University of Alberta

# Mass Spectrometric Analysis of Protein Sequences and Posttranslational Modifications

by

Hongying Zhong ©

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

Department of Chemistry

Edmonton, Alberta

Fall 2004

*Your file Votre référence*
*ISBN: 0-612-96046-3*
*Our file Notre référence*
*ISBN: 0-612-96046-3*

*To my dear father, mother and my husband*
*Especially to my lovely daughter Anji*
*With love and respect*

*For the beautiful dream in the heart*

# Acknowledgements

I would like to sincerely thank my supervisor Professor Liang Li for his excellent guidance, encouragement, discussion and fruitful criticism in the long journey of my graduate study. His enthusiasm, dedication and inspiration will continue to benefit my future career.

I would also like to thank my committee members, Dr. Charles A. Lucy, Dr. George Kotovych, Dr. John-Bruce Green, Dr. Andrew Shaw and Dr. George R. Agnes for constructive suggestions on this thesis.

I would like to thank my collaborators. Without their assistance, my research maybe more difficult. Especially I acknowledge Dr. Sandra Marcus (cell culture lab, Department of Chemistry, University of Alberta) for her kindness, stimulating discussions and performing cell culture of *E. coli* K12 and human cancer cell line MCF7 and HT29; Dr. Christina Benishin (Department of Physiology, University of Alberta) for bioassay experiments; Dr. Dave Wishart and Ms. Haiyan Zhang, Mr. Nelson Young (Department of Pharmacy and Pharmaceutical Sciences, University of Alberta) for their collaboration in the $C^{++}$ computer program for peptide *de novo* sequencing; CV Technology Inc. (Edmonton, Alberta) for providing shark cartilage and performing part of the bioassay experiments.

I would like to thank the Department of Chemistry and University of Alberta for giving me the opportunity to study in this beautiful university and providing me the facilities. I would like to thank the financial support from Natural Sciences and Engineering Research Council of Canada, Genome Canada, Canada Foundation for Innovation, The Protein Engineering Network of Centres of Excellence, and Alberta Cancer Board.

I would like to thank the kind help and immediate assistance from the machine shop, electronic shop, glass shop, IT group, general office and copy room in the Department of Chemistry. I would also like to thank Dr. Randy Whittal and Dr. Don Morgan for their support as I used the instruments in the Mass Spectrometry Lab of Chemistry Department. I appreciate Dr. Paul Semchuk and Dr. Lorne Burke for allowing me use their ESI-Q-TOF and HPLC system in the Institute of Biomolecular Design of the University of Alberta.

I would also thank Dr. Andrew Shaw and Mr. Alan Mak in the Cross Cancer Institute for allowing and helping me use their electroblot equipment. My thanks also go to Dr. Monica Palcic in the chemistry department for allowing me use the lyophilizer, microplate reader and centrifuge in her lab.

In the past four and half years, I have enjoyed so much fun and help from the research group that I have been working with. My thanks then will go to my colleagues: Dr. Alan Doucette, Dr. David Craft, Mr. Chris Macdonald, Mr. Jacek Stupak, Dr. Bernd Keller, Dr. Talat Yalcin, Dr. Ken Yeung, Dr. Zhengping Wang, Dr. Ying Zhang, Dr. Nan Li, Dr. Dawn Richards, Dr. Boyan Zhang, Mr. Chengjie Ji, Mrs. Jing Zheng, Mr. Kevin Dunlop, Ms. Xinlei Yu, Mr. Mulu Gebre, Mr. Zhihui Wen, Ms. Lidan Tao, Ms. Nan Guo, Ms.

Tien Quach, Ms. Nan Wang. Especially I would like to thank Mr. Chengjie Ji and Mr. Jiang Jiang for the great friendship of sending me food in a very cold winter night when I had to stay in the lab. My special thanks will also go to Dr. Zhengping Wang and her family for their providing me accommodation and driving me around when I first came to this new city.

Finally, I would like to thank my dear parents and my sisters for their love, support and encouragement. In the deep heart, I would like to thank my dear husband Yangfan Gao for taking lots of housework and taking care of our daughter. Because of his sacrifice I can then focus on my studies and research. Especially, I owe too much thanks to my lovely daughter Anji. She is so understanding at her young age. I appreciate so much that she is always waiting for me to read story and play piano for her without any complaints. Her innocence and her smile inspire and brighten my life.

## Table of Contents

# List of Tables

## List of Figures

## List of Abbreviations

| | |
|---|---|
| 2D HPLC | 2 Dimensional high performance liquid chromatography |
| 2D-MS | 2 Dimensional mass spectrometry |
| ACN | Acetonitrile |
| ATCC | American Type Culture Collection |
| BLAST | Basic local alignment search tool |
| BR | Bacteriorhodopsin |
| BRCA1 | Breast cancer type 1 susceptibility protein |
| BSA | Bovine serum albumin |
| CAD | Collision activated dissociation |
| CE | Capillary electrophoresis |
| CHAPS | 3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate |
| CID | Collision induced dissociation |
| CNBr | Cyanogen bromide |
| cSNPs | SNPs occurring in the coding regions |
| CYC | Cytochrome C |
| Da | Dalton |
| DHB | 2, 5-dihydroxybenzoic acid |
| DNA | Deoxyribonucleotide |
| DTT | Dithiothreitol |
| ECD | Electron capture dissociation |
| *E. coli* | *Escherichia coli* |

| | |
|---|---|
| EDTA | Ethylenediamine tetraacetate |
| ESI | Electrospray ionization |
| FTICR | Fourier transform ion cyclotron resonance |
| G3P2 | Glyceraldehyde-3-phosphate dehydrogenase |
| GC | Gas chromatography |
| GRAVY | Grand average of hydrophobicity |
| HCCA | α-Cyano-4-hydroxycinnamic acid |
| HH | Hereditary hemochromatosis |
| HPLC | High performance liquid chromatography |
| HSPs | High score pairs |
| ICAT | Isotope-coded affinity tag |
| IRMPD | Infrared multiphoton dissociation |
| LC | Liquid chromatography |
| QTOF | Quadrupole time of flight |
| MAAH | Microwave assisted acid hydrolysis |
| MALDI | Matrix assisted laser desorption/ionization |
| MAPs | Mass analysis of polypeptide ladders |
| MCAT | Mass-coded abundance tag |
| MCP | Multichannel plate |
| MI | Microwave irradiation |
| MS | Mass spectrometry |
| MudPIT | Multidimensional protein identification technique |
| MW | Molecular weight |

| | |
|---|---|
| m/z | mass to charge |
| NCBI | The National Center for Biotechnology Information |
| OMPA | Outer membrane protein A |
| ORF | Open reading frame |
| PBS | Phosphate-buffered saline |
| pI | Isoelectric point |
| PMSF | Phenylmethyl sulfonyl fluoride |
| ppm | part(s) per million |
| PTMs | Posttranslational modifications |
| RF | Radio frequency |
| RNA | Ribonucleotide |
| RP | Reversed phase |
| SCX | Strong cation ion exchange |
| SDS | Sodium dodecylsulfate |
| SDS PAGE | Sodium dodecylsulfate polyacrylamide gel electrophoresis |
| S/N | Signal to noise ratio |
| SNP | Single-nucleotide polymorphism |
| TFA | Trifluoroacetic acid |
| TMD | Transmembrane domain |
| TOF | Time of flight |
| Tris | Tris (hydroxymethyl) aminomethane |
| Triton® X-100 | t-Octylphenoxypolyethoxyethanol |
| Ubi | Ubiquitin |

| | |
|---|---|
| m | milli- ($10^{-3}$) |
| μ | micro- ($10^{-6}$) |
| n | nano- ($10^{-9}$) |
| p | pico- ($10^{-12}$) |

# Part I. Mass Spectrometry and Functional Proteomics

In the post-genomics era, more and more DNA sequences accumulate in the databases. High throughput and large-scale protein identification has been achieved by searching protein and DNA databases directly using data produced by mass spectrometry [1-5]. However, merely having complete sequences of genomes is not sufficient to elucidate biological function. The dynamic nature of the proteome of a cell or a tissue provides great justification for studying gene expression directly at the proteomic level, especially in disease [6-8]. Capturing this dynamic state that is not encoded in DNA sequences represents a challenge to the established technologies. To understand disease pathogenesis and develop effective strategies for early diagnosis and treatment, new technologies are needed to complement genomic analysis.

## I. 1. From Expression Proteomics to Functional Proteomics

DNA contains genes for specific proteins in both prokaryotes and eukaryotes. As shown in Figure I-1-(A) and (B), the synthesis of proteins in biological organisms proceeds by transcription of DNA (deoxyribonucleotide) sequences to messenger RNA (mRNA). mRNA is then translated into proteins by ribosomes. As the results of posttranslational modifications, a bewildering number of gene products will be generated. Therefore, the existence of an open reading frame (ORF) in genomic data does not necessarily imply the existence of the corresponding functional gene. Even with the advances in bioinformatics, it is still difficult to accurately predict genes, especially small genes [9-

13]. Proteomic methods thus become a necessary step in annotating the genome. A powerful proteomic approach is *de novo* analysis of proteins. As demonstrated by Peter *et al.*, proteomics complements genomics in showing which genes are really expressed [14]. They discovered the expression of six genes of *Mycobacterium tuberculosis* not predicted by genomics. Their discovery was supported by evidence from two-dimensional electrophoresis, matrix-assisted laser desorption ionization and *de novo* peptide sequencing using MS/MS data generated from nano-electrophoresis mass spectrometry. Similarly, Valerie *et al.* studied small genes or gene products in *Escherichia coli* k 12 by Edman protein degradation micro-sequencing and found only 14/42 isolated small proteins (with Mr 6-15 KDa) were expressed as annotated [15]. Posttranslational modifications such as cleavage of methionine start codon, loss of a signal peptide, internal and N-terminal fragmentation of much larger proteins have all been detected. These results highlight the necessity to complement genomic analysis with detailed proteomic analysis in order to obtain a better understanding of cellular molecular biology. This is especially important for large size genomes such as the human genome [16-18]. As shown in Figure I-1-(B), the coding sequences of eukaryotic genes are usually not continuous but consist of protein coding sequences called exons that are interrupted by non-coding sequences known as introns. In a process known as RNA splicing, the introns are removed and the exons are linked together to form a mature mRNA that can be translated into protein. So it is obvious that the final RNA molecule may be different due to the utilization of alternative exon combinations. Additionally, almost all proteins are posttranslationally modified in processes that may range from simple proteolytic cleavage to covalent modification of specific amino acid residues. As many as 200 types of covalent modification may exist and their biological function are still poorly understood [19-21]. Thus the same DNA region can in many cases lead to many proteins with different structures.

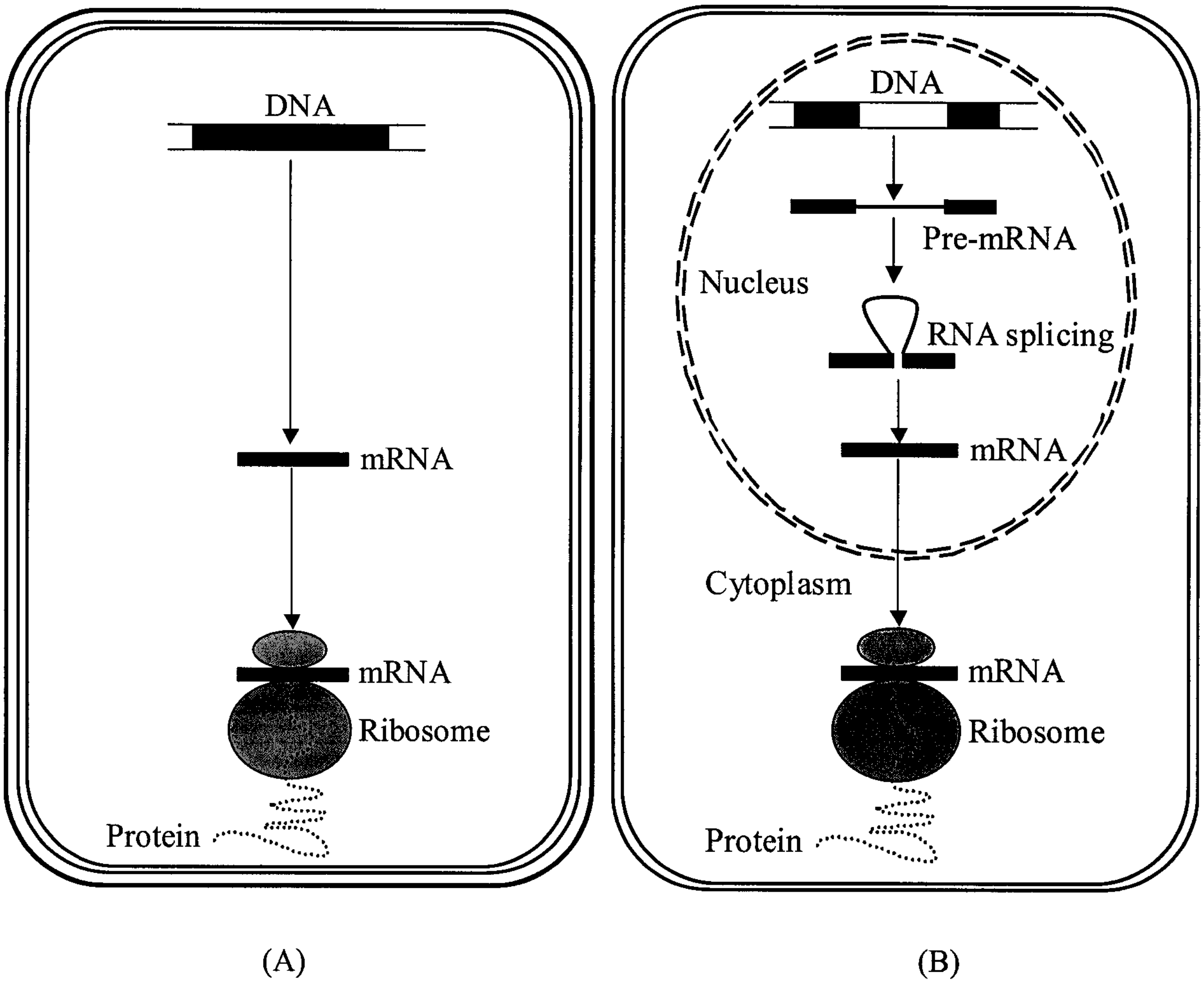


Figure I-1. Gene structure and the flow of genetic information in (A) a prokaryote cell and (B) an eukaryote cell. In prokaryote cells, the genetic information is stored as a continuous segment of DNA, and the messager RNA can immediately direct the synthesis of the corresponding protein. In eukaryote, the gene is usually split, and the precursor messenger RNA has to be processed by splicing before it can be translated into a protein.

