SDS fraction. For future work, to shorten the analysis time, the SDS fraction should be analyzed in high priority whiles the buffer protein solubilization step may be omitted. In addition, if no protein fractionation is carried out, the SDS-assisted protein solubilization and digestion method should be applied. If protein fractionation is needed to simplify the proteome, the protein pellet can be subjected to methanol solubilization, followed by SDS-assisted solubilization. Analysis of these two fractions should cover a wide range of the proteome.

It is worth commenting on the protein level separation vs. peptide level separation for proteome analysis. Both separations are aimed at reducing the complexity of the proteome sample to an extent that the final individual peptide samples introduced to the LC MS/MS system will not cause severe under-sampling problem. Under-sampling is due to the limitation of a mass spectrometer to sequence all peptide ions eluted from LC, i.e., only a fraction of the peptide ions are sequenced. The more complex the peptide mixture, the smaller proportion of the peptides sampled or sequenced. Protein separation based on solubility difference of proteins in different reagents such as buffer, methanol, SDS, or other solution is a crude method, but can be quite effective in reducing the complexity of a proteome sample, as demonstrated in this work. However, the protein fraction generated in each reagent is still very complex. Fractionation of the peptides generated in each protein fraction may further reduce the complexity of the final peptide samples. But, this requires more analysis time. At this stage, it is difficult to ascertain which method, protein fractionation or peptide fractionation, is more efficient for comprehensive proteome profiling. Considering the complexity of a proteome sample such as the zebrafish tissue, one may need to combine both protein and peptide fractionations to reduce the sample complexity for LC MS/MS.

A total of 5671 proteins identified from the membrane component represent the most comprehensive proteome profile generated to date for the zebrafish. Since the membrane proteins were enriched in this fraction, a large portion of the identified proteins are belonging to the membrane or membrane associated proteins. Among the membrane proteins, the integral membrane proteins are the most difficult to analyze by current proteomic technologies. To gauge the applicability of our method to analyze the integral membrane proteins, we extracted the membrane classification information and the results are shown in Figure 6. 5A. This figure plots the number of proteins identified as a function of the number of transmembrane domains (TMDs) for the three protein samples. It is clear that higher percentages of integral membrane proteins including those with high number of TMDs were identified, particularly in the methanol and SDS fractions. If we compare the proteins according to their hydrophobicity property gauged by the GRAVY index (see Figure 6.5B), many more hydrophobic proteins were identified from the methanol or SDS fractions than the buffer fraction. These results make sense considering that hydrophobic proteins are more soluble in methanol or SDS than the buffer solution. Unfortunately, for the integral membrane proteins which are the most difficult class of proteins to be identified, we do not know the distribution of their TMDs for the entire zebrafish proteome and thus we cannot gauge the overall performance of our method in terms of the detectability of the integral membrane proteins (i.e., if any proteins with a certain TMD such as TMDs > 13 are underrepresented).

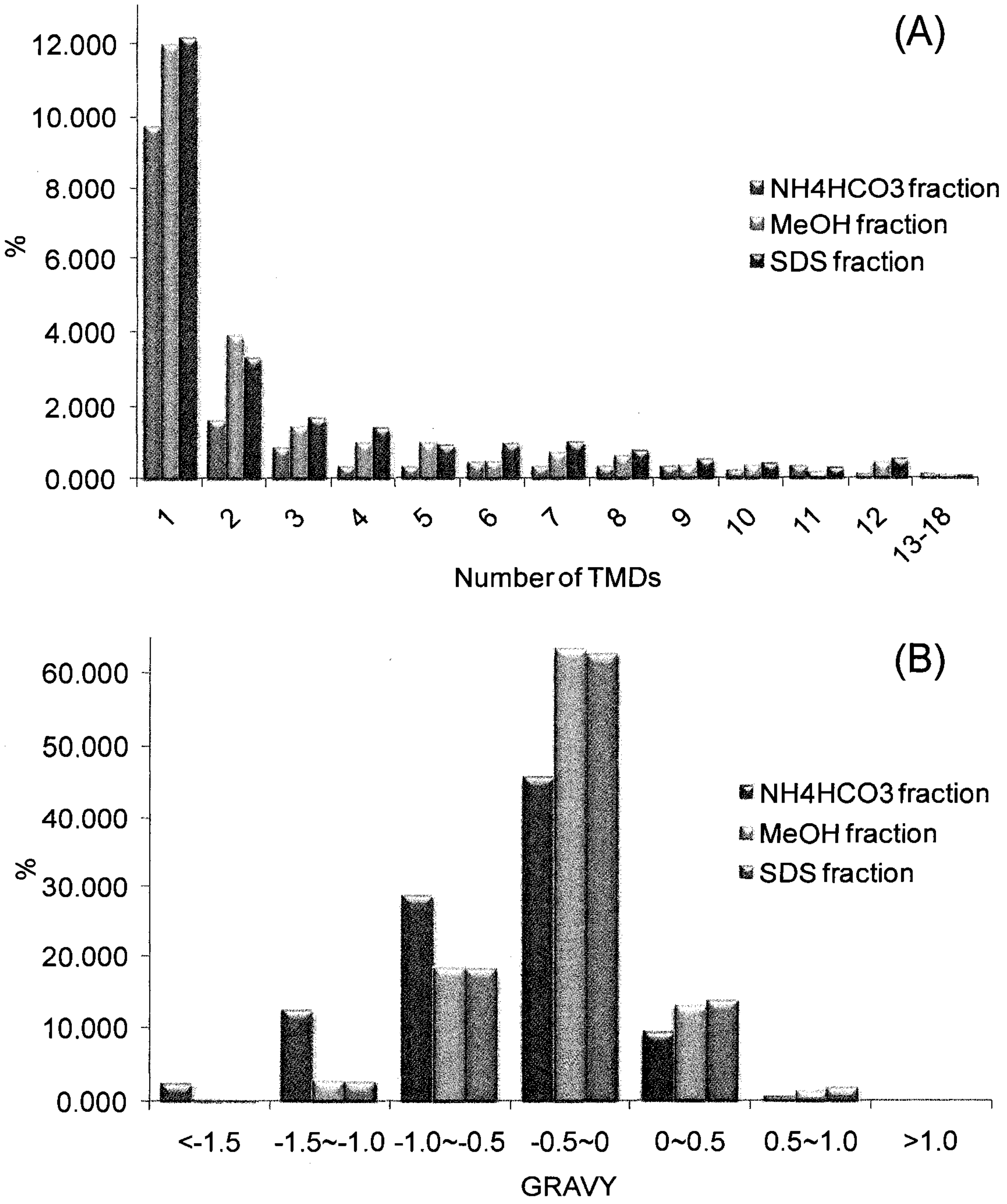


Figure 6.5 Distribution of membrane proteins as a function of (A) the number of the transmembrane domains (TMDs), (B) GRAVY.

Nevertheless, the fact that we identified many integral membrane proteins suggests that our method is still effective in dealing with membrane proteins including integral membrane proteins with many TMDs.

As indicated earlier, after SDS solubilization of the protein pellet, some precipitates still remained. We applied the MAAH method to degrade the precipitates and the resulting hydrolysate was analyzed by LC-ESI MS/MS. This work is still on the way. But, some preliminary results are shown in Table 6.1. In this case, we only ran 1D LC-ESI QTOF on the hydrolysate to gauge if there were any new proteins identified from this protein fraction. As Table 6.1 shows, a total of 265 proteins were identified. Among them, 39 proteins were unique to this fraction, bringing in the total number of proteins identified from the membrane component of the zebrafish liver to be 5710. Future work on this fraction will be involved in SCX fractionation of the hydrolysate, followed by RP-LC ESI MS/MS. We expect additional unique proteins will be identified from this protein fraction.

6.4 Conclusions

In this work, we have combined our sensitive off-line 2D LC-ESI QTOF MS technique with sequential protein solubilization and digestion for the analysis of the proteome from the membrane component of the zebrafish liver. By analyzing the soluble fractions with trypsin digestion, a total of 5671 proteins were identified including many integral membrane proteins. It was found that SDS was most effective in solubilizing the proteins isolated from the membrane component. Two independent SCX runs with different salt gradient conditions resulted in a total of 52 samples for LC MS/MS. The analysis of these SCX fractions resulted in the identification of a total of 4476 unique proteins, representing 78.9% of the identified proteome. The analysis of the buffer-solubilized protein fraction identified only 234 unique proteins, while the methanol fraction generated 916 unique proteins. Even after the use of SDS to dissolve the protein pellet, some precipitates were still observed. Analysis of these precipitates by using MAAH followed by 1D LC-ESI MS/MS resulted in the identification of an additional 46 unique proteins. Analysis of this fraction by using 2D LC-ESI MS/MS is currently underway.

Now that we demonstrate that a large number of proteins can be identified using the method described, our future work will focus on the combination of this method with peptide isotope labeling for quantitative proteome analysis of the zebrafish liver. In addition, more efficient protein and peptide fractionation methods will be developed to improve the proteome coverage as well as protein identification efficiency.

Table 6.1 Unique proteins identified from the MAAH fraction.

#	Accesion ID	Protein Description	Peptide Score	MASCOT Score for Identity	Peptide Sequence	Modification
1	gi|113678366	hypothetical protein LOC559475 [Danio rerio]	47	41	IPPVFAIIAR	
			52	50	AEPSVALVSSLAGALR	
			51	43	FVQLVQLLR	
			50	45	FVQLVQLLR	2 Deamidated (NQ)
			45	43	PLLAAEVR	
			55	45	PSVALVSSLAGALR	
			55	53	SPFNEIHGAAMMEAK	Deamidated (NQ)
			58	51	VEALPVELPEHIA	
			79	50	YEALLLGGLPQEGLAR	
			77	52	YEALLLGGLPQEGLAR	Deamidated (NQ)
			62	52	YEALLLGGLPQEGLAR	
2	gi|116487517	Unknown (protein for IMAGE:7054100) [Danio rerio]	36	30	LLDIAELNR	Deamidated (NQ)
3	gi|12597402	ATP synthase lipid binding protein p3 precursor [Danio rerio]	62	52	DIDTAAKFIGAGAATVGVAG	
			52	52	SGAGIGTVFGSLIIGYAR	
4	gi|125807758	PREDICTED: hypothetical protein [Danio rerio]	33	28	MSLILR	
5	gi|125824303	PREDICTED: hypothetical protein [Danio rerio]	55	51	VVVTMEHSAK	
6	gi|125825951	PREDICTED: hypothetical protein [Danio rerio]	48	20	LSPIVILFPK	
7	gi|125841397	PREDICTED: similar to P2Y-	31	31	MIRTPR	Oxidation (M)

		like G-protein coupled receptor [Danio rerio]				
8	gi|125842385	PREDICTED: hypothetical protein [Danio rerio]	64	53	IQDYVMSYPFVR	Deamidated (NQ)
			70	53	TIEYLEEVAVEAAR	
9	gi|126632422	novel protein similar to vertebrate SNAP25-interacting protein (SNIP) [Danio rerio]	52	51	HGVMVGSLKT	Oxidation (M)
10	gi|131889490	transmembrane protein 168 [Danio rerio]	55	49	LSSFNLLVA	Deamidated (NQ)
11	gi|148726556	ribosomal protein, large, P0 [Danio rerio]	59	51	IIQLLDDYPK	Deamidated (NQ)
12	gi|152012733	Unknown (protein for MGC:174137) [Danio rerio]	74	51	YEALLLGGLPQEGLAR	Deamidated (NQ)
			62	52	YEALLLGGLPQEGLAR	
13	gi|18858947	keratin 4 [Danio rerio]	37	31	AVYEAELR	
			82	52	NKYEDEINKR	2 Deamidated (NQ)
14	gi|29126846	Hydroxysteroid (17-beta) dehydrogenase 12a [Danio rerio]	36	32	AFVDFFSR	
15	gi|33284843	novel protein similar to human member of RAS oncogene family (RAB35) [Danio rerio]	40	25	LLIIGDSGVGK	
16	gi|33468618	novel protein similar to human and rodent member RAS oncogene family RAB7 (RAB7) [Danio rerio]	40	25	LILIGNSGVGK	Deamidated (NQ)
17	gi|34784792	Selenoprotein T, 1b [Danio rerio]	41	30	VFEEYTR	
18	gi|41055102	H2A histone family, member X [Danio rerio]	53	41	GVLPNIQAVLLPK	Deamidated (NQ)
			68	42	GVLPNIQAVLLPK	2 Deamidated (NQ)

19	gi|4504279	H3 histone, family 3A [Homo sapiens]	66	48	STELLIR	
20	gi|47085823	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain [Danio rerio]	63	49	GITFEDVVIPK	
			42	32	IYQIYEGTAQIQR	3 Deamidated (NQ)
21	gi|50080185	alcohol dehydrogenase 8a [Danio rerio]	53	52	IMLDEFITHK	
22	gi|50539696	hypothetical protein LOC436590 [Danio rerio]	40	25	LLLLGDSGVGK	
23	gi|125843107	PREDICTED: similar to Uncharacterized protein C10orf30 [Danio rerio]	34	23	AILLELR	
24	gi|125847453	PREDICTED: similar to Arylsulfatase B precursor (ASB) (N-acetylgalactosamine-4-sulfatase) (G4S), partial [Danio rerio]	39	32	VEVELLGQK	Deamidated (NQ)
25	gi|125850623	PREDICTED: hypothetical protein [Danio rerio]	74	51	YEALLLGGLPQEGLAR	Deamidated (NQ)
			62	52	YEALLLGGLPQEGLAR	
26	gi|125854078	PREDICTED: similar to tetratricopeptide repeat-containing hedgehog modulator 1 [Danio rerio]	55	48	LAHVDLALT	
27	gi|125864643	PREDICTED: similar to LOC398653 protein [Danio rerio]	37	27	YVAAYLLAALGGK	
28	gi|125875288	PREDICTED: hypothetical protein, partial [Danio rerio]	64	50	PTLPAQYFHLLR	Deamidated (NQ)
29	gi|125881244	PREDICTED: similar to Retinol	74	53	TIVTDFNIVESTLHR	

		dehydrogenase 1, like, partial [Danio rerio]				
			85	53	TIVTDFNIVESTLHR	Deamidated (NQ)
30	gi|125888945	PREDICTED: hypothetical protein, partial [Danio rerio]	67	52	PVNYYVDTAVR	
31	gi|57864779	vitellogenin 2 [Danio rerio]	54	52	AEAGLLGEFPAFR	
			69	52	EAYPGDVFYLHSR	
32	gi|62955563	hypothetical protein LOC550493 [Danio rerio]	70	53	YYVTIIDAPGHR	
33	gi|66911667	LOC559407 protein [Danio rerio]	54	53	ILEQIGAQERNISQ	3 Deamidated (NQ)
34	gi|81294186	Aldh2b protein [Danio rerio]	53	51	ANYISHGLR	Deamidated (NQ)
			61	50	STDVGHLIQR	Deamidated (NQ)
			91	53	TFVQESIYDEFVER	Deamidated (NQ)
			73	54	TFVQESIYDEFVER	
			65	53	TFVQESIYDEFVER	Deamidated (NQ)
			77	48	VAEQTPLTALYIASLIK	Deamidated (NQ)
			57	47	VAEQTPLTALYIASLIK	
			57	48	VAEQTPLTALYIASLIK	Deamidated (NQ)
			61	53	YGLAGAVFTQDIDK	Deamidated (NQ)
			56	51	YYAGWADKWEGK	
35	gi|94732716	novel protein [Danio rerio]	54	49	KMLSKKGSP	
36	gi|94732999	novel protein similar to vertebrate eukaryotic translation elongation factor 2 (EEF2) [Danio rerio]	39	31	GLPEANLALHR	Deamidated (NQ)
37	gi|94733107	novel protein similar to vertebrate hydrocephalus inducing	33	32	NSVVMEK	

		(HYDIN) [Danio rerio]				
38	gi|94733551	novel protein (zgc:77752) [Danio rerio]	65	50	LNIKSIINMQLPG	Deamidated (NQ)
39	gi|9857942	chaperonin 10 [Danio rerio]	43	33	VMLEDKDYFLFR	

6.5 Literature Cited

(1) Hinton, D.; Kullman, S.; Hardman, R.; Volz, D.; Chen, P.; Carney, M.; Bencic, D. *Marine Pollution Bulletin* **2005**, *51*, 635-648.

(2) Langheinrich, U. *Bioessays* **2003**, *25*, 904-912.

(3) Rubinstein, A. *Expert Opinion on Drug Metabolism & Toxicology* **2006**, *2*, 231-240.

(4) Yoshizawa, K.; Heatherly, A.; Malarkey, D.; Walker, N.; Nyska, A. *Toxicologic Pathology* **2007**, *35*, 865-879.

(5) Shrader, E.; Henry, T.; Greeley, M.; Bradley, B. *Ecotoxicology* **2003**, *12*, 485-488.

(6) Bosworth, C.; Chou, C.; Cole, R.; Rees, B. *Proteomics* **2005**, *5*, 1362-1371.

(7) Link, V.; Shevchenko, A.; Heisenberg, C. *Molecular & Cellular Proteomics* **2005**, *4*, S253-S253.

(8) Link, V.; Shevchenko, A.; Heisenberg, C. *Bmc Developmental Biology* **2006**, *6*, 1-9.

(9) Ziv, T.; Gattegno, T.; Chapovetsky, V.; Wolf, H.; Bamea, E.; Lubzens, E.; Admon, A. *Comparative Biochemistry and Physiology D-Genomics & Proteomics* **2008**, *3*, 12-35.

(10) Lucitt, M.; Price, T.; Pizarro, A.; Wu, W.; Yocum, A.; Seiler, C.; Pack, M.; Blair, I.; Fitzgerald, G.; Grosser, T. *Molecular & Cellular Proteomics* **2008**, *7*, 981-994.

(11) Lemeer, S.; Pinkse, M.; Mohammed, S.; Van Breukelen, B.; Den Hertog, J.; Slijper, M.; Heck, A. *Journal of Proteome Research* **2008**, *7*, 1555-1564.

(12) Wang, N.; Mackenzie, L.; De Souza, A.; Zhong, H.; Goss, G.; Li, L. *Journal of Proteome Research* **2007**, *6*, 263-272.

(13) Speers, A.; Wu, C. *Chemical Reviews* **2007**, *107*, 3687-3714.

(14) Han, D.; Eng, J.; Zhou, H.; Aebersold, R. *Nature Biotechnology* **2001**, *19*, 946-951.

(15) Zhang, N.; Chen, R.; Young, N.; Wishart, D.; Winter, P.; Weiner, J.; Li, L. *Proteomics* **2007**, *7*, 484-493.

(16) Li, N.; Shaw, A.; Zhang, N.; Mak, A.; Li, L. *Proteomics* **2004**, *4*, 3156-3166.

Chapter 7

Unraveling the Complete Proteome of *Escherichia coli* by Mass Spectrometry*

7.1 Introduction

Genomics technology development in the past two decades including rapid DNA sequencing and microarray techniques has allowed bioscience researchers to carry out functional genomics work efficiently.[1-8] The scale of DNA microarrays has rapidly increased in the recent years while the cost of analysis is coming down. For example, microarrays composed of oligonucleotide complements of all predicted genes of an organism such as *E. coli*, yeast, human, etc. are being routinely used for many genomics applications including searching for biomarkers of diseases.[4] By comparison, proteome analysis has not reached the same scale as the genome analysis. In most proteomics applications, only a fraction of the proteome is examined. This disparity is mainly due to the complexity of the proteome and the difficulty of characterizing proteins, compared to genome and DNA analysis. The goal of our research is to develop techniques ultimately useful to examine the entire proteome of a given biological sample. The first step towards realizing this goal is to develop a technique that will be able to identify all of the proteins present in a proteome sample. This should facilitate future work in whole-proteome quantification, posttranslational modification characterization and protein-protein interaction studies –the hallmarks of proteomics for linking proteome analysis with biological functional studies or systems biology.[9]

As described in the previous chapters, our lab has developed several improved proteome analysis techniques and record numbers of proteins have been identified from organisms such as zebrafish. However, one question still remains: can we use our techniques to identify all of the proteins present in a proteome sample? To address this

*Xiaoxia Ye contributed partially to sample preparation, data acquisition and data processing of this work. Dr. Joe Weiner performed bioinformatic characterization of the *E. coli* proteins.

important question, we have attempted to analyze the proteome of a relatively simple microorganism, *E. coil* K12. This model system was chosen because there are only about 4300 genes predicted from the genome of *E. coli.*[10-14] Since gene splicing does not occur for *E. coli*, an equal number or about 4300 unique proteins or protein groups can be potentially expressed in *E. coli* cells. A protein group is defined as proteins having the same sequences, but with sequence truncation (e.g., mature form vs. a protein with a signal peptide) or modification (e.g., phosphorylated vs. non-phosphorylated protein). The actual number of proteins and the amount of each protein expressed in the cells should be dependent on the cell culture conditions. Unfortunately, under a certain culture condition, the exact number of proteins present in a cell cannot be predicted and currently is not known, as there is no technique which can detect all the proteins. Nevertheless, the upper limit of 4300 proteins provides a metric from which the proteome coverage by a technique can be gauged.

In this work, we describe our attempts of unraveling the complete proteome of the *E. coli* cells grown under a commonly used culture condition, i.e., in a rich media to the stationary growth phase. To increase the proteome coverage, cellular proteins were devided into three fractions, namely cytoplasm, peripheral membrane and integral membrane fractions. Shotgun analysis of these protein fractions was then carried out by using two-dimensional LC-ESI QTOF MS. Additional experiments, including the use of low molecular weight cutoff filters to enrich low molecular weight proteins from cell lysates followed by shotgun analysis, were done to improve the proteome coverage. The results generated from these experiments will be described in detail.

7.2 Experimental

7.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide, phenylmethylsulfonyl fluoride (PMSF), trifluoroacetic acid (TFA), heparin, sodium bicarbonate, potassium chloride (KCl), LC-MS grade formic acid and SDS were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, LC-MS grade water, acetone, methanol (MeOH), and acetonitrile (ACN) were from Fisher Scientific Canada

(Edmonton, Canada). Roche mini protease inhibitor cocktail was obtained from Roche Applied Science (Indianapolis, IN). The BCA assay kit was from Pierce, Rockford, IL..

7.2.2 Cell Culture and Protein Extraction.

Escherichia coli K-12 (*E. coli*, ATCC 47076) was from the American Type Culture Collection (Manassas, VA). A single *E. coli* K12 colony was used to inoculate 50 mL of LB broth (BBL, Becton Dickinson). The culture was incubated overnight with shaking at 37 °C. Cells were harvested in the stationary phase by centrifugation in a Beckman SX4250 rotor at 3 200 *g* for 15 min at 4 °C. Afterwards the cells were washed in PBS buffer, and collected by centrifugation at 3 200 *g* for 15 min at 4 °C. The cells were then resuspended in 30 mL water. The suspension was passed twice through a French press (Aminco Rochester, NY) at 20 000 psi after adding 3 tablets of Roche mini protease inhibitor cocktail. The lysate was centrifuged at 3 200 *g* for 15 min to pellet unbroken cells. The supernatant was transferred and centrifuged again in Type55.2Ti rotor for 55 min at 118 000 g. The supernatant (cytoplasm) was collected and stored in -80 °C for future use.

The pellets were suspended in 6 mL 50 mM ammonium bicarbonate and transferred to a 250 mL beaker in an ice bath. The original bottles were rinsed using ~ 5 mL buffer. Then 100 mL of 0.1 M sodium carbonate (pH 11.0) was slowly added. The solution was stirred slowly in an ice bath for 1 h to extract membrane proteins. The extract was divided equally into two tubes and centrifuged in a Beckman Type 45Ti rotor for 60 min at 38 400 rpm (115 000 *g*). The supernatant (peripheral membrane proteins) was aspirated, and the pellet was gently rinsed with 5 mL of water. Each pellet was suspended in 7 mL of 50 mM MOPS buffer (pH 7.3) and transferred to two 8-mL tubes. The tubes were centrifuged in a Beckman Type 70.1Ti rotor at 40 000 rpm (115 000 *g*) for 25 min. The pellets were kept as integral membrane proteins. All protein concentrations were measured by BCA protein assay using bovine serum albumin as the standard. The protein extracts were stored at -80 °C for later analysis.

7.2.3 Sequential Protein Precipitation, Re-solubilization and In-solution Digestion.

The cytoplasm and peripheral membrane fractions were first subjected to disulfide bond reduction and alkylation. Briefly, 10 mg of protein sample was reduced in 20 mM DTT solution for 1 h at 37 °C. Free thiol groups were blocked by reaction with a double concentration of iodoacetamide for 1 h at room temperature in the dark. The cloudy solutions were then centrifuged at 14 000 rpm for 10 min (4 °C). The supernatants were transferred to two new vials. And the protein pellets were stored for further analysis. Acetone precipitation was performed for the supernatants to purify the proteins and remove detergent, unreacted DTT, and iodoacetamide. Acetone, precooled to -80 °C, was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The mixture was kept at -20 °C overnight and then centrifuged (14 000 rpm, 10 min, 4 °C). The supernatant was decanted and properly disposed. Acetone was evaporated at room temperature.

All the protein pellets from the above steps were subjected to the sequential solubilization and digestion protocol described in Chapter 3 (see Figure 7.1). Firstly, ammonium bicarbonate (50 mM, pH 8.0) was used to solubilize the membrane protein pellet with intermittent vortexing applied. The vial was then centrifuged at 14 000 rpm for 5 min at 4 °C. Trypsin solution was added into the supernatant for an enzyme/protein ratio of 1:45, and digestion was conducted at 37 °C overnight.

The pellets remaining following ammonium bicarbonate treatment were resuspended in 60% MeOH, with sufficient vortexing. Trypsin was added at an enzyme/protein ratio of 1:30, and the solution was incubated at 37 °C overnight. The solution was then centrifuged (14 000 rpm, 5 min, 4 °C), and the supernatant was transferred to a different vial. MeOH in the supernatant was evaporated using a SpeedVac (Thermo Savant, Milford, MA). After this MeOH-assisted solubilization and digestion, the undissolved pellet was redissolved in 2% SDS (i.e., SDS-assisted solubilization), followed by 40-fold dilution. Trypsin was added to achieve a final enzyme/protein ratio of 1:40. The sample was incubated at 37 °C for 2 days with 10% (of

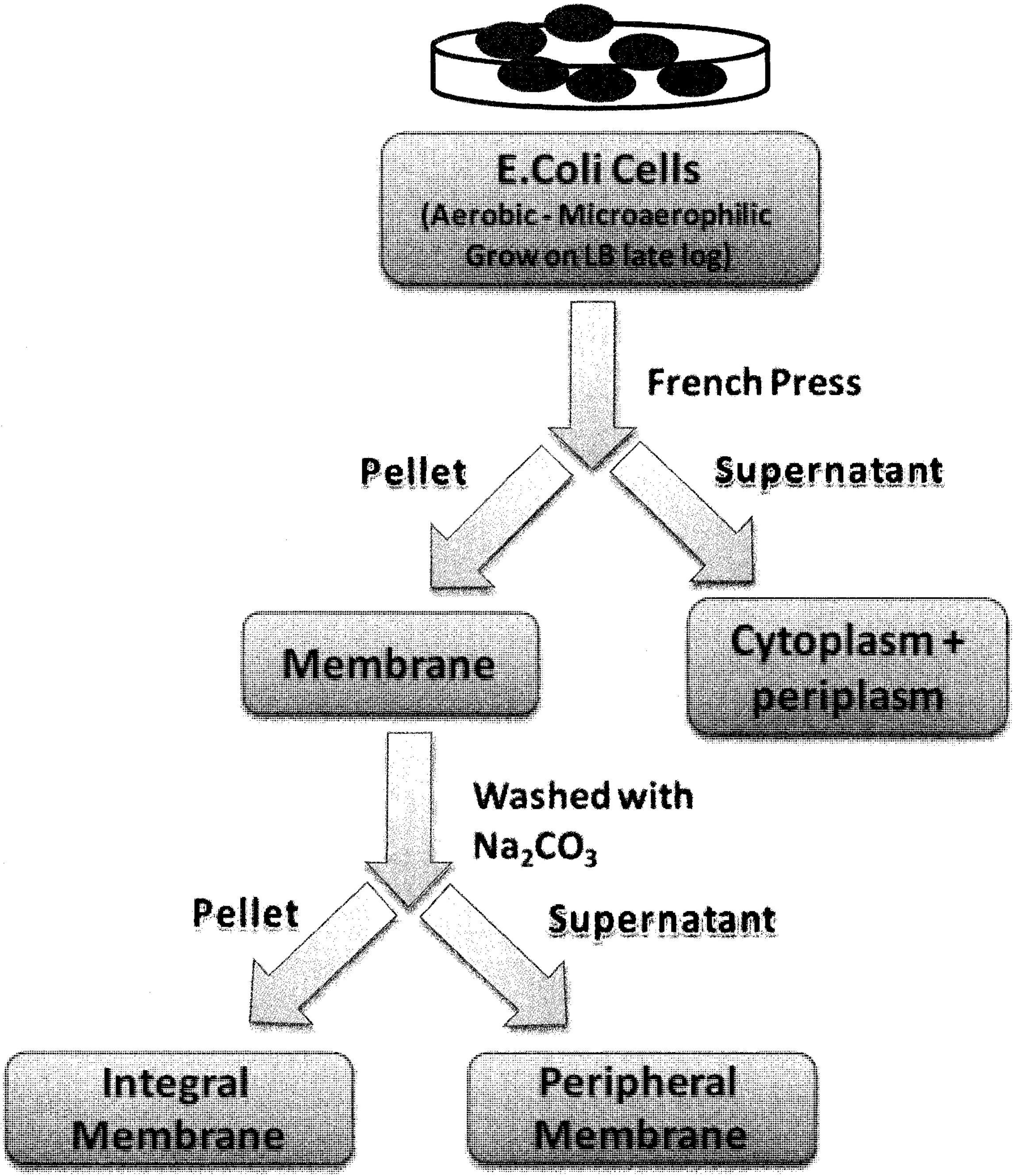


Figure 7.1 Workflow of *E. coli* protein extraction and fractionation.

amount added for the 1st overnight) more fresh trypsin added before the 2nd overnight digestion.

The integral membrane protein pellet was subjected to the removal of lipids (de-lipid) in a cold MeOH/Acetone solution at -20 °C overnight. The solution was centrifuged (14 000 rpm, 15 min, 4 °C) and the pellet were then subjected to sequential solubilization using ammonium bicarbonate, 60% MeOH and 2% SDS, as described above. Standard reduction and alkylation were carried out before trypsin digestion was performed. The digestion was stopped by acidifying the solution to pH 2.

7.2.4 Cation Exchange Chromatography

Peptide mixtures from each step were separated by strong cation exchange (SCX) chromatography on an Agilent 1100 HPLC system (Palo Alto, CA) using a 2.1 × 250 mm highly hydrophilic polysulfoethyl A column at ambient temperature (particle size of 5 μm diameter and 300 Å pore, PolyLC Inc. U.S.). Gradient elution was performed with mobile phases A (10 mM KH_2PO_4, pH 2.76) and B (10 mM KH_2PO_4, pH 2.76, 500 mM KCl). Protein digests from each method were loaded separately onto the SCX column, and peptides were eluted using linear gradients (0 min: 0% B, 5 min: 0% B, 30 min: 20% B, 40 min: 60% B, 50 min: 100% B, 60 min: 100% B) at 0.20 mL/min, with fraction collection at 1 min intervals. In all, 24~26 fractions were collected from each run based on the chromatography UV absorption signals recorded at 214 nm.

7.2.5 Peptide Desalting and Quantification by RPLC.

Desalting and quantification were carried out in an Agilent 1100 HPLC system (Palo Alto, CA) using the newly developed method described in Chapter 5. Desalting of the tryptic peptides was performed on a Polaris C18 A column (Varian, USA). After loading of the peptide sample, the column was flushed with mobile phase A (0.1% TFA in water) and the salts were effectively removed. Subsequently, the concentration of mobile phase B (0.1% TFA in ACN) in the mobile phase was increased to 85% in 3 s to ensure complete elution of the peptide fractions from the column. During the peptide elution process, a chromatographic peak was produced and based on the peak area the amount of peptides was determined. Four-protein digests of various amounts were used as standards for the generation of a linear calibration between the peak area and the injected peptide amount. The calibration curve was generated as y=430.0x-269.6, where y refers to the peak area of the peptide sample, and x refers to the peptide amount analyzed in μg.

7.2.6 LC-ESI QTOF MS and MS/MS Analysis.

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, the desalted and quantified digests were concentrated using a SpeedVac (Thermo Savant, Milford, MA) to ~5 μL and reconstituted to a specific concentration using 0.1% formic acid. Then 1-μg of peptides was injected onto a 75 μm × 100 mm Atlantis dC18 column with 3 μm particle and 300 Å (Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using a 120 min gradient (2-6% Solvent B for 2 min, 6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min; the column was pre-equilibrated at 2% Solvent B before each sample run) and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 0.8 s each. For both dynamic and precursor ion mass exclusion, a mass tolerance window of 80 mDa was applied. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants was infused at a flow rate of 300 nL/min, and a 1 s MS scan was acquired every 1 min throughout the run.

Peptide precursor ion exclusion (PIE) strategy was applied to exclude relatively high-abundance peptides identified from the adjacent two SCX fractions to enable additional and less abundance peptides to be analyzed and identified. An exclusion list was generated based on MASCOT (Matrix Science, London, U.K.) searching results of peptides with a score 10 points equal to or higher than the identification threshold.

7.2.7 Protein Database Search.

Raw search data were lock-mass-corrected, and converted to peak list files by ProteinLynx Global Server 2.2.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the Mascot search program. Database searching was restricted to *E. coli* K12 in the Swiss-Prot database. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, ±30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, *N*-Acytyl (Protein), oxidation (M), Pyro_glu (N-term Q), Pyro_glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides with scores lower than the MASCOT threshold score for identity were then deleted from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

7.2.8 Hydropathy Calculation, Transmembrane Domain Prediction and Annotation of Localization.

All proteins identified were examined using the ProtParam program available at the ExPASy Web site: http://ca.expasy.org/tools/protparam.html, which allows for calculation of the grand average of hydropathy (GRAVY). Proteins exhibiting positive values were considered hydrophobic, and those that exhibit negative values were considered hydrophilic. The transmembrane domains for all the proteins identified in all 3 fractions were predicted using the TMHMM server 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0). The subcellular locations for the enlisted proteins were categorized according to the information acquired from http://www.geneontology.org.

7.3 Results and Discussion

Although *E. coli* is a simple microorganism, proteins are still expected to be present in a cell at a wide concentration dynamic range from a few copies to hundreds of thousands of copies.[10] To increase the probability of detecting low abundance proteins, we fractionated the proteins from the cell lysates according to their cellular properties. Figure 7.1 shows the workflow to pre-fractionate the cell lysates into three samples. As Figure 7.1 illustrates, the cultured cells were lysed by French press, followed by centrifugation to separate the cell lysate into a pellet and a supernatant. The supernatant was considered to contain mainly cytoplasm and periplasm proteins. This cytoplasm fraction was further fractionated using the workflow shown in Figure 7.2. The pellet was subjected to Na_2CO_3 washing and, after the centrifugation of the sample, the supernatant was considered to contain mainly peripheral membrane proteins and the remaining pellet was the integral membrane fraction. These two fractions were also subjected to the further fractionation as shown in Figure 7.2.

As shown in Figure 7.2, the cytoplasm and peripheral membrane fractions were processed in the same manner while the integral membrane fraction was processed differently. In the case of cytoplasm or peripheral membrane fraction, the proteins were first reduced and alkylated. Some precipitates were observed during reduction and alkylation. The proteins in the supernatant were precipitated from cold acetone. Both of the protein precipitates were then subjected to sequential protein solubilization and trypsin digestion in NH_4HCO_3, methanol and SDS, followed by off-line 2D-LC ESI QTOF MS analysis. For the integral membrane fraction, the protein pellet was subjected to each level of solubilization, followed by reduction and alkylation and then trypsin digestion. The resultant peptide samples in the NH_4HCO_3, methanol and SDS solutions were individually analyzed by 2D-LC MS/MS.

Figure 7.3 shows the summary of the numbers of unique proteins or protein groups identified from the three major fractions. Shotgun proteome analysis of the cytoplasm fraction resulted in the identification of 3224 unique proteins, representing 75.0% of the possible 4300 proteins in *E. coli*. From the peripheral membrane fraction, 2983 unique proteins were identified. In addition, a total of 1938 proteins were identified from the integral membrane fraction. All together, a total of 3659 unique proteins were identified,

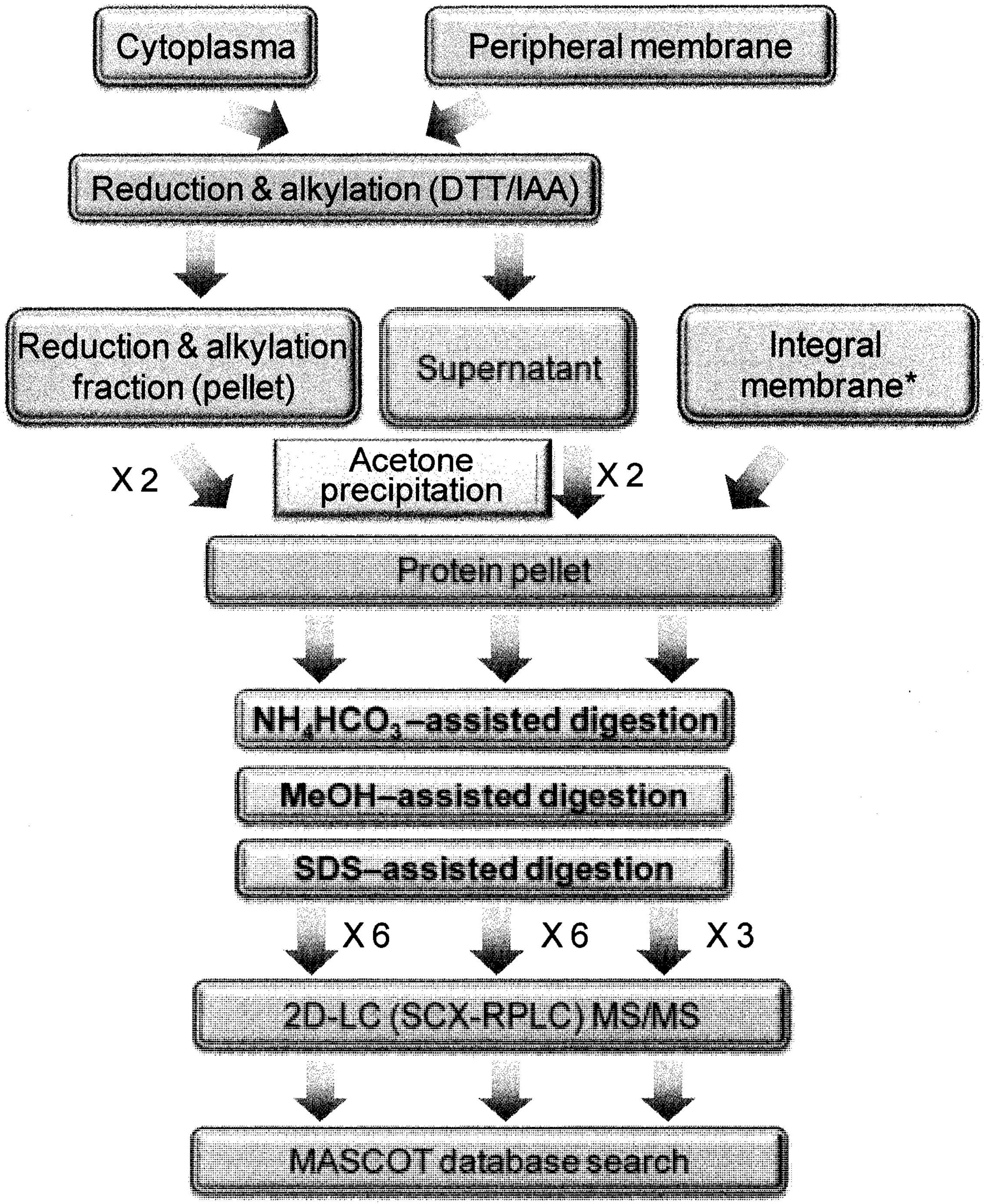


Figure 7.2 Workflow of sequential protein fractionation, solubilization, digestion and downstream 2D-LC ESI QTOF MS/MS analysis of the digests. *For integral membrane protein fraction, reduction & alkylation were done after the sequential solubilization and digestion steps.

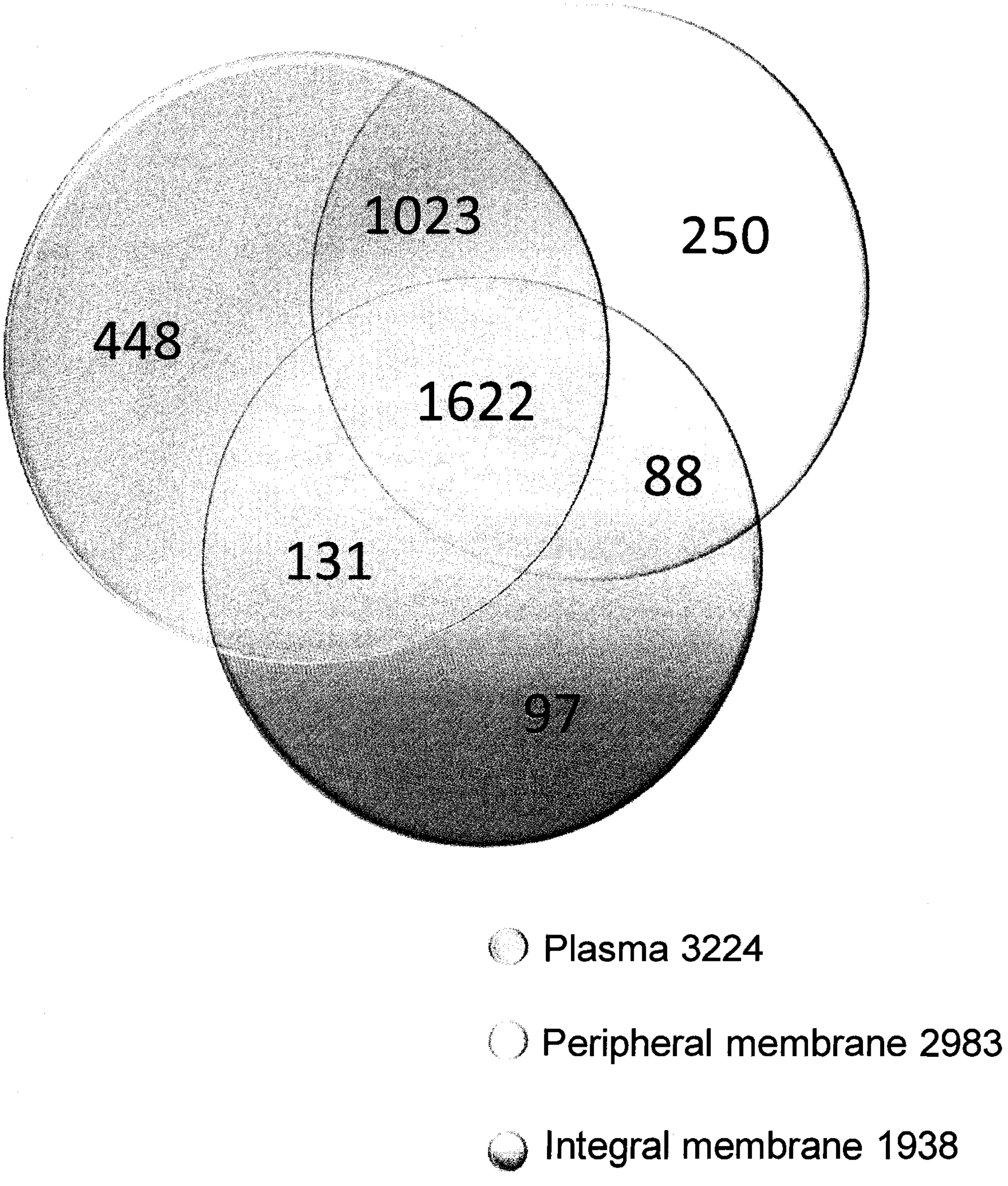


Figure 7.3 Venn diagram of the unique proteins identified from the three major fractions.

covering 85.1% of the proteome. A total of 1622 common proteins were found in all three fractions. There were many more common proteins identified from the cytoplasm and peripheral membrane fractions (i.e., 1023+1622) than those identified from the cytoplasm and integral membrane (i.e., 97+1622). This result indicates that the proteins from the cytoplasm and peripheral membrane fractions have more similar properties than those from the cytoplasm and integral membrane fractions. The total number and unique number of proteins identified in the integral membrane fraction were smaller than those from the other fractions. This may reflect the actual protein composition of these fractions, i.e., the integral membrane fraction contained a smaller number of different proteins or less complex than the other fractions. This may also be the result of technical difficulty in identifying the proteins present in the integral membrane fraction. Identifying integral membrane proteins is known to be more challenging than analyzing other more soluble proteins.

To further examine the difference of proteins identified in the three fractions, the proteins identified in each fraction were grouped according to their hydrophobicity and number of transmembrane domains (TMDs) (see Figure 7.4). The general trend is that proteins with increasing hydrophobicity and number of TMDs were identified in the integral membrane fraction while the proteins identified from the other two fractions are similar. These results match with the expectation that the proteins be fractionated according to their degree of association with the cell membrane – the very hydrophobic proteins are present in the integral membrane fraction.

A total of 3659 unique proteins identified from the cell extracts represent the most comprehensive proteome coverage of *E. coli*. *E. coli* is an important model organism for many biological studies. Proteome analysis of *E. coli* has been reported in a number of papers and the field of proteomics related to *E. coli* has been recently reviewed.[15-17] All kinds of proteome analysis methods including gel-based and solution-based techniques have been applied to the analysis of *E. coli* proteome.[15-17] For example, Hunt et al. identified a total of 1147 proteins from a membrane fraction of *E. coli* using 2D-LC MS/MS[18], which was, to our knowledge, the largest number of proteins identified in *E. coli* in the literature. Our proteome coverage is much greater; but about 641 predicted

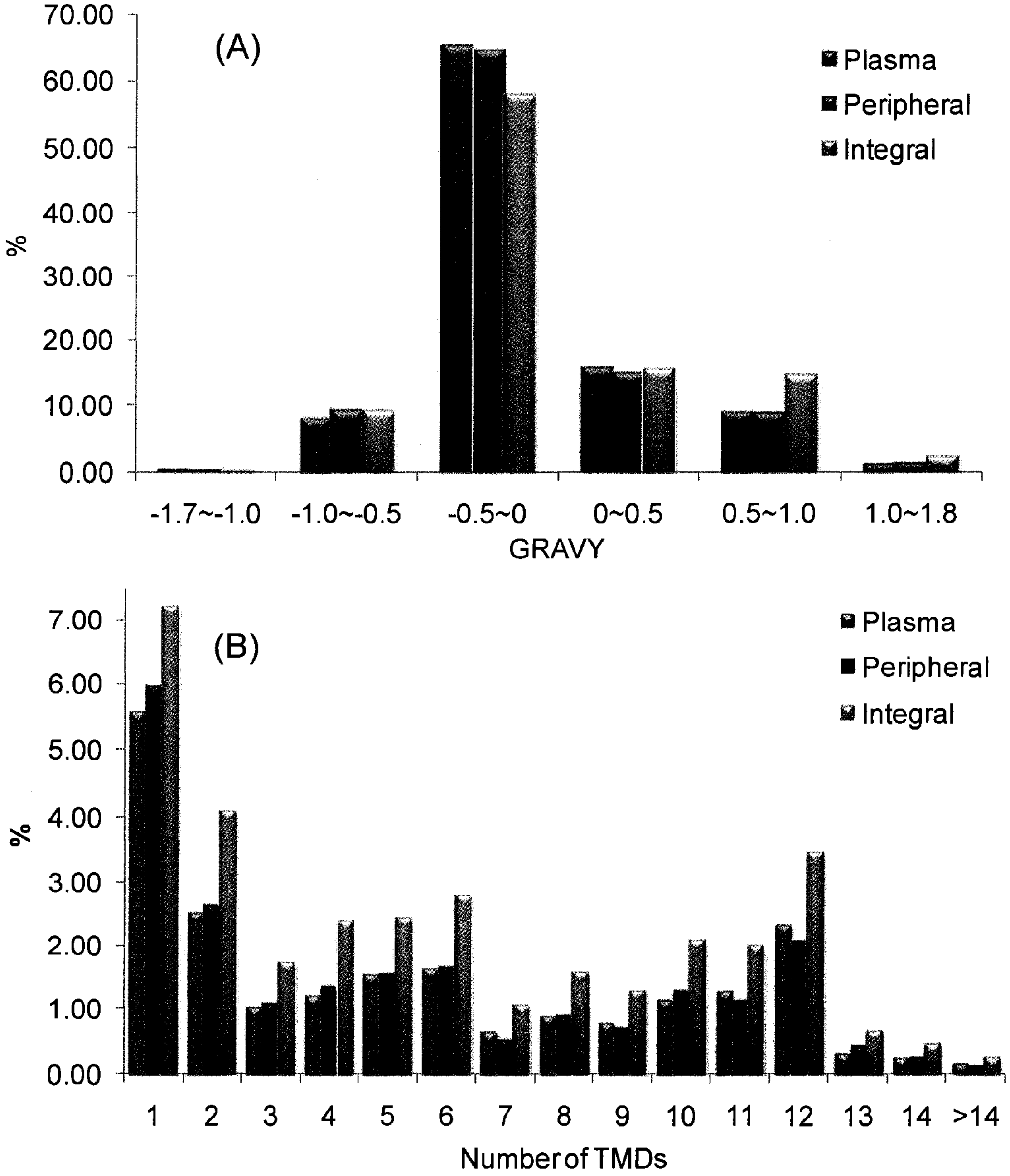
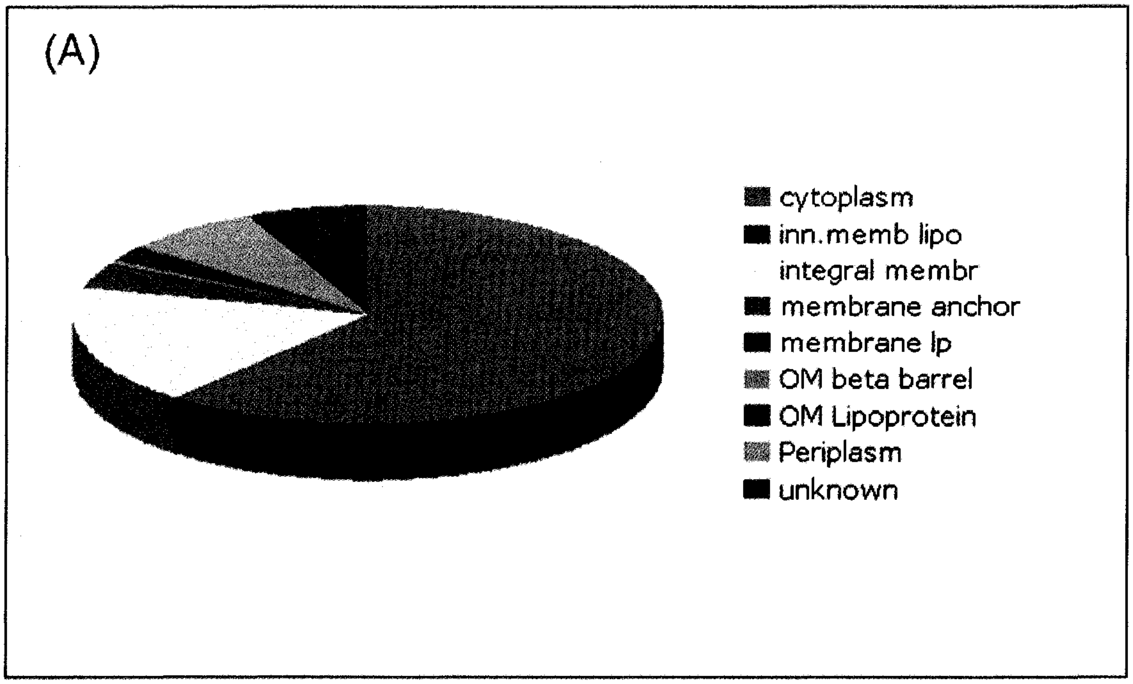


Figure 7.4 Distributions of the proteins identified in three fractions according to their (A) hydrophobicity and (B) number of transmembrane domains (TMDs).

proteins were still not found. The incomplete proteome may reflect the true state of the proteome in the cells cultured under the specific growth condition or the limitation of the applied technique. To gauge if there was any technical bias towards the detection of certain groups of proteins while under-detecting the others, we grouped the identified proteins against the predict proteins according to physical and chemical properties. For example, Figure 7.5 shows the cellular distributions of the identified proteins (3659) and the predict proteins (4300). There is no significant difference in the cellular distributions, indicating that the cellular proteins are identified with no apparent bias towards a certain cellular group or groups.

One surprising finding is shown in Figure 7.6. Figure 7.6A plots the number of proteins identified from the three fractions as well as the number of proteins predicted from the *E. coli* genome as a function of protein molecular weight (MW). An important observation is that the high MW proteins (MW>60 kDa) were well represented in the identified proteome, compared to the predicted proteome, while the low MW proteins were under-represented. This is also clearly shown in Figure 7.6B where the percentage of identified vs. predicted proteins within a MW window is plotted as a function of protein molecular weight. What is striking is that the very low MW proteins (<20 kDa) were severely under-represented: about 30% of the predicted 833 proteins in the MW range of 10-20 kDa were not identified and near 50% of the predicted 333 proteins with less than 10 kDa were not found.

To understand the potential source of this identification bias, we re-examined the workflow shown in Figure 7.2. In the cytoplasm or peripheral fraction, acetone precipitation was first applied to generate a protein pellet which was then washed to reduce impurities, followed by sequential protein solubilization and digestion. In dealing with many cell extract samples, we always find that acetone precipitation is very effective in precipitating proteins from cell extracts. However, it is possible that, during the acetone precipitation process, a fraction of the low MW proteins remains in solution while the majority of the high MW proteins are precipitated out. To find out whether this was the case for the *E. coli* sample, the supernatant from the cytoplasm fraction after acetone precipitation was analyzed by the shotgun method. After trypsin digestion, the



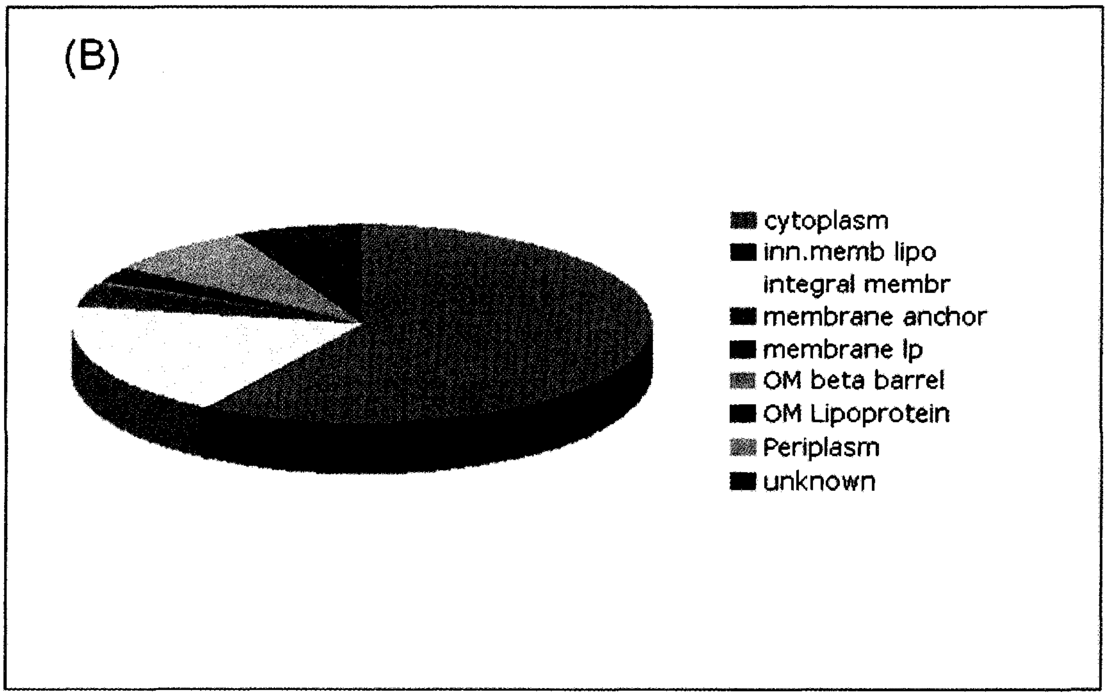


Figure 7.5 (A) Distribution of proteins identified in three fractions according to their cellular location and (B) Distribution of proteins annotated from the genome according to their cellular location.

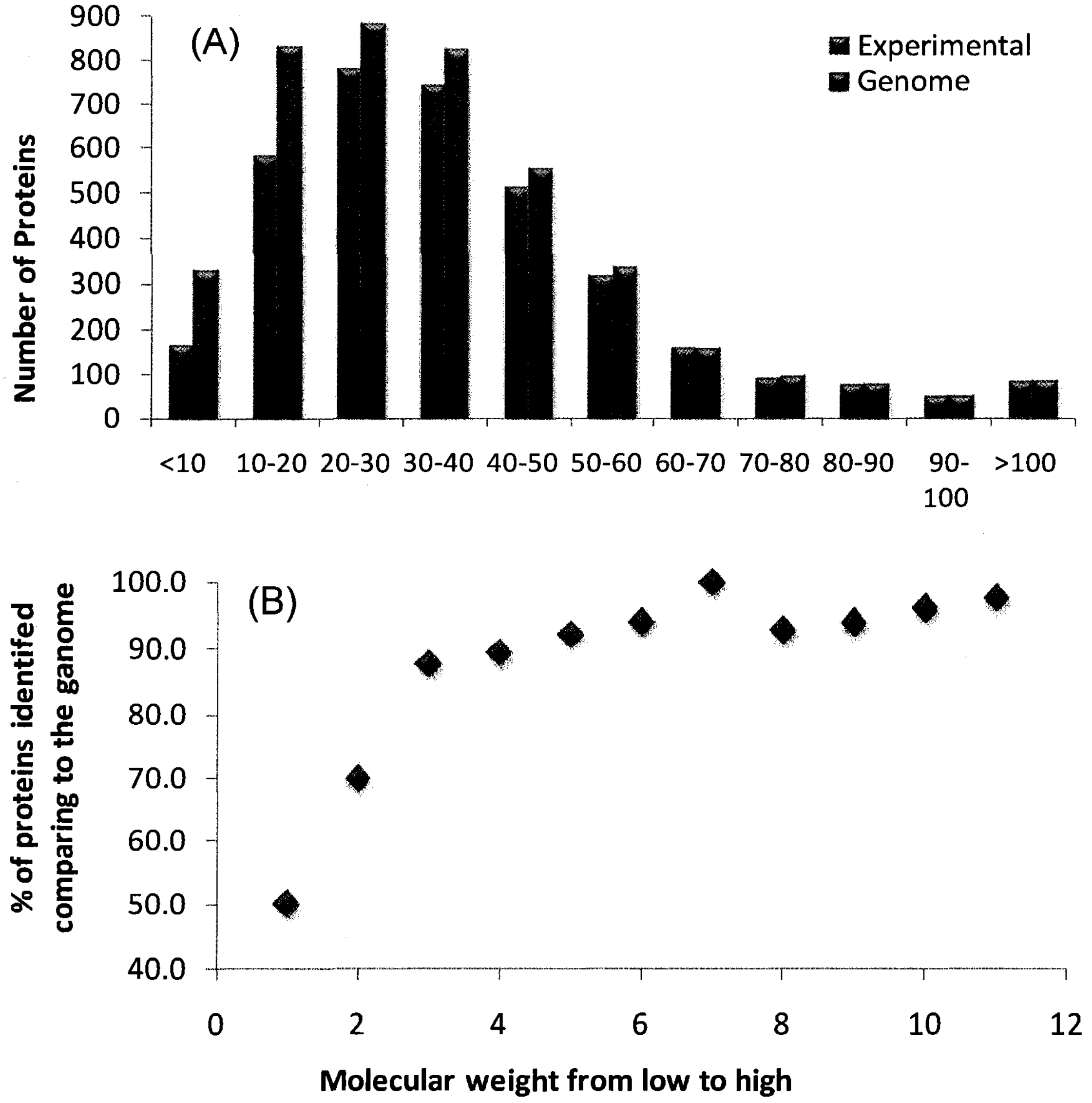


Figure 7.6 Distribution of (A) the identified proteins from the cellular protein fractions and the genome-predicted proteins as a function of protein molecular weight, (B) the percentage of proteins identified comparing to the genome as a function of molecular weight.

digest was desalted with concomitant measurement of peptide concentration by RPLC with UV detection. An optimal amount of the de-salted digest (i.e., 1 μg) was injected into the capillary LC-ESI QTOF instrument for peptide sequencing. Only 61 proteins were identified from this run. Compared to an average of about 300-500 proteins routinely identified from a peptide mixture with modest complexity (e.g., whole cell lysate digest or an SCX fraction from the NH_4HCO_3 sample of the cytoplasm fraction), this low number of proteins identified from a sample without SCX fractionation indicates that the protein composition in the supernatant was much less complex than the acetone-precipitated protein sample from the cytoplasm fraction. More importantly, 58 out of the 61 proteins identified from the supernatant have already been found in the acetone-precipitated pellet. Only 3 low MW proteins were detected uniquely to the supernatant.