The importance of studying alternative splicing and posttranslational modifications is directly reflected in medical aspects. The well-known disease called beta-thalassemia is caused by errors in the splicing process, resulting in the synthesis of poorly functioning beta-hemoglobin [22, 23]. In this case, the life-span of the red blood cells is shortened resulting in anemia. Mutations in a gene called *BRCA1* have been found in human breast or ovarian cancer [24-25]. It was found that the loss of the last 11 amino acids at the protein's C terminus is associated with aggressive, early-onset breast cancer. Another example is provided by studies of heart disease [26-28]. Although only a small proportion of the proteome has been analyzed, huge changes in the composition of the cardiac proteome has been found, affecting proteins with diverse functions. Protein expression studies have uncovered proteins that exhibited new disease-related posttranslational modifications with predicted functional relevance. For instance, novel phosphorylation of myosin light chain I was found in the failing heart.

So it is obvious that proteome alterations in disease may occur in many different ways that are not predictable from genomic analysis. It is clear that a better understanding of these alterations will likely have a significant impact on medicine. Further technological innovations are needed to increase sensitivity, reduce sample requirement, increase throughput and more effectively uncover various types of protein alterations such as posttranslational modifications, and amino acid mutations in order to meet the need for better diagnostics and to shorten the path for developing effective therapies.

There is substantial interest in applying mass spectrometry to proteomics [29-35]. It has the potential to yield comprehensive profiles of peptides and proteins in biological fluids with high throughput and high sensitivity. In the past few years, MS has enjoyed tremendous success and has become an indispensable technology for the interpretation of the information encoded in genomes.

## I. 2. Principles and Instrumentation of Mass Spectrometry

A mass spectrometer usually has seven major components: a sample introduction interface; an ion source; a mass analyzer; a detector; a vacuum system; an instrument control system; and a data processing system. The throughput, sensitivity, accuracy, resolution and the ability to generate information-rich spectra from protein or peptide fragments of a mass spectrometer are largely dependent on the sample introduction interface, the ionization efficiency within the ion source and the mass analyzer.

### I. 2. 1. Sample Introduction

One of the advantages of mass spectrometry is the ability to couple the sample inlet with separation technologies such as microfluidic chips, CE and HPLC. This feature enables mass spectrometry be used with very complex biological samples. Figure 1-2 illustrates interfaces for mass spectrometer to couple with HPLC separation. Panel A summaries the interfaces that are often used in HPLC-ESI coupling [36-44]. The simplest one is sheathless interface that consists of a sole capillary (Figure I-2-A-(1)). The high voltage can be applied via a metallic junction connected upstream. If incompatible solvents are used such as high surface tension or inadequate pH, a sheath liquid can be supplied via a coaxial capillary to sustain the spray stability and efficiency (Figure I-2-A-(2)). The outer capillary can be coated with conductive metal such as gold so that high voltage can be applied to the solution. The principle is the same as liquid junction (Figure I-2-A-(4)) except that the voltage connection is located far upstream. If used in high-flow range, it is necessary to flow a sheath gas over the sheath liquid capillary to enhance the solvent evaporation (Figure I-2-A-(3)). The gas flow also destabilizes the Taylor cone and therefore helps with the formation of droplets. Similarly, an ultrasonic transducer can also be used. Among these configurations, sheathless or liquid junction designs are usually employed for nanospray interface coupled with capillary HPLC system. Decreasing in dimension and flow rates allows the formation of smaller droplets and then a more efficient solvent evaporation during the gas phase ion formation and thus an enhanced ionization efficiency [45-50]. Sheathless design was used in the following research project. Newly developed MudPIT technology [51] (Figure I-2-A-(5)) for shotgun proteomics designs a biphasic microcapillary column packed with strong cation-exchange and reverse-phase material connected to a microcross. The microcross splits the flow from HPLC column and also serve as the connection for the electrospray voltage.

Panel B summarizes the interfaces that are often used in HPLC and MALDI coupling [52-56]. Piezoelectric flow–through microdispenser applies a short voltage pulse at a given frequency and pressure pulses were generated to eject a burst of droplets from the continuous flow (Figure I-2-B-(1)). The droplets are collected on the MALDI sample target. Continuous on-line LC-MALDI-TOF interface deposit the effluent mixed with a suitable matrix on a rotating quartz wheel and transport to the repeller where laser desorption takes place (Figure I-2-B-(2)). Electric field-driven droplet deposition is another choice for LC-MALDI interface (Figure I-2-B-(3)). The flexibility of this method allows the co-deposition of eluents with MALDI matrixes that range from hydrophobic to hydrophilic. Heated droplet interface is designed to handle high flow rate LC (Figure I-2-B-(4)) and it was used in the following research projects. A newly developed LC/MALDI interface is called laser spray (Figure I-2-B-(5)). With this method, the solvent in the LC effluent acts as a matrix and an infrared laser was used for vaporization and irradiation. Compared with other HPLC-MALDI interfaces, this interface is much more simple and should become a versatile interface for coupling mass spectrometry with on-line separation techniques.

### I. 2. 2. Ionization Methods

Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) as shown in Figure I-3 and Figure I-4 are the two techniques that are commonly used to volatize and ionize proteins and peptides or other small molecules for mass spectrometric analysis. ESI ionizes analyzes out of a solution and is readily coupled to liquid-based separation tools. There are four major processes in ESI [57-59]. (1) Charged droplets are produced at the ESI capillary tip. Effluent from LC or other source continuously build up at the liquid surface and cause the formation of a Taylor cone. Liquid filament is then subsequently formed due to the instability of the Taylor cone in the presence of the high electric field at the capillary tip. As the liquid filament becomes more and more unstable, charged droplets were ejected out of the capillary tip. (2) By solvent evaporation, these charged droplets shrink and result in smaller charged droplets. (3) Charged droplets are then further split into offspring droplets through uneven fission. As solvent continuously

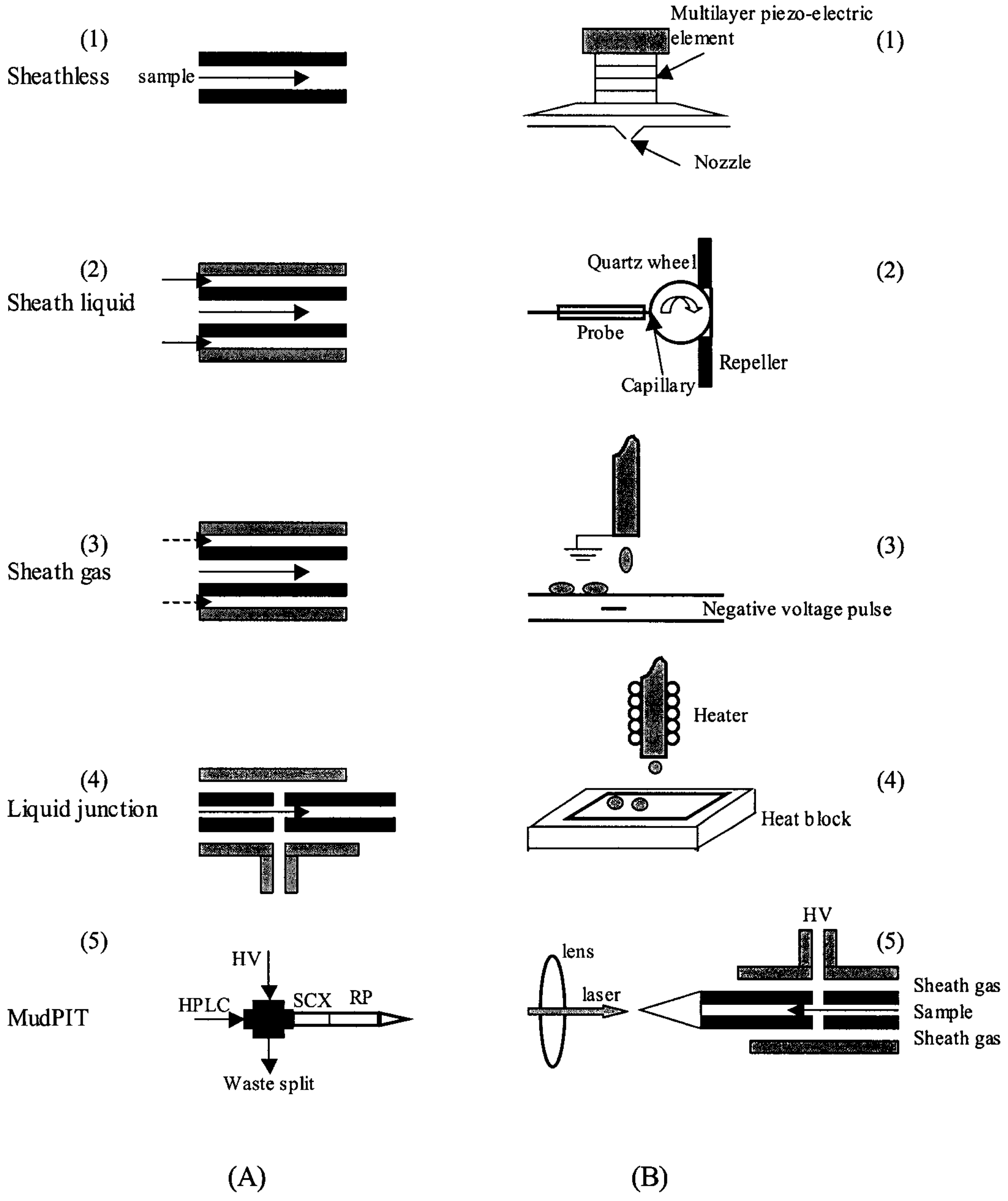


Figure I-2. Interfaces used to couple HPLC with mass spectrometer. (A) HPLC-ESI interfaces. (B) HPLC-MALDI interfaces.

evaporates, the droplet radius become smaller and coulombic repulsion forces Q become sufficient to overcome surface tension forces $KR^3$ (K is a constant for a specific solvent and R is the radius of droplets) then fission occurs. Typically, as R<10 nm, gas-phase ion formation is expected to occur. (4) Consequent solvent evaporation and fission lead to extremely small droplets that contain only one ion and final conversion to a gas ion.

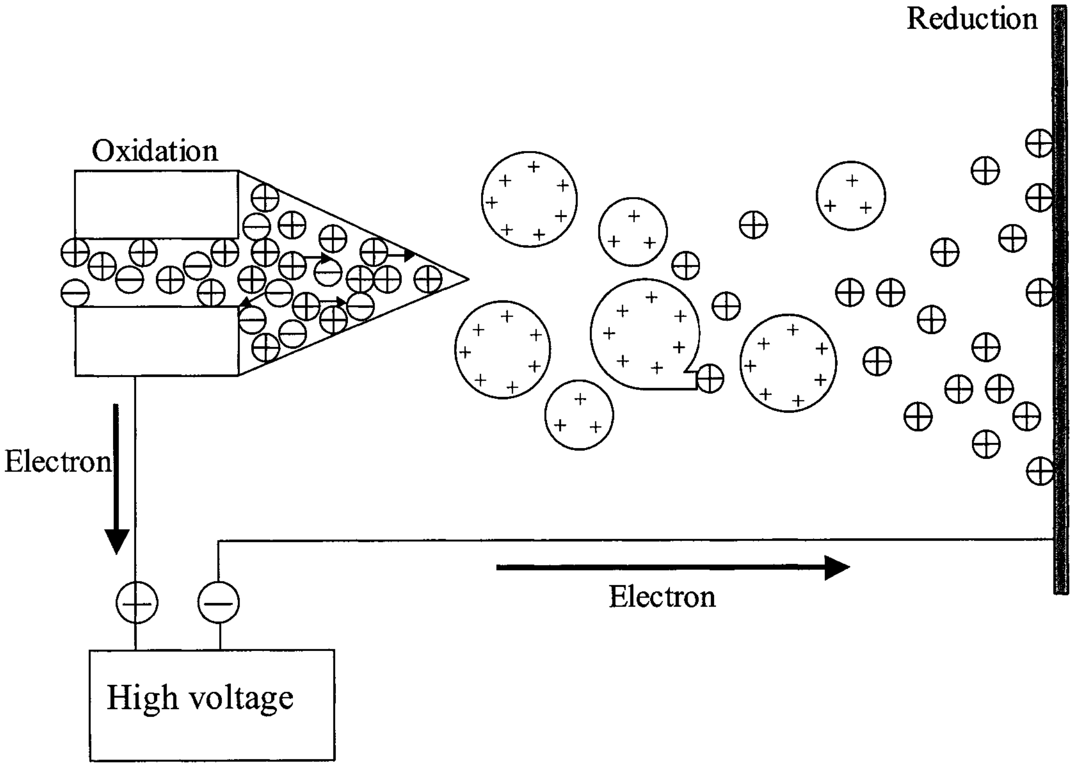


Figure I-3. Schematic representation of the processes in electrospray ionization (ESI)

In contrast to ESI, MALDI sublimates and ionizes samples out of a dry, crystalline matrix via laser pulses as shown in Figure I-4 [60-62]. In this technique, a compound that has absorption at laser wavelength is used as the matrix for entrapment, isolation, vaporization and ionization of analyzes. The processes in MALDI are: (1) Matrix that co-crystallized with sample absorbs laser energy through vibrational or electronic excitation. (2) Sample molecules close to the impact point are sputtered from the surface. (3) The primary energy is dissipated and converted to vibrational excitation. (4) The vibrational excitation results in desorption of small fragment ions and finally intact molecular ions near the fringes of the interaction zone.

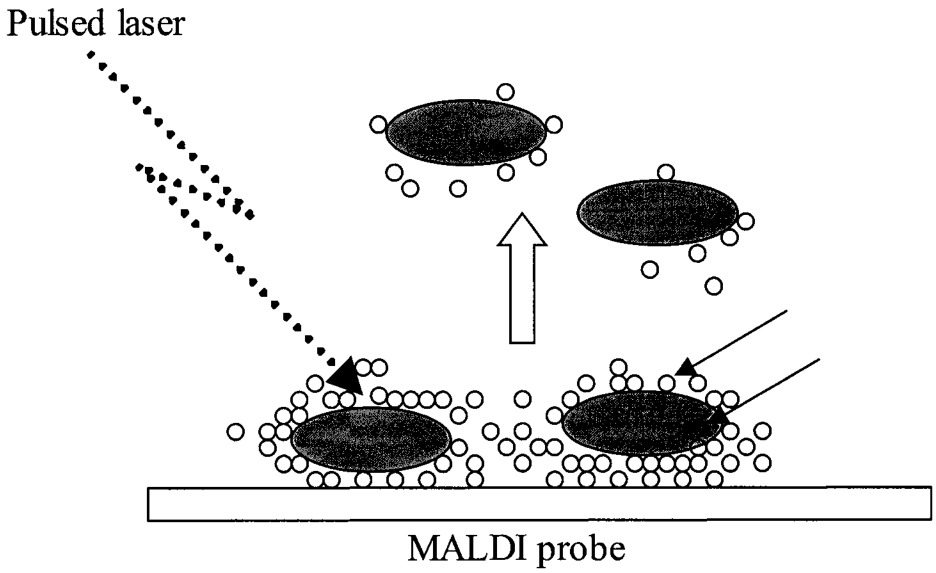


Figure I-4. Principle of matrix-assisted laser desorption/ionization (MALDI)

In order to avoid thermal decomposition of the thermally labile molecules, the energy is transferred during very short time period. Lasers with pulse width in the 1-100 ns range are used. MALDI-MS was originally used for relatively simple mixtures. With the development of integrated liquid chromatography MALDI-MS systems, MALDI-MS is beginning to handle very complex biological samples with better sensitivity and resolution than that of ESI-MS.

### I. 2. 3. Mass Analyzer

There are four basic types of mass analyzers currently used in proteomics research. These are quadrupole, ion trap, time-of-flight (TOF), and Fourier transform ion cyclotron (FTICR-MS). They are very different in design and performance with their own strengths and weaknesses. These analyzers are often put together in tandem to take advantage of the strengths of each. In tandem mass spectrometry, an ion with a specific m/z is isolated first and then analyzed by fragmenting the mass-selected ion and by determining the m/z of fragment ions in a second stage of mass analysis. Therefore, a specific ion in a complex mixture can be selectively studied and structural information can be obtained from the interpretation of the resultant spectra. Depending on how the experiment is performed, tandem mass spectrometers can be classified into two types: tandem-in-time and tandem-in-space.

An Ion trap mass spectrometer as shown in Figure I-5 is a notable example of tandem-in-time instruments [63, 64]. In this method, a radio frequency voltage is applied to the ring electrode while the two end-cap electrodes are held at ground potential. The ions are first captured or trapped for a certain time interval and then subjected to MS or MS/MS analysis. For MS analysis, the radio frequency voltage is increased with time so that ions of successively greater mass-to-charge ratios develop unstable trajectories and are ejected through perforations in an end-cap. For MS/MS analysis, unwanted ions are removed from the trap and specific ions of a very narrow mass range remain trapped. These ions then undergo CAD (Collision Activated Dissociation) reaction in which a gas induces fragmentation and the products of this CAD reaction are subjected to mass analysis to yield a daughter MS/MS spectrum. Ion trap mass spectrometers are robust, sensitive and relatively inexpensive thus lots of proteomics data have been reported in the literature using this type of tandem mass spectrometer. The major disadvantage of an ion trap mass analyzer is its relatively low mass accuracy and low resolution due to the limited number

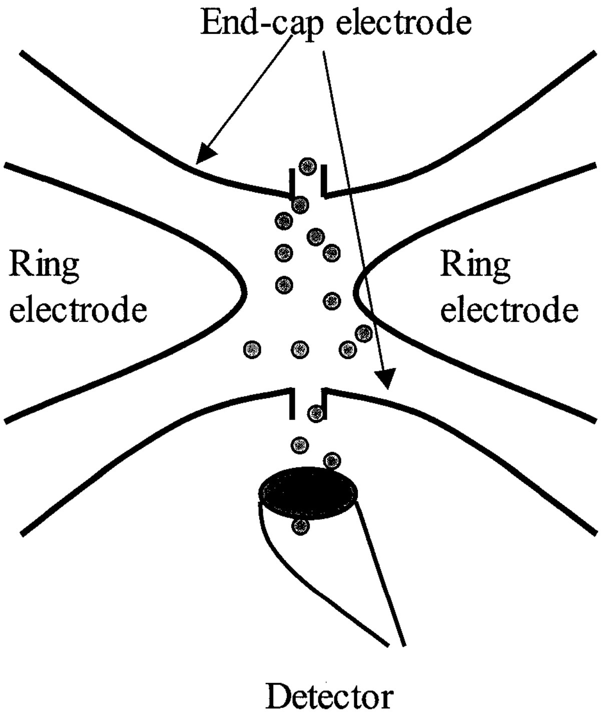


Figure I-5. Schematic diagram of the ion trap mass analyzer.

of ions that can be trapped before space-charging distort their distribution. The linear or two dimensional ion trap is a newly developed technique. It stores ions in a cylindrical volume that is considerably larger than that of the traditional three-dimensional ion traps. This greater volume results in much increased sensitivity, resolution and mass accuracy. FT-MS is also a trapping instrument. It captures ions under high vacuum in a high magnetic field. It has high resolution, sensitivity, mass accuracy and dynamic range. However, the expense, operational complexity and low peptide-fragmentation efficiency limit its routine use in proteomics research.

Tandem-in-space mass spectrometers have more than one mass analyzer. Each of them performs separately to accomplish different stages of the experiments. The Quadrupole-time-of-flight tandem mass spectrometer (QTOF) is one of the most popular instruments currently used in proteomics research [65]. It includes 3 quadrupoles and 1 TOF mass analyzer as shown in Figure I-6. The RF-only quadrupole ($Q_0$) and stubbies (ST) transfer ions from the vacuum interface into the mass filter quadrupole ($Q_1$). The mass filter quadrupole ($Q_1$) separates ions based on their mass-to-charge ratio. In the collision cell quadrupole ($Q_2$), precursor ions are fragmented by collisionally activated dissociation (CAD) reaction with neutral gas molecules. Precursor ions and fragment ions are analyzed by a time-of-flight analyzer and detected by a microchannel plate (MCP). The flexibility of this type of tandem mass spectrometer to couple with both ESI and MALDI source offers extensive advantages to handle complex biological samples. It is well suited for high throughput, high sensitivity, high resolution and high mass accuracy proteomics research.

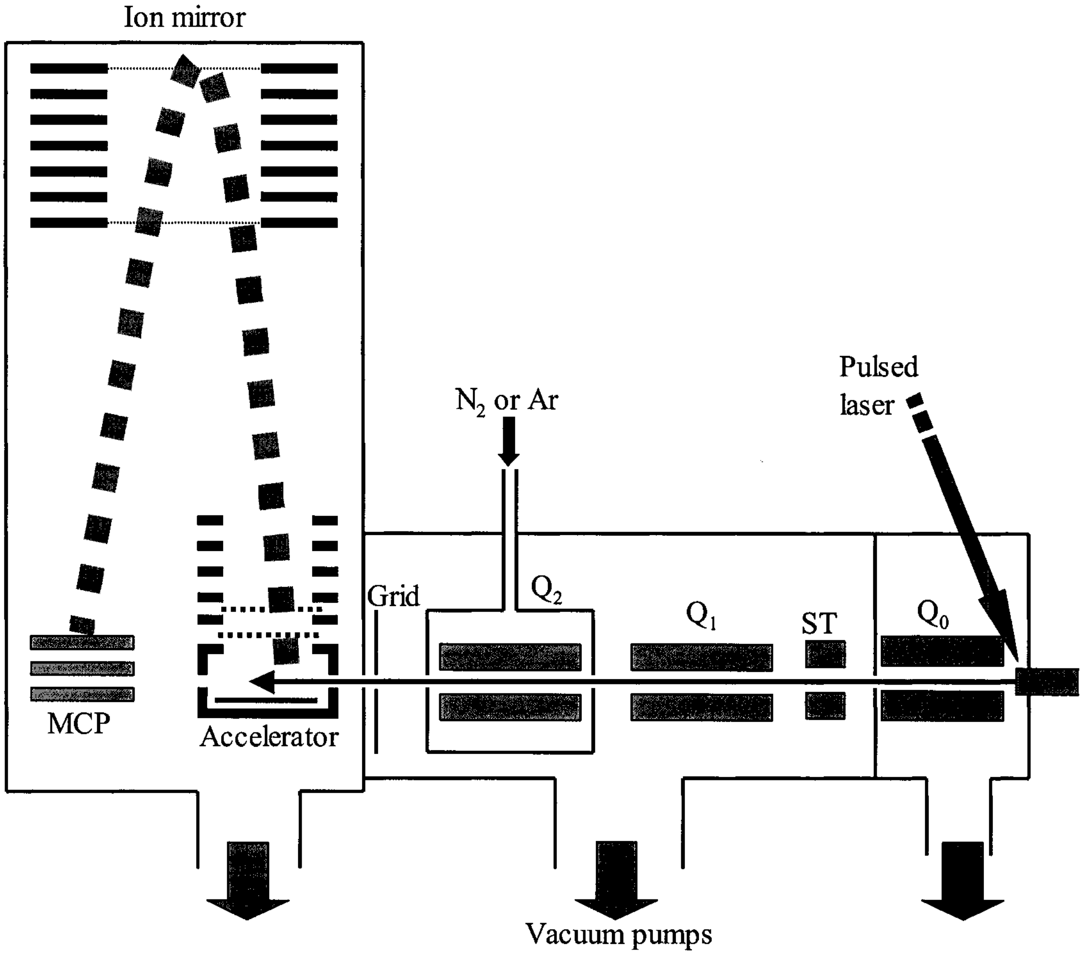


Figure I-6. Schematic diagram of the MALDI/ESI-Q-TOF

### I. 2. 4. Fragmentation of Peptides and Proteins

In mass spectrometric sequencing, the amino acid sequence of a peptide or protein is often contained in a product ion spectrum produced by a tandem mass spectrometer. In the past few years, several techniques based on gas phase reactions have been developed. Examples include conventional collisionally activated dissociation (CAD) [5, 66-68], infrared multiphoton dissociation (IRMPD) [69], and electron capture dissociation (ECD) [70-72]. Table I-1 summarizes the main types of fragment ions. Under positive ion operation of CAD, peptide ions are protonated at the basic sites within peptides such as N-terminal amine, side group of lysine, arginine and histidine residues. In the gas phase, a proton associated with strong basic sites are highly attached and remains associated at those sites even during collisional activation. However, the proton on the less basic N-terminal amine may migrate among the amide linkages. The proton migration produces peptide ions with different protonation sites that follow different fragmentation reactions. As a result, a series of product ions are formed that reveal the whole sequence of peptides. CAD can provide full or significant sequence coverage for proteolytic peptides and has been implemented on most types of instruments. However, only low efficiency has been observed for large multiply charged ions compared with small fragment ions.

Infrared Multiphoton Dissociation (IRMPD) provides far greater efficiency and selectivity than CAD for fragmentation of large multiple charged ions [69]. The more intense y, b, and internal fragments obtained make IRMPD especially useful for sequence verification. In this approach, irradiation levels affect the product ions and spectrum quality. Increasing irradiation time lowers the product yield because the dissociation of primary products produces secondary products that do not provide sequence information and are not measurable. A technique to remove product ions on formation from the laser beam should further improve the present efficiency.

Table I-1. Fragment ions generated in tandem mass spectrometer.

| Ion type | Ion masses[a] | Ion type | Ion masses | Ion type | Ion masses |
|---|---|---|---|---|---|
| a | [N]+[M]-CO | $b^o$ | b-$H_2O$ | x | [C]+[M]-CO |
| $a^*$ | a-$NH_3$ | $b^{++}$ | (b+H)/2 | y | [C]+[M]+$H_2$ |
| $a^o$ | a-$H_2O$ | c | [N]+[M]+$NH_3$ | $y^*$ | y-$NH_3$ |
| $a^{++}$ | (a+H)/2 | d | a-partial side chain | $y^o$ | y-$H_2O$ |
| b | [N]+[M] | v | y-complete side chain | $y^{++}$ | (y+H)/2 |
| $b^*$ | b-$NH_3$ | w | z-partial side chain | z | [C]+[M]-NH |

a. [N] is the mass of the N-terminal group, [C] is the mass of the C-terminal group, [M] is the mass of the sum of the neutral amino acid residue masses.