The above result was disappointing from the point view of trying to detect the "missing" low MW proteins. But, it also raised an important question about the shotgun proteome analysis approach in general. Is the current approach biased towards the detection of high MW proteins in a proteome sample? To address this question, we attempted to fractionate the proteome sample by enriching the low MW proteins using molecular weight cutoff filters. Prior to applying this technique to the *E. coli* sample, a mixture of protein standards including insulin, ubiquitin, cytochrome c and BSA was used to test the efficiency of this method for enriching low MS proteins. It was found, by comparing the gel electrophoresis images of the protein mixture before and after applying 10 kDa or 30 kDa MW filtration, that this method was effective. We then applied this method to the *E. coli* sample using the workflow shown in Figure 7.7. In this case, the cytoplasm sample was combined with the peripheral membrane fraction. This combined solution was fractionated using two molecular weight cutoff filters into three samples: the 10 kDa sample containing proteins of less than 10 kDa, the 30 kDa sample containing proteins of less than 30 kDa, and the 10-30 kDa sample containing proteins with a MW range of 10 to 30 kDa. These three low MW samples were analyzed by the shotgun proteome analysis method.

Figure 7.8 shows the distribution of the numbers of proteins identified in the three low MW samples. A total of 2383 unique proteins were identified from the three

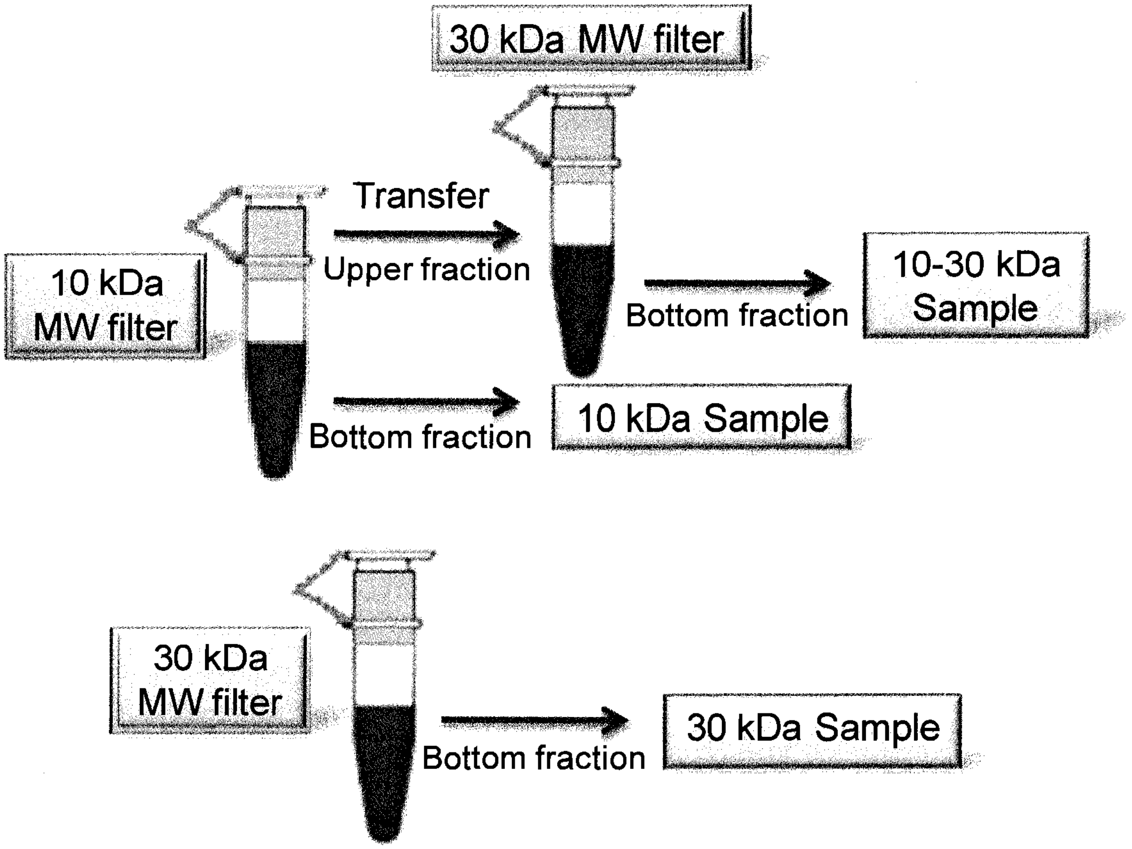


Figure 7.7 Workflow for enrichment of low molecular weight proteins from the *E. coli* cell lysate.

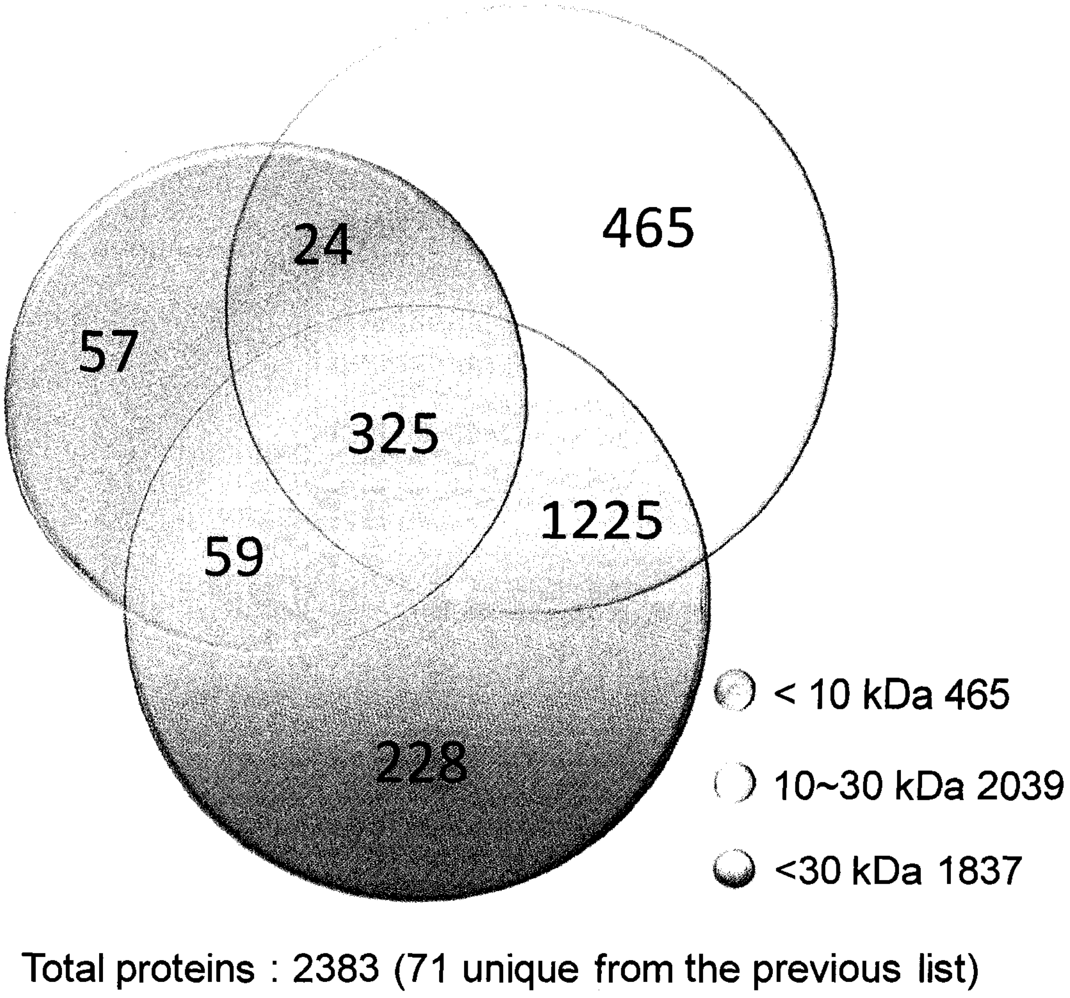


Figure 7.8 Venn diagram of the unique proteins identified from the low MW fractions.

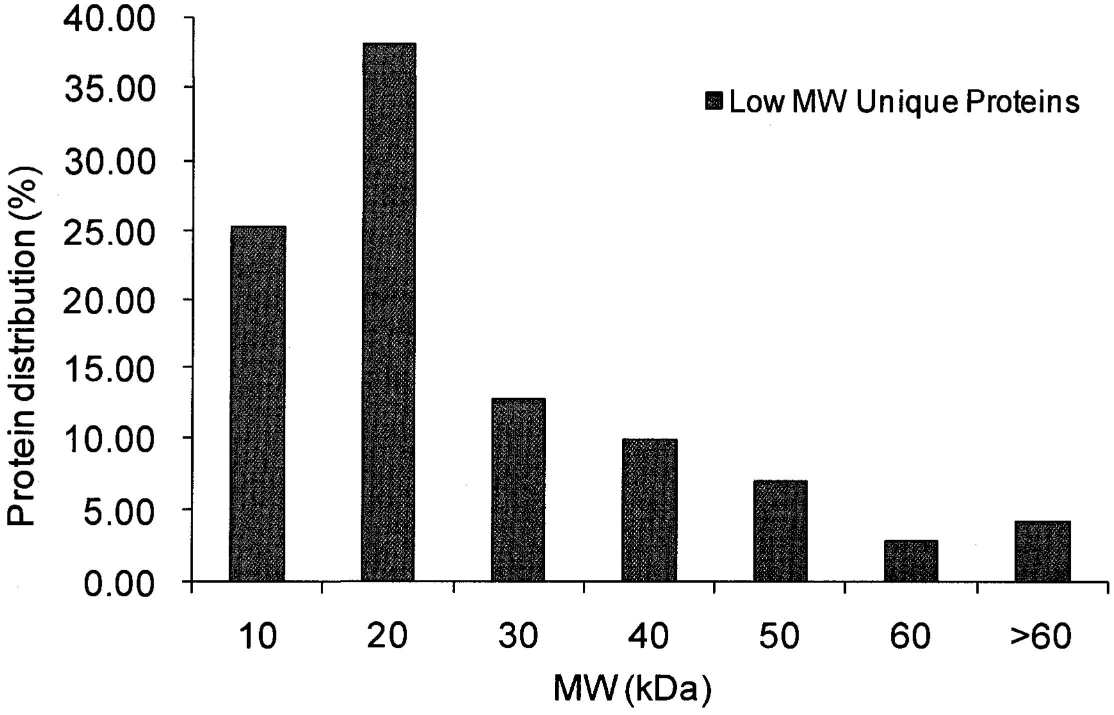


Figure 7.9 Distribution of the percentage of unique proteins identified in the low MW fractions.

fractions combined. However, only 71 new proteins were identified and all the remaining proteins have been previously identified in the cytoplasm, peripheral or integral membrane fractions. The distribution of these proteins as a function of molecular weight is shown in Figure 7.9. Comparing this distribution with that shown in Figure 7.6A, it is clear that a much greater proportion of low MW proteins were detected from the low MW samples. Some high MW proteins were still identified from these samples. This is not surprising considering that proteins of >30 kDa can still pass through a 30-kDa filter, albeit at a reduced efficiency. In addition, some high MW proteins may degrade in *vivo* or in *vitro* into smaller proteins which can pass through the low MW cutoff filters. Shotgun method, based on sequence match of one or a few peptides to a protein for protein identification, cannot readily differentiate a protein fragment from the intact protein.

Combining the 71 new proteins identified from the three low MW fractions with the 3659 proteins identified earlier, a total of 3730 unique proteins have been identified from the *E. coli* cells. This represents 86.7% proteome coverage of the 4300 predicted proteins. Figure 7.10 shows the distribution of 3730 proteins according to their molecular weights. It is clear that low MW proteins are still largely under-represented. Examining the sequences of some of 570 missing proteins, we did not find any unique characters of these missing proteins that prevent them from detection. For example, many proteins have A or K spaced in the amino acid sequence. One would expect trypsin digestion of these proteins generate a set of peptides detectable by MS/MS. Many of them are not hydrophobic so they should be readily amendable for the shotgun method. Thus, from the technical point of view, we believe that most of proteins present in the cells have been identified. The missing proteins are likely not present in the cells, i.e., they were not expressed under the culture conditions used to grow the cells. Additional experiments, such as the use of Western blot or RNA analysis, are planned to confirm whether these proteins are present in the cells grown using the rich media.

7.4 Conclusions

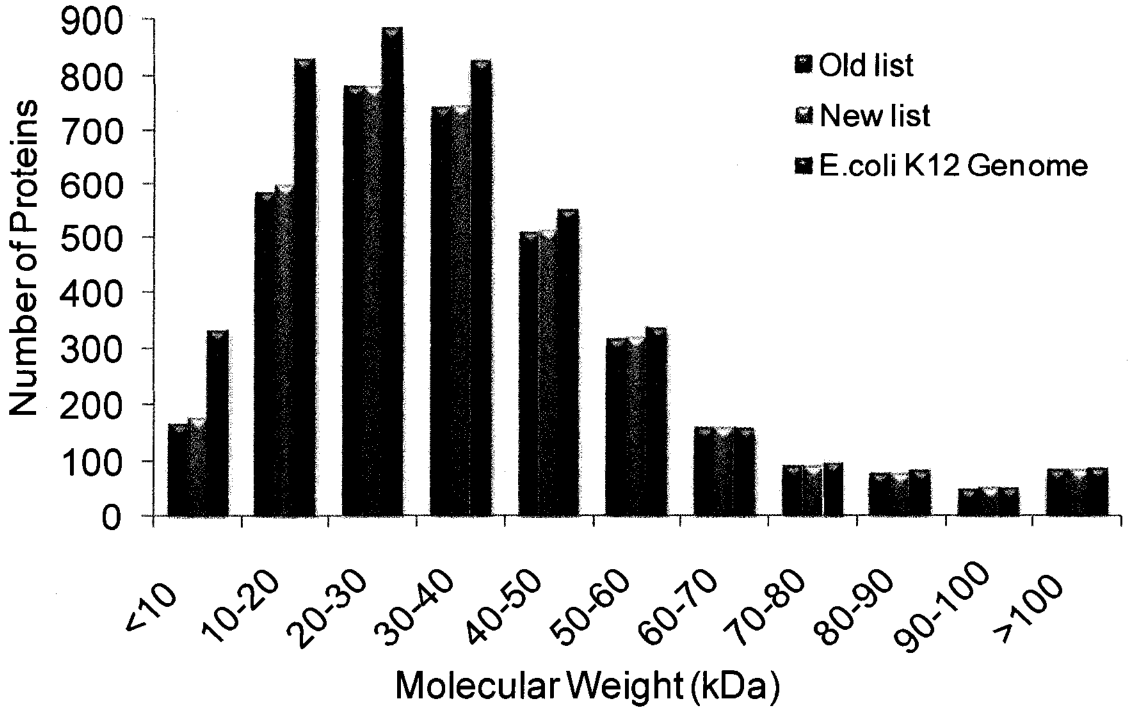


Figure 7.10 Distribution of all the identified proteins and genome-predicted proteins as a function of protein molecular weight.

For the first time, a comprehensive proteome profile of *E. coli* has been generated by using a combination of several techniques including cell lysate pre-fractionation, sequential protein solubilization and digestion, 2D-LC ESI MS/MS and proteome database search. A total of 3730 unique proteins or protein groups have been identified, representing 86.7% of the 4300 predicted proteins. Interestingly, 570 proteins not identified from this work are mainly relatively low molecular weight proteins (MW<60kDa). The percentage of missing proteins increases as the protein molecular weight decreases, with an extreme that about 50% of the predicted proteins with MW<10 kDa were not identified in this study. Whether these missing proteins were present in the cells grown in the rich media remains to be investigated. Nevertheless, this work illustrates that it is now possible to generate proteome coverage of about 86.7% for *E. coli.* Future work will be focusing on simplifying the experimental procedures to increase sample throughput for proteome analysis.

7.5 Literature Cited

(1) Zhang, J.; Yang, M.; Puyang, X.; Fang, Y.; Cook, L.; Dovichi, N. *Analytical Chemistry* **2001**, *73*, 1234-1239.

(2) Hudson, M. *Molecular Ecology Resources* **2008**, *8*, 3-17.

(3) Barbulovic-Nad, I.; Lucente, M.; Sun, Y.; Zhang, M.; Wheeler, A.; Bussmann, M. *Critical Reviews in Biotechnology* **2006**, *26*, 237-259.

(4) Ewis, A.; Zhelev, Z.; Bakalova, R.; Fukuoka, S.; Shinohara, Y.; Ishikawa, M.; Baba, Y. *Expert Review of Molecular Diagnostics* **2005**, *5*, 315-328.

(5) Hall, N. *Journal of Experimental Biology* **2007**, *210*, 1518-1525.

(6) Michels, E.; De Preter, K.; Van Roy, N.; Speleman, F. *Genetics in Medicine* **2007**, *9*, 574-584.

(7) Stoughton, R. *Annual Review of Biochemistry* **2005**, *74*, 53-82.

(8) Verducci, J.; Melfi, V.; Lin, S.; Wang, Z.; Roy, S.; Sen, C. *Physiological Genomics* **2006**, *25*, 355-363.

(9) Aebersold, R.; Goodlett, D. *Chemical Reviews* **2001**, *101*, 269-295.

(10) Loo, R.; Cavalcoli, J.; Vanbogelen, R.; Mitchell, C.; Loo, J.; Moldover, B.; Andrews, P. *Analytical Chemistry* **2001**, *73*, 4063-4070.

(11) Vijayendran, C.; Burgerneister, S.; Friehs, K.; Niehaus, K.; Flaschel, E. *Biochemical and Biophysical Research Communications* **2007**, *363*, 822-827.

(12) Watt, R.; Wang, J.; Leong, M.; Kung, H.; Cheah, K.; Liu, D.; Danchin, A.; Huang, J. *Nucleic Acids Research* **2007**, *35*, -.

(13) Maillet, I.; Berndt, P.; Malo, C.; Rodriguez, S.; Brunisholz, R.; Pragai, Z.; Arnold, S.; Langen, H.; Wyss, M. *Proteomics* **2007**, *7*, 1097-1106.

(14) Vanbogelen, R.; Abshire, K.; Moldover, B.; Olson, E.; Neidhardt, F. *Electrophoresis* **1997**, *18*, 1243-1251.

(15) Molloy, M.; Herbert, B.; Slade, M.; Rabilloud, T.; Nouwens, A.; Williams, K.; Gooley, A. *European Journal of Biochemistry* **2000**, *267*, 2871-2881.

(16) Anon *Molecular & Cellular Proteomics* **2005**, *4*, S487-S487.

(17) Han, M.; Lee, S. *Microbiology and Molecular Biology Reviews* **2006**, *70*, 362-439.

(18) Corbin, R.; Paliy, O.; Yang, F.; Shabanowitz, J.; Platt, M.; Lyons, C.; Root, K.; Mcauliffe, J.; Jordan, M.; Kustu, S.; Soupene, E.; Hunt, D. *Proceedings of The National Academy of Sciences of The United States of America* **2003**, *100*, 9232-9237.

Chapter 8

Development of a Mass Spectrometric Method for the Analysis of Proteomes from Thousands of Cancer Cells*

8.1. Introduction

As we have demonstrated in the previous chapters, current mass spectrometric technology can identify thousands of proteins from a proteome sample and is on the verge of becoming a powerful tool for mapping the entire proteome. This large scale proteome profiling work is generally carried out using hundreds of micrograms or milligrams of starting materials. To produce this quantity of sample, millions or billions of cells are used. For cultured cells, the use of this large number of cells is usually not a major issue. However, in many other studies, the number of cells available for proteome analysis can be quite limited. For example, in a tissue sample containing both normal and transformed (e.g., cancer) cells, the number of cancerous cells may be very limited.[1] This is particularly true for tissue samples from patients at an early stage of cancer development.[2, 3] Another example is the analysis of proteome from a small number of circulating cancerous cells in a blood sample of a patient with early sign of tumor in a specific organ.[4]

Our research goal is to develop new technologies that can generate as large of proteome coverage as possible from a small number of cells. Our initial target is to analyze the proteome of about 1000 cells. Adequate coverage of the proteome from this number of cells may lead to several important applications. For example, 1000 cells may be collected from a patient blood containing rare circulating cancerous cells from an early stage of metastasis of a solid tumor.[5-9] Analyzing the proteome of these cells may be used as a fingerprint for diagnosis or prognosis of a cancer. Another example is that about 1000 cells may be procured from a tissue section using laser capture microdissection

*Mingguo Xu contributed partially to sample preparation, data acquisition and data processing of this work. The flow cytometry work was done at the Cross Cancer Institute, Edmonton.

(LCM) within a couple of hours. Analyzing these cells may assist in identifying specific protein markers for disease diagnosis. The ultimate goal of this research is to analyze single cell proteome.[10] Unfortunately, this is a huge challenge at this moment for mass spectrometry based technologies due to limited sensitivity. Developing and applying techniques for analyzing the proteome of thousands of cells is a more realistic goal. However, very few studies of proteome analysis from a few thousands of cells have been reported.[11-15]

In this chapter, a shotgun proteome analysis method is described for analyzing proteomes of MCF-7 cells ranging from 500 to 5000 cells. MCF-7 cells, derived from breast cancer, are representative of many different types of cancerous cells in terms of size and proteome complexity. Thus, the method developed from analyzing MCF-7 cells should be applicable to other cancerous cells. The performance of this method in terms of the numbers of peptides and proteins identifiable from small numbers of cells is reported. This method is then applied to a model system where a small number of MCF-7 cells are added to a human blood to mimic a patient blood sample containing cancerous cells. These cells are captured by the combination of antibody attachment to the cells and flow cytometry for cell sorting. The captured cells are analyzed by the shotgun method.

8.2 Experimental

8.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), iodoacetamide, trifluoroacetic acid (TFA), sodium bicarbonate were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). Sequencing grade modified trypsin, HPLC grade formic acid, LC-MS grade water, acetone, and acetonitrile (ACN) were from Fisher Scientific Canada (Edmonton, Canada). The BCA assay kit was from Pierce (Rockford, IL).

8.2.2 Cell culture, Cell isolation from human blood sample and Cell sorting by flow cytometry

The MCF7 breast cancer cells (ATCC® number: HTB-22™) were cultured in 15 cm diameter plates at 37 °C in DMEM Gibco medium supplemented with 10% fetal bovine serum. The plates were then washed twice with ice-cold 25 mL PBS^{++} buffer (0.68 mM $CaCl_2$, 0.5 mM $MgCl_2$, 1.4 mM KH_2PO_4, 4.3 mM Na_2HPO_4, 2.7 mM KCl, and 137 mM NaCl). The cells were harvested by scraping from the plates into the PBS^{++} buffer and centrifugation at 100 *g* for 8 min at 4 °C. The cell numbers were first roughly counted by an Axiovert 25 hemocytometer (Carl Zeiss, Inc. Minneapolis, MN).

The fresh whole blood provided by a healthy donor was first diluted by PBS buffer (1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM KH_2PO_4, pH 7.4) in a 1:10 ratio (v:v). The MCF-7 human breast cancer cells (2 million) were then spiked into the diluted blood solution. Density separation was then conducted to remove red blood cells by using Ficoll-Hypaque (GE Healthcare) (See Figure 8.2). In brief, 10 mL diluted blood sample containing MCF-7 cells blood was slowly added into 4 mL Ficoll solution. The solution was spun down at 2 000 rpm, 4 °C for 20 min. Considering the density of MCF-7 cells, the cancer cells preferentially aggregated with peripheral blood leukocytes (PBL) at the layer called buffy coat after centrifugation. The buffy coat was isolated, washed and re-suspended in PBS buffer. Afterwards, the cell mixture, was incubated with a FITC–conjugated mouse anti–human HEA antibody (Miltenyi Biotec number: 130-080-301) in a 1:100 (v:v) ratio on ice for 15 min. Therefore, most MCF-7 cells were fluorescently stained, while PBL were not.

Both the unstained MCF-7 cells and the stained cell mixtures were introduced into the flow cytometer (Beckman Coulter EPICS Altra) for counting, according to the cell size and their fluorescence response. Then 500, 1000, 2500 or 5000 MCF-7 cells were collected into 0.6 mL low retention microcentrifuge vials (Fisher Scientific).

8.2.3 Protein extraction, Purification and Trypsin digestion

The cells in each vial were mixed with 5 to 10 μL Nonidet-P40 (NP40) lysis buffer (1%) and sonicated in ice-water ultrasonic bath for 5 min. The protein solutions were then reduced with 20 mM (0.4 to 0.75 μL) dithiothreitol (DTT) and alkylated with the same volume of 40 mM iodoacetamide. Acetone (precooled to -80 °C) was added gradually (with intermittent vortexing) to the protein extract to a final concentration of 80% (v/v). The solution was then incubated at -20 °C for 60 minutes and centrifuged at 14 000 rpm

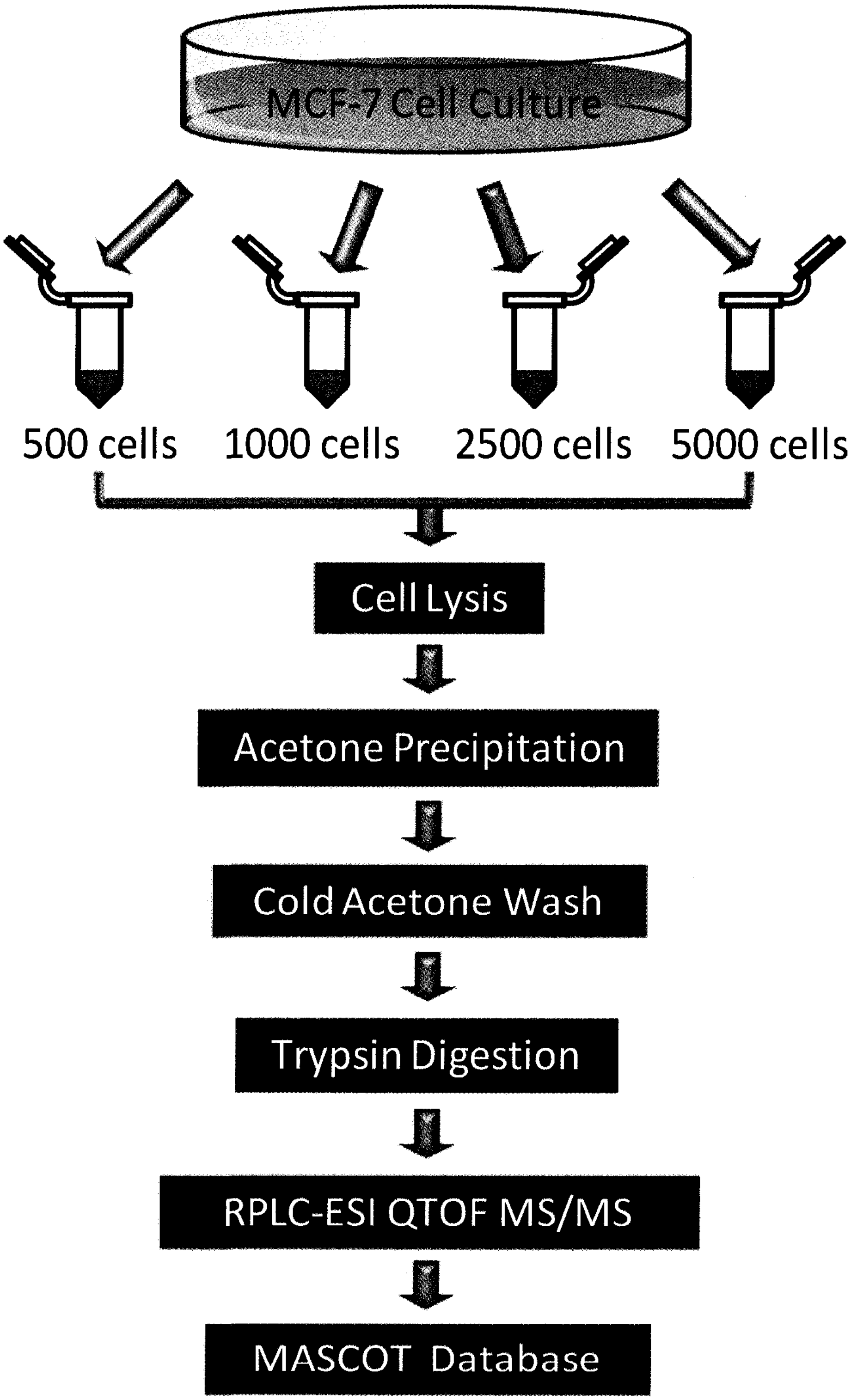


Figure 8.1 Workflow for both method development and application.