Besides the collision induced or irradiation induced fragmentation methods, a complementary new technique, electron capture dissociation (ECD), has been recently developed [70-72]. Electron capture occurs at a protonated site to release an energetic $H^•$ atom that can be captured at a high affinity site such as a backbone amide or –S-S- to cause nonergonic dissociation. Posttranslational modifications such as glycosylation, carboxylation and sulfation are less easily lost in ECD than with other fragmentation methods thus it can nearly complete the MS sequence coverage. Fragment ions such as y, a, c and z are usually observed in ECD spectrum.

## I. 3. Mass spectrometry in Proteomics Research and Challenges

Functional proteomics in general investigates the large-scale identification of gene and cellular function directly at the protein level. Mass spectrometry based proteomics has become an indispensable method for the analysis of complex protein samples [73-75]. It has essentially replaced the traditional technique of Edman Degradation because of its much improved sensitivity and high throughput. Also the ability couple MS to separation tools allows it to deal with protein mixtures. Currently, mass spectrometry based protein analysis can be mainly classified into four areas: protein identification by primary sequence; detection of posttranslational modifications; differential-display of proteomics using limited or no separation followed by mass spectrometric quantification; and analysis of protein-ligand interactions. In the past few years, mass spectrometry has been successfully applied to small set of proteins or even larger number of proteins expressed in a cell due to the newly developed experimental approaches.

### I. 3. 1. Identification of Proteins

There are mainly two approaches for protein identification by mass spectrometry. One is called Peptide Mass Mapping [76-78]. Proteins are first digested into peptides by a specific enzyme such as trypsin and the peptide mixtures are then analyzed by MALDI-TOF. The resultant mass spectra represent the fingerprint of the protein being studied. This approach has been automated to identify hundreds of protein spots with database searching. With the increased database size, it is rarely sufficient to identify proteins unambiguously by MALDI mass mapping [79-81]. This becomes even more serious when only a limited number of peptides are available due to the low abundance of proteins or the low ionization efficiency of some peptides. Increased mass accuracy can improve the identification efficiency. However, even with an error of only 50 ppm [82], a single peptide will still match to several hundreds of proteins even for a moderate size of proteome such as the nematode worm.

The second approach for protein identification relies on the fragmentation of peptides in a mixtures to obtain sequence information using a tandem mass spectrometer such as ESI-Ion trap MS/MS [83-85] or MALDI/ESI Quadrupole-Time-of-Flight MS/MS [34, 65, 86]. Using this technique, peptides are fragmented in a predictable manner and sequences from the databases can be used to predict an expected fragmentation pattern and match the expected pattern to that observed in the spectrum. It is obvious that more specific sequence information can be obtained by this approach than that of peptide mass mapping. Consequently, matching one or more tandem mass spectra to sequences in the same protein can provide a high level of confidence in the identification.

For both approaches, protein identification is achieved through database searching. The efficiency of protein identification ultimately requires that the acquired mass spectra be accurately matched to protein sequences from the corresponding database entries inferred from known (or predicted) DNA sequences using bioinformatics tools. This process is facilitated by the growing number of completely sequenced genomes. However, more and more researchers are realizing that mere completed sequences of genomes are not enough to investigate all biological functions. Much of genomic annotation relies on the similarity of assembled sequence to other gene and protein database entries to detect open reading frames (ORFs) and thereby assign functions. Despite the success of bioinformatics, it is still difficult to predict genes accurately from genomic data. Small genes or genes with little or no homology to other known genes may be entirely missed. Though the genomic sequences of DNA specifies the variation in amino acid composition, sequence, and size of proteins, lots of novel proteins resulting from co/posttranslational modifications or gene products undetected by the genomics approach are not encoded in DNA sequence. Homology searching can only identify the similar parts. *De novo* interpretation of MS/MS data is needed for unambiguously protein identification. Even already predicted by DNA sequence, it is still challenge to unambiguously identify proteins with low sequence coverage or incomplete tandem MS/MS spectra. A single peptide or an incomplete MS/MS spectrum may match several peptides in the database. Additionally, background peaks resulting from protease autolysis can also reduce the database searching efficiency. Thus additional information is needed to eliminate false positive identifications.

### I. 3. 2. Detection of Posttranslational Modifications

One of the most interesting features of proteomics is the ability to analyse posttranslational modifications. Identification of posttranslational modifications is important for protein function. because they can determine the activity, stability, localization and turnover of a protein [87-89]. Table I-2 summarizes some common and important posttranslational modifications. More comprehensive list of modifications can be found from the web site of the association of biomolecular resource facilities (http://www.abrf.org/index.cfm/dm.home). Though mass spectrometry is often sufficient to identify proteins with as few as one or two peptides, it is far more difficult to determine modifications than to determine protein identity. As for the commonly used "bottle-up" approach [90-92], proteins are cleaved into peptides with optimal sizes (500~3000 Da) amendable to tandem MS/MS analysis by specific enzymes and all the peptides that do not have expected molecular masses are further analyzed. This is a time-consuming process to analyze these overlapped and short peptides. Additionally, peptides with molecular weight beyond the optimal range of tandem mass spectrometer will often produce low quality MS/MS spectra that cannot be unambiguously identified by database searching. Affinity techniques have been applied to select peptides bearing some modifications such as phosphorylation [93-98] and glycosylation [99-102]. Up to now, there is still no rapid and universal technique to detect all other huge amount of modifications. Even if the modified peptides can be selected, subsequent identification can still be problematic. Under this situation, the observed MS/MS data cannot match the mass calculated from the database. Therefore, making an unambiguously identification is very difficult. The database search program can be informed of some possible modifications (such as oxidized Methionine and reduction of disulfide bonds). However with so many modification possibilities, the number of database sequences with correct mass also increases and then the possibility of obtaining a false positive match also increases [103, 104].

The classic technique of Edman degradation is still the method of choice to obtain the characterization of both sequence and modifications. It is mature, reliable and automated but relatively poor sensitivity and very slow [105]. Combination of multiple steps of wet degradation with a final, single-step mass spectrometric readout of the amino acid

Table I. 2. Some common and important posttranslational modifications.

| PTM type | Δmass (Da) | Stability | functions |
|---|---|---|---|
| Phosphorylation | | | Reversible, activation /inactivation of enzyme activity, modulation of molecular interactions, signaling |
| pTyr | +80 | Stable | |
| pSer, pThr | +80 | Labile/moderate | |
| Acetylation | +42 | Stable | Protein stability, protection of N-terminus, regulation of protein-DNA interactions (histones) |
| Methylation | +14 | Stable | Regulation of gene expression |
| Acylation | | | Cellular localization and targeting signals, membrane tethering, mediator of protein-protein interactions |
| Farnesyl | +204 | Stable | |
| Myristoyl | +210 | Stable | |
| Palmitoyl | +238 | Labile/moderate | |
| Glycosylation | | | Excreted proteins, cell-cell recognization / signaling , reversible, regulatory functions |
| N-linked | >800 | Labile/moderate | |
| O-linked | 203, >800 | Labile/moderate | |
| GPI anchor | >1, 000 | moderate | Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane |
| Hydroxyproline | +16 | stable | Protein stability and protein-ligand interactions |
| Sulfation (sTyr) | +80 | labile | Modulator of protein-protein and receptor-ligand interactions |
| Disulfide bond | -2 | moderate | Intra- and intermolecular crosslink, protein stability |
| Deamidation | +1 | stable | Possible regulator of protein-ligand and protein-protein interactions, also a common chemical artifact |
| Pyroglutamic acid | -17 | stable | Protein stability, block N-terminus |
| Ubiquitination | >1, 000 | Labile/moderate | Destruction signal. After tryptic digestion, ubiquitination site is modified with Gly-Gly dipeptide |
| Nitration of tyrosine | +45 | Labile/moderate | Oxidative damage during inflammation |

Sequence leads to greatly improved sensitivity and high sample throughput [106]. However, degradation blockage from N-terminal modification and modifications on some internal amino acids requires additional treatment of proteins in order to be sequenced by Edman degradation.

In-source fragmentation has ever been developed for analysis of protein sequences and posttranslational modifications. Though it is fast and easy to handle, the poor sensitivity and co-existed multiple types of fragment ions limit its further application to real biological samples [107-109].

The recently developed "top-down" approach is based on intact protein fragmentation using high resolution and high mass accuracy FT-ICR mass spectrometer. It circumvents this problem to some extent [110-115]. Characterization of any modification of a large protein up to 29 kDa can be done within one residue. Co-existed c, z, a, y or other secondary product ions and their different charge-state ions make the spectrum very complex. The assignment of sequence becomes difficult with the increased protein mass especially with the limited detectable mass range and limited sensitivity,.

### I. 3. 3. Differential-Display of Proteomics and Mass Spectrometric Quantification

The comparative two-dimensional gel electrophoresis [116-119] is a classic technique to identify proteins that are up- or downregulated in a disease-specific manner for use as diagnostic markers or therapeutic targets. Challenges related to this technique include the display of hydrophobic and large proteins that enter the second dimension of the gel poorly; Also the limited dynamic range makes it difficult to visualize low abundant proteins. Especially, in body fluids such as serum, serum albumin and globulin are more than 99% of total proteins so it is difficult to define normal protein expression pattern. Additionally, biological variation from one person to another person or one age to another age makes it even more difficult to find the diagnostic markers.

Protein chips based on the immobilization of antibodies in an array format onto specially treated surfaces is another choice for large scale profile of two different cell states [120-122]. Cell lysates are labeled by different fluorophores and mixed so that the color can be used as an indicator of the protein abundance bound to the specific antibody. This system depends on the specific antibodies and there are still a lot of technical problems that need to be overcome.

Differential-display proteomics can also be performed using a shotgun approach combined with isotope dilution mass spectrometric quantification. Quantification is achieved by isotopic labeling of one of the two states. Proteins are enzymatically digested to form small peptides. The labeled and unlabelled peptides that have similar chromatographic retension and ionization efficiency are then quantified by comparison of the peak areas. Two very different stable isotope coding strategies are being used for differential-display proteomics. One is the external isotope labeling technique. $^{18}O$-labeled water is widely used in protease digestion to introduce an $^{18}O$ tag through the hydrolysis reaction to label all the proteolytic peptides at the C- terminus uniformly except the peptide originating from the carboxy-terminus of the protein [123, 124]. Derivatization of primary amino acids either at the C-terminus or N-terminus through acylation or esterification is also a global internal standard technique [125]. The major disadvantage of the global isotope labeling technique is the complexity of the samples. Many types of cells contain more than 10,000 proteins and the corresponding proteolytic peptides can number greater than $10^5$. Those are far beyond the separation capacity of commonly used chromatographic methods. Moreover, protein abundance in bacteria cells or human cells is about $10^6$ or $10^9$ respectively. Thus signal suppression from high abundant proteins prevents the low abundant proteins from detection [126]. Obviously, with complex proteomes, targeting structural features is an attractive strategy for simplifying mixtures. A number of amino acids such as cysteine, lysine, N-terminal serine/threonine, tryptophan and methionine have been targeted by various *in vitro* methods [127-131].

Incorporation of stable isotopes into metabolic products is another well-established techniques. Replacement of $^{13}C$ for $^{12}C$, $^{15}N$ for $^{14}N$, or $^{2}H$ for $^{1}H$, in proteins can generate characteristic mass shifts in their isotopic distribution patterns without affecting their chemical or structural properties. However, sample complexity remains a problem in this technique. Therefore, quantitative incorporation of specific essential and nonessential amino acids into cell lines combined with affinity technique is promising to obtain better results [132, 133]. The limitation of this approach is that the amino acid can only be at the end of a metabolic pathway. Otherwise the conversion to another amino acids results in unpredictable isotope patterns.

### I. 3. 4. Mass Spectrometric Analysis of Protein-Ligand and Protein-Protein Interactions.

The mapping of protein-ligand interaction is key to understanding protein function and discovering cellular pathways. Compared with two-hybrid and microchip-based methods, MS-based approaches have the advantages that the fully processed and modified proteins can serve as the bait (that is a substrate to bind with specific antibody), that the interactions take place in the native environment and cellular location, and that multicomponent complexes can be isolated and analyzed in a single operation [134]. Usually MS-based approaches have essentially three steps: bait presentation; affinity purification of the complex; and analysis of the bound proteins. If an antibody or other reagent exists to allow specific isolation of the protein with its bound partners, endogenous proteins can be used as the bait. However, there is no enough antibodies and many of them do not immunoprecipitate well or lack of specificity. So a more generic method is to tag the interested protein that can be recognized by a specific antibody [73]. ESI-MS/MS is currently used to characterize the stoichiometry and the structural assignment of protein complexes [135-137]. Combined with isotope labeling such as H/D exchange, MS-based approach can also quantitatively measure protein stability changes upon ligand binding [138].

## I. 4. Objective of the Thesis

Rapid and efficient identification of protein primary sequence and posttranslational modification is fundamental and critical in proteomics research. Although proteomics already has had tremendous success as previously discussed, this field still faces significant challenges. This thesis will focus on the development of mass spectrometric methods for analysis of protein sequence and posttranslational modifications.