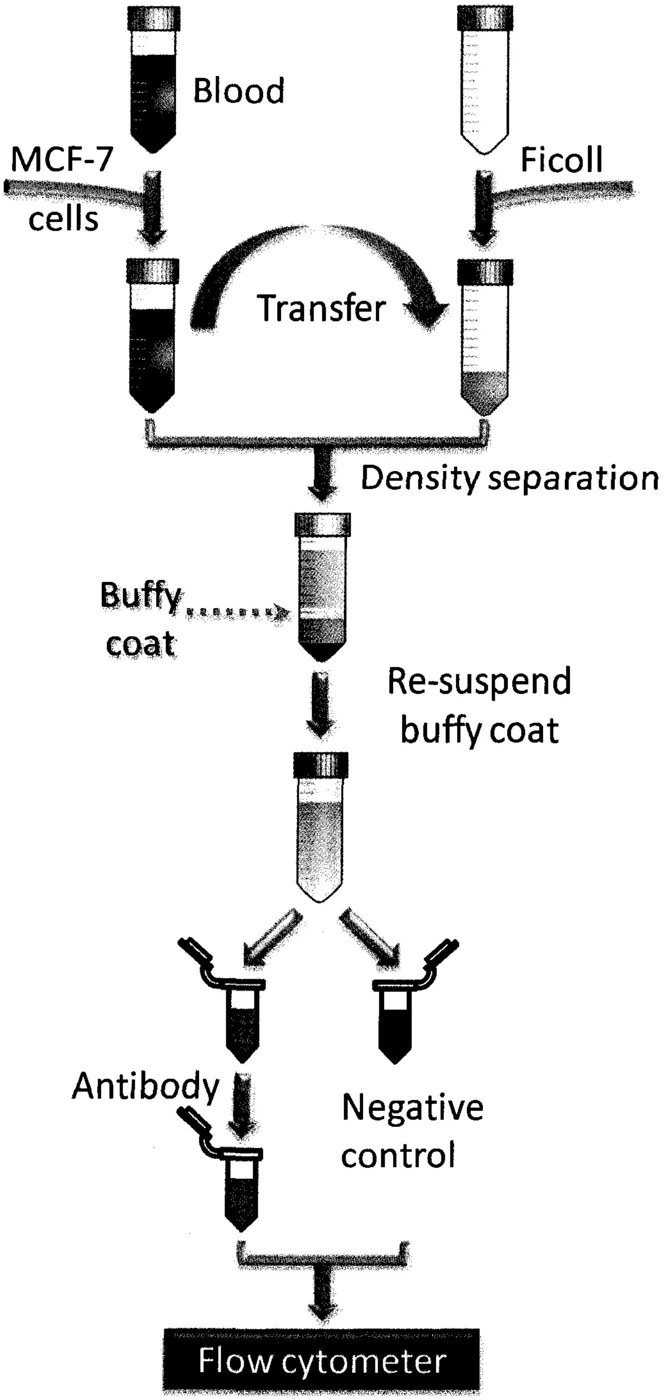


Figure 8.2 Workflow for the enrichment of MCF-7 cells in a blood sample.

for 10 min. The supernatant was decanted. The pellet was carefully washed once using cold acetone to ensure the efficient removal of NP40 detergent (See Figure 8.1). The residual acetone was evaporated at ambient temperature. 50 mM ammonium bicarbonate was used to sufficiently redissolve the pellet in the vial. Trypsin digestion was then carried out in a final enzyme concentration of 8 ng/μl (5 to 20 uL) at 37 °C for 4 hours.

8.2.4 Peptide Desalting and Quantification by RPLC

The desalting and quantification setup consisted of an Agilent 1100 HPLC system (Palo Alto, CA) with a UV detector. The desalting of tryptic peptides was performed on a 4.6 mm × 50 mm Polaris C18 A column with 3 μm particle and 300 Å pore (Varian, CA). After loading all the digests of each sample, the column was flushed at 1 mL / min with 97.5% mobile phase A (0.1% TFA in water) for 3 min and then 85% of mobile phase B (0.1% TFA in acetonitrile) for 5 min to ensure the complete elution of the peptide fractions from the column.

8.2.5 LC-ESI QTOF MS and MS/MS Analysis

The desalted digests were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, the desalted and quantified digests were concentrated using a SpeedVac (Thermo Savant, Milford, MA) to ~1 μl and reconstituted to a specific concentration using 0.1% formic acid. Then the intended amount of digest solution was injected onto a 75 μm × 100 mm Atlantis dC18 column (Waters, Milford, MA). For the digests from 500 and 1000 cells, multiple injections were applied for each sample to make sure the loading of maximum amount of peptides. Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were separated using their optimal lengths of solvent gradients ranging from 90 min to 270 min and electrosprayed into the mass spectrometer fitted with a nanoLockSpray source at a flow rate of 300 nL/min. A survey MS scan was acquired from m/z 350-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from m/z 50-1900 for 0.8 s each. A mixture of leucine enkephalin and (Glu1)-fibrinopeptide B, used as mass calibrants (i.e., lock-mass), was infused at a flow rate of 300 nL/min, and a 1 s MS scan was acquired every 1 min throughout the run.

8.2.6 Protein Database Search

Raw LC-ESI data were lock-mass corrected, de-isotoped, and converted to peak list files by using ProteinLynx Global Server 2.2.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program (version 1.8). Database searching was restricted to *Homo sapiens* (human) in the SWISSPROT database (October 4, 2007) and 17317 entries were searched. The following search parameters were selected for all database searching: enzyme, trypsin; missed cleavages, 1; peptide tolerance, 30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); fixed modification, Carbamidomethyl (C); variable modifications, acetyl (Protein), oxidation (M), pyro-Glu (N-term Q) and pyro-Glu (N-term E). The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files using in-house software. All the identified peptides with scores lower than the MASCOT threshold score for identity at a confidence level of 95% were then removed from the protein list. The redundant peptides for different protein identities were deleted, and the redundant proteins identified under the same gene name but different access ID numbers were also removed from the list.

To gauge the false positive peptide matching rate in our analysis, we applied the target-decoy search strategy (see Chapter 5) by searching the MS/MS spectra against the forward and reversed human proteome sequences.

8.3 Results and Discussion

Shotgun proteome analysis is a relatively sensitive technique, compared to other methods such as gel-based proteome analysis. For example, as illustrated in other chapters, about 1 μg of a cell extract digest injected to LC-ESI MS/MS can result in the identification of about 300 to 500 proteins. In the shotgun method, the sample workup process includes cell lysis, protein extraction, protein digestion and injection of peptides into the LC-MS/MS system for analysis. Any one of the steps can potentially involve the loss of some proteins. In working with a large quantity of samples, this sample loss may not be very significant so long as the sample loss is not biased towards a particular group of proteins. If a bias (i.e., selective sample loss) does occur, that group of proteins will be under-represented in the final results. If the sample loss is un-biased, as long as we have sufficient amounts of peptides in the end for LC-MS/MS analysis (e.g., 1 μg per injection), the same proteome coverage would be expected. However, in handling small numbers of cells, sample loss of any type can be detrimental to the proteome coverage. The reason is that the amount of sample generated from a small number of cells will be limited and it will often not meet the optimal sample amount required for peptide sequencing in LC-MS/MS (e.g., < 1 μg). As it was demonstrated in Chapter 6, the amount of sample injection is very important in determining the outcome of peptide and protein identification. Injection of a smaller amount of sample results in less number of peptides and proteins identified.

With the above considerations in mind, we developed a sample analysis protocol as shown in Figure 8.1. The cultured MCF-7 cells were sorted into tubes containing different numbers of cells using a flow cytometer. The cells were lysed using a lysis solution containing NP-40 detergent (see Experimental section for details). This lysis step was chosen after we examined several reagents including SDS, acid labile surfactant (ALS) from Waters, Tris and water. It was found that NP-40 lysis was the most efficient for thousands of cells; the efficiency was gauged by measuring the protein amounts using BCA assay. Gel electrophoresis was also used to monitor the protein composition during the cell lysis optimization experiments. However, one major problem encountered in using this polyethylene glycol based detergent for cell lysis was that, after acetone precipitation of proteins from the lysate, the pellet still contained a small amount of NP-40, causing severe interference in LC-ESI MS/MS analysis of the cell lysate protein digest. To eliminate this interference, the pellet was also washed three times with cold acetone. This simple step was found to be very effective in reducing the NP-40 content to a level that did not cause interference in LC-ESI MS/MS. As Figure 8.1 shows, the cold-acetone washed pellet was dissolved in NH_4HCO_3, followed by trypsin digestion.

The tryptic digest was desalted, quantified and then injected into the LC-ESI QTOF instrument for MS/MS sequencing of the peptides.

The amount of peptides produced from a cell lysate was determined using the LC-UV system as described in Chapter 6. The average amount (n=3) of peptides from the 5000-cell sample was found to be 1.40±0.12 μg. And the average amount of the 2500-cell sample was 0.83±0.12 μg, which is not exactly half of the amount of peptides produced from the 5000-cell sample. But, within the experimental errors, the amount of peptides produced appears to proportionally decrease as the cell number decreases. If this proportionality held true for the 1000- or 500-cell sample, then the amount of peptides produced would be less than 0.28 μg for the 1000-cell sample and 0.14 μg for the 500-cell sample. The lower limit of the UV-LC system used to measure the peptide concentration is about 0.25 μg (see Chapter 6). We attempted to measure the peptide amounts for the 1000- and 500-cell samples and the results were not reliable as they generated UV signals with intensities similar to that of the blank. The failure to quantify the 1000-cell sample suggests that the amount of peptides produced from this sample must be less than 0.25 μg. Thus, sample loss may be more severe for these two samples, compared to the 2500- or 5000-cell sample. For future work, a simple and accurate quantification method to determine nanograms of peptides or proteins in each step of the workflow shown in Figure 8.1 should facilitate the optimization process. One approach is to modify our current LC-UV system using a capillary column, instead of a 1 mm column, to shift the linear calibration curve to the nanogram region. Work along this direction is planned for the future.

Besides sample preparation, optimization of LC-ESI MS/MS is also critical in analyzing samples of a few cells. Using a relatively dilute solution and performing multiple injection to minimize the amount of sample remaining in the vial after injection, we can introduce about 90% of the sample to the column for analysis. After sample injection, peptides are separated by a solvent gradient optimized for chromatographic resolution. However, the gradient speed can significantly affect the detectability of peptides in LC-MS/MS. If a fast gradient is used, a peptide elutes quickly to form a fast rising peak in an ion chromatogram, resulting in intense signals in both MS and MS/MS spectra. But, in this case, only a few MS and MS/MS spectra can be acquired within the peak elution time. If a slow gradient is used, the same peptide would elute out more slowly to form a broader peak and the mass spectral signal of the peptide would be less intense. If a sufficient amount of sample is injected, the peptide signal intensity may be adequate to generate a database-searchable MS/MS spectrum. One major advantage of using a slow gradient for peptide elution is that a greater number of MS and MS/MS spectra can be acquired over this broad peak. For the analysis of a complex peptide sample, co-elution of different peptides cannot be avoided and one always tries to sequence as many co-eluting peptides as possible; slow gradient provides this opportunity. However, if the amount of sample injected is small, the peptide signal may not be sufficiently intense to produce a database-searchable MS/MS spectrum. Thus, the gradient speed needs to be optimized according to the sample amount injected to the LC-MS/MS instrument. We have investigated how the gradient speed affects the number of peptides identified by LC-ESI MS/MS.

Figure 8.3 shows a plot of the number of peptides identified as a function of gradient time used to elute the peptides from a capillary reversed-phase LC-ESI MS/MS. In this study, a peptide sample was first prepared from 50,000 MCF-7 cells using the workflow shown in Figure 8.1. This sample was then diluted to produce a peptide sample with a calculated amount equivalent to that of a smaller number of cells (i.e., 500, 1000, 2500, or 5000 cells). To avoid confusion, we refer these samples prepared by dilution of the 50,000-cell sample (i.e., stock solution) as aliquoted samples. For example, a 5000-cell aliquoted sample refers to a sample prepared by diluting 10-fold of the peptide digest from 50,000 cells, while a 500-cell sample refers to a sample prepared from 500 cells as the starting material. As Figure 8.3 illustrates, the optimal gradient time or speed is different for the aliquoted samples of 500-, 1000-, 2500-, and 5000-cell. Generally speaking, the optimum gradient time increases as the number of cells in a sample increases. In addition, within a group of samples (e.g., the 500-cell aliquoted samples), there is an optimal gradient time for detecting peptides. Too long of a gradient can result in the identification of fewer peptides. Thus, for the subsequent experiments, the gradient time was adjusted according to the number of cells used for proteome analysis. Specifically, for the 500-cell samples, a 90-min gradient was used. The

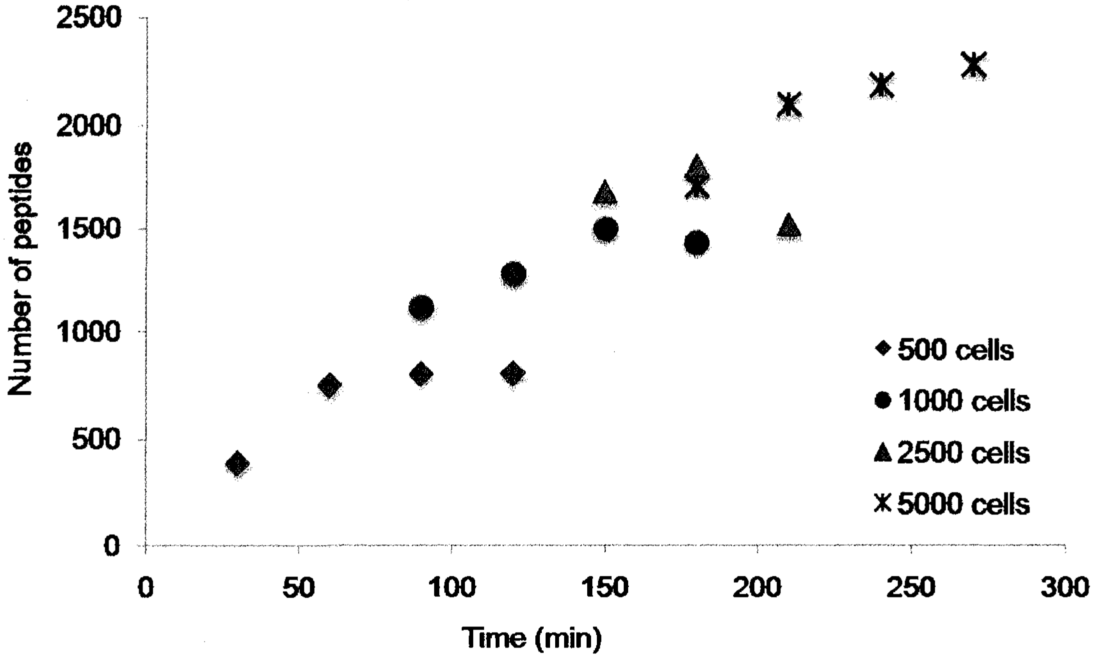


Figure 8.3 Number of peptides identified as a function of gradient time in RPLC-MS analysis of different numbers of cells.

gradient time was increased to 150 min for the 1000-cell samples. The gradient time was 180 min for the 2500-cell samples and 270 min for the 5000-cell samples. Table 8.1 summarizes the peptide and protein identification results from the samples of 500-, 1000-, 2500- and 5000-cell. In each group, three replicate experiments were carried out. The numbers of peptides and proteins identified from these samples are plotted in Figure 8.4. As Table 8.1 and Figure 8.4 show, both the numbers of peptides and proteins increase as the cell number increases and the number change is not in linear proportion to cell numbers. For example, an average of 1891±266 peptides or 619±59 proteins (n=3) were identified from the 5000-cell sample, while 381±11 or 167±21 proteins were identified from the 500-cell sample. Although the cell number decreases by 10-fold, the number of peptides and proteins identified decreases by only about 5.0- and 3.7-fold, respectively. However, the peptide/protein ratio decreases from 3.05 for the 5000-cell sample to 2.14 for the 500-cell sample.

The above results indicate that we can identify an average of 167 proteins from 500 cells, 237 proteins from 1000 cells, 491 proteins from 2500 cells, and 619 proteins from 5000 cells. In all cases, the run-to-run reproducibility was good, indicating that the experimental protocol used in this study can be used to generate reproducible results from as few as 500 cells. However, comparison of the number of peptides identified in the aliquoted samples (see Figure 8.3) with those directly prepared from the small number of cells (see Table 8.1 or Figure 8.4) indicates that further optimization in the experimental protocol, particularly the sample preparation process, may identify even a greater number of peptides and proteins. Figure 8.3 shows that the number of peptides identified was about 2278, 1807, 1496, or 798 peptides for the aliquoted 5000-, 2500-, 1000- or 500-cell sample, respectively. In comparison, an average of 1891±266, 1308±251, 513±53, or 381±11 peptides was identified from the 5000-, 2500-, 1000-, or 500-cell sample. Except the 5000-cell sample, the number of peptides identified from the other samples is significantly less than those of the corresponding aliquoted samples. Since the number of peptides identified in a sample is related to the amount of sample, it is reasonable to conclude that the amount of peptides injected to LC-MS/MS from the sample of 2500, 1000 or 500 cells was less than that of the corresponding aliquoted sample. This finding indicates that there was some sample loss during the sample preparation process. This

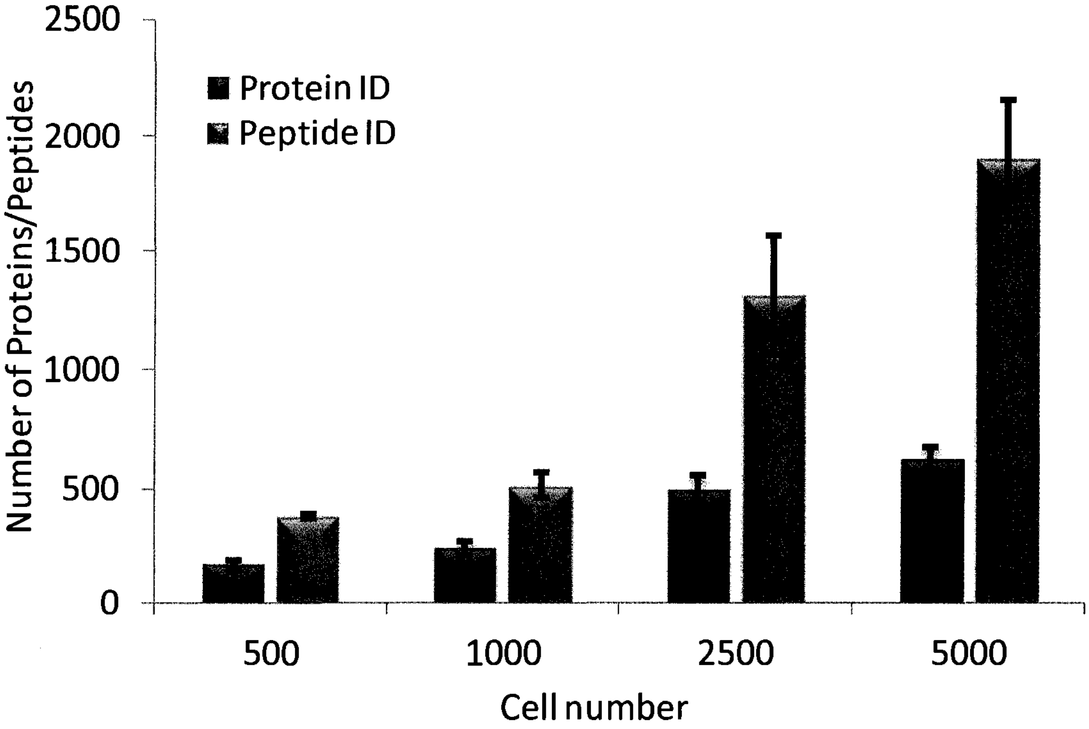


Figure 8.4 Protein and peptide identification results under optimized sample preparation and LC-MS/MS conditions.

notion was also suggested from the LC-UV peptide quantification results as discussed earlier. It should be noted that, since sample loss did occur, the gradient speed used to analyzing the 1000- or 500-cell sample might not be optimal – a faster gradient might result somewhat better results. This needs to be confirmed in the future.

While further optimization of the protocol will likely increase the number of peptides and proteins identified, the ability of detecting hundreds of proteins from as few as 500 cells using the current protocol opens the possibility of studying the proteome of a small number of cells. One potential area of application is to use a proteome profile as a signature or fingerprint to identify a specific type of cancer cells in the human blood for cancer diagnosis. The hypothesis is that, when tumor metastasis starts from a specific organ, the cancer cells entered into the blood stream will have a similar proteome profile to that of the cancer cells found in the organ. Tissue biopsy allows procurement of tissue samples from which the cancer cells can be isolated using antibody recognition combined with either LCM from the tissue sections or flow cytometry after dissolving the connecting tissue to release the cells. The circulating cancer cells from the patient blood sample can also be isolated using antibody recognition combined with flow cytometry.[4, 6, 7, 14, 16] The proteome profiles of cancer cells from the blood and tissue samples will be compared to determine whether they are the same type and, if so, the metastasis site is positively identified. It should also be noted that future work on proteome profiling of different types of cancer cells may result in a proteome profile database from which the proteome profile generated from the cancer cells isolated from the blood may be directly compared to cancer diagnosis, without the need of taking a tissue sample from a patient. This non-invasive method would also be useful for cancer prognosis.

To mimic the above scenario, we used a model system where MCF-7 cells were spiked to normal human blood, followed by isolation of these cells using antibody recognition and flow cytometry. The proteome profile of the isolated cells was then generated by the method described above and compared to those of the MCF-7 cell lines. The entire workflow for the isolation of the MCF-7 cells in blood is shown in Figure 8.2 and has been described in the Experimental section. Figure 8.5 shows the numbers of peptides and proteins identified from different numbers of cells isolated from the blood

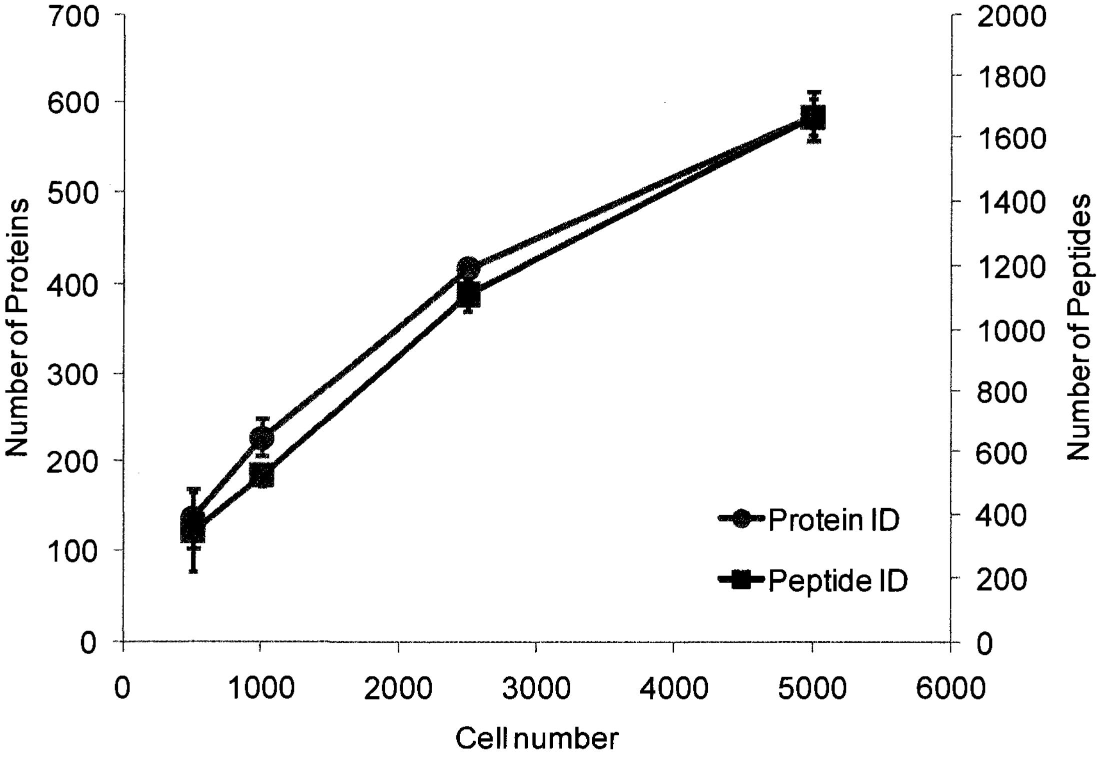


Figure 8.5 Protein and peptide identification results of MCF-7 cells isolated from a blood sample.

samples. The numbers are very similar to those obtained from the samples prepared directly from the cultured cells. Moreover, the proteome profiles are very similar, judging from the common proteins obtained from the two comparative samples (see Table 8.2). In Table 8.2, the results of intra- and inter-sample comparison (i.e., percent of common proteins found in two samples) are listed. For example, in the case of 500 cells, three replicate experiments were carried for the 500-cell samples (Table 8.2 refers them as A, B, and C). Likewise, three replicate experiments were done for the 500-cell samples from blood spiked with MCF-7 cells (Table 8.2 refers then as A', B' and C'). Within the dataset of A, B, and C, the average percent of common proteins found in two samples is 57.0%±10.2%. For the A', B' and C' samples, the average is 64.6%±10.6%. The average common protein percentage from the comparison of A vs. A', B vs. B', and C vs. C' is 60.3%±13.5%. The difference of these data is not significant. Thus, these proteome profiles are considered to be indistinguishable. This example illustrates that it is possible to generate a proteome profile from as few as 500 cells isolated from a blood sample and the proteome profile may be used for cell typing.

8.4 Conclusions

A shotgun proteome analysis method has been developed for protein identification from thousands of cells. This method is based on the use of a detergent (NP-40) to lyse the cells, followed by acetone precipitation. After washing the pellet with cold acetone to remove any residual detergent, the pellet was dissolved in NH_4HCO_3 and the solubilized proteins were subjected to trypsin digestion. By optimizing the sample volume, about 90% of the digest solution was injected into a capillary LC-ESI MS/MS system for analysis. The resultant MS/MS spectra were searched against a proteome database for protein identification. In analyzing the MCF-7 cells, this method was demonstrated to be capable of identifying an average of 167±21, 237±30, 491±63, and 619±59 proteins from 500, 1000, 2500, and 5000 cells, respectively. This method was then applied to the analysis of proteome profiles of small numbers of cells isolated from a blood sample spiked with the MCF-7 cells. It was shown that the proteome profiles generated from the cells isolated in the blood sample were similar to those of the MCF-7 cells. We envisage that this method will be useful in proteome profiling of small numbers of cells for disease diagnosis and prognosis. In addition, further optimization in the sample preparation process to reduce sample loss may result in identification of even more proteins from thousands of cells.

Table 8.1 Unique proteins and peptides identified from different numbers of cells.

Number of Cells	Unique Peptides	Unique Proteins
500	369	168
	386	187
	389	145
1000	574	271
	485	226
	481	215
2500	1036	422
	1531	546
	1358	504
5000	1630	552
	2161	665
	1883	640

Table 8.2 Summary of protein identification reproducibility from different runs.

	500 cells			1000 cells			2500 cells			5000 cells		
Sample	Overlap (%)*		Average	Overlap (%)*		Average	Overlap (%)*		Average	Overlap (%)*		Average
A&B	64	58	57±10	59	70	63±6	72	78	72±7	74	76	71±6
B&C	57	73		60	63		64	77		63	74	
C&A	41	48		56	70		60	78		64	77	
A'&B'	49	75	65±11	69	58	60±5	69	70	69±2	72	75	72±3
B'&C'	73	67		54	61		71	70		72	76	
C'&A'	51	72		59	56		67	66		67	73	
A&A'	48	47	60±14	51	66	58 ±5	57	74	68±7	66	72	72±4
B&B'	49	82		59	53		62	76		68	73	
C&C'	62	73		61	60		70	70		77	77	

*Percent of common proteins found in two comparative runs. A, B, and C refer to the samples of three replicate experiments from the MCF-7 cells. A', B', C' refer to the samples of three replicate experiments from the cells isolated from blood spiked with the MCF-7 cells.

8.5 Literature Cited

(1) Espina, V.; Wulfkuhle, J.; Calvert, V.; Vanmeter, A.; Zhou, W.; Coukos, G.; Geho, D.; Petricoin, E.; Liotta, L. *Nature Protocols* **2006**, *1*, 586-603.

(2) Hutter, G.; Sinha, P. *Proteomics* **2001**, *1*, 1233-1248.

(3) Ladanyi, A.; Sipos, F.; Szoke, D.; Galamb, O.; Molnar, B.; Tulassay, Z. *Cytometry Part A* **2006**, *69a*, 947-960.

(4) De Roos, B.; Duthie, S.; Polley, A.; Mulholland, F.; Bouwman, F.; Heim, C.; Rucklidge, G.; Johnson, I.; Mariman, E.; Daniel, H.; Elliott, R. *Journal of Proteome Research* **2008**, *7*, 2280-2290.

(5) Schneider, T.; Moore, L.; Jing, Y.; Haam, S.; Williams, P.; Fleischman, A.; Roy, S.; Chalmers, J.; Zborowski, M. *Journal of Biochemical and Biophysical Methods* **2006**, *68*, 1-21.

(6) Swerts, K.; Ambros, P.; Brouzes, C.; Navarro, J.; Gross, N.; Rampling, D.; Schumacher-Kuckelkorn, R.; Sementa, A.; Ladenstein, R.; Beiske, K. *Journal of Histochemistry & Cytochemistry* **2005**, *53*, 1433-1440.

(7) Allan, A.; Vantyghem, S.; Tuck, A.; Chambers, A.; Chin-Yee, I.; Keeney, M. *Cytometry Part A* **2005**, *65a*, 4-14.

(8) Chosy, E.; Nakamura, M.; Melnik, K.; Comella, K.; Lasky, L.; Zborowski, M.; Chalmers, J. *Biotechnology and Bioengineering* **2003**, *82*, 340-351.

(9) Utz, P. *Immunological Reviews* **2005**, *204*, 264-282.

(10) Harwood, M.; Christians, E.; Fazal, M.; Dovichi, N. *Journal of Chromatography A* **2006**, *1130*, 190-194.

(11) Umar, A.; Luider, T.; Foekens, J.; Pasa-Tolic, L. *Proteomics* **2007**, *7*, 323-329.

(12) Sitek, B.; Sipos, B.; Schulenborg, T.; Marcus, K.; Schmiegel, W.; Hahn, S.; Kloppel, G.; Meyer, H.; Stuhler, K. *Molecular & Cellular Proteomics* **2006**, *5*, S147-S147.

(13) Marko-Varga, G.; Berglund, M.; Malmstrom, J.; Lindberg, H.; Fehniger, T. *Electrophoresis* **2003**, *24*, 3800-3805.

(14) Seshi, B. *Proteomics* **2007**, *7*, 1984-1999.

(15) Wang, H.; Qian, W.; Mottaz, H.; Clauss, T.; Anderson, D.; Moore, R.; Camp, D.; Khan, A.; Sforza, D.; Pallavicini, M.; Smith, D. *Journal of Proteome Research* **2005**, *4*, 2397-2403.

(16) Righetti, P.; Castagna, A.; Antonucci, F.; Piubelli, C.; Cecconi, D.; Campostrini, N.; Rustichelli, C.; Antonioli, P.; Zanusso, G.; Monaco, S.; Lomas, L.; Boschetti, E. *Clinica Chimica Acta* **2005**, *357*, 123-139.