### I. 4. 1. *De novo* Peptide Sequencing for Protein Identification and Detection of Posttranslational Modification

As for protein identification, a *de novo* peptide sequencing approach based on ESI-Ion trap MS/MS was developed. This was used to identify bioactive proteins isolated from shark cartilage that are not in the present database. To improve the accuracy of the protein identification, a 2D mass spectrometric method using a whole cell *in vivo* stable isotope labeling technique was developed. Peptide *de novo* sequencing and identify proteins with low sequence coverage or incomplete tandem MS/MS spectrum resulting from the fragmentation of weak peptide ion signal. 2D MS/MS spectrum provides the correlation between fragment ion masses and their corresponding nitrogen number. The specificity of peptide fragmentation signals in a tandem MS/MS spectrum is thus greatly enhanced, resolving a high degree of mass degeneracy of proteolytic peptides from a complex proteome. Microwave-assisted chemical derivatization was also investigated to enhance peptide *de novo* sequencing by tagging the peptide at either N-or C-terminus. Comparison of the fragmentation between the tagged peptide and untagged peptide generates the unique identification of peptide sequence.

### I. 4. 2. Intact Protein *de novo* Sequencing for Detection of Posttranslational Modifications.

Though the identification of proteins is already well established, the currently used methods are not efficient to detect all modification sites due to a variety of reasons. A new technique for deciphering different forms of proteins has been developed and is poised to greatly facilitate the studying of protein functions, which can lead to rapid drug development and identification of protein biomarkers for disease diagnosis. While a DNA or gene sequence provides a blueprint for making a protein with a specific amino acid sequence, it does not predict what the final form of the protein will be. Proteins are often modified during and after production in a cell. Proteins with different modifications can have totally different functions. The new technique allows for rapid determination of the protein sequence and its modifications. In this technique, generation of sequence specific ladders from the N- and C-terminus by microwave assisted acid hydrolysis of intact proteins followed by mass analysis of resultant polypeptides has been investigated as a method for rapid analysis of sequence and posttranslational modifications. The high cleavage specificity to amide bonds results from the specificity of hydrolysis. Therefore the resultant MALDI-TOF spectrum of the polypeptide ladders is essentially two series of sequence specific ladders from the N- and the C-terminus of the studied protein. Mass analysis of the polypeptide ladders by mass spectrometry allows for direct reading of the amino acid sequence and modifications of the protein. Further secondary ion fragmentation is controlled by microwave irradiation. As examples, this technique was applied to determination of posttranslational modifications (PTMs) such as acetylation, covalently bounded heme group, oxidization, disulfide bonds and protein phosphorylation sites. Proteins isolated from human breast cancer cell line were purified by two-dimensional HPLC separation and analyzed by this method.

### I. 4. 3. Analysis of Sequence and Posttranslational Modifications of Membrane Proteins

Membrane proteins are challenging because of their poor solubilization, low abundance and extensive posttranslational modifications. General mass spectrometric methods that analyze peptides generated by proteolysis developed for analysis of soluble proteins are not amenable to membrane proteins. Transmembrane domains often escape digestion because they are either not accessible to a specific enzyme or lack the proteolytic cleavage sites. Hydrophobic residues existing in the transmembrane domains prevent the attack of hydrophilic water to the polypeptide bonds. In order to map the complete sequence and modifications, it is critical to develop techniques that are not only compatible with mass spectrometric detection but also can cleave inside hydrophobic domains. This thesis focuses on the microwave-assisted chemical cleavage. It is found that the presented technique can provide characteristic cleavage that is useful for efficient for identifying protein. One cleavage pattern is the N-or C-terminal ladder that is efficient to identify posttranslational modifications such as signal peptide cleavage from N-terminus. Another one is cleavage from both sides of Glycine that results in the adjacent peaks in the corresponding mass spectra. Especially important, the two kinds of cleavages are also observed in hydrophobic transmembrane domains that provide the basis to detect sequence errors or modifications inside these kinds of domains.

### I. 5. Cited Literature.

(1) Yates, J. R., III. *Trends Genet.* **2000,** *16*, 5-8.

(2) Pandey, A.; Mann, M. *Nature* **2000,** *405*, 837-846.

(3) Aebersold, R. H.; Goodlett, D. R. *Chem. Rev.* **2001,** *101*, 269-295.

(4) Hunt, D. F. et al. *Science* **1992,** *255*, 1261-1263.

(5) Kinter, M.; Sherman, N. E. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, Desiderio, D. M.; Nibbering, N. M. M., Ed.; John Wiley and Son's Ltd.: New York, 2000.

(6) Wilkins, M. R.; William, K. L.; Apple, R. D.; Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics,* Springer: Berlin, 1997.

(7) Burley, S. K. et al. *Nature Genet.* **1999,** *23,* 151-157.

(8) Hanash, S. *Nature* **2003,** *422,* 226-232.

(9) Krogh, A. In *Guide to Human GenomeComputing*, Bishop, M. J., Ed.; Academic: San Diego, 1998.

(10) Dunham, I. et al. *Nature* **1999,** *402,* 489-495.

(11) Claverie, J. M. *Hum. Mol. Genet.* **1997,** *6,* 1735-1744.

(12) Paney, A.; Lewitter, F. *Trends Biochem. Sci.* **1999,** *24,* 276-280.

(13) Brenner, S. E. *Trends Genet.* **1999,** *15,* 132-133.

(14) Jungblut, P. R.; Muller, E. C.; Mattow, J.; Kaufmann, S. H. E. *Infection and Immunity* **2001,** *69,* 5905-5907.

(15) Wasinger, V. C.; Smith, I. H. *FEMS Microbio. Letters* **1998,** *169,* 375-382.

(16) Garcia-Blanco, M. A.; Baraniak, A. P.; Lasda, E. L. *Nat. Biotech.* **2004**, *22,* 535-546.

(17) Faustino, N. A.; Cooper, T. A. *Genes Dev.* **2003,** *17,* 419-437.

(18) Goldstrohm, A. C.; Greenleaf, A. L.; Garcia-Blanco, M. A. *Gene* **2001,** *277,* 31-47.

(19) Anderson, L. B. et al. *J. Am. Chem. Soc.* **2004,** *126,* 8399-8405.

(20) Yates, J. R., III; Eng, J. K.; McConnack, A. L.; Schieltz, D. *Anal. Chem.* **1995,** *67,*

1426-1436.

(21) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R., III. *Anal. Chem.* **2000,** *72,* 757-763.

(22) Nelson, K. K.; Green, M. R. *Proc. Natl. Acad. Sci. U. S. A.* **1990,** *87,* 6253-6257.

(23) Antonarakis, S. E. et al. *Proc. Natl. Acad. Sci. U. S. A.* **1984,** *81,* 1154-1158.

(24) King, M. C.; Marks, J. H.; Mandell, J. B. *Science* **2003,** *302,* 643-646.

(25) Scully, R. et al. *Science* **1996,** *272,* 123-126.

(26) Arrell, D. K.; Neverova, I.; Fraser, H.; Marban, E.; Van Eyk, J. E. *Circ. Res.* **2001,** *89,* 480-487.

(27) Heinke, M. Y. et al. *Electrophoresis* **1998,** *19,* 2021-2030.

(28) Westbrook, J. A.; Yan, J. X.; Wait, R.; Welson, S. Y.; Dunn, M. J. *Electrophoresis* **2001,** *22,* 2865-2871.

(29) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001,** *73,* 5683-5690.

(30) Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., III. *Anal. Chem.* **2003,** *75,* 2470-2477.

(31) le Coutre, J.; Whitelegge, J. P.; Gross, A.; Turk, E.; Wright, E. M.; Kaback, H. R.; Faull, K. F. *Biochem.* **2000,** *39,* 4237-4242.

(32) Bakhtiar, R.; Thomas, J. J.; Siuzdak, G. *Acc. Chem. Res.* **2000,** *33,* 179-187.

(33) Natsume, T.; Yamauchi, Y.; Nakayama, H.; Shinkawa, T.; Yanagida, M.; Takahashi, N.; Isobe, T. *Anal. Chem.* **2002,** *74,* 4725-4733.

(34) Shevchenko, A.; Loboda, A.; Shevchenko, A.; Ens, W.; Standing, K. G. *Anal. Chem.* **2000,** *72,* 2132-2141.

(35) Wu, S.-L.; Choudhary, G.; Ramstrom, M.; Bergquist, J.; Hancock, W. S. *J.*

*Proteome Res.* **2003,** *2,* 383-393.

(36) Figeys, D.; Gygim, S. P.; McKinnon, G.; Aebersold, R. *Anal. Chem.* **1998,** *70,* 3728-3734.

(37) Li, J. J.; Kelly, J. F.; Chemushevich, I.; Harrision, D. J.; Thibault, P. *Anal. Chem.* **2000,** *72,* 599-609.

(38) Licklider, L.; Wang, X. Q.; Desai, A.; Tai, Y. C.; Lee, T. D. *Anal. Chem.* **2000,** *72,* 367-375.

(39) Bings, N. H.; Wang, C.; Skinner, C. D.; Colyer, C. L.; Thibault, P.; Harrison, D. J. *Anal. Chem.* **1999,** *71,* 3292-3296.

(40) Lazar, I. M.; Ramsey, R. S.; Sundberg, S.; Ramsey, J. M. *Anal. Chem.* **1999,** *71,* 3627-3631.

(41) Figeys, D.; Ning, Y.; Aebersold, R. *Anal. Chem.* **1997,** *69,* 3153-3160.

(42) Zhang, B.; Liu, H.; Karger, B. L.; Foret, F. *Anal. Chem.* **1999,** *71,* 3258-3264.

(43) Chan, J. H.; Timperman, A. T.; Aebersold, R. *Anal. Chem.* **1999,** *71,* 4437-4444.

(44) Li, J. J.; Thibault, P.; Bings, N. H.; Skinner, C. D.; Wang, C.; Colyer, C.; Harrision, D. J. *Anal. Chem.* **1999,** *71,* 3036-3045.

(45) Wahl, J. H.; Goodlet, D. R.; Udseth, H. R.; Smith, R. D. *Electrophoresis* **1993,** *14,* 448-457.

(46) Adren, P. E.; Emmett, M. R.; Caprioli, R. M. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 867-869.

(47) Juraschek, R.; Dulcks, T.; Karas, M. *J. Am. Soc. Mass. Spectrom.* **1999,** *10,* 300-308.

(48) Wilm, M.; Mann, M. *Anal. Chem.* **1996,** *68,* 1-8.

(49) Fligge, T. A.; Bruns, K.; Przybylski, M. *J. Chromatogr. B* **1998,** *706,* 91-100.

(50) Schmit, A.; Karas, M.; Dulcks, T. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 492-500.

(51) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotech.* **2001,** *19,* 242-247.

(52) Miliotis, T. et al. *J. Mass Spectro.* **2000,** *35,* 369-377.

(53) Preisler, J.; Foret, F.; Karger, B. L. *Anal. Chem.* **1998,** *70,* 5278-5287.

(54) Ericson, C. et al. *Anal. Chem.* **2003,** *75,* 2309-2315.

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

(56) Hiraoka, K. *J. Mass Spectrom.* **2004,** *39,* 341-350.

(57) Kebarle, P.; Tang, L. *Anal. Chem.* **1993,** *65,* 972A-986A.

(58) Kebarle, P.; Ho, Y. In *Electrspray Ionization Mass Spectrometry*, Cole, R. B. Ed.; John Wiley and Son's Ltd.: New York, 1997.

(59) Yamashita, M.; Fenn, J. B. *J. Phys. Chem.* **1984,** *88,* 4451-4459.

(60) Sunner, J. A.; Kulatunga, R.; Kebarel, P. *Anal . Chem.* **1986,** *58,* 1312-1316.

(61) Schrouk, L. R.; Cotter, R. J. *Biomed. Environ. Mass Spectrom.* **1986,** *13,* 395-399.

(62) Tanaka, K. et al. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 151-153.

(63) McLuckey, S. A.; Van Berkel, G. J.; Goeringer, D. E.; Glish, G. L. *Anal. Chem.* **1994,** *66*, 689A-695A.

(64) Louris, J. N. et al. *Anal. Chem.* **1987,** *59,* 1677-1685.

(65) Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **2000, 14,** 1047-1057.

(66) Loo, J. A.; Udseth, H. R.; Smith, R. D. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 207-210.

(67) Gauthier, J. W.; Trautman, T. R.; Jacobsen, D. B. *Anal. Chim. Acta* **1991,** *246,* 211-225.

(68) Senko, M. W.; Speir, J. P.; McLafferty, F. W. *Anal. Chem.* **1994,** *66,* 2801-2808.

(69) Little, D. P.; Speir, J. P.; Senko, M. W.; O Connor, P. B.; McLafferty, F. W. *Anal. Chem.* **1994,** *66,* 2809-2815.

(70) Zubarev, R. A. et al. *Anal. Chem.* **2000,** *72,* 563-573.

(71) Ge, Y. et al. *J. Am. Chem. Soc.* **2002,** *124,* 672-678.

(72) Sze, S. K.; Ge, Y.; Oh, H.; McLafferty, F. W. *Anal. Chem.* **2003,** *75,* 1599-1603.

(73) Aebersold, R.; Mann, M. *Nature,* **2003,** *422,* 198-207.

(74) McLafferty, F. W.; Fridriksson, E. K.; Horn, D. M.; Lewis, M. A.; Zubarev, R. A. *Science,* **1999,** *284,* 1289-1290.

(75) Dell, A.; Morris, H. R. *Science* **2001,** *291,* 260-262.

(76) Doucette, A.; Craft, D.; Li, L. *Anal. Chem.* **2000,** *72,* 3355-3362.

(77) Russell, W. K.; Park, Z.-Y.; Russell, D. H. **2001,** *73,* 2682-2685.

(78) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000,** *72,* 2482-2489.

(79) Cao, P.; Moini, M. *Rapid Commun. Mass Spectrom.* **1998,** *12,* 864-870.

(80) Green, M. K.; Johnston, M.V.; Larsen, B. S. *Anal. Biochem.* **1999,** *275,* 39-46.

(81) Lehmann, W. D.; Bohne, A.; Von Der Lieth, C. W. *J. Mass. Spectrom.* **2000,** *35,* 1335-1341.

(82) Sidhu, K. S. et al. *Proteomics* **2001,** *1,* 1368-1377.

(83) Le Bihan, T.; Pinto, D.; Figeys, D. *Anal. Chem.* **2001,** *73,* 1307-1315.

(84) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002,** *1,* 21-26.

(85) Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2002,** *1,* 211-215.

(86) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001,** *73,* 1917-1926.

(87) Cohen, P. *Trends Biochem. Sci.* **2000,** *25,* 596-601.

(88) Tyers, M.; Jorgensen, P. *Curr. Opin. Genet. Dev.* **2000,** *10,* 54-64.

(89) Mann, M.; Jensen, O. N. *Nat. Biotech.* **2003,** *21,* 255-261.

(90) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R., III. *Anal. Chem.* **2000,** *72,* 757-763.

(91) Shang, C.; Shibahara, T.; Hanada, K.; Iwafune, Y.; Hirano, H. *Biochem* **2004,** *43,*

6281-6292.

(92) Bateman, R. H. et al. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 792-803.

(93) Andersson, L.; Porath, J. *Anal. Chem.* **1986,** *154,* 250-254.

(94) Posewitz, M. C.; Tempst, P. *Anal. Chem.* **1999,** *71,* 2883-2892.

(95) Ficarro, S. B. et al. *Nat. Biotechnol.* **2002,** *20,* 301-305.

(96) Salomon, A. R. et al. *Proc. Natl. Acad. Sci. USA.* **2003,** *100,* 443-448.

(97) Zhou, H.; Watts, J. D.; Aebersold, R. A. *Nature. Biotechnol.* **2001,** *19,* 375-378.

(98) Oda, Y.; Nagasu, T.; Chait, B. T. *Nature. Biotechnol.* **2001,** *19,* 379-382.

(99) Li, Y.; Ogata, Y.; Freeze, H. H.; Scott, C. R.; Turecek, F.; Gelb, M. H. *Anal. Chem.* **2003,** *75,* 42-48.

(100) Bundy, J. L.; Fenselau, C. *Anal. Chem.* **2001,** *73,* 751-757.

(101) Tseng, K.; Wang, H.; Lebrilla, C. B.; Bonnell, B.; Hedrick, J. *Anal. Chem.* **2001,** *73,* 3556-3561.

(102) Ghosh, D.; Krokhin, O.; Antonovici, M.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *J. Proteome Res.* **2004,** *ASAP,* web released.

(103) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997,** *11,* 1067-1075.

(104) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594-2604.

(105) Edman, P.; Henschen, A. In *Protein Sequence Determination.* Needleman, S. B. Ed.; Springer-Verlag: Berlin, 1975; pp232-279.

(106) Chait, B. T.; Wang, R.; Beavis, R. C.; Kent, S. B. *Science* **1993**, *262*, 89-92.

(107) Reiber, D. C., Brown, R. S., Weinberger, S., Kenny, J., Bailey, J. *Anal. Chem.* **1998,** *70,* 214-1222.

(108) Lennon, J. J., Walsh, K. A. *Protein Sci.* **1997,** *6,* 2446-2453.

(109) Reiber, D. C.; Grover, T. A.; Brown, R. S. *Anal. Chem.* **1998,** *70,* 673-683.

(110) Fridriksson, E. K. et al. *Biochem.* **2000,** *39,* 3369-3376.

(111) Reid, G. E.; McLafferty, S. A. *J, Mass Spectrom.* **2002,** *37,* 663-675.

(112) Kruppa, G. H.; Schoeniger, J.; Young, M. M. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 155-162.

(113) Nemeth-Cawley, J. F.; Tangarone, B. S.; Rouse, J. C. *J. Proteome Res.* **2003,** *2,* 495-505.

(114) Taylor, G. K. et al. *Anal. Chem.* **2003,** *75,* 4081-4086.

(115) Ge, Y. et al. *Protein Sci.* **2003,** *12,* 2320-2326.

(116) Ostergaard, M.; Wolf, H.; Orntoft, T. F. *Electrophoresis* **1999,** *20,* 349-354.

(117) Page, M. J. et al. *Proc. Natl. Acad. Sci. USA* **1999,** *96,* 12589-12594.

(118) Gauss, C. et al. *Electrophoresis* **1999**, *20,* 575-600.

(119) Aicher, L. et al. *Electrophoresis* **1998,** *19,* 1998-2003.

(120) Lueking, A.; Horn, M.; Eickhoff, H.; Lehrach, H.; Walter, G. *Anal. Biochem.* **1999,** *270,* 103-111.

(121) Davies, H.; Lomas, L.; Austen, B. *Biotechniques* **1999,** *27,* 1258-1261.

(122) Nelson, R. W. *Mass Spectrom. Rev.* **1997,** *16,* 353-376.

(123) Dancik, D.; Addona, T.; Clauser, K.; Vath, J.; Pevzner, P. *J. Comput. Biol.* **1999,** *6,* 327-342.

(124) Kosaka, T.; Takazawa, T.; Nakamura, T. *Anal. Chem.* **2000,** *72,* 1179-1185.

(125) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001,** *73,* 2836-2842.

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

(127) Kuyama, H. et al. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 1642-1650.

(128) Gygi, S. P. et al. *Nat. Bitechnol.* **1999,** *17,* 994-999.

(129) Shen, M. et al. *Mol. Cell. Proteomics* **2003,** *2,* 315-324.

(130) Hale, J. E.; Butler, J. P.; Knierman, M. D.; Becker, G. W. *Anal. Biochem.* **2000,** *287,* 110-117.

(131) Chelius, D.; Shaler, T. A. *Bioconjugate Chem* **2003,** *14,* 205-211.

(132) Cagney, G.; Emili, A. *Nature. Biotechnol.* **2002,** *20*, 163-170.

(133) Chen, X.; Smith, L. M.; Bradbury, E. M. *Anal. Chem.* **2000,** *72*, 1134-1143.

(134) Rappsilber, J.; Siniossoglou, S.; Hurt, E. C.; Mann, M. *Anal. Chem.* **2000,** *72,* 267-275.

(135) McCammon, M. G.; Hernandez, H.; Sobott, F.; Robinson, C. V. *J. Am. Chem. Soc.* **2004,** *126,* 5950-5951.

(136) Keetch, C. A. et al. *Anal. Chem.* **2003,** *75,* 4937-4941.

(137) Benesch, J. L. P.; Sobott, F.; Robinson, C. V. *Anal. Chem.* **2003,** *75,* 2208-2214.

(138) Powell, K. D. et al. *J. Am. Chem. Soc.* **2002,** *124,* 10256-10257.

# Part II. *De novo* Peptide Sequencing for Protein Identification and Detection of Posttranslational Modifications

# Chapter 1. A method for *De Novo* Peptide Sequencing by Low Energy Collision-Induced Dissociation and the Application to the Identification of Unknown Proteins Isolated from Shark Cartilage[1]

Mass spectrometry (MS) has been widely used for protein identification. However, difficulties still exist with rapid identification techniques for unknown proteins from species that do not have available complete genome and proteome databases. Even for species that already have available genome databases, these same difficulties are also present for novel proteins that are not in the current database due to co/ posttranslational modifications. Structurally and genetically distinct types of proteins such as collagen are present in many species. A rapid and reliable identification method is needed to deal with all these situations. *De novo* interpretation combined with the BLAST (Basic Logic Alignment Search Tool) program is a practical protocol. However, it is still crucially important whether or not the *de novo* interpretation can rapidly provide a limited number of peptide sequence candidates with inclusion of the correct sequences. In this work, a rapid *de novo* protein sequencing method based on ESI-Quadrupole-Ion-Trap MS/MS experiments is developed. Using a known protein, bovine Serum Albumin (ALBU-BOVIN) as an example, it is illustrated that this prototype *de novo* interpretation approach is rapid and efficient. This method combined with the BLAST program is then applied to search homologies and predict posttranslational modifications of an unknown protein isolated from shark cartilage that do not have available complete genome and

[1] A version of this chapter will be submitted for publication as:
Hongying Zhong and Liang Li, "A method for *De Novo* Peptide Sequencing by Low Energy Collision-Induced Dissociation and the Application to the Identification of Unknown Proteins Isolated from Shark Cartilage".

proteome databases. PROSITE (a database of protein families and domains) program is applied to search the functions of this unknown protein.

### II. 1. 1. Introduction

Shark cartilage is known for being highly resistant to infections, for possessing remarkable wound-healing ability, and for its stimulation of the immune system to cancer [1]. Early studies [2-4] indicate that the proteins and other compounds that make up the entire endoskelton of shark may be an excellent source of such bioactivities. It is clear that isolation and identification of individual components that bear specific biological functions from shark cartilage can be of great value for the development of potential drugs for pharmaceutical applications.

Mass spectrometry combined with computer algorithms for database searching has been widely recognized as a powerful technique for protein identification because of its high sensitivity, high speed and the versatility to couple with many kinds of separation apparatus [5, 6]. Identification of proteins using both MALDI-TOF peptide mass mapping and partial sequencing by tandem mass spectrometry requires that the acquired mass spectral data can be accurately matched to those derived from protein sequences in a database. This process is greatly facilitated by the rapidly growing protein entries in a database due to the increase in the number of genome sequences being completed. However, there are many cases that the experimental data cannot be correlated with any database sequences. Protein entries for many important species in a database are still very limited. If the interested proteins are not present in a database and share only moderate sequence similarity with their known homologues, the enzymatic digestion cannot yield a sufficient number of identical peptide masses to produce a statistically reliable hit through database searching by MALDI-TOF peptide mapping [7]. Computer programs that can be more tolerant of sequence variants have been developed to identify proteins by tandem mass spectrometry data [8-12]. Unfortunately, these error-tolerant searching programs are unlikely to hit peptides that have multiple amino acid substitutions compared with the relevant sequences in a database. In addition to the unknown proteomes of species with unsequenced genomes, identification of known proteomes with available sequenced genome databases is also challenged to some extent by difficulties in direct database searching. There are lots of novel proteins that are not in the present databases resulting from co/posttranslational modifications [13] or small gene products that are difficult to be detected by the genomic approach [14]. The need to deal with these situations requires the development of new methods.

Historically, automated Edman degradation has been extensively used for protein sequencing. However, this approach requires very long running time and large amounts of highly purified proteins [15]. Sequencing by mass spectrometry has rapidly evolved and has been proven to be faster and require less protein sample than Edman techniques [16-17]. In the past few years, different techniques combined with mass spectrometry have been developed to derive the amino acid sequence of a peptide. Examples include C- or N-terminal sequencing using a specific enzyme digestion [18-19], peptide-ladder sequencing using a ladder generating chemistry [20-21] and chemical derivatization of peptides or proteins on either the N- or C- terminus [22-26]. Extra background peaks, unwanted side reactions and the need for highly purified samples limit their wide applications. Isotope labeling of C-terminus of peptides by performing proteolytsis in a mixture of $H_2{}^{16}O$ and $H_2{}^{18}O$ buffer [27] is known to help in correctly assigning the fragmentation pattern in which the C-terminus of a peptide exhibit a doublet spaced by 2 mass units. However, the precursor and fragment ion intensities are reduced.

Automated *de novo* protein sequencing by peptide fragmentation through collision-induced dissociation MS/MS experiments has been the subject of intense interest. Towards this end, several automated *de novo* interpretation programs of MS/MS spectra have been developed to derive peptide sequences. These have achieved varying degrees of success [28-39]. For example, a two-dimensional fragment correlation mass spectrometry based on $MS^3$ experiments was recently developed for automated data analysis to facilitate *de novo* sequencing [28]. It needs to acquire a large number of spectra. For an online HPLC separation of complicated peptide mixtures, the narrow time window present during peak elution may compromise the number of peptides that can be sequenced.

Over the years, there have been a lot of algorithms developed to interpret $MS^2$ spectra for sequencing known or unknown peptides. Sakurai et al. [29] calculated all possible amino acid combinations for a given peptide mass and then determine which sequence can best match for the ions found in the MS/MS spectrum. Another less computationally intensive approach is to build sequences beginning with small "subsequences" which seem to account for some of the observed fragment ions and then the small sequences are extended one residue by one residue. At each step, the predicted fragmentations are compared to the observed ions [30-34]. A graph theory approach has been applied to assign ions onto a "Sequencing Spectrum" and determines the sequence possibility from this graph [35-39]. The first step of this approach is to import data and then convert the ions mathematically into their corresponding transformed spectrum that contains a single type ion masses (b-type or y type). In the absence of additional information, it is impossible to assign ion types to each ion. To make the sequence graph, it would seem that each of the observed ions has to be assumed to be all of the possible ion types. For high-energy CID data, Hines et. al. [38] classified the peaks in the spectrum by using a pattern-based algorithm before generation of sequences. Each of the more intense of the original peaks is hypothetically assigned in turn to each of nine sequence-specific ion categories in which six major ion types result from cleavages along the peptide backbone and three other ions types are from side chain losses. Nine category scores are computed by using a simple function that correlates postulated ion masses with those actually observed. Obviously, there are not only large amounts of calculation but also peaks with relative low abundance below a constant threshold will be excluded and effectiveness of this method largely depends on the ability of the score function to correctly classify the original ions. Scarberry et. al. [39] developed a different approach, in which they applied artificial neutral networks to classify observed fragment ions into specific ion types before the data was transformed into sequence spectra. It is a viable approach to rapid computer interpretation of peptide CID spectra. The artificial neural networks used by this program is a nonlinear fashion with an n-element input vector and an m-category classification output vector. The parameters are determined by exposure to training data to learn the implicit associations between data elements and classification categories. Much large data set (i.e., hundreds of spectra) is needed to guarantee the correctness of classification. Additionally, peptides with different kinds of properties and modifications make the training step more complicated. Taylor and Johnson [35-36] took the advantage of the predictable fragmentation pattern in low-energy CID data and determined the N- and C-terminal evidence in which ions are not preselected as either b- or y- ions but all possibilities are considered. Since both b and y ions can produce sub-fragment ions with loss of water (-18) or ammonium (-17), even more in many cases these sub-fragment ions are absent in the spectra because b- or y- ions are underrepresented in the spectra in some mass regions, the determination of N- and C-terminal evidence is challenged which means that an ion has to be assumed to be both b-type and y-type. Once the sequence graph has been determined, sequences are generated by starting at the N-terminus of the graph and jumping from node to node in increments corresponding to amino acid monoisotopic masses. Several thousands of completed sequences are generated for the subsequent scoring, sorting and ranking.