Chapter 9

Microwave-assisted Acid Hydrolysis Combined with Liquid Chromatography Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry for Mapping Protein Sequences

9.1 Introduction

Protein sequence mapping is different from protein identification. Sequence mapping is commonly used to study post-translational modifications of a protein or amino acid substitutions from point mutations in the genome. Ideally, the entire amino acid sequence of a protein should be mapped to pinpoint where a modified amino acid or a substitution is located. Understanding protein modification is very important in studying biological functions of a protein. For example, protein phosphorylation and de-phosphorylation plays an essential role in cell signaling.[1-4] Determination of the phosphorylation site(s) of a protein can often provide the vital information for understanding the signaling process.[2]

Mass spectrometry (MS) has become an indispensable tool for protein sequence mapping. This is commonly done by using a top-down or a bottom-up proteomic approach.[5-8] In the top-down method, a protein ion is dissociated in a tandem mass spectrometer (MS/MS) and the fragment ions generated are interpreted to generate a stretch of amino acid sequence information which can then be used to search a proteome database for protein identification. Spectral interpretation can be automated;[8] but quite often manual interpretation is required as the fragment ion spectrum is often very complex. Sequence coverage by this method is dependent on the nature of the protein, ranging from a few residues to a full sequence.[6] In general, full sequence information is difficult to obtain for proteins with molecular weights of above 20,000 Da. The bottom-up method described in the previous chapters is a robust method for protein identification based on sequencing one or more peptides generated from chemical or enzymatic degradation of a protein by MS/MS.[9] For a digest of a protein such as that produced by trypsin digestion, one or a few peptides are sequenced, resulting in only a partial coverage of the protein sequence. To increase the sequence coverage of a protein by the bottom-up method, multiple enzyme or chemical degradation experiments can be done to generate complementary sequence information.[10] Of course, this is a time consuming process and there is no guarantee that the peptides produced from the multiple digestions will cover the entire sequence of a protein. In some cases, the combination of top-down and bottom-up methods is used to increase sequence coverage.[6]

Recently, our laboratory has developed an alternative MS technique for protein sequence mapping. It is based on the use of microwave-assisted acid hydrolysis (MAAH) to degrade the protein into peptides, followed by MS analysis.[11, 12] When a strong acid such as 6 N HCl is used for MAAH and a microwave irradiation time is kept at less than 2 min, the hydrolysate generated from a protein consists of mainly terminal peptides.[11] Analysis of the peptide mixture by matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) MS results in a spectrum composed of peaks from both the N- and C-terminal peptides. Deconvolution of these peaks into two series of peptide peaks and determination of the mass difference from adjacent peaks of N- or C-terminal peptide series (i.e., sequence ladders) allow us to map the protein sequence. A complete sequence can be read from a protein with a molecular weight of up to about 18,000 Da (human hemoglobin). The sensitivity of MALDI-TOF at the high mass region determines the upper mass limit of a protein that can be fully sequenced. Even with state-of-the-art TOF instruments, the detection sensitivity at >18,000 Da is not sufficient to generate any signals from a protein hydrolysate. However, for higher mass proteins, partial sequences from the N- and C-terminus can still be obtained. One major shortcoming of this method is that it requires a relatively pure sample, i.e., the protein to be sequenced must be present at >80% in a mixture.

If HCl is replaced by 25% trifluoroacetic acid (TFA) for MAAH and the irradiation time is increased to about 8 to 10 min, proteins can be degraded into small peptides.[12] These small peptides have molecular weights of up to about 3000 Da which are ideal sizes for MS/MS using collision-induced dissociation (CID) in a tandem mass spectrometer. These peptides consist of both terminal and internal peptides and, thus, unambiguous sequence ladder information cannot be obtained. However, LC-MS/MS analysis of the hydrolysate, followed by proteome database search, can result in identification of these peptides. In the work by Zhong et al, it was shown that the sequence of bacteriorhodopsin (MW 24000) can be mapped with 95% coverage.[12] This original work was carried out using LC-MALDI MS/MS in a quadrupole time-of-flight (QTOF) mass spectrometer.

Recent advances in LC electrospray ionization (ESI) QTOF, as described in Chapters 3-5, have resulted in a highly sensitive method for sequencing peptides by MS/MS. In this work, we report the combination of TFA MAAH with LC-ESI QTOF MS/MS and demonstrate that it is a facile technique for protein sequence mapping. TFA MAAH is done using an inexpensive household microwave oven and highly reproducible results can be obtained. We demonstrate that, using one-dimensional LC-ESI MS/MS for the analysis of MAAH hydrolysates, several protein standards including α-casein (a mixture S1 and S2 forms), β-casein, bovine serum albumin (BSA) (66,500 Da) can be sequence-mapped with almost complete coverage.

9.2 Experimental

9.2.1 Chemicals and Reagents

Dithiolthreitol (DTT), trifluoroacetic acid (TFA), LC-MS grade formic acid, bovine serum albumin, α-casein and β-casein protein standards were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). LC-MS grade water, acetonitrile (ACN) and the micro centrifuge tube holders were from Fisher Scientific Canada (Edmonton, Canada).

9.2.2 Microwave acid hydrolysis

10 μL (1 μg/μL) of the protein standard stock solution was mixed with an equal volume of 20 mM DTT in a 1.5 mL polypropylene centrifuge vial and incubated at 60 °C for 20 min. 20 μL 50% TFA was added to the sample solution after incubation. The vial

was then capped, sealed with Teflon tape and then placed in a domestic 900W (2450 MHz) microwave oven (Panasonic). For our traditional microwave method, 100 mL of water in a loosely covered container was placed besides the sample vial to absorb extra microwave energy. For the new and improved microwave method, the sample vial was placed on a Scienceware* round bubble rack (Fisher Scientific, Edmonton, Canada) and floated in a plastic beaker which contained 100 ml water (see Figure 9.1). The beaker was placed in the centre of the rotating plate in the microwave oven. For the location dependence study (see Results and Discussion), the plastic beaker was placed off-center in the microwave in order to make sure the positions selected on the bubble rack were randomly distributed in the microwave oven. The volume of the sample including the acid was limited so that the relatively large sample vial could tolerate the vapor pressure produced when the samples was microwave irradiated. After microwave irradiation for a period indicated in the Results and Discussion, the sample vial was taken from the microwave and the solution was dried in a SpeedVac to remove the acid. The protein digest was re-suspended in 50 μL of 0.1% formic acid aqueous solution and centrifuged at 14 000 rpm for 5 min to remove any possible residual particles. A portion of the hydrolysate was injected into the LC-MS/MS system for separation and sequencing.

9.2.3 LC-ESI QTOF MS and MS/MS

The hydrolysates from the protein standards were analyzed using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). In brief, 5 μL of peptide solution was injected onto a 75 μm × 100 mm Atlantis dC18 column with 3 μm particle size (Waters, Milford, MA). Solvent A consisted of 0.1% formic acid in water, and Solvent B consisted of 0.1% formic acid in ACN. Peptides were first separated using 120 min gradients (2-6% Solvent B for 2 min, 6-25% Solvent B for 95 min, 30-50% Solvent B for 10 min, 50-90% Solvent B for 10 min, 90-5% Solvent B for 5 min; column was pre-equilibrated at 2% Solvent B for 20 min) and electrosprayed into the mass spectrometer (fitted with a nanoLockSpray source) at a flow rate of 350 nL/min. Mass spectra were acquired from *m/z* 300-1600 for 0.8 s, followed by 4 data-dependent MS/MS scans from *m/z* 50-1900 for 0.8 s each. The collision energy used to perform MS/MS was varied

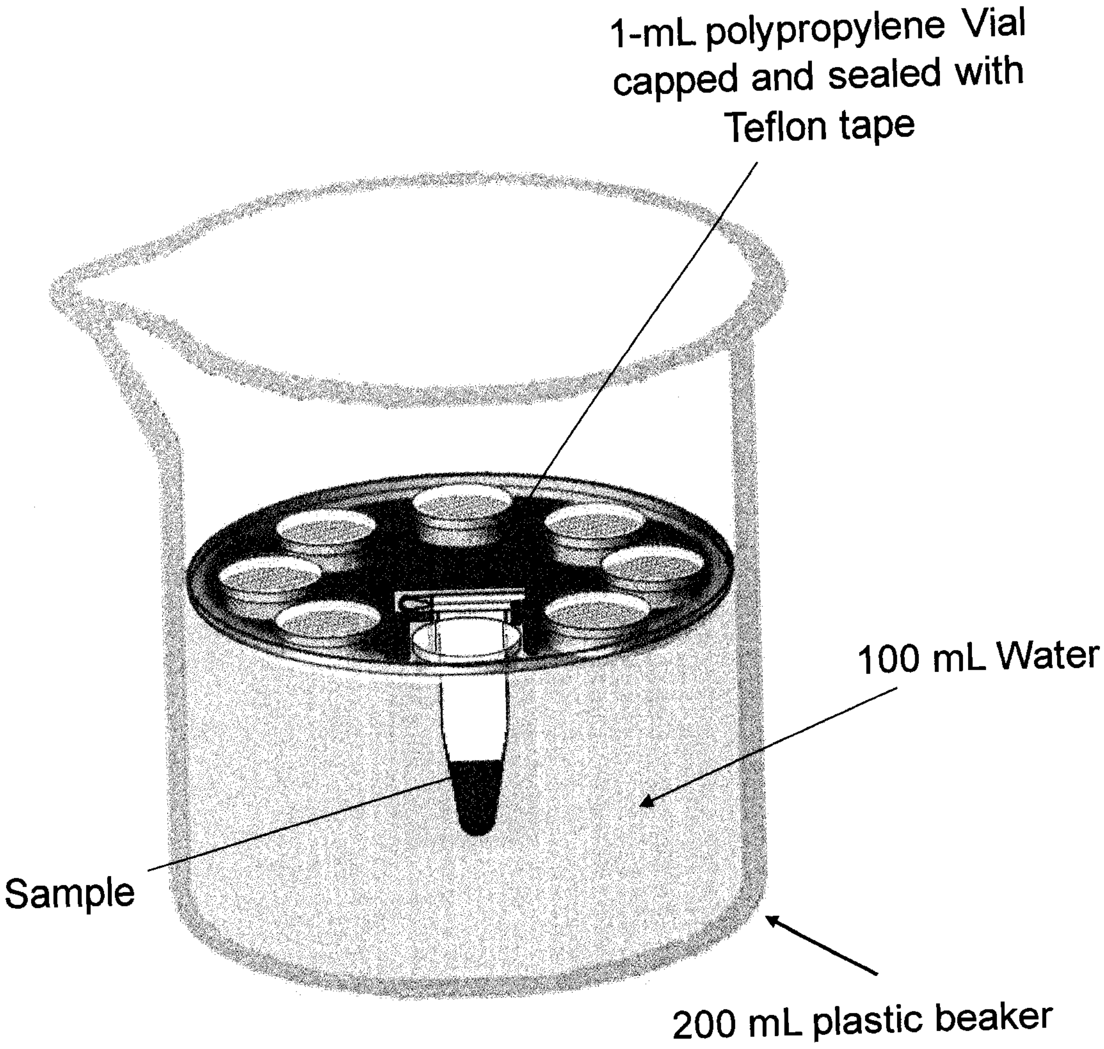


Figure 9.1 Schematic diagram of the experimental setup for microwave-assisted acid hydrolysis.

according to the mass and charge state of the eluting peptide. Leucine Enkephalin and (Glu1)-Fibrinopeptide B, a mixed mass calibrant (i.e., lock-mass), was infused at a rate of 300 nL/min, and an MS scan was acquired for 1 s every 1 min throughout the run.

9.2.4 Protein database search

Raw MS and MS/MS data were lock-mass-corrected, de-isotoped, and converted to peak list files by ProteinLynx Global Server 2.2.5 (Waters). Peptide sequences were identified via automated database searching of peak list files using the MASCOT search program (http://www.matrixscience.com). Database search was restricted to the protein sequences of the corresponding protein standards downloaded from the SwissProt database. The following search parameters were selected for all database searching: enzyme, non-specified; missed cleavages, 1; peptide tolerance, ±30 ppm; MS/MS tolerance, 0.2 Da; peptide charge, (1+, 2+, and 3+); variable modifications, oxidation (M), deamidation of asparagine and glutamine. The search results, including protein names, access IDs, molecular mass, unique peptide sequences, ion score, MASCOT threshold score for identity, calculated molecular mass of the peptide, and the difference (error) between the experimental and calculated masses were extracted to Excel files. All the identified peptides with scores lower than the MASCOT identity threshold scores for identity were then deleted from the protein list.

9.3 Results and Discussion

For TFA MAAH, a household microwave oven was used. Such a device is inexpensive and readily available. However, one major concern in using the oven was whether it could generate reproducible results for mass spectrometric analysis. LC-ESI MS/MS is a sensitive technique for sequencing peptides and, thus, small variations of the peptide composition in the hydrolysates generated from the TFA MAAH process can affect the final results. Household microwave ovens are known to have "hot" spots where the microwave irradiation is unevenly distributed.[13, 14] To examine the effect of sample location inside the microwave oven on data reproducibility, we placed four vials of a BSA solution containing TFA at different locations on a sample rack in the rotating plate (see Figure 9.2A for illustration) and irradiated the samples for 10 min. Each

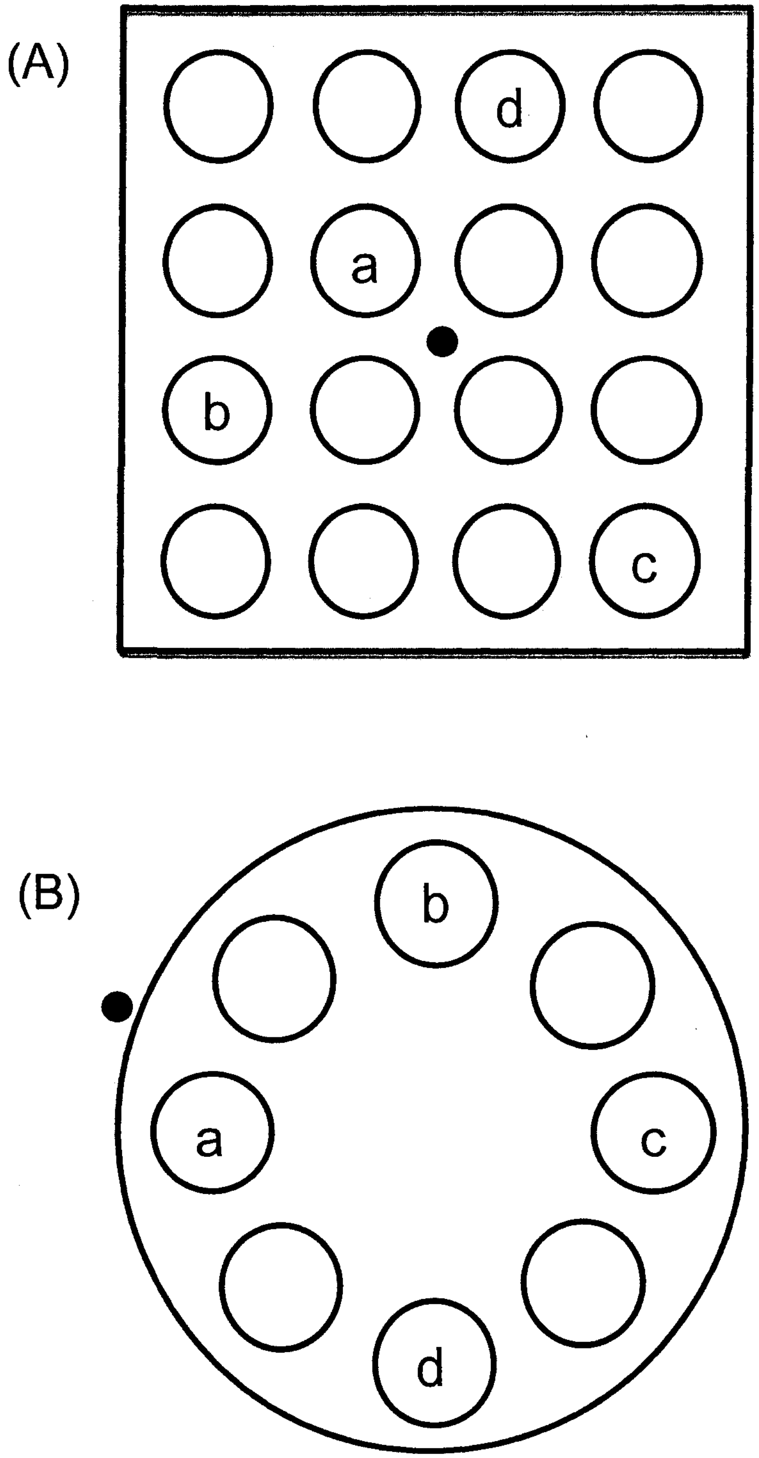


Figure 9.2 Diagram of the micro centrifuge tube positions tested. "." indicates the position corresponding to the centre of the rotating plate in the microwave oven.

hydrolysate was analyzed by LC-ESI QTOF MS and the resulting chromatograms are shown in Figure 9.3. As Figure 9.3 shows, the ion chromatograms generated are somewhat different, indicating that sample location inside the oven affects the results.

The location dependence as shown in Figure 9.3 is most likely related to uneven heating of the vials by the microwave irradiation. Note that, in this traditional microwave experiment, about 100-mL water was placed in a beaker inside the oven to absorbs most of the radiation. The sample vial containing about 40 μL of solution only absorb a small fraction of the energy. We speculate that the solution temperatures of the different vials during the microwave irradiation process might be different. Thus, to create a more uniform heating, we placed the sample vial inside a beaker filled with water (see the schematic diagram shown in Figure 9.1). The relatively larger volume of water in the beaker was heated to boiling during the microwave process. The sample solution placed inside a sealed sample vial appeared not to be boiled to a great extent, as the majority of the solution still remained at the bottom of the sample vial after the microwave experiment. It is likely that the pressure inside the vial might be greater than 1 atm, thus increasing the boiling point of the solution. Figure 9.4 shows the LC-ESI ion chromatograms of the BSA hydrolysates produced by using this new microwave heating method. The ion chromatograms from the samples of different locations were almost identical, indicating that the sample heating was uniform in this new setup. We also found that day-to-day reproducibility was very good using this method. From this work, we can conclude that acid hydrolysis can be reproducibly carried out using an inexpensive household microwave oven with the new setup shown in Figure 9.1.

The sequence coverage of the hydrolysate on BSA was investigated. Figure 9.5 shows the BSA amino acid sequence including the signal peptide of the protein. As Figure 9.5 shows, the entire sequence of the mature form of BSA (the whole sequence minus the signal peptide) was covered by the peptides detected from the LC-ESI MS/MS run of the hydrolysate. In our initial MS/MS search using the SwissProt database, we missed one amino acid, residue 214, which was listed as A (alanine) in this database. This was quite odd, as all other amino acids were covered by the peptides detected. It

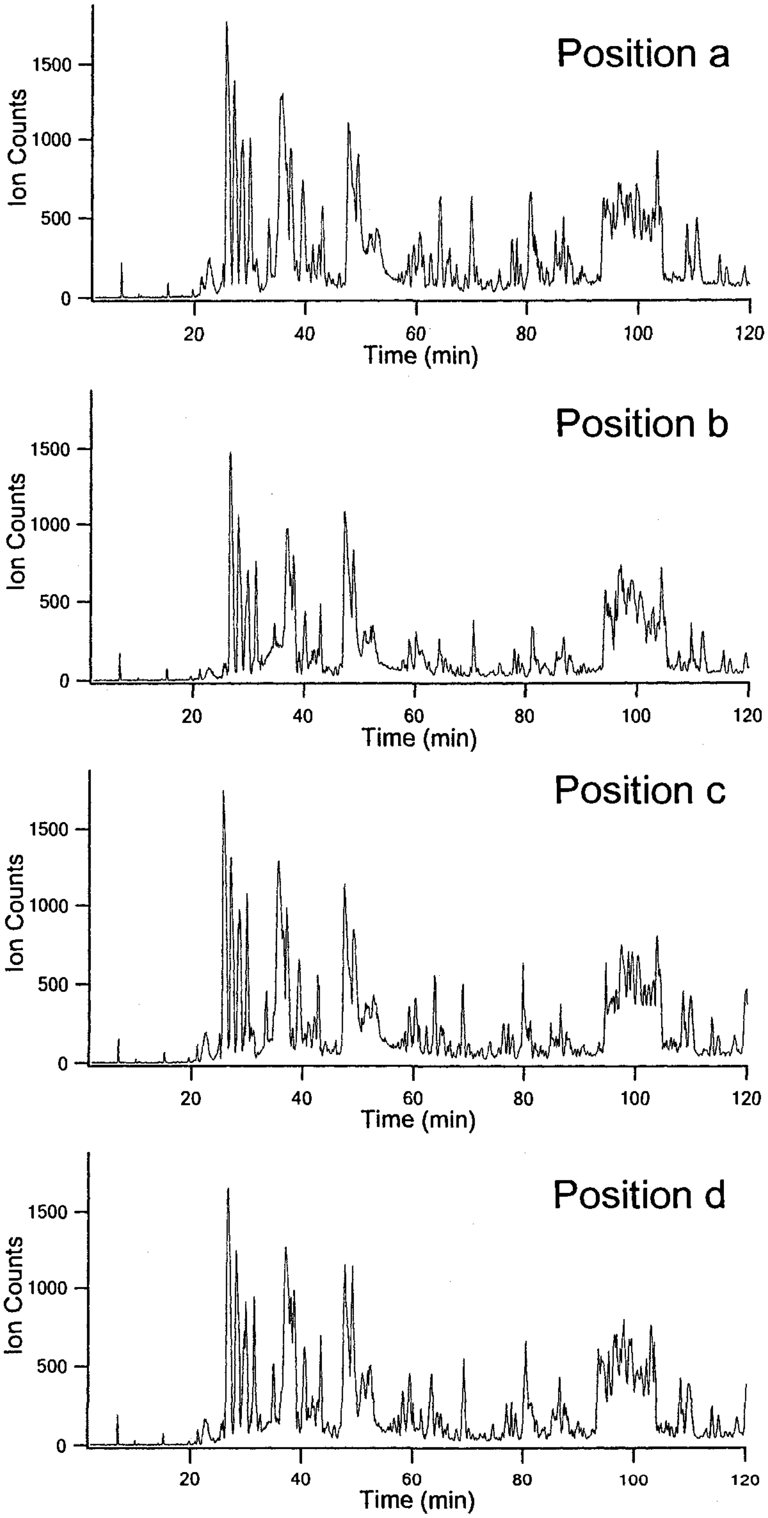


Figure 9.3. LC-ESI MS ion chromatograms of BSA hydrolysates obtained from the samples prepared using the traditional microwave heating method at four different locations on the rotating plate of the microwave oven (see Figure 9.2A for the corresponding locations). The irradiation time was 10 min.

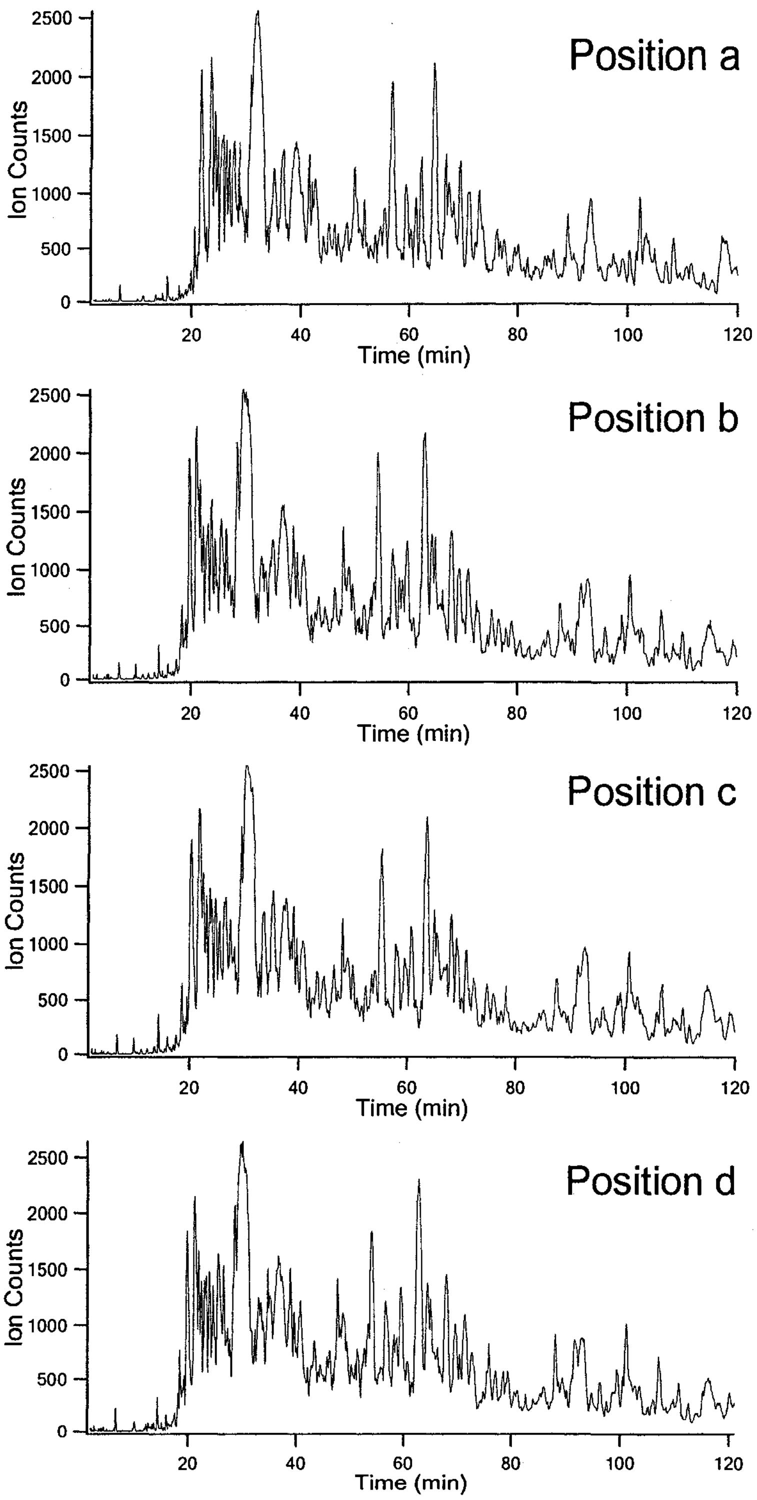


Figure 9.4 LC-ESI MS ion chromatograms of BSA hydrolysates obtained from the samples prepared using the improved microwave heating method at four different locations on the rotating plate of the microwave oven (see Figure 9.2B for the corresponding locations). The irradiation time was 10 min.

(P02769) Bovine Serum Albumin

MKWVTFISLLLLFSSAYSRGVFRR**DTHKSEIAHRFKDLGEEHFKGL**

VLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHT

LFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDL

PKLKPDPNTLCDEFKADEKKFWGKYLYEIARRHPYFYAPELLYYA

NKYNGVFQECCQAEDKGACLLPKIETMREKVLTSSARQRLRCASI

QKFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKECCH

GDLLECADDRADLAKYICDNQDTISSKLKECCDKPLLEKSHCIAEV

EKDAIPENLPPLTADFAEDKDVCKNYQEAKDAFLGSFLYEYSRRH

PEYAVSVLLRLAKEYEATLEECCAKDDPHACYSTVFDKLKHLVDE

PQNLIKQNCDQFEKLGEYGFQNALIVRYTRKVPQVSTPTLVEVSRS

LGKVGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVT

KCCTESLVNRRPCFSALTPDETYVPKAFDEKLFTFHADICTLPDTE

KQIKKQTALVELLKHKPKATEEQLKTVMENFVAFVDKCCAADDKE

ACFAVEGPKLVVSTQTALA

------ Signal peptide

Letter in bold black indicates the sequence covered

Letter underlined by "=" indicates natural variant site

Figure 9.5 Amino acid sequence of BSA including the signal peptide underlined with a dashed line. The letters in bold indicate the sequence covered by the peptides detected from the LC-ESI MS/MS analysis of the BSA hydrolysates. The natural variant site, residue 214, is indicated by "="

turned out that this amino acid has been documented to be T (threonine), a natural variant site (see SwissProt database, P02769 (ALBU_BOVIN)). In our experiment, we positively confirmed that this sample was BSA with T in residue 214. This is best illustrated in the three MS/MS spectra shown in Figure 9.6. MASCOT database search using the BSA sequence with T-214 resulted in the matches of the mass spectra to three peptide sequences shown in Figure 9.6. In contrast, when T was replaced by A, no matching was found. Comparison of the three MS/MS spectra shown in Figure 9.6 also provides unambiguous sequence assignment to the last few amino acids in the C-terminus. For example, the fragment ions, y11 in Figure 9.6A, y12 in Figure 9.6B and y13 in Figure 9.6C, belong to the same series that together indicate the extension of the sequence from ACLLPKIETMREKVL, to ACLLPKIETMREKVLT, to ACLLPKIETMREKVLTS.

There were a total of 1292 peptides (669 of them were unique) detected from the BSA hydrolysate using LC-ESI MS/MS. The peptides detected and their sequences along with MASCOT scores are listed in Table 9.1. These peptides are mainly from the internal peptides produced during the MAAH process. Although many of the peptides do not contain arginine or lysine at the C-terminus as in the case of tryptic peptides, these peptide ions can be readily dissociated, generating good quality MS/MS spectra for database search. For this dataset, we do not see any significant bias towards the detection of peptides with a particular amino acid at either C or N terminus. Thus, it appears that there is no specificity in bond breakage during the acid hydrolysis using TFA.