The corresponding CIDentify program [35-36] was constructed which uses a modified FASTA sequence comparison algorithm to deal with the peculiarities and ambiguities of sequences obtained by *de novo* sequencing. With the rapid growth of sequence databases, the throughput of this approach is limited by the very long running time. Compared with FASTA program, BLAST (Basic Local Alignment Search Tool) programs [40] achieve much of their speed by avoiding the calculation of optimal alignment scores for all but only a handful of related sequences. It effectively identifies alignment "seeds" and extends around then a majority of database sequences is discarded without aligning with the queried sequence. Elsewhere, conventional BLAST programs require a single unambiguous sequence as input. Otherwise the similar sequences may match the same region of the protein sequence in a database, thus, resulting in a false positive high score. Shevchenko et. al. [41] developed a modified faster strategy called MS BLAST to help one to achieve homology-based database search which can tolerate a very high level of information noise resulting from the intrinsic ambiguities in *de novo* peptide sequencing produced by mass spectrometry. It overcomes the difficulty by selecting the best matching High Score Pairs (HSPs) from a number of redundant sequences and sorting the hits according to their total scores. By this way, the top hit protein is the one that has been matched to multiple peptide sequences.

In addition to the techniques that have been developed to deal with the many similar alternative sequences deduced from one spectrum, it is still critically important whether or not the spectra interpretation software can provide correct sequence candidates. To this end, there are still difficulties associated with many of sequencing programs. First, the difficulty to distinguish b-or y-ions in the spectra make the calculation steps very complicated. Without extra information, each ion present in the spectrum has to be assumed as all ion types. Based on this principle, many alternative sequences can be deduced from one MS/MS spectrum. The similar candidate sequences produced from *de novo* interpretation are not wanted for the following homology-based sequence database search. Secondly, in addition to the distinguishing of b-and y- ions in the spectra, other fragment ions resulting from unexpected cleavage also make the sequences construction difficult or completely wrong if they are assigned to be one of the b-or y- ions in the sequences. Thirdly, the situation will become more complicated if more than one peptide with very close mass elutes in the same time window and the obtained MS/MS spectra are essentially mixtures. Therefore, developing methods to extract the useful data from a massive raw data file can simplify the calculations, discard unexpected fragment ions, purify experimental data, calculate peptide masses and eventually obtain correct sequence candidates.

In eukaryotic organisms, after the mRNA is transcribed from DNA sequences, it is processed to remove introns and then translated on the ribosomal complex to synthesize the protein. Almost all protein sequences are posttranslationally modified by simple proteolytic cleavage or covalent modification of specific amino acids. There are more than 200 covalent modifications and their functions are still poorly understood. Moreover, these posttranslational modifications cannot be inferred from nucleotide sequence. Difficulties also are associated with searching homologous nonidentical matches to database-derived sequences even if the database search program has been informed of some modifications. The more modification possibilities are considered, the greater the number of database sequences with correct mass that will be obtained and the greater the possibility to get a false positive match. Therefore, it is expected that *de novo* MS/MS interpretation software can provide information that can predict modifications.

In this work, a new data interpretation protocol for *de novo* sequencing of peptides based on the μLC-ESI-MS/MS data is reported. The main purpose of this protocol is to reduce calculation, decrease the number and increase the accuracy of candidate sequences. It is also expected to predict posttranslational modifications. Experimental data was first pre-purified by an extracting step in which many unwanted fragment ions are excluded from calculations, retaining only b- and y- ions. By this approach, many calculations employed to classify fragment ions in the raw data file are discarded and also the possibility to obtain a wrong assignment of b- or y- ions is avoided. Additionally, there are only a few sequence candidates are provided by this method, so the follow up steps for scoring, sorting and ranking are greatly reduced.

MS/MS spectra acquired from an ESI-Quadrupole-Ion-Trap mass spectrometer usually contain continuous series of b- and y- ions. In this work, the raw data file was pre-purified before sequence generation. The most useful data used for making up the main frame of sequence generation was extracted from the massive raw data file based on the complementary relationship between b- and y- ions. Only the pairs whose sum is equal to a constant (peptide mass +2) were selected to construct the sequence graph. Accordingly, the fragment ions resulting from side chain cleavage, loss of $H_2O/NH_3$, multiple-charged fragment, and other unexpected fragmentation are excluded from calculations. This decreases the number and increases the accuracy of the candidate sequences. Sequence generation is achieved using a "two pass approach". First, the sequence graph containing both b- and y- ions is used to generate sequences by jumping from the smallest node to the next node on the limited amount of extracted data. The connectivity among the nodes are restricted by the proposed rules. When there are gaps, reasonable combinations of amino acids are needed to produce complete sequences and undefined gaps is represented as X. The obtained raw sequences are then compared with the raw data file. The highly matched sequences are highly scored. In this work, a modified preliminary score [42] is used to evaluate the *de novo* interpretation-derived sequences by comparing the experimental data with the predicted fragment ion values calculated for each sequence candidate. After this "first pass step" most of the possible sequence candidates have been obtained. Sequences of as many peptides as possible are merged to input into BLAST program (http://dove.embl-heidelberg.de/Blast2/) to search for homologies. In the "second pass step", the mass difference between the database-derived sequences and the *de novo* interpretation-derived sequences are compared for each high scoring pairs to retrieve a complete sequence or predict a potential modifications for each identified peptides. Peptides whose MS/MS spectrum do not yield any significant protein hit in BLAST searching are also chosen to do further posttranslational modification analysis. The high score pairs are submitted to PROSITE to predict the functions of the protein.

In this work, the protein isolated from shark cartilage extraction that does not have available complete genome database was identified as a collagen like protein and two posttranslational modifications were predicted. One is the hydroxyproline and another one is the glycosylation. These two modifications are very common in collagen [43] and they provide another evidence for the identification of collagen like protein.

### II. 1. 2. Experimental Section

**Materials and Reagents.** Unless otherwise noted, all chemicals were purchased from Sigma and were of analytical grade. For HPLC separation, mass spectrometric analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific). Dialysis tubing with a 500 Da molecular weight cut-off was purchased from Spectrum Laboratories, Inc.

**Sample Preparation.** The shark fin purchased from a local store was extracted using 95°C hot water for 2 hours and precipitated under 40% saturated $(NH_4)_2SO_4$. The resultant white precipitate was dialyzed overnight against pure water with a molecular weight cut-off membrane of 500 Da. The dialyzed sample was then lyophilized. Protein assay shows positive result. Further separation was carried out on an Agilient LC 1100 system. The sample was injected onto the C4 reversed-phase column (Vydac 214TP54, 250mm×4.6mm i.d., 5 µm, 300 A°) with a guard column (Vydac 214GK54). Mobile phase A was water/0.1% TFA (trifluoroacetic acid) and mobile phase B was acenotrile/0.1%TFA. A gradient was run from 5% to 60% B over a period of 50 minutes and then from 60% to 85% B over a period of 20 minutes at a flow rate of 1 ml/minute. A UV detector was used at 214 nm. The effluent was fractionated every one minute using a fraction collector. Fraction #33 was taken for further identification.

**Peptide Mass Mapping.** 5 µl of the collected HPLC fraction $33^{\#}$ was first reduced by mixing with 10 µl of 0.1 M $NH_4HCO_3$ and 5µl of 45mM DTT. The solution was reacted for 20 minutes at 37°C. Then, it was carbamidomethylated by adding 5 μl of 100 mM iodoacetamide and allows it to react at room temperature in the dark for 15 minutes. Finally, it was digested using trypsin at 37°C. The tryptic peptides were prepared for MALDI analysis, using the two-layer method. A first layer of matrix (0.09 M HCCA in 33% methanol/acetone) was spotted and allowed to dry followed by a thicker second layer of matrix (30% methanol/water saturated with HCCA) mixed with the sample by 1:1 ratio. MALDI mass spectra were obtained on an Applied Biosystems Voyager Elite TOF MS.

**Acquisition of MS/MS Spectra.** All MS/MS data used for sequence generation were acquired on a μLC-ESI-MS/MS system with data-dependent experiments and zoom scan data-dependent experiments (LCQ DECA, Thermal Finnigan). The tryptic peptides were injected into a C8 column (Micro-Tech Scientific, 15cm × 150μm i.d., 5 μm) that is directly coupled to a mass spectrometer. Mobile phase A was water/0.5% acetic acid and mobile phase B was acetonitrile/0.5% acetic acid. A linear gradient was run from 5% to 60% B over a period of 60 minutes. The flow from the pump was reduced from 300 μl/min to 1 μl/min using a splitting tee and a length of restriction tubing made from fused silica. The flow rate was measured by collecting the effluent from the end of the PicoTip (New Objective) using a 5μl graduated glass capillary during a period of time. The post-translational modification of glycosylation was characterized by MALDI-MS/MS (Qq-TOF, Sciex). The peptides were deposited using the dried droplet method with 0.5M DHB as matrix.

**Data Interpretation and Database Searching.** Database searching using a peptide mass mapping or MS/MS spectra was performed by either MASCOT (http://www.matrixscience.com) or SEQUEST (LC Q DECA built in software). Prediction of functional sites was searched by PROSITE at PBIL (Pole Bio-Informatique Lyonnais, http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html). Homologies searching was performed by either conventional BLAST or ungapped MS BLAST program provided at EMBL (Heidelberg, http://dove.embl-heidelberg.de/Blast2/). For conventional BLAST searching using peptide sequences with the highest preliminary score Sp, the search parameters were set as: blastp program; SwissPort database; PAM30MS matrix; 100 expect; both strand; cut off 50. No species is defined during searching. For ungapped MS BLAST searching using all peptide sequence proposals, the advanced options were set as: -nogap -hspmax 100 -sort_by_totalscore -span1. Symbols for editing sequences were defined as described in ref. 41:

(a) L stands for both leucine and isoleucine residues. Z stands for glutamine and lysine residues if undistinguishable in the spectra.

(b) X stands for an undefined amino acid residue.

(c) B stands for a putative trypsin cleavage site if the peptide is complete.

(d) All sequence proposals are spaced with the minus symbol (-) and are merged into a single string.

**Statistical Evaluation.** The sequences derived from *de* novo interpretation were evaluated against the raw data file by the preliminary score described in ref. 19. It mainly considers the matched intensities, the number of matched fragment ions, the type b- and y-ion continuity, the presence of immonium ions in the spectrum and the total number of predicted fragment ions. Statistical analysis of the conventional BLAST searching result was as described in references17 and 18. This BLAST searching was performed on the sequences which have the highest preliminary score Sp for each peptide. Regions of high local sequence similarity between individual peptides in the query and a protein sequence from the database entry were referred to as High Scoring Pairs (HSPs). The number of random HSPs scores equal or greater than score S is described by the Poisson distribution. This is the P value associated with the score S. Highly significant scores have P values close to zero. Ungapped MS BLAST searching was performed on all complete and partial peptide sequence proposals that have been merged together by the minus symbol (-). If the score is higher than the suggested threshold [41], the match is considered as a statistically significant match.

### II. 1. 3. Results and Discussion

**The *de novo* Sequence Analysis Protocol Using ESI Tandem Mass Spectrometry** is outlined in Fig. II-1-1. First, proteins were digested with trypsin and the tryptic peptides mass mapping and MS/MS spectra were acquired. A database searching was performed using MASCOT or SEQUEST. If no protein candidate has been hit significantly, the MS/MS spectra of unknown peptides are further interpretated *de novo*. In a MS/MS spectrum, there are many types of fragment ions that result from unexpected fragmentation, multiple-charged fragment ions or fragmentation ions with loss of $H_2O/NH_3$ in addition to b- and y- ions. These unexpected fragment ions need to be Discarded, otherwise they may produce complete wrong sequences. As discussed above, accurate classification of these ions not only results in huge calculations but also may lead to an incorrect assignment. It is thus necessary to have a method that can appropriately extract the useful data from the massive data files. The method presented

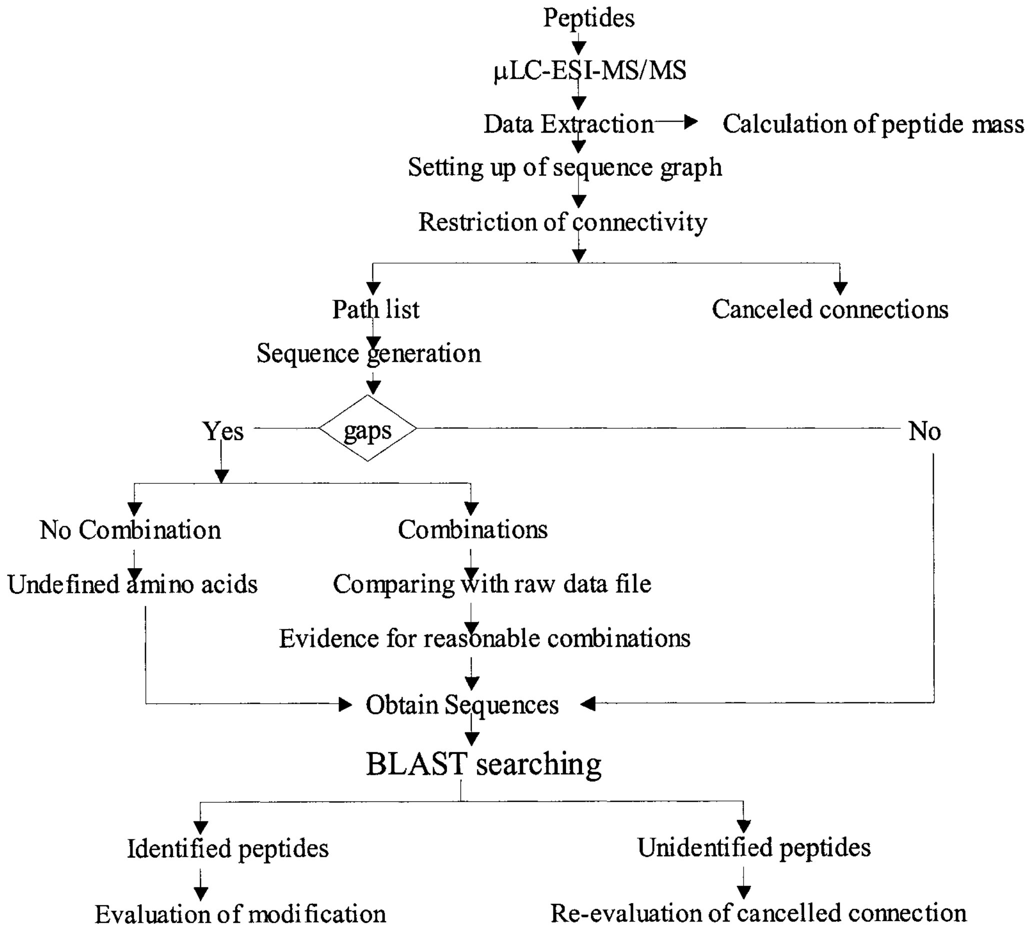


Figure II-1-1. Flow chart of *de novo* interpretation using ESI tandem mass spectrometry

makes use of the complementary relationship between b- and y- ions. The MS/MS spectrum of BSA (scan#1517 as shown in Fig. II-1-2 (a)) will be used to explain this *de novo* sequencing approach. Because b plus y is equal to a constant M+2 (peptide mass +2), only those data that have a complementary ion pair can be extracted as shown in Fig. II-1-2 (a) (X + X* = M + 2, X and X* are complementary ions). X is assigned as the smaller one for each data pairs. No distinguishment of b- and y- ions was made. Using this extracting step, the data for the main frame of sequences was rapidly purified and the calculation was greatly reduced. Through the data extraction step, the complicated experimental data can be differentiate by only extracting the complementary b- and y- ions under one parent ion mass and the correct precursor peptide mass can be calculated according to equation (1). In the meantime, random errors resulting from temperature fluctuation or other random factors are also reduced through this multiple points average.

$$[\mathrm{MH}] = \frac{\sum_{i=1}^{n}(X_n + X_n^*)}{n} - 1.01 \qquad (1)$$

n: number of pairs of complementary ions

In this case, the experimental peptide mass $[\mathrm{MH}]^+$ calculated from the average of the extracted data was 1479.69 u (monoisotopic mass) with an error of 0.0074 %. This result also indicates that the extraction step has correctly selected the data to be used for further sequence generation. After the data extraction, the spectrum was idealized into a sequence graph that contains two types of fragment ions, as shown in Figure II-1-2 (b). Sequences can be generated by jumping from the smallest node to the next node.

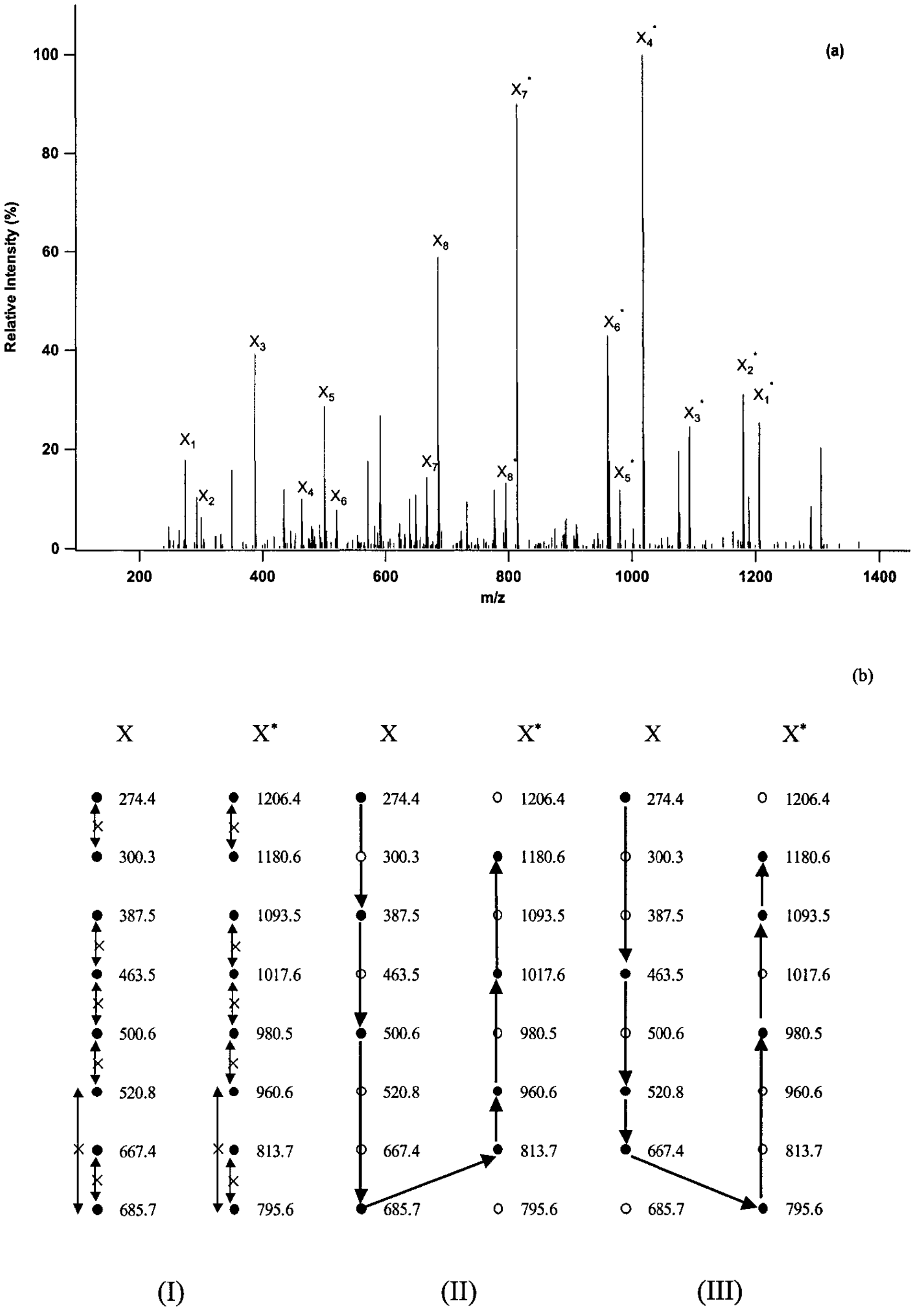


Fig. II-1-2. Process of sequence generation. (a) Data extraction from a MS/MS spectrum of ESI tandem mass spectrometer. $X_i$ and $X_i^*$ (i=1 to 8) are complementary ions. (b) Setting up of sequence graph. There are 8 pairs, so the possible paths are 64. Each path is rooted at the smallest node. The connectivity of this graph is restricted by the proposed mass rules. There are 6 apparent illegal small connections indicated as crossed arrows. According to the restriction, the final paths are shown as II and III.

Because X and X$^*$ are complementary ions, each path contains either X or X$^*$. If there are n pairs, the number of possible paths (R) is:

$$R = \left(C_2^1\right)^{n-1} = 2^{n-1} \qquad (2)$$

In order to further limit the number of sequence generated, the connectivity among the nodes are restricted. The first restriction is to cancel the apparent small illegal connections according to the masses of amino acid residues. So if there are m illegal connections, the final reasonable jumping routes R is:

$$R = \frac{\left(C_2^1\right)^{n-1}}{2^m} = 2^{n-m-1} \qquad (3)$$

n: pairs of complementary ions, m: apparent illegal connections

In situations where the pathe is cancelled due to its unmatched amino acid mass, there may be some modifications. But *de novo* interpretation will be performed on as many peptides as possible, so a few modified peptides will not affect the identification of highly hit proteins. Actually, because the cancelled paths are known, MS/MS spectra with unidentified peptides will be re-evaluated and analyzed if they are from potential modifications or from other proteins in the "second pass". The second restriction is to consider the non-apparent connections. If the gaps do not fit the first restriction and also do not match any combination of amino acids, they are replaced by undefined amino acids symbol X.

In Fig. II-1-2 (b) (I), the 6 apparent small illegal connections are shown as crossed arrows. Therefore, there are two reasonable paths rooted at the smallest node as calculated by equation (3). These are shown as a solid arrow in (II) and (III). Then sequences are generated according to these two paths. The dark dots are selected and the white dots are rejected during the process of sequence generation. In the path shown as Figure II-1-2 (b) (III), there is no amino acid combination that matches the mass difference between 274.4 and 463.5, it was then replaced by an undefined amino acid symbol X and a modification possibility was proposed for this path. The obtained sequences are compared with raw data file. The highly matched one will be highly scored.

A preliminary score [19] was modified to evaluate the sequences obtained from *de novo* interpretation in order to compare the information contained in the raw data file with the predicted fragment ion values calculated for each sequence derived from the *de novo* interpretation. The following formula was used to calculate the preliminary score $S_P$ [42]:

$$S_P = (\sum i_m) n_i (1+\beta)(1+\delta)(1+\varepsilon)/\eta_\tau \qquad (4)$$

where $i_m$ is the matched intensities, $n_i$ is the number of matched fragment ions, $\beta$ is the type b- and y-ion continuity, $\delta$ is the presence of immonium ions, $\varepsilon$ is the presence of fragment ions with loss of $H_2O$ or $NH_3$ and $\eta_\tau$ is the total number of predicted fragment ions.

Finally the sequences reported for these peptides are YGFQNALLVR and SLNAQFGXVR. In total ten peptide MS/MS spectra acquired from in-solution tryptic digestion of BSA were used for *de novo* sequence interpretation. Two kinds of BLAST searching results are summarized in Table II-1-1. It is encouraging to find that conventional BLAST searching using sequences which have the highest preliminary score Sp for each peptides shows the top six significantly hit proteins are all serum albumin. Further ALBU-BOVIN from bovine is correctly on the top 1 position with distinguishable P values. Compared with the result of un-gapped MS BLAST search using all sequence proposals, the same proteins are identified and 9 out of 10 sequences with the highest Sp are identified by MS BLAST. It means that this *de novo* interpretation approach can correctly provide sequence candidates and achieve satisfactory searching for BLAST program.

**Identification of Proteins from Shark Cartilage.** We applied the proposed protocol to identify a HPLC fraction from the hot water extraction of shark fin that does not have a genome database. A peptide mass mapping acquired from an in-solution tryptic digestion of this fraction using MALDI-TOF is shown in Figure II-1-3. MASCOT searching did not yield any significant hit. Subsequently, peptides were injected into a μLC-ESI-MS/MS system with data-dependent experiments (LCQ DECA, Thermal Finnigan) to acquire MS/MS spectra. The data was input into MASCOT and SEQUEST. Again this did not yield significant hit. *De novo* interpretation was run on the 11 most intense peaks. Sequences are generated according to this *de novo* interpretation approach as described above. Figure II-1-4 shows the assignment of b-, y- ions, multiple charged ions and other

Table II-1-1. Candidate sequences determined by *de novo* interpretation.

| Peptides ($MH^+$) | *De novo* sequences | High Score Pairs | BLAST | MS BLAST |
|---|---|---|---|---|
| 1479.80 | **YGFQNALLVR-**<br>SLNAQFGXVR | Query: 154 LGSFLYEYSR 163<br>LGSFLYEYSR<br>Sbjct: 350 LGSFLYEYSR 359 | P02769/ALBU-BOVIN, P=1.3E-15<br>P14639/ALBU-SHEEP, P=1.3E-10<br>P49064/ALBU-FELCA, P=3.4E-10<br>P02768/ALBU-HUMAN, P=1.3E-09<br>P49822/ALBU-CANFA, P=3.0E-09<br>P08835/ALBU-PIG, P=4.0E-09 | P02769/ALBU-BOVIN, S=507<br>P14639/ALBU-SHEEP, S=423<br>P49064/ALBU-FELCA, S=400<br>P02768/ALBU-HUMAN, S=350<br>P49822/ALBU-CANFA, S=345<br>P08835/ALBU-PIG, S=335 |
| 1163.63 | **LVNELT-**<br>PDNELNGP-<br>LVQDLNGP | Query: 23 LVNELT 28<br>LVNELT<br>Sbjct: 66 LVNELT 71 | | |
| 1014.62 | **ZTALVE**-NDAVLEV | not identified | | |
| 927.49 | **YLYELAR-**<br>YLYELVK | Query: 63 YLYELAR 69<br>YLYE+AR<br>Sbjct: 161 YLYEIAR 167 | | |