Acid type and concentration have been investigated in a previous study and the use of 25% TFA was found to be optimal for generating peptides for MS/MS.[12] In the present study, we found that the microwave irradiation time during the MAAH process could affect the peptide identification results. This is illustrated in Figure 9.7. Figure 9.7 shows a series of LC-ESI MS ion chromatograms of the BSA hydrolysates prepared by using different irradiation times (i.e., 2.5, 5, 7.5 and 10 min). The ion chromatograms are different, indicating that the peptide compositions of the hydrolysates were noticeably different. Relatively speaking, larger peptides are found at the late elution chromatographic peaks (i.e., retention time > 50 min) and smaller peptides tend to elute

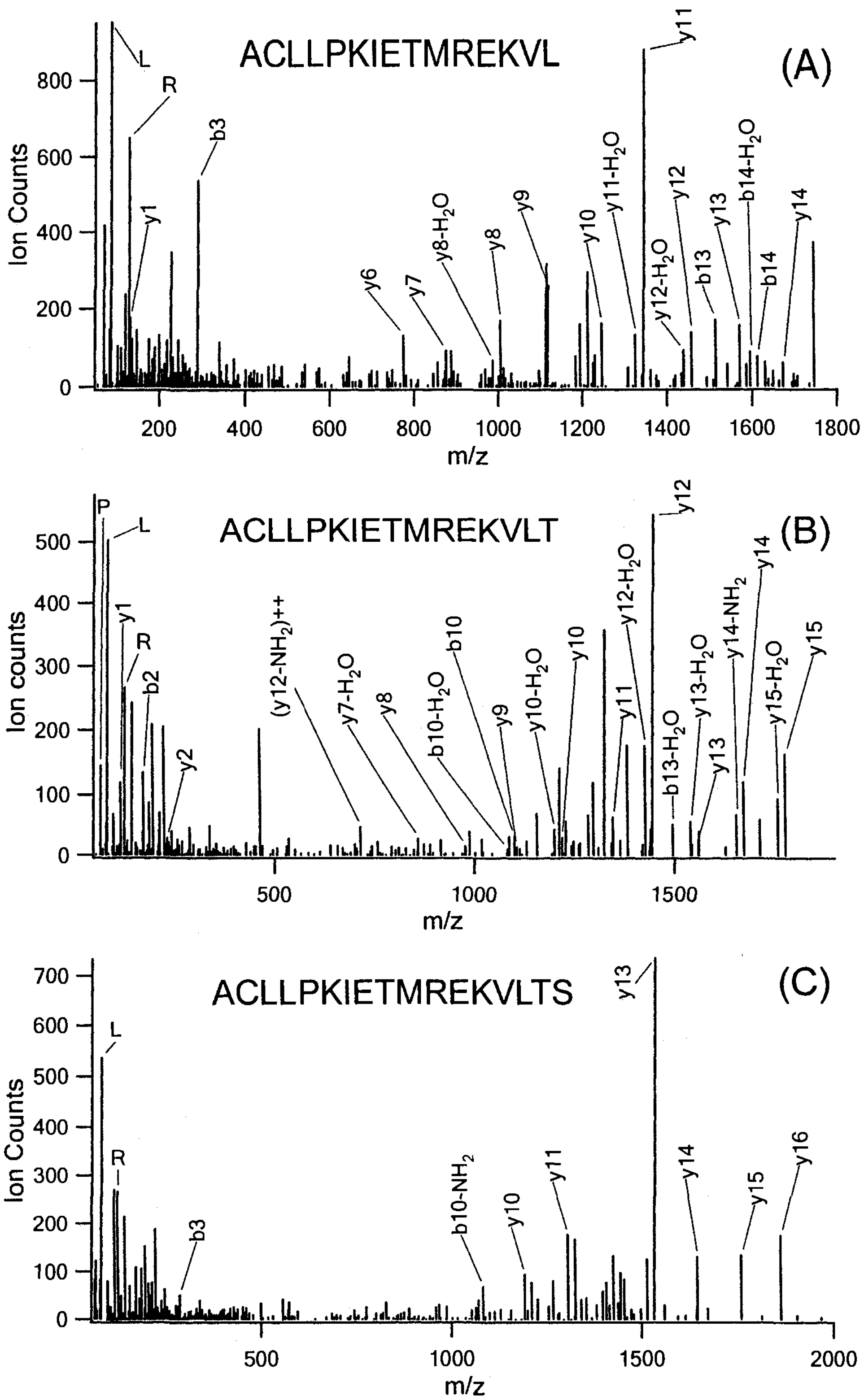


Figure 9.6 MS/MS spectra of three peptides matched with sequences near the residue 214, a natural variant site of BSA. The matched peptide sequences are shown in the spectra along with fragment ion peak assignments.

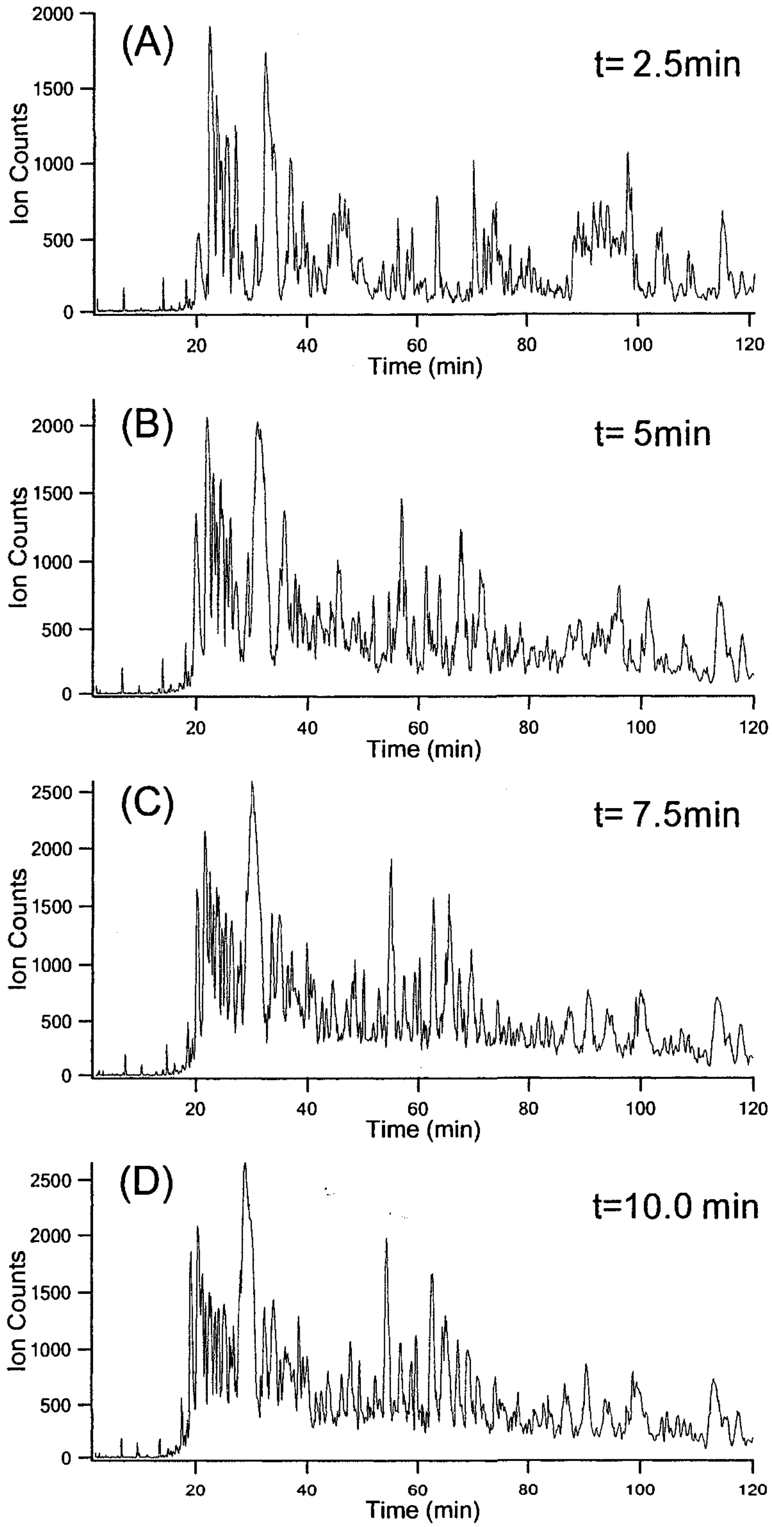


Figure 9.7 Ion chromatograms of BSA hydrolysates prepared by using different irradiation times: (A) 2.5 min, (B) 5 min, (C), 7.5 min, and (D) 10 min.

earlier in the reversed phase separation, although charge states and other interaction forces affect the retention properties of peptides. When the irradiation time increases, shorter peptides are generated. This can be seen by comparing the relative peak intensities of the ion chromatograms, e.g., (A) vs. (C). However, the chromatograms shown in Figures 9.7C (7.5 min irradiation time) and 9.7D (10 min irradiation time) appear to be similar. To further investigate this trend, Figure 9.8 shows the comparison of the number of unique peptides identified and sequence coverage obtained from the LC-ESI MS/MS chromatograms shown in Figure 9.7. The number of identified peptides or sequence coverage was the highest when 7.5 min irradiation time was used. The 10-min run produced similar results as those of the 7.5-min run. While we did not perform the replicate experiments for this plot, our experience in working with BSA and other proteins indicates that irradiation time of between 7.5 to 10 min is optimal in generating the peptides for LC-ESI MS/MS. It should be noted that these results were obtained from a 900-W household microwave oven. For a higher power microwave oven, the hydrolysis time likely needs to be reduced and should be optimized to achieve the best performance.

The minimum amount of sample required to generate near complete sequence coverage was investigated using BSA as a standard. Figure 9.9 shows the number of peptides identified and sequence coverage as a function of the injection amount of the BSA hydrolysate (i.e., 0.5, 1, 2 and 4 μg). Three replicates were done for each amount of injection. As expected, using a smaller amount sample injection, the sequence coverage is reduced. When the injection amount is equal to or above 1 μg, the sequence coverage of greater than 97% can be obtained.

The above work indicates that TFA MAAH and LC-ESI MS/MS can be used to cover near complete sequence of a protein and thus it is quite suitable for investigating any amino acid sequence variation, such as the alternation of an amino acid in the case of a natural variant of BSA. We believe that this technique should work for any size of proteins. However, for a larger protein than BSA, more separation at the peptide level (e.g., using 2D-LC instead of 1D-LC) may be required to identify a greater number of peptides to cover the entire protein sequence. Our work indicates that, for a protein size

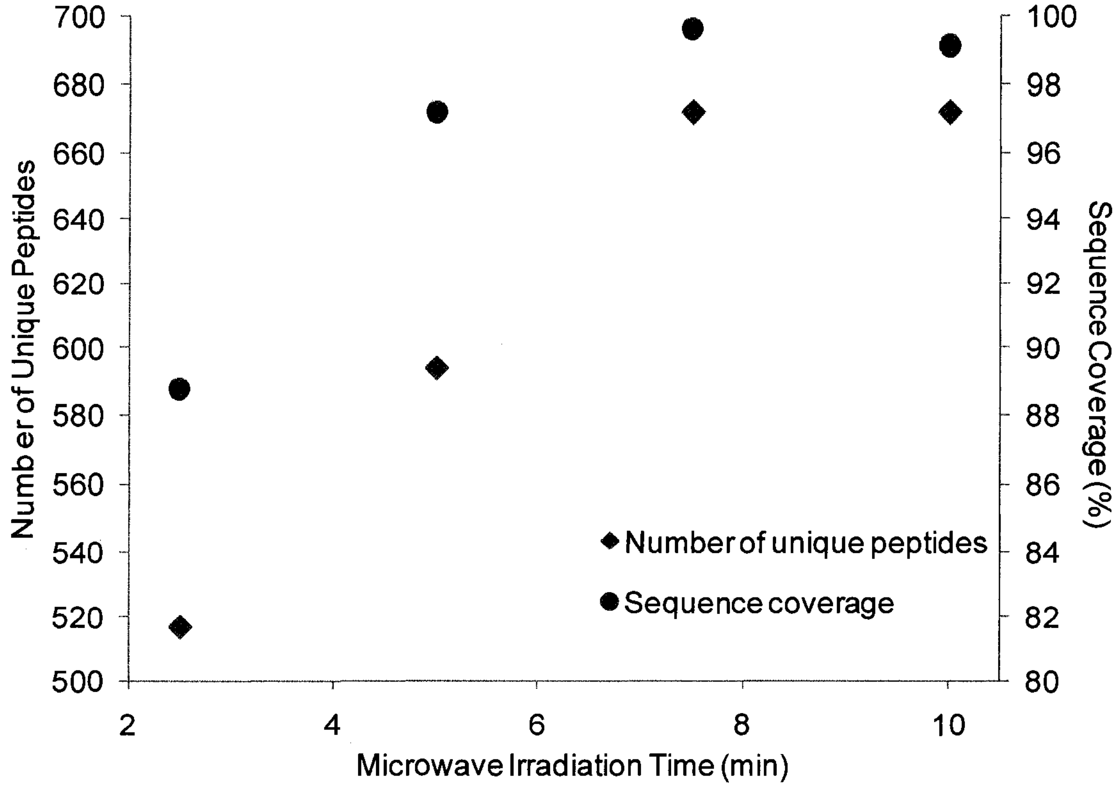


Figure 9.8 Number of unique peptides identified and BSA sequence coverage as a function of microwave irradiation time. The data were taken from the LC-ESI MS/MS analysis of the BSA hydrolysates with the corresponding ion chromatograms shown in Figure 9.7.

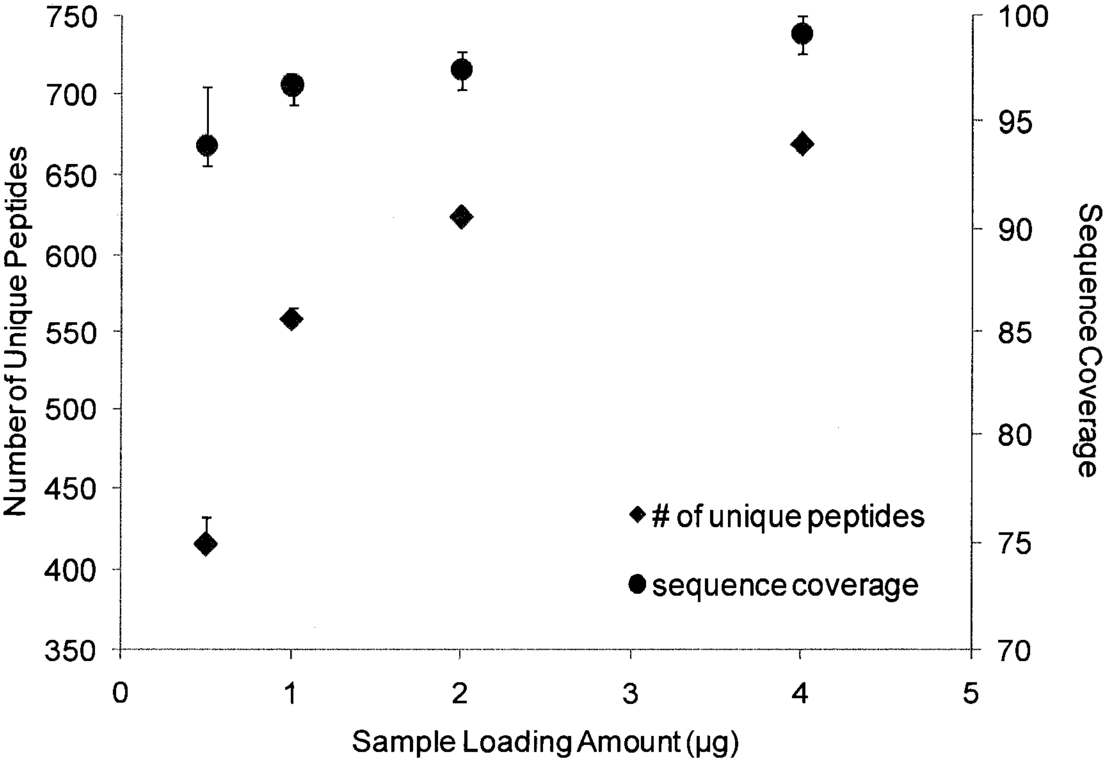


Figure 9.9 Number of unique peptides identified and sequence coverage as a function of the injection amount of the BSA hydrolysate. The error bar represents 2 standard derivations of the results from triplicate runs.

of up to BSA, 1D-LC separation combined with ESI MS/MS is sufficient to cover the entire sequence.

A more challenging application of this technique is in the area of characterizing post-translational modifications (PTMs) of proteins. The technique should work well to generate information on PTM analysis. One example of PTM analysis is shown for the characterization of phosphoproteins, α-casein and β-casein. These two protein samples were studied by TFA MAAH and LC-ESI MS/MS. In the case of α-casein, the sample contains two variants, S1 and S2. S1 is a predominant form, ~80% content in the mixture. For α-S1-casein, a total of 3262 MS/MS spectra were collected, resulting in the matching of 1414 peptides (573 unique ones) with the protein sequence. These peptides cover the entire sequence of α-S1-casein as shown in Figure 9.10A. In the case of α-S2-casein, the protein content in the mixture is smaller so the number of peptides identified from this protein is smaller than that of α-S1-casein. The failure of covering the entire sequence for this small protein is due to the limited concentration dynamic range of the 1D-LC ESI MS/MS experiment. Many peptides generated from this protein could not be detected due to low abundance and/or low ionization efficiency. To increase the sequence coverage of this minor protein in the mixture, additional separation at the peptide level, which is expected to increase the concentration dynamic range, may address this problem. Alternatively, protein level separation of the two forms of phosphoproteins may be carried out, followed by TFA MAAH of the purified individual proteins. However, separating these two forms of proteins may not be trivial using liquid chromatography. Other high resolution separation methods such as capillary electrophoresis may be better suited for pre-fractionation of the proteins.[15]

In terms of the number of phosphopeptides identified, for α-S1-casein, there are 10 known phosphorylation sites and this method identified 8 sites. For α-S2-casein, 3 out of 12 known phosphorylation sites are mapped. For β-casein, 5 sites out of 5 known sites are found. Note that, no phosphopeptide enrichment was carried out and, thus, some of the phosphopeptides were not detected. This is because phosphopeptides are generally not readily ionizable and, without enrichment, other non-phosphopeptides in the mixture α-S1-casein (P02662)

MKLLILTCLVAVALA**RPKHPIKHQGLPQEVLNENLLRFFVAPFPEVF**

GKEKVNELSKDIGSESTEDQAMEDIKQMEAESIS**SSEEIVPNSVEQ**

KHIQKEDVPSERYLGYLEQLLRLKKYKVPQLEIVPNSAEERLHSM

KEGIHAQQKEPMIGVNQELAYFYPELFRQFYQLDAYPSGAWYYVP

LGTQYTDAPSFSDIPNPIGSENSEKTTMPLW

α-S2-casein (P02663)

MKFFIFTCLLAVALAKNTMEHVSSSE**ESIISQETYKQEKNMAINPSK**

ENLCSTFCKEVVRNANEEEYSIGSSSEESAEVA**TEEVKITVDDKHY**

QKALNEINQF**YQKFPQYLQYLYQGPIVLNPWDQVKRNAVPITPTLN**

REQLSTSEENSKK**TVDMESTEVFTKK**TKLTEEEKN**RLNFLKKISQR**

YQKFA**LPQYLKTVYQHQKAMKPWIQPKTKVIPYVRYL**

β-casein (P02666)

MKVLILACLVALALA**RELEELNVPGEIVESLSSSEESITRINKKIEKFQ**

SEEQQQTEDELQDKIHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQT

PVVVPPFLQPEVMGVSKVKEAMAPKHKEMPFPKYPVEPFTESQS

LTLTDVENLHLPLPLLQSWMHQPHQPLPPTVMFPPQSVLSLSQSK

VLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV

------ Signal peptide ═══ Known phosphorylation site

Letter in bold black indicates the sequence covered

Letter in red indicates the identified phosphorylation site

Figure 9.10 Sequence coverage of α-S1-casein, α-S2-casein and β-casein. The phosphorylation sites identified are highlighted. The known phosphorylation sites are underlined with "=".

can cause ion suppression, rendering the difficulty of detecting phosphopeptide ions. The MAAH LC-ESI MS technique produces many peptides and, to effectively detect the phosphopeptides, some enrichment prior to LC-ESI MS analysis is clearly required. Many techniques have been developed for phosphopeptide enrichment including immobilized metal ion affinity chromatography (IMAC) purification.[16] We will explore the use of these techniques in combination with MAAH for phosphopeptide analysis in the future.

Many other modifications such as methylation, acetylation, deamination, oxidation, etc, do not affect the peptide detectability significantly and, therefore, we expect that these modifications can be identified from the direct analysis of the hydrolysate by LC-ESI MS/MS. However, glycosylation sites in a glycoprotein cannot be characterized by the technique, because glycans would be hydrolyzed from the amino acid side chains resulting in peptides indistinguishable from the unmodified peptides. It is interesting to note that MAAH has been reported to be useful to generate oligosaccharide ladders for mass spectrometric sequencing.[17-19]

9.4 Conclusions

We have developed a new microwave acid hydrolysis setup that generates reproducible hydrolysates using a simple domestic microwave oven. We have combined TFA MAAH with one dimensional LC-ESI QTOF tandem MS for mapping protein sequences. Complete sequence coverage can be obtained for proteins with molecular weights of up to 67,000 Da (i.e., BSA). For BSA, only about 1 μg of hydrolysate was required to generate 97% sequence coverage. Protein modifications including phosphorylation can be characterized. In the case of α-casein, two proteins in the mixture (S1 and S2 forms) can be mapped. This sequence mapping technique should be, in principle, applicable to any size of protein. Sequence mapping of a protein mixture with modest complexity such as strongly bound protein complex of several proteins should also be possible. In these cases, the hydrolysate is expected to contain many peptides, which likely requires multidimensional LC-ESI MS/MS analysis, instead of 1D LC-ESI MS/MS shown in this work.

Table 9.1 Identified peptide sequences for bovine serum albumin (P02769).

Peptide Sequence	Observed m/z	Mr (calc)	Mr (error)	Peptide Score	MASCOT Score for Identity
ACLLPKIETMREKVL	436.75	1742.98	-0.02	51	13
ACLLPKIETMREKVLT	462.01	1844.03	-0.02	27	15
ACLLPKIETMREKVLTS	483.77	1931.06	-0.02	22	16
ACLLPKIETMREKVLTSSARQRLR CA	744.41	2973.60	0.00	20	17
ACYSTVFDKLKHLVD	580.30	1737.88	0.00	39	18
ADDKEACFAVE	599.26	1196.50	0.01	38	16
ADLAKYICD	506.25	1010.47	0.01	15	15
AEDKDVCK	454.22	906.41	0.01	21	14
AEDKDVCKNYQEAKD	586.58	1756.75	-0.02	63	14
AFDEKLF	435.22	868.43	0.00	23	14
AFDEKLFTFHAD	480.89	1439.67	-0.02	39	17
AFLGSFLYEYSR	726.87	1451.71	0.02	52	17
AFLGSFLYEYSRR	536.94	1607.81	-0.01	54	17
AFLGSFLYEYSRRHPE	658.00	1970.96	0.00	34	17
AGCEKSLHTLFGDELCKVASLRET YG	707.61	2826.37	0.04	21	21
AIPENLPPLTADFAEDK	920.98	1839.93	0.02	40	18
AIPENLPPLTADFAEDKD	978.49	1954.95	0.01	70	19
AKEYEATLEECCAKDD	606.59	1816.75	0.00	36	13
ALIVRY	367.73	733.45	0.00	26	13
ALIVRYTRKVPQ	482.28	1443.86	-0.03	15	13
ALIVRYTRKVPQV	515.31	1542.92	-0.01	23	13
ALIVRYTRKVPQVS	544.33	1629.96	0.00	14	13
ALIVRYTRKVPQVSTPTLVEVSR	654.13	2612.50	0.00	37	13
ALKAWSVARL	372.23	1113.67	-0.01	58	13
APELLYYANK	591.32	1180.61	0.01	27	15
APELLYYANKY	673.34	1344.66	0.01	65	16
APELLYYANKYN	730.87	1459.69	0.04	51	16
ASIQKFGERALKAWSVARL	533.80	2131.19	-0.02	65	16
ATEEQLKTVME	640.31	1278.60	0.00	20	16
ATLEECCAKDD	599.25	1196.47	0.01	54	15
AVEGPKLVVSTQTAL	505.28	1512.84	-0.02	42	16
AVEGPKLVVSTQTALA	528.63	1582.89	-0.03	42	14
AVSVLLRLAKEYEA	521.30	1560.89	0.00	18	14
AVSVLLRLAKEYEATLEECCAK	813.76	2438.25	0.01	62	19
AVSVLLRLAKEYEATLEECCAKD	639.32	2553.28	-0.03	49	20
AVSVLLRLAKEYEATLEECCAKD	668.08	2668.31	-0.03	30	20

D					
CCEKQEPERNECFLSHKDD	578.49	2309.94	-0.01	30	13
DAFLG	522.26	521.25	0.00	26	15
DCCEKQEPERNECFLSHKD	578.49	2309.94	-0.02	33	13
DCCEKQEPERNECFLSHKDD	607.25	2424.96	-0.01	49	13
DDRADLAKYIC	641.82	1281.60	0.02	32	16
DDRADLAKYICD	699.33	1396.63	0.02	39	16
DDSPDLPKLKPD	447.23	1338.67	-0.01	28	16
DEFKADEKKFWG	500.57	1498.71	-0.01	26	16
DEHVKLVNELTEFAK	591.65	1771.90	0.02	20	18
DEKKFWGKYLYEIARR	526.28	2101.11	-0.01	64	19
DELCKVASLR	378.53	1132.59	-0.01	75	16
DELCKVASLRE	421.55	1261.63	-0.02	72	17
DELCKVASLRETY	509.59	1525.74	0.00	33	17
DELCKVASLRETYG	528.59	1582.77	-0.01	40	18
DELCKVASLRETYGD	566.94	1697.79	-0.01	38	18
DELCKVASLRETYGDMAD	672.64	2014.90	0.00	46	17
DFAEDKDVCK	585.27	1168.51	0.02	72	14
DKPLLEKSHCIAEVEK	460.50	1837.96	-0.01	106	17
DKPLLEKSHCIAEVEKD	489.25	1952.99	-0.02	32	18
DLLECADDRA	560.76	1119.49	0.02	36	15
DLLECADDRADL	674.82	1347.60	0.02	45	15
DLLECADDRADLA	710.33	1418.63	0.02	42	15
DLLECADDRADLAK	516.58	1546.73	0.00	29	16
DLLECADDRADLAKYIC	963.97	1925.89	0.03	71	19
DLLECADDRADLAKYICD	1021.48	2040.91	0.02	121	18
DLLECADDRADLAKYICDN	1078.99	2155.94	0.03	69	17
DLPKLKPD	463.27	924.53	-0.01	43	13
DMADCCEKQE	586.71	1171.38	0.02	28	13
DMADCCEKQEPER	518.87	1553.58	0.00	59	13
DMADCCEKQEPERN	557.21	1668.61	0.00	42	13
DMADCCEKQEPERNE	605.22	1812.66	-0.02	17	13
DMADCCEKQEPERNECFLSHKD	657.76	2627.04	-0.03	45	13
DMADCCEKQEPERNECFLSHKDD	686.52	2742.07	-0.03	37	13
DPHACYSTVFDKLKHLVD	522.76	2087.01	-0.02	20	18
DQFEKLGEYG	593.77	1185.52	0.01	15	14
DRADLAKYIC	584.30	1166.58	0.01	22	16
DRADLAKYICD	641.82	1281.60	0.02	34	16
DRADLAKYICDN	699.33	1396.63	0.01	42	15
DRADLAKYICDNQD	820.87	1639.71	0.02	50	16
DRADLAKYICDNQDTISSKLKE	632.81	2527.21	0.00	20	19

DSPDLPKLK	506.79	1011.56	0.01	32	13
DSPDLPKLKP	555.32	1108.61	0.00	41	13
DSPDLPKLKPD	408.88	1223.64	-0.02	67	15
DTHKSEIAHR	398.54	1192.59	0.01	35	15
DTHKSEIAHRF	447.56	1339.66	-0.02	39	15
DTHKSEIAHRFK	367.94	1467.76	-0.02	18	17
DTHKSEIAHRFKD	528.60	1582.79	-0.01	31	18
DTHKSEIAHRFKDL	424.97	1695.87	-0.01	49	16
DTHKSEIAHRFKDLG	439.22	1752.89	-0.03	22	18
DTHKSEIAHRFKDLGE	471.48	1881.93	-0.02	33	18
DTHKSEIAHRFKDLGEE	503.75	2010.98	0.00	33	18
DTHKSEIAHRFKDLGEEH	717.02	2148.03	0.00	82	19
DTHKSEIAHRFKDLGEEHF	574.78	2295.10	-0.02	101	19
DTHKSEIAHRFKDLGEEHFK	606.80	2423.20	-0.02	63	19
DTHKSEIAHRFKDLGEEHFKG	827.74	2480.22	-0.02	28	19
DTHKSEIAHRFKDLGEEHFKGL	649.32	2593.30	-0.04	68	20
DTHKSEIAHRFKDLGEEHFKGLVL	702.36	2805.46	-0.05	29	19
DTHKSEIAHRFKDLGEEHFKGLVLI	730.63	2918.54	-0.03	21	19
DTHKSEIAHRFKDLGEEHFKGLVLIA	997.53	2989.58	0.00	52	19
DTHKSEIAHRFKDLGEEHFKGLVLIAF	785.17	3136.65	-0.01	63	20
DTHKSEIAHRFKDLGEEHFKGLVLIAFS	806.93	3223.68	0.01	46	20
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQ	1118.59	3352.72	0.01	41	21
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYL	908.23	3628.87	0.02	63	21
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQ	940.49	3757.91	0.02	66	22
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQ	1296.66	3886.95	0.01	88	22
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQC	998.26	3988.98	0.03	57	22
DTISSKLKEC	562.29	1122.56	0.01	17	16
DTISSKLKECCD	671.31	1340.60	0.02	30	16
DTISSKLKECCDKPLLEKSHCIAEVEKD	791.14	3160.55	0.00	31	21
DYLSLILNRL	610.84	1219.68	-0.01	56	13
DYLSLILNRLCVLH	558.30	1671.90	-0.01	70	16
DYLSLILNRLCVLHE	601.32	1800.94	-0.01	59	17
DYLSLILNRLCVLHEK	644.01	1929.04	-0.02	64	17
EATLEECCAKDD	663.78	1325.51	0.03	67	14
ECFLSHKDDSPDLPKLKPD	546.77	2183.06	-0.02	21	19

EDYLSLILNRL	675.37	1348.72	0.00	57	16
EDYLSLILNRLCVLHEK	515.52	2058.08	-0.04	48	17
EFKADEKKFWG	462.23	1383.68	-0.01	29	15
EFKADEKKFWGKYLYEIAR	606.07	2420.25	0.00	25	18
EFVEVTKLVTD	640.34	1278.67	-0.01	29	17
EFVEVTKLVTDLTKVHKECCHG	629.57	2514.26	0.01	62	19
EHVKLVNEL	541.29	1080.58	-0.01	16	13
EHVKLVNELTEFAK	552.97	1655.89	0.00	96	16
EKKFWGKYLYEIAR	458.50	1829.98	-0.02	18	16
EKKFWGKYLYEIARR	497.52	1986.08	-0.03	82	16
EKKFWGKYLYEIARRH	531.79	2123.14	0.00	43	17
EKLFTFHA	496.76	991.51	-0.01	21	14
EKLFTFHAD	554.28	1106.54	0.00	39	16
EKQEPERNECFLSHKD	498.23	1988.89	-0.01	23	16
EKQEPERNECFLSHKDD	526.99	2103.92	-0.01	28	15
ELCKVASLRE	574.32	1146.61	0.02	42	16
ELCKVASLRETYG	734.88	1467.74	0.00	40	17
ELTEFAKTCVADESHAG	603.28	1806.81	0.00	45	16
EPERNECFLSHKDD	573.59	1717.74	0.01	55	13
EPERNECFLSHKDDSPDLPKLKPD	703.08	2808.34	-0.03	43	20
EPQNLIKQ	486.26	970.50	0.01	16	13
EPQNLIKQN	543.28	1084.54	0.01	32	14
EPQNLIKQNCDQFEKL	650.98	1949.89	0.01	22	18
EPQNLIKQNCDQFEKLG	1004.47	2006.91	0.01	31	18
ERALKAWSVA	565.83	1129.62	0.01	42	13
ERALKAWSVARL	467.27	1398.81	-0.01	47	13
ERALKAWSVARLS	496.28	1485.84	-0.01	34	13
ERALKAWSVARLSQ	539.30	1614.88	-0.01	35	14
ERALKAWSVARLSQKF	473.51	1890.05	-0.02	31	14
ERMPCTEDYLSLILNRLCVLHEK	694.85	2775.38	-0.02	24	20
ESHAGCEKSLHTLFG	539.25	1614.75	-0.01	19	16
ETYVPKAF	477.75	953.49	-0.01	22	13
ETYVPKAFDEKL	480.58	1438.73	-0.01	27	16
ETYVPKAFDEKLF	529.61	1585.80	-0.01	25	16
ETYVPKAFDEKLFTFHA	681.68	2042.01	0.00	41	17
ETYVPKAFDEKLFTFHAD	540.26	2157.04	-0.02	36	18
EVEKDAIPE	515.26	1028.50	0.01	26	16
EVEKDAIPEN	572.78	1143.53	0.02	28	16
EVSRSLGKVG	516.30	1030.58	0.00	30	14
EYAVSVLLRLAKEYEA	618.68	1852.99	0.01	40	16
EYAVSVLLRLAKEYEATLEECCAKDD	987.81	2960.41	0.01	30	20

EYGFQNALIVRY	737.37	1472.73	0.00	19	15
EYSRRHPEYAVSVLLRLAKEYEA	695.63	2778.45	0.04	20	18
FAEDKDVCKNYQEAKD	476.72	1902.83	0.00	53	16
FAEDKDVCKNYQEAKDAFLG	764.69	2291.04	0.00	23	18
FAVEGPKLVVS	573.34	1144.65	0.01	29	14
FAVEGPKLVVSTQTALA	577.99	1730.95	-0.01	20	16
FDEKLF	399.71	797.40	0.00	21	13
FDEKLFTFHA	418.87	1253.61	-0.02	68	17
FDEKLFTFHAD	457.22	1368.64	-0.01	41	17
FDKLKHL	450.77	899.52	0.00	30	14
FDKLKHLV	500.30	998.59	0.00	31	13
FDKLKHLVD	557.82	1113.62	0.01	38	13
FEKLGEY	443.22	884.43	0.00	24	14
FEKLGEYG	471.74	941.45	0.02	18	16
FGDELCKVA	491.24	980.46	0.00	40	14
FGDELCKVASLRE	489.58	1465.72	0.00	49	17
FGDELCKVASLRETYG	596.63	1786.86	0.00	62	18
FGERALKAW	359.86	1076.58	-0.02	25	15
FGERALKAWSVARL	535.31	1602.90	0.01	55	13
FKADEKKFWG	419.22	1254.64	-0.01	33	15
FKDLGEEHFKG	436.22	1305.64	-0.01	38	16
FLSHKDDSPDLPKLKPD	488.75	1951.01	-0.02	64	17
FPKAEFVEVTKLVTDLTKVH	576.08	2300.28	0.01	99	13
FPKAEFVEVTKLVTDLTKVHKE	640.36	2557.42	-0.01	53	15
FQNALIVR	481.27	960.54	-0.01	48	13
FQNALIVRY	563.30	1124.59	0.00	49	15
FQNALIVRYTR	461.26	1380.75	0.00	22	15
FQNALIVRYTRKVPQVS	674.71	2021.09	0.01	16	15
FVAFVDKCCA	551.76	1101.50	0.01	36	13
FVAFVDKCCAA	587.28	1172.54	0.01	15	14
FVDKCCAADD	543.72	1085.42	0.01	21	13
FVEVTKLVTDL	632.37	1262.71	0.01	27	13
FVEVTKLVTDLTKVHKECCHG	597.31	2385.22	0.00	73	19
FWGKYLYEIARRH	435.48	1737.91	-0.02	40	18
FWGKYLYEIARRHPY	500.51	1998.03	-0.02	45	19
FYAPELLYYANK	746.88	1491.73	0.03	46	17
FYAPELLYYANKY	828.42	1654.79	0.03	45	18
FYAPELLYYANKYN	885.44	1768.83	0.03	48	18
GACLLPKIETMREKVL	451.00	1800.00	-0.04	32	15
GCEKSLHTLF	567.78	1133.55	0.00	17	16
GCEKSLHTLFG	596.30	1190.58	0.00	16	15

GCEKSLHTLFGDELCKVASLRE	609.55	2434.20	-0.01	44	20
GCEKSLHTLFGDELCKVASLRETYG	689.85	2755.33	0.04	29	20
GDELCKVASLRE	660.33	1318.66	0.00	70	16
GDELCKVASLRETYG	547.60	1639.79	0.00	58	18
GDLLECADDRA	589.26	1176.51	0.00	30	15
GDLLECADDRAD	646.78	1291.54	0.01	32	13
GDLLECADDRADLAKYICD	1049.99	2097.93	0.04	71	18
GDMADCCEKQEPERN	581.54	1741.62	-0.03	16	13
GERALKAWSVA	594.34	1186.65	0.02	44	13
GERALKAWSVARL	486.28	1455.83	-0.01	37	13
GERALKAWSVARLSQK	451.00	1800.00	-0.03	46	15
GERALKAWSVARLSQKF	487.77	1947.07	-0.01	37	16
GFQNALIVRY	591.32	1180.62	0.00	47	15
GLVLIAF	732.48	731.46	0.02	22	13
GPKLVVSTQ	465.27	928.52	0.01	48	14
GPKLVVSTQT	515.80	1029.57	0.01	46	13
GPKLVVSTQTA	551.32	1100.61	0.02	89	13
GPKLVVSTQTAL	607.36	1212.71	0.00	53	13
GPKLVVSTQTALA	428.92	1283.74	0.00	34	13
GSFLYEYSRRHPEYAVSVLLRLAKEYEA	837.45	3345.71	0.04	65	20
GSFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAKDD	1114.28	4453.14	-0.03	50	23
GVFQECCQAEDK	679.77	1357.52	0.02	43	13
GVFQECCQAEDKG	707.78	1413.55	-0.01	48	13
GVFQECCQAEDKGA	743.31	1484.59	0.02	48	13
HCIAEVEKDA	557.77	1113.51	0.02	57	14
HKDDSPDLPKLKPD	401.96	1603.82	-0.02	74	18
HKSEIAHRFKDLG	513.27	1536.82	-0.03	18	17
HKSEIAHRFKDLGEEH	483.99	1931.96	-0.04	49	17
HPEYAVSVLLRLAKEY	630.02	1887.03	0.00	37	16
HPEYAVSVLLRLAKEYE	673.04	2016.07	0.02	34	17
HPEYAVSVLLRLAKEYEA	696.71	2087.11	0.01	41	16
HPEYAVSVLLRLAKEYEATLEECCAK	742.11	2964.47	-0.04	30	20
HPEYAVSVLLRLAKEYEATLEECCAKD	770.87	3079.50	-0.04	36	20
HPEYAVSVLLRLAKEYEATLEECCAKDD	799.63	3194.53	-0.03	36	20
HPYFYAPELLYYA	823.91	1645.78	0.03	35	18
HPYFYAPELLYYAN	881.42	1760.81	0.02	24	17
HPYFYAPELLYYANKY	1027.01	2051.97	0.03	76	19

HPYFYAPELLYYANKYN	723.34	2166.99	0.01	19	19
HRFKDLGEEHFK	386.44	1541.77	-0.03	47	17
HRFKDLGEEHFKG	400.70	1598.80	-0.03	34	15
HTLFGDELCKVA	666.84	1331.65	0.02	56	17
HVKLVNELTEFAK	510.28	1527.83	0.00	19	15
IQKFGERALKAWSVARL	494.28	1973.12	-0.02	41	15
KDAFLGSFLYEY	726.85	1451.70	0.00	31	17
KDAIPENLPPL	604.32	1206.65	-0.02	31	13
KDDSPDLPKLKPD	489.93	1466.76	0.00	91	16
KDVCKNYQEAKDAFLG	610.63	1828.87	-0.01	63	17
KEYEATLEECCAKDD	582.91	1745.71	0.00	30	13
KFGERALKAW	402.56	1204.67	-0.02	17	14
KFGERALKAWSVARL	433.75	1730.99	-0.02	38	14
KFWGKYLYEIARR	433.24	1728.95	-0.03	47	17
KFWGKYLYEIARRH	467.50	1866.01	-0.02	24	16
KGACLLPKIETMREKVL	483.02	1928.10	-0.03	22	15
KLGEYGFQ	471.74	941.45	0.03	21	15
KLVVSTQTALA	566.34	1130.65	0.00	13	13
KPLLEKSHCIAEVE	532.62	1594.84	0.01	69	14
KPLLEKSHCIAEVEK	431.74	1722.93	-0.01	90	15
KPLLEKSHCIAEVEKD	460.49	1837.96	-0.03	55	17
KPLLEKSHCIAEVEKDA	478.25	1909.00	-0.04	35	19
KPLLEKSHCIAEVEKDAI	506.52	2022.08	-0.02	33	17
KPLLEKSHCIAEVEKDAIPE	563.05	2248.18	-0.02	102	19
KPLLEKSHCIAEVEKDAIPEN	591.81	2363.20	0.00	51	18
KPLLEKSHCIAEVEKDAIPENLPPL	696.88	2783.48	0.01	25	18
KQEPERNECFLSHKDD	494.72	1974.87	-0.03	26	15
KSLHTLFGDELCKVA	554.31	1659.87	0.04	29	15
KSLHTLFGDELCKVASLRETYG	617.57	2466.26	-0.02	23	20
KVGTRCCTKPESERMP	456.22	1820.87	-0.04	33	17
KVPQVSTPTLVEVSRSLG	633.36	1897.05	0.00	27	15
KVPQVSTPTLVEVSRSLGKVG	546.31	2181.24	-0.03	23	16
KYLYEIARR	404.57	1210.68	0.00	15	13
KYLYEIARRHPY	536.96	1607.86	0.00	19	17
KYLYEIARRHPYF	439.73	1754.93	-0.03	19	17
KYLYEIARRHPYFY	640.34	1917.99	0.01	21	18
KYLYEIARRHPYFYAPELLYYANKY	812.17	3244.64	0.01	23	21
KYLYEIARRHPYFYAPELLYYANKYN	840.68	3358.68	0.01	24	22
LAKYICD	413.22	824.41	0.01	20	13
LAKYICDN	470.73	939.44	0.01	25	15

LCKVASLRE	509.79	1017.56	0.00	28	15
LEKSHCIAEVEKD	500.91	1499.73	-0.03	36	17
LFGDELCKVA	547.78	1093.55	0.00	26	15
LFGDELCKVASLRE	527.28	1578.81	0.00	43	16
LGEEHFKGLVLIA	713.42	1424.80	0.01	35	13
LKAWSVARL	522.32	1042.63	0.00	19	13
LLECAD	663.31	662.29	0.01	20	15
LLECADDR	467.73	933.42	0.02	42	14
LLECADDRA	503.24	1004.46	0.00	41	15
LLECADDRAD	560.76	1119.49	0.01	70	14
LLECADDRADLA	652.81	1303.61	-0.01	23	15
LLECADDRADLAKYICD	963.96	1925.89	0.01	55	19
LLEKSHCIAEVEKD	538.61	1612.81	-0.01	45	17
LLRLA	585.40	584.40	-0.01	22	13
LLYYANKYN	582.29	1162.55	0.02	28	16
LPKLKPD	405.76	809.50	0.00	38	13
LPPLTA	611.38	610.37	0.00	33	13
LPPLTADFA	472.76	943.50	0.00	16	14
LPPLTADFAEDK	658.85	1315.67	0.03	86	14
LPPLTADFAEDKD	716.36	1430.69	0.01	20	17
LPPLTADFAEDKDVCK	587.96	1760.87	0.00	78	18
LPPLTADFAEDKDVCKNYQEAK	624.81	2495.19	0.00	27	20
LPPLTADFAEDKDVCKNYQEAKD	653.56	2610.22	0.00	77	20
LSHKDDSPDLPKLKPD	451.99	1803.94	-0.01	90	16
LTPDETYVPKA	617.32	1232.63	0.00	18	16
LVDEPQNLIKQN	707.36	1412.70	0.01	25	16
LVELLKH	426.27	850.53	0.00	44	13
LVEVSRSLG	480.28	958.54	0.01	31	13
LVEVSRSLGKVG	415.25	1242.73	-0.01	25	13
LVLIA	528.39	527.37	0.01	25	13
LVLIAF	675.46	674.44	0.02	31	13
LVLIAFSQYLQQ	713.38	1424.74	0.00	20	15
LVVSTQTALA	502.29	1002.56	0.00	20	13
LYEYSRR	493.76	985.50	0.00	20	13
MADCCEKQEP	577.71	1153.41	0.00	23	13
MADCCEKQEPE	642.23	1282.45	-0.01	34	13
MADCCEKQEPER	480.52	1438.55	-0.02	38	13
MADCCEKQEPERN	518.86	1553.58	-0.03	72	13
MADCCEKQEPERNE	561.54	1681.64	-0.03	65	13
MADCCEKQEPERNECFLSHKDD	657.76	2627.04	-0.03	49	13
MPCTEDYLSLILNRLCVLH	745.37	2233.09	-0.01	39	18

MPCTEDYLSLILNRLCVLHEK	623.31	2489.25	-0.05	28	20
NALIVRY	424.75	847.49	0.00	25	13
NALIVRYTRKVPQVS	582.34	1744.00	-0.01	28	13
NCDQFEKLG	527.24	1052.46	0.01	14	14
NECFLSHKDDSPDLPKLKPD	575.28	2297.10	-0.01	138	19
NLPPLTADFAEDK	716.36	1430.69	0.02	20	17
NLPPLTADFAEDKD	773.38	1544.74	0.00	59	17
NQDTISSKLKECCDKPLLEK	574.04	2292.13	-0.01	45	19
NYQEAKDAFLG	628.30	1254.59	0.00	26	17
PCTEDYLSLILNRLCVLHEK	590.80	2359.19	-0.03	50	19
PDLPKLK	405.76	809.50	0.00	17	13
PDLPKLKP	454.28	906.55	-0.02	21	13
PDLPKLKPD	511.79	1021.58	-0.01	13	13
PELLY	634.34	633.34	-0.01	15	13
PELLYYANKY	637.33	1272.64	0.01	36	16
PELLYYANKYN	694.85	1387.67	0.02	44	17
PERNECFLSHKDD	398.43	1589.68	-0.01	19	13
PERNECFLSHKDDSPDLPKLKPD	670.82	2679.30	-0.05	64	20
PEYAVSVLLRLAKE	529.97	1586.90	-0.01	60	15
PEYAVSVLLRLAKEY	584.33	1749.97	0.00	61	15
PEYAVSVLLRLAKEYE	627.35	1879.01	0.01	88	15
PEYAVSVLLRLAKEYEA	651.02	1950.05	0.00	37	17
PEYAVSVLLRLAKEYEATL	722.39	2164.18	-0.02	20	17
PEYAVSVLLRLAKEYEATLEEC	842.77	2525.27	0.01	38	20
PEYAVSVLLRLAKEYEATLEECCA	900.78	2699.32	0.00	30	20
PEYAVSVLLRLAKEYEATLEECCAK	707.86	2827.41	0.00	99	21
PEYAVSVLLRLAKEYEATLEECCAKD	736.62	2942.44	0.01	77	21
PEYAVSVLLRLAKEYEATLEECCAKDD	765.37	3057.47	-0.01	23	21
PFDEHVKLV	361.86	1082.58	-0.02	45	15
PFDEHVKLVN	399.87	1196.62	-0.02	35	15
PFDEHVKLVNE	442.89	1325.66	-0.01	32	16
PFDEHVKLVNEL	480.91	1439.73	-0.01	50	16
PFDEHVKLVNELTEF	606.31	1815.90	0.00	86	17
PFDEHVKLVNELTEFA	629.99	1886.94	0.00	27	18
PFDEHVKLVNELTEFAK	505.01	2016.02	0.00	51	18
PFDEHVKLVNELTEFAKT	530.27	2117.07	-0.01	44	19
PFDEHVKLVNELTEFAKTCVA	797.42	2389.20	0.03	22	19
PFDEHVKLVNELTEFAKTCVAD	627.06	2504.23	-0.01	40	20
PFDEHVKLVNELTEFAKTCVADE	659.33	2633.27	0.01	38	20

PFDEHVKLVNELTEFAKTCVADES	681.08	2720.30	-0.02	23	19
PFDEHVKLVNELTEFAKTCVADESH	715.35	2857.36	0.02	87	21
PFDEHVKLVNELTEFAKTCVADESHA	733.11	2928.40	0.00	41	21
PFDEHVKLVNELTEFAKTCVADESHAG	747.36	2985.42	0.01	65	21
PFDEHVKLVNELTEFAKTCVADESHAGCEK	837.65	3346.55	0.02	27	20
PHACYSTVF	512.73	1023.45	0.00	36	15
PHACYSTVFD	570.25	1138.48	0.00	26	14
PHACYSTVFDKL	690.84	1379.65	0.01	67	17
PHACYSTVFDKLK	754.88	1507.75	0.00	35	17
PHACYSTVFDKLKH	412.20	1644.81	-0.03	30	16
PHACYSTVFDKLKHL	440.47	1757.89	-0.02	44	17
PHACYSTVFDKLKHLV	465.24	1856.96	-0.03	28	17
PHACYSTVFDKLKHLVD	494.00	1971.99	-0.03	28	18
PHACYSTVFDKLKHLVDE	526.26	2101.03	-0.02	50	19
PHACYSTVFDKLKHLVDEPQ	582.54	2326.14	-0.02	25	19
PHACYSTVFDKLKHLVDEPQN	815.06	2442.15	0.01	23	19
PHACYSTVFDKLKHLVDEPQNLIK	699.87	2795.43	0.03	81	19
PHACYSTVFDKLKHLVDEPQNLIKQ	732.38	2925.46	0.02	27	21
PHACYSTVFDKLKHLVDEPQNLIKQN	760.64	3038.52	0.01	25	21
PKAEFVEVTKLVTDLTKVH	539.31	2153.21	0.00	54	15
PKAFDEKLF	547.80	1093.58	0.00	36	13
PKLVVSTQTA	522.80	1043.59	0.01	60	13
PKLVVSTQTAL	579.34	1156.67	0.00	24	13
PKLVVSTQTALA	409.91	1226.72	-0.02	34	13
PLLEKSHCIAEVEKD	570.96	1709.87	-0.01	43	16
PLTADFAEDKD	611.28	1220.56	-0.01	23	15
PNTLCDEFK	533.75	1065.48	0.00	35	16
PNTLCDEFKA	569.26	1136.52	-0.02	22	15
PNTLCDEFKAD	626.79	1251.54	0.01	43	16
PNTLCDEFKADE	691.31	1380.59	0.02	61	16
PNTLCDEFKADEK	504.24	1509.67	0.03	46	16
PNTLCDEFKADEKK	546.60	1636.78	0.01	80	18
PNTLCDEFKADEKKF	446.96	1783.85	-0.02	43	18
PNTLCDEFKADEKKFW	657.98	1970.91	0.01	50	18
PNTLCDEFKADEKKFWG	676.66	2026.95	0.00	33	19
PNTLCDEFKADEKKFWGK	540.01	2156.02	0.00	36	18
PNTLCDEFKADEKKFWGKYLYEIA	727.86	2907.42	0.01	24	21

PNTLCDEFKADEKKFWGKYLYEIAR	766.89	3063.52	0.03	120	21
PNTLCDEFKADEKKFWGKYLYEIARR	806.17	3220.60	0.04	47	20
PNTLCDEFKADEKKFWGKYLYEIARRH	840.43	3357.66	0.02	33	21
PPLTADFAEDK	602.30	1202.58	0.00	69	16
PPLTADFAEDKD	659.81	1317.61	0.00	65	16
PPLTADFAEDKDVCK	550.27	1647.78	-0.01	37	18
PQNLIKQN	479.74	957.47	0.00	33	15
PQVSTPTLVEVSRSLG	835.96	1669.89	0.01	58	15
PTLVEVSR	450.76	899.51	0.01	47	13
PTLVEVSRSL	550.82	1099.62	0.01	57	13
PTLVEVSRSLG	579.34	1156.65	0.02	54	13
PTLVEVSRSLGKVG	481.28	1440.83	-0.01	52	13
PYFYAPELLYYA	755.37	1508.72	0.01	32	17
PYFYAPELLYYAN	812.40	1622.77	0.02	25	18
PYFYAPELLYYANK	876.45	1750.86	0.02	48	18
PYFYAPELLYYANKY	957.98	1913.92	0.03	47	18
PYFYAPELLYYANKYN	677.33	2028.95	0.02	27	19
PYFYAPELLYYANKYNG	1044.01	2085.97	0.03	36	18
PYFYAPELLYYANKYNGVF	1167.56	2333.09	0.01	23	18
QEPERNECFLSHKDD	616.60	1846.78	0.00	32	13
QFEKLGEY	507.75	1013.47	0.01	17	15
RADLAKYIC	526.78	1051.55	0.00	23	16
RADLAKYICD	584.29	1166.58	-0.01	35	17
RADLAKYICDN	641.81	1281.60	0.00	58	16
RADLAKYICDNQD	763.35	1524.69	0.00	80	16
RALKAWSVARL	424.26	1269.77	-0.02	22	13
RCASIQKFGERALKAW	467.00	1863.98	-0.03	19	18
RHPEYAVSVLLRLAKEY	511.79	2043.13	-0.01	62	16
RHPEYAVSVLLRLAKEYE	544.05	2172.17	0.00	49	18
RHPEYAVSVLLRLAKEYEA	561.81	2243.21	0.01	40	16
RHPEYAVSVLLRLAKEYEATLEE	679.86	2715.42	0.00	31	19
RHPEYAVSVLLRLAKEYEATLEECCAKD	809.90	3235.60	-0.02	28	21
RHPEYAVSVLLRLAKEYEATLEECCAKDD	1117.88	3350.63	-0.02	33	21
RHPYFYAPELLYY	866.45	1730.85	0.04	26	17
RHPYFYAPELLYYANKYN	775.38	2323.09	0.02	27	19
RKVPQVSTPTLVEVSRSLGKVG	585.34	2337.34	-0.01	31	14
RLRCASIQKFG	427.24	1278.69	0.00	18	15
RMPCTEDYLSLILNRLCVLHEK	662.59	2646.33	-0.02	55	21

RNECFLSHKDDSPD	554.91	1661.71	0.00	48	15
RNECFLSHKDDSPDLPKLKPD	614.29	2453.20	-0.05	41	20
RRHPEYAVSVLLRLAKEY	550.81	2199.23	-0.02	26	14
RRHPEYAVSVLLRLAKEYEA	600.83	2399.31	-0.01	31	17
RRHPEYAVSVLLRLAKEYEATLEECCAKDD	877.69	3506.73	0.02	36	21
RRHPYFYAPELLYYANKYN	827.41	2479.20	0.00	53	19
SALTPDETYVPKAFD	827.41	1652.79	0.00	28	18
SARQRLRCASIQKF	417.22	1664.88	-0.02	18	17
SARQRLRCASIQKFG	431.48	1721.90	-0.02	35	18
SEIAHRFKDL	405.89	1214.64	0.01	15	14
SEIAHRFKDLG	636.84	1271.66	0.01	73	14
SEIAHRFKDLGE	467.91	1400.70	0.01	43	15
SEIAHRFKDLGEE	510.93	1529.75	0.01	37	17
SEIAHRFKDLGEEH	556.61	1666.81	0.01	21	16
SEIAHRFKDLGEEHFK	486.49	1941.97	-0.02	87	18
SEIAHRFKDLGEEHFKG	500.75	1998.99	-0.02	51	19
SEIAHRFKDLGEEHFKGLVLIAF	664.86	2655.42	0.00	65	17
SERMPCTEDYLSLILNRL	718.69	2153.05	-0.01	38	19
SERMPCTEDYLSLILNRLCVL	823.74	2468.21	-0.01	59	19
SERMPCTEDYLSLILNRLCVLH	652.07	2604.29	-0.04	41	20
SERMPCTEDYLSLILNRLCVLHE	912.44	2734.31	0.00	29	21
SERMPCTEDYLSLILNRLCVLHEK	955.15	2862.41	0.02	72	20
SFLYEYSR	532.76	1063.50	0.00	26	15
SFLYEYSRR	407.53	1219.60	-0.02	33	17
SFLYEYSRRHPE	528.59	1582.75	0.00	38	18
SFLYEYSRRHPEY	582.95	1745.82	0.00	34	16
SFLYEYSRRHPEYA	606.62	1816.85	-0.01	30	17
SFLYEYSRRHPEYAVSVLLR	622.08	2484.29	-0.01	44	18
SFLYEYSRRHPEYAVSVLLRLA	668.11	2668.41	-0.02	24	18
SFLYEYSRRHPEYAVSVLLRLAK	700.14	2796.51	0.00	55	17
SFLYEYSRRHPEYAVSVLLRLAKE	976.20	2925.55	0.03	39	18
SFLYEYSRRHPEYAVSVLLRLAKEY	773.17	3088.61	0.02	55	20
SFLYEYSRRHPEYAVSVLLRLAKEYE	805.43	3217.66	0.02	42	20
SFLYEYSRRHPEYAVSVLLRLAKEYEA	823.19	3288.69	0.03	40	20
SFLYEYSRRHPEYAVSVLLRLAKEYEATL	876.72	3502.82	0.01	22	19
SFLYEYSRRHPEYAVSVLLRLAKEYEATLE	908.97	3631.87	-0.01	35	21
SFLYEYSRRHPEYAVSVLLRLAKEYEATLEE	941.23	3760.91	-0.01	22	22

SFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAK	1042.52	4166.06	0.00	25	22
SFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAKD	1071.28	4281.09	-0.02	95	22
SFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAKDD	1100.03	4396.11	-0.01	45	23
SHAGCEKSLHTLF	477.23	1428.68	-0.01	38	17
SHAGCEKSLHTLFGDELCKVA	562.02	2244.07	0.00	33	18
SHCIAEVEKDA	601.29	1200.54	0.01	48	15
SHCIAEVEKDAIPE	770.87	1539.72	0.00	92	17
SHKDDSPDLPKLKP	394.96	1575.83	-0.01	29	16
SHKDDSPDLPKLKPD	423.72	1690.85	-0.01	78	18
SIQKFGERAL	575.32	1148.62	0.01	26	13
SIQKFGERALK	426.58	1276.71	-0.01	19	15
SIQKFGERALKA	449.93	1346.77	0.00	18	13
SIQKFGERALKAW	384.46	1533.83	-0.02	48	16
SIQKFGERALKAWSVA	598.00	1790.97	0.00	26	16
SIQKFGERALKAWSVAR	487.77	1947.07	-0.02	36	16
SIQKFGERALKAWSVARL	515.79	2059.17	-0.03	55	16
SIQKFGERALKAWSVARLS	537.80	2147.19	-0.02	45	16
SIQKFGERALKAWSVARLSQ	570.06	2276.23	-0.02	65	18
SIQKFGERALKAWSVARLSQK	602.09	2404.32	-0.01	18	15
SIQKFGERALKAWSVARLSQKF	638.85	2551.39	0.00	24	16
SKLKECCDKPLLEK	409.22	1632.86	-0.01	46	16
SKLKECCDKPLLEKSH	465.23	1856.95	-0.05	49	17
SLGKVGTRCCTKPESERM	496.23	1980.95	-0.06	23	18
SLHTLFG	387.71	773.41	0.00	21	15
SLHTLFGD	445.22	888.43	0.00	29	16
SLHTLFGDEL	566.29	1130.56	0.00	29	16
SLHTLFGDELCKVA	511.59	1531.77	-0.01	58	17
SLHTLFGDELCKVASLRE	673.35	2017.03	0.01	77	18
SLHTLFGDELCKVASLRETY	571.29	2281.14	-0.02	62	20
SLHTLFGDELCKVASLRETYG	585.54	2338.16	-0.02	70	19
SLILNRLCVLHEK	513.63	1537.87	-0.01	30	13
SLRETYG	413.20	824.40	-0.01	23	14
SLRETYGDMA	571.76	1141.51	0.01	55	16
SLRETYGDMAD	637.28	1272.53	0.01	51	15
SLVNRRPCF	364.53	1090.57	-0.01	35	16
SPDLPKL	385.22	768.44	-0.01	41	13
SPDLPKLKPD	370.54	1108.61	-0.02	87	14
SRRHPEYAVSVLLRLAK	499.54	1994.15	-0.02	44	13
SRRHPEYAVSVLLRLAKE	531.81	2123.20	0.00	41	14

SRRHPEYAVSVLLRLAKEY	572.57	2286.26	-0.02	68	17
SRRHPEYAVSVLLRLAKEYE	604.83	2415.30	-0.01	26	17
SRRHPEYAVSVLLRLAKEYEA	622.59	2486.34	-0.01	78	17
SRRHPEYAVSVLLRLAKEYEATLE	708.39	2829.51	0.00	27	18
SRRHPEYAVSVLLRLAKEYEATLEE	740.65	2958.56	0.03	43	18
SRRHPEYAVSVLLRLAKEYEATLEECCAK	841.94	3363.71	0.01	50	20
SRRHPEYAVSVLLRLAKEYEATLEECCAKD	870.70	3478.73	0.02	95	21
SRRHPEYAVSVLLRLAKEYEATLEECCAKDD	899.45	3593.76	0.02	108	22
SSKLKECCD	506.73	1011.44	0.01	31	14
STPTLVEVSRSL	644.86	1287.70	0.01	36	14
STPTLVEVSRSLG	673.38	1344.72	0.01	40	13
STPTLVEVSRSLGKVG	543.98	1628.91	0.00	30	13
STVFDKLK	469.27	936.53	0.00	28	13
STVFDKLKHL	396.56	1186.67	-0.01	65	13
STVFDKLKHLV	429.58	1285.74	-0.02	47	13
STVFDKLKHLVD	467.92	1400.77	-0.02	60	15
STVFDKLKHLVDE	510.94	1529.81	-0.02	55	17
SVARLSQKFPKA	444.92	1331.76	-0.01	36	13
SVARLSQKFPKAEFVEVTKLVTD	649.11	2592.42	0.00	25	16
SVARLSQKFPKAEFVEVTKLVTDL	677.39	2705.50	0.03	39	13
TADFAEDKDVC	607.26	1212.50	0.01	24	14
TADFAEDKDVCK	447.88	1340.59	0.02	60	16
TADFAEDKDVCKN	486.21	1455.62	0.00	80	15
TADFAEDKDVCKNYQEAK	519.73	2074.92	-0.02	39	14
TADFAEDKDVCKNYQEAKD	548.73	2190.93	-0.03	46	13
TADFAEDKDVCKNYQEAKDAFLG	860.39	2578.15	0.01	21	16
TALVELLK	443.79	885.55	0.00	39	13
TALVELLKH	512.32	1022.61	0.01	49	13
TALVELLKHK	384.57	1150.71	-0.02	42	13
TALVELLKHKPK	459.62	1375.86	-0.01	29	13
TALVELLKHKPKA	362.72	1446.89	-0.03	46	13
TALVELLKHKPKATEEQLK	545.06	2176.25	-0.02	69	13
TCVADESHAG	495.20	988.39	0.00	39	13
TCVADESHAGCEKSLH	562.90	1685.71	-0.02	61	14
TEDYLSLILNRL	725.90	1449.77	0.01	67	15
TEDYLSLILNRLCVLH	634.67	1901.01	-0.02	50	17
TEDYLSLILNRLCVLHE	678.02	2031.03	0.00	57	19
TEDYLSLILNRLCVLHEK	540.54	2158.15	-0.03	49	19
TEEQLKTVMENFVAFVDK	710.68	2129.02	-0.01	90	17

TEFAKTCVAD	542.76	1083.49	0.02	25	14
TEFAKTCVADESHAG	522.57	1564.68	0.01	111	16
TEFAKTCVADESHAGCEK	482.21	1924.83	-0.03	19	14
TEKQIKKQTALVELLK	468.78	1871.10	-0.03	39	13
TEKQIKKQTALVELLKH	502.79	2007.17	-0.03	37	15
TESLVNRRPCF	441.22	1320.66	-0.01	25	16
TESLVNRRPCFSAL	531.94	1592.80	-0.01	21	18
TFHADICTLPD	616.79	1231.55	0.02	30	15
THKSEIAHRFKD	367.94	1467.76	-0.02	17	17
THKSEIAHRFKDL	396.21	1580.84	-0.03	22	17
THKSEIAHRFKDLG	546.95	1637.86	-0.05	37	18
THKSEIAHRFKDLGE	442.73	1766.91	-0.02	54	17
THKSEIAHRFKDLGEE	474.99	1895.95	-0.01	21	18
THKSEIAHRFKDLGEEH	509.25	2033.01	-0.04	96	18
THKSEIAHRFKDLGEEHFK	578.04	2308.17	-0.03	28	19
THKSEIAHRFKDLGEEHFKG	592.30	2365.19	-0.03	52	19
THKSEIAHRFKDLGEEHFKGLVLIAF	756.41	3021.62	0.00	57	18
TISSKLKECC	556.28	1110.54	0.00	26	15
TISSKLKECCD	613.80	1225.57	0.02	74	16
TISSKLKECCDK	452.23	1353.66	0.01	52	15
TISSKLKECCDKPLL	559.98	1676.88	0.02	17	16
TISSKLKECCDKPLLE	602.98	1805.93	0.01	20	17
TISSKLKECCDKPLLEK	484.51	1934.02	-0.02	96	18
TISSKLKECCDKPLLEKSH	540.53	2158.11	-0.02	65	18
TISSKLKECCDKPLLEKSHCIAEVEK	733.63	2930.49	0.00	121	20
TISSKLKECCDKPLLEKSHCIAEVEKD	762.38	3045.52	-0.01	27	21
TLCDEFKAD	521.23	1040.45	0.00	23	13
TLCDEFKADEK	433.53	1297.59	-0.03	28	13
TLCDEFKADEKK	476.23	1425.68	0.00	48	16
TLEECCAKDD	563.72	1125.43	-0.01	49	13
TLFGDELCKVA	598.31	1194.60	0.01	26	16
TLFGDELCKVASLR	517.94	1550.81	0.00	58	17
TLFGDELCKVASLRE	560.95	1679.86	-0.02	40	17
TLFGDELCKVASLRETY	649.00	1943.97	0.01	59	17
TLFGDELCKVASLRETYG	668.00	2000.99	0.00	18	17
TLPDTEKQIKKQTALVELLK	575.33	2297.31	0.00	59	13
TLPDTEKQIKKQTALVELLKH	609.60	2434.37	-0.01	45	14
TLVEVSRSL	502.29	1002.57	0.00	27	13
TLVEVSRSLGKVG	448.93	1343.78	-0.02	50	13

TPDETYVPK	525.27	1048.51	0.01	32	16
TPDETYVPKAFD	691.83	1381.64	0.01	67	15
TPDETYVPKAFDE	756.34	1510.68	-0.01	53	15
TPDETYVPKAFDEK	547.26	1638.78	-0.02	41	17
TPDETYVPKAFDEKLF	633.99	1898.93	0.01	42	19
TPDETYVPKAFDEKLFTFH	572.03	2284.11	-0.01	20	20
TPDETYVPKAFDEKLFTFHA	589.79	2355.14	-0.01	26	19
TPDETYVPKAFDEKLFTFHAD	618.55	2470.17	-0.02	26	19
TPTLVEVSR	501.29	1000.56	0.00	24	14
TPTLVEVSRSL	601.35	1200.67	0.02	69	13
TPTLVEVSRSLG	629.86	1257.69	0.00	26	14
TPTLVEVSRSLGK	462.94	1385.79	0.00	41	14
TPTLVEVSRSLGKV	495.96	1484.86	-0.01	50	13
TPTLVEVSRSLGKVG	514.97	1541.88	0.00	42	13
TPTLVEVSRSLGKVGTR	450.76	1799.03	-0.03	50	14
TPTLVEVSRSLGKVGTRCCTKPE	616.08	2460.28	-0.01	24	18
TPVSEKVTKCC	398.86	1193.58	-0.02	34	15
TPVSEKVTKCCTE	475.57	1423.67	0.01	72	17
TPVSEKVTKCCTES	504.58	1510.70	0.01	51	16
TPVSEKVTKCCTESL	542.27	1623.78	0.01	27	17
TPVSEKVTKCCTESLVN	613.64	1837.88	0.01	37	17
TPVSEKVTKCCTESLVNR	665.34	1993.00	0.01	83	18
TPVSEKVTKCCTESLVNRR	538.28	2149.10	-0.02	36	19
TPVSEKVTKCCTESLVNRRPCF	625.31	2497.21	0.02	74	20
TPVSEKVTKCCTESLVNRRPCFS	647.07	2584.24	0.00	50	20
TPVSEKVTKCCTESLVNRRPCFSAL	692.85	2767.38	-0.01	74	20
TRCCTKPESERMP	513.23	1536.68	-0.03	21	14
TRCCTKPESERMPCT	581.24	1740.74	-0.03	25	14
TRCCTKPESERMPCTE	624.26	1869.78	-0.03	32	14
TRCCTKPESERMPCTED	662.61	1984.81	-0.02	56	13
TRKVPQVSTPTLVEVSR	475.26	1897.06	-0.04	36	17
TRKVPQVSTPTLVEVSRS	662.36	1984.10	-0.04	32	17
TRKVPQVSTPTLVEVSRSL	525.30	2097.18	0.00	34	13
TRKVPQVSTPTLVEVSRSLG	539.55	2154.20	-0.01	40	16
TRKVPQVSTPTLVEVSRSLGK	571.58	2282.30	-0.02	23	13
TRKVPQVSTPTLVEVSRSLGKVG	610.35	2437.40	-0.01	41	13
TVFDKLKHLVD	657.88	1313.73	0.01	45	13
TVMENFVAFVDK	700.86	1399.67	0.03	44	16
TVMENFVAFVDKCCA	839.38	1676.72	0.01	61	16
TVMENFVAFVDKCCAAD	932.41	1862.79	0.02	22	14
TYVPKAF	413.23	824.44	0.00	16	13

TYVPKAFDE	535.26	1068.51	0.00	16	14
TYVPKAFDEK	399.88	1196.61	0.00	46	15
TYVPKAFDEKL	437.57	1309.69	-0.01	60	16
TYVPKAFDEKLF	486.59	1456.76	-0.02	38	15
TYVPKAFDEKLFT	520.28	1557.81	0.00	15	15
TYVPKAFDEKLFTF	569.29	1704.88	-0.02	21	17
TYVPKAFDEKLFTFH	461.49	1841.94	-0.02	90	18
TYVPKAFDEKLFTFHA	479.24	1912.97	-0.03	48	18
TYVPKAFDEKLFTFHAD	508.00	2028.00	-0.02	48	19
VCKNYQEAKDAFL	510.58	1528.72	-0.02	27	16
VCKNYQEAKDAFLG	529.26	1584.76	-0.01	24	17
VEGPKLVVSTQ	579.33	1156.63	0.01	33	13
VEGPKLVVSTQTALA	505.28	1512.84	-0.01	22	16
VEVSRSLGKVG	377.55	1129.65	0.00	25	13
VFDKLKHL	500.31	998.59	0.01	38	13
VFQECCQAEDK	651.26	1300.50	0.02	62	13
VFQECCQAEDKG	679.77	1357.52	0.01	54	13
VFQECCQAEDKGA	715.29	1428.55	0.00	78	13
VPKAFDEKLF	398.55	1192.65	-0.01	66	15
VPQVSTPTLVEVSR	756.93	1511.82	0.03	62	14
VPQVSTPTLVEVSRSLG	885.49	1768.96	0.02	55	15
VPQVSTPTLVEVSRSLGKVG	685.39	2053.14	0.01	59	13
VSTPTLVEVSR	594.33	1186.66	0.00	60	14
VSTPTLVEVSRSL	694.40	1386.77	0.01	50	14
VSTPTLVEVSRSLG	722.91	1443.79	0.01	60	13
VSTPTLVEVSRSLGKVG	576.99	1727.98	-0.02	41	14
VSVLLRLAKEYEA	497.62	1489.85	-0.01	26	13
VSVLLRLAKEYEATLEECCAKDD	650.32	2597.27	-0.02	44	20
WSVARL	366.21	730.41	-0.01	14	13
WSVARLSQKFPK	483.27	1446.80	-0.02	16	15
YAPELLYYANK	673.35	1344.66	0.02	51	15
YAPELLYYANKY	754.38	1506.74	0.01	43	17
YAPELLYYANKYN	811.90	1621.77	0.01	37	18
YAVSVLLRLAKEY	508.96	1523.87	-0.01	42	13
YAVSVLLRLAKEYE	551.98	1652.91	0.00	37	15
YAVSVLLRLAKEYEA	575.66	1723.95	0.01	40	13
YAVSVLLRLAKEYEATLEECCAK	651.33	2601.32	-0.04	44	20
YAVSVLLRLAKEYEATLEECCAKD	680.08	2716.35	-0.04	86	20
YAVSVLLRLAKEYEATLEECCAKDD	944.80	2831.37	-0.01	51	21
YLSLILNRL	553.33	1104.65	0.00	54	13

YLSLILNRLCVL	710.91	1419.82	-0.01	26	13
YLSLILNRLCVLH	519.97	1556.87	0.01	59	13
YLSLILNRLCVLHE	562.98	1685.92	0.01	27	15
YLSLILNRLCVLHEK	454.50	1814.01	-0.03	66	14
YQEAKDAFL	543.26	1084.51	0.00	15	14
YQEAKDAFLG	571.28	1140.55	0.00	23	16
YSRRHPEYAVSVLLRLAKEYEA	663.36	2649.40	0.00	43	18
YSTVFDKLKHLVD	782.92	1563.83	-0.01	20	17
YVPKAFDEKLF	452.91	1355.71	-0.01	43	14

9.5 Literature Cited

(1) Hattori, S.; Iida, N.; Kosako, H. *Expert Review of Proteomics* **2008**, *5*, 497-505.

(2) Kruger, M.; Kratchmarova, I.; Blagoev, B.; Tseng, Y.; Kahn, C.; Mann, M. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 2451-2456.

(3) Marcantonio, M.; Trost, M.; Courcelles, M.; Desjardins, M.; Thibault, P. *Molecular & Cellular Proteomics* **2007**, *6*, 33-33.

(4) Tekirian, T.; Thomas, S.; Yang, A. *Expert Review of Proteomics* **2007**, *4*, 573-583.

(5) Borchers, C.; Thapar, R.; Petrotchenko, E.; Torres, M.; Speir, J.; Easterling, M.; Dominski, Z.; Marzluff, W. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 3094-3099.

(6) Loo, R.; Yang, Y.; Dunsmore, J.; Williams, K.; Boontheung, P.; Sondej, M.; Mouttakki, H.; Wong, D.; Mcinerney, M.; Gunsalus, R.; Loo, J. *Molecular & Cellular Proteomics* **2006**, *5*, S298-S298.

(7) Mclafferty, F.; Breuker, K.; Jin, M.; Han, X.; Infusini, G.; Jiang, H.; Kong, X.; Begley, T. *Febs Journal* **2007**, *274*, 6256-6268.

(8) Parks, B.; Jiang, L.; Thomas, P.; Wenger, C.; Roth, M.; Boyne, M.; Burke, P.; Kwast, K.; Kelleher, N. *Analytical Chemistry* **2007**, *79*, 7984-7991.

(9) Fournier, M.; Gilmore, J.; Martin-Brown, S.; Washburn, M. *Chemical Reviews* **2007**, *107*, 3654-3686.

(10) Mcdonald, C.; Li, L. *Analytica Chimica Acta* **2005**, *534*, 3-10.

(11) Zhong, H.; Zhang, Y.; Wen, Z.; Li, L. *Nature Biotechnology* **2004**, *22*, 1291-1296.

(12) Zhong, H.; Marcus, S.; Li, L. *Journal of the American Society for Mass Spectrometry* **2005**, *16*, 471-481.

(13) Giberson, R.; Demaree, R. *Microscopy Research and Technique* **1995**, *32*, 246-254.

(14) Kok, L.; Boon, M.; Smid, H. *Scanning* **1993**, *15*, 100-109.

(15) Yassine, M.M.; Guo, N.; Zhong, H.; Li, L.; Lucy, C.A. *Analytica Chimica Acta* **2007**, *597*, 41-49.

(16) Liu, H.; Stupak, J.; Zheng, J.; Keller, B.; Brix, B.; Fliegel, L.; Li, L. *Analytical Chemistry* **2004**, *76*, 4223-4232.

(17) Lee, B.; Krishnanchettiar, S.; Lateef, S.; Lateef, N.; Gupta, S. *Rapid Communications in Mass Spectrometry* **2005**, *19*, 2629-2635.

(18) Sandoval, W.; Arellano, F.; Arnott, D.; Raab, H.; Vandlen, R.; Lill, J. *International Journal of Mass Spectrometry* **2007**, *259*, 117-123.

(19) Warrand, J.; Janssen, H. *Carbohydrate Polymers* **2007**, *69*, 353-362.

Chapter 10

Conclusions and Future Work

In a proteomics application, ideally all the proteins or the entire proteome in a sample should be profiled. However, this is a very challenging task. The overall goal of my thesis research was to develop new mass spectrometric techniques for comprehensive proteome analysis. By comprehensive, I mean we wanted to detect the majority of the proteins present in a proteomic sample. Several techniques have been described in this thesis along with the illustration of their analytical performances for proteome profiling of complex samples.

After a brief introduction to a number of key techniques and methods related to my thesis work in Chapter 1, I described a work on the development of mass spectrometric methods for proteome analysis of human tear fluids in Chapter 2. This work is significant as tear proteome profiling may generate useful information for the understanding of the interaction between an eye and its contacting objects, such as a contact lens or a lens implant. This is important for designing improved eye-care devices and maintaining the health of an eye. Proteome profiles of tear fluids may also be used for disease diagnosis and prognosis. However, only a small volume of tear fluid (<5 μL) can be collected in a clinical laboratory under normal operational conditions, which makes proteome profiling a challenge. In our work, as described in Chapter 2, we applied several proteomic analysis techniques, including gel-based and solution-based approaches with LC-ESI and LC-MALDI MS and MS/MS to gauge the relative merits of producing proteome profiles and to generate as broad a coverage of the tear proteome as possible from this small amount of sample. It was shown that a total of 54 proteins could be confidently identified using less than 5 μL of tear fluid. Of these, 44 proteins could be detected by LC-MALDI MS alone with a consumption of 2 μL of tear fluid. Furthermore, LC-MALDI could be used to determine post-translational modifications (PTMs), such as glycosylation and phosphorylation, without any sample enrichment or treatment. At the time we published this work in *Journal of Proteome Research* in 2005, this work represented one of the most extensive proteome profiles (i.e., proteins identified and PTMs characterized) generated from tear fluids using clinically relevant amounts of sample.

In Chapter 3, I described my research on the analysis of zebrafish proteome using a shotgun method based on protein digestion and LC-ESI MS/MS analysis of the resultant peptides. Zebrafish has been widely used as a model system for biological studies ranging from developmental biology to toxicology. At the time I started to work on this project in 2005, in collaboration with Professor Greg Goss's lab, Biological Sciences, U of A, the zebrafish genome had just been sequenced and annotated. The availability of the genome database opened the possibility of high throughput proteomic analysis. In a paper published in *Journal of Proteome Research* in 2006 which is adapted for Chapter 3, we reported for the first time a proteomic subset of zebrafish liver, an important organ for metabolising toxins. Using a then newly developed analytical procedure, we identified 1204 proteins from the cytosolic component of a zebrafish liver tissue sample. Our methods involved cell-compartment fractionation of liver tissue samples, four levels of protein digestion, and off-line two-dimensional liquid chromatography (2D-LC) separations of resultant peptides. Proteins were identified using an electrospray ionization quadrupole time-of-flight tandem mass spectrometer, which provides high resolution and high accuracy mass measurement of peptide ions and their fragment ions. We demonstrated that greater proteome coverage could be achieved by combining the results obtained from four methods of protein digestion: three tryptic digests (one in buffer, one in methanol, and another in SDS), and a microwave-assisted acid hydrolyte of the protein extracts. Identified proteins - which included several groups of established protein biomarkers - were functionally classified. In Chapter 3, I have also discussed the functions and implications of these biomarkers within the context of zebrafish toxicology.

In shotgun proteome analysis by LC-MS/MS, not all coeluting peptides at a given retention time are subjected to MS/MS due to the limitation of spectral acquisition speed of a mass spectrometer. In Chapter 4, precursor ion exclusion (PIE) in an ESI quadrupole time-of-flight (QTOF) mass spectrometer was explored as a means of mitigating the under-sampling problem, and this work has been published in *Analytical Chemistry* in 2008. This strategy is based on running replicates of the sample where the precursor ions detected in the initial run(s) are excluded for MS/MS in the subsequent run. Four PIE methods as well as running replicates without PIE were investigated and compared for their effectiveness in identifying peptides and proteins. In the analysis of a MCF7 breast cancer cell lysate digest by three replicate 2-h gradient LC-ESI runs, the first PIE method used a list of precursor ions detected in the initial run(s) for exclusion and identified a total of 572 proteins from the three runs combined with an average of 3.59 peptides matched to a protein. The second PIE method involved in the generation of a list of m/z values of precursor ions along with their retention time information from the initial run(s), followed by entering these ions with retention times into the ion exclusion program of the QTOF control software for exclusion at a predefined retention time window (i.e., ±150 s). Compared to the first PIE method, this method reduced the possibility of excluding different peptide ions of the same m/z (within a mass tolerance window) eluted at different retention windows. A total of 657 proteins were identified with an average of 3.75 peptides matched to a protein. The third PIE method studied relied on the exclusion of the precursor ions of peptides identified through database search of the MS/MS spectra generated in the initial run(s). This method identified a total of 681 proteins with an average of 3.68 peptides matched to a protein. The final PIE method investigated involves the expansion of the selective PIE list by including nonidentifiable peptide ions found in the database search. This complete PIE method identified a total of 726 proteins with an average of 3.66 peptides per protein. In the case of three replicate runs without PIE, a total of 460 proteins were identified with an average of 3.51 peptides matched to a protein. Thus, the use of an optimal precursor ion exclusion strategy significantly increased the number of proteins identified from replicate runs (i.e., 726 vs 460 or a 58% increase).

In Chapter 5, I described a strategy of maximizing the performance of RPLC MS/MS by optimizing the sample loading to the instrument in an off-line 2D-LC tandem MS platform. To determine the quantity of peptides present in a proteome digest or fractionated peptides from strong-cation exchange (SCX) separation, an automated system based on RPLC with rapid step solvent gradient for peptide elution and ultraviolet

(UV) detection was developed. This system also allowed the purification of the peptides by getting rid of salts and other impurities present in a sample. It was found that controlling the amount of peptides injected into a RPLC MS/MS system was critical to achieve the maximum efficiency in peptide and protein identification. Using off-line 2D-LC MS/MS, peptide fractions from the 1st dimension of separation were desalted and quantified, followed by injecting the optimal amount of the sample into RPLC MS/MS for peptide sequencing. The application of this strategy was demonstrated in the proteome profiling of breast cancer MCF-7 cells. From the analysis of 28 SCX fractions with each injecting 1 μg of sample into a 75 μm × 100 mm C18 column interfaced to a QTOF mass spectrometer, a total of 2362 unique proteins or protein groups were identified with a false positive peptide identification rate of 0.19%, as determined by target-decoy proteome sequence searches. Replicate 2-h runs of individual fractions with the exclusion of precursor ions of peptides already identified in the first runs resulted in the identification of an additional 549 unique proteins or protein groups with a false positive identification rate of 0.60%. Finally, the biological significances of some of 2911 proteins identified in MCF-7 cells were discussed within the context of reported putative breast cancer biomarkers.

After I have developed the sequential protein solubilization and digestion method (Chapter 3), the effective precursor ion extraction (PIE) technique for running LC-MS/MS experiments (Chapter 4), and the maximum sample loading strategy to LC-MS/MS for improving peptide and protein identification efficiency (Chapter 5), I applied these new tools to generate a comprehensive proteome profile of zebrafish liver. This work was described in Chapter 6. In our experiment, the liver sample was first fractionated according to their cellular compartments. The membrane fraction was analyzed in this study. The proteins in the membrane fraction of the liver extracts were first acetone precipitated. The protein pellet was then subjected to sequential solubilization in the order of NH_4HCO_3, methanol and SDS. The solubilized proteins in each sample were digested by trypsin. The digests were analyzed by the off-line 2D LC-ESI QTOF MS technique with PIE and maximum sample loading to RPLC-ESI MS/MS. A total of 5671 unique proteins or protein groups were identified from the three samples. This represents the most comprehensive proteome coverage of the zebrafish liver to date.

Compared to the cytosolic fraction analysis where 1204 proteins were identified (Chapter 3), the membrane fraction was significantly more difficult to analyze, as it contained a large portion of membrane proteins. Analyzing membrane proteins was a challenge. However, with our sequential solubilization method, most of the proteins (4476 or 78.9% of the 5671 proteins) could be identified from the SDS fraction. The fact that a much greater number of proteins could be identified from a more challenging sample, i.e., the membrane fraction, using our newly developed methods indicates that our methods have provided a substantial improvement in proteome analysis over those used only a couple of years ago. However, the predicted number of proteins from the zebrafish genome is around 17,000. While we do not know how many proteins are actually expressed or present in the liver, we cannot determine whether our method has identified all the proteins present in the sample. However, it is safe to say that further work (i.e., analyzing the nuclear and cytoskeletal components of the liver and re-analyzing the cytosolic component with the improved method) will allow the identification of additional proteins from the liver sample.

To gauge the overall performance of our methods in terms of proteome coverage, we carried out a comprehensive proteome analysis of *E. coli*, a simple microorganism with a known proteome size of about 4300 proteins. This work was described in Chapter 7. In this work, *E. coli* cells were lysed with French press and the lysed cells were fractionated into cytoplasm, peripheral and integral membrane fractions. Each fraction was then subjected to sequential solubilization and digestion, followed by 2D-LC ESI MS/MS analysis of the individual digests. In addition, to enrich the low molecular weight proteins (<30 kDa), two molecular weight cutoff filters (10 kDa and 30 kDa filters) were used to fractionate the soluble proteins in the cytoplasm and peripheral fractions. The low MW fractions were also analyzed by trypsin digestion and 2D-LC MS/MS. Combined all the data generated from the multiple fractions, a total of 3730 unique proteins or protein groups were identified. Thus, about 86.7% of the 4300 predicted proteins were detected in this study. About 570 proteins not identified in this work were mainly low molecular weight proteins (MW<60 kD), about half of them having molecular weights of less than 30 kD. The cause of missing identification of these proteins is unknown. It could be a technical limitation, although there was no evidence suggesting this. We speculate that it is likely that these proteins were not expressed and hence not present in the sample from the cells grown in a rich media.

While the *E. coli* results described in Chapter 7 illustrate that a good progress on comprehensive proteome analysis has been made, it was done with a lot of starting materials: millions or billions of cells. From an analytical chemistry point of view, an even more challenging task is to generate a comprehensive proteome profile from a limited number of cells, such as a few cancer cells in a tumor tissue. Some preliminary work has been carried out along this direction. In Chapter 8, I described a shotgun proteome analysis method and its performance for protein identification from thousands of cells. Since a small number of cells were used, cell lysis was done using a detergent (NP-40) instead of French press. The lysed cells were subjected to acetone precipitation, followed by washing with cold acetone and then solubilizing in NH_4HCO_3. After trypsin digestion of the solubilized proteins, the digest was analyzed by using LC-ESI MS/MS. Gradient speed in running LC-MS/MS was optimized according to the sample amount injected into the column. It was shown that this method could identify an average (n=3) of 167±21, 237±30, 491±63, and 619±59 proteins from 500, 1000, 2500, and 5000 MCF-7 cells, respectively. To demonstrate the potential use of this method for generating proteome profiles from cancer cells isolated from human blood, MCF-7 cells were spiked to a healthy human blood sample and this mixture was processed and then subjected to antibody tagging of the MCF-7 cells. The tagged cells were sorted and collected using flow cytometry. The proteome profiles of small numbers of cells isolated in this way were found to be similar to those of the MCF-7 cells. This work illustrated that we could potentially do proteome profiling of a small number of cells isolated from blood and then compare the profile with a standard profile to do cell typing, which may prove to be useful for cancer diagnosis or prognosis.

Aside from protein identification, proteomics work often requires the characterization of protein modifications such as identification of the modification groups and sites of proteins. Microwave-assisted acid hydrolysis (MAAH) originally developed in Professor Li's group by a former PhD student, Hongying Zhong, is one of the promising techniques for analyzing protein modifications. In Chapter 9, I described my research on the development of an improved MAAH setup for efficient and reproducible protein degradation into peptides and the combination of this improved setup with the sensitive LC-ESI MS/MS method described in Chapters 4 and 5 for mapping protein sequences. It was demonstrated that, for BSA, a protein with a molecular weight of about 67,000 Da, the entire protein sequence could be covered by the peptides produced in TFA MAAH using about 2 μg of sample. In the analysis of α- and β-casein, we illustrated that this method could be used for mapping phosphorylation sites in phosphoproteins. Future work will involve in applying this technique to characterize real world sample.

As summarized above, my thesis work was mainly focused on the development of the LC-MS/MS technology for proteome analysis. I have developed some tools that facilitated the proteome analysis, particularly for increasing proteome coverage. However, there are still a lot of work remained to be done to realize the ultimate goal of detecting the entire proteome of a cell or tissue sample. For example, in the analysis of the zebrafish proteome, we need to complete the analysis of different cellular fractions of the liver extracts and combine the identification results to generate a more complete map of the liver proteome. In addition, MAAH should be applied to the analysis of the remaining pellets after sequential solubilization. The improved MAAH setup as described in Chapter 9 should facilitate the generation of hydrolysates from a small amount of pellets.

In the *E. coli* work, we need to investigate the cause(s) of missing identification of about 570 proteins. Perhaps antibody based protein identification methods and/or RNA expression analysis may be used to confirm whether the missing proteins are present in the cultured cells. If these proteins were indeed present in the cells, it would indicate that our method had some limitation and we would need to further develop techniques to address the problem. We also need to improve the efficiency of proteome analysis to significantly shorten the analysis time. In our current work, many proteins were identified multiple times in different protein fractions. Better resolution in protein fractionation may reduce the redundancy. Alternative, proteins may be fractionated for solubilization and digestion and the resultant digests may be pooled to form one complicated peptide sample. This sample may be subjected to multidimensional LC separation followed by MS/MS. These strategies are being pursed in our laboratory.

To generate a proteome profile of a small number of cells, we need to improve the sample preparation protocol to minimize sample loss. One approach may be to perform all the sample handling steps in one apparatus such as inside a capillary or micro-vial and the resultant sample may be directly introduced into LC MS/MS. In addition, further improvement in MS detection may be needed to handle the small amount of sample generated from a few cells. For example, using a capillary column with ID of less than 75 μm may improve the detection limit in absolute amounts. Another important question we need to address is how many proteins are required to form a unique proteome profile from which one cell type can be unambiguously differentiated from the others? The use of proteome profiling for cancer diagnosis and prognosis is new and has yet been demonstrated to be effective. Suffice it to say that a lot of work remains to be done in this area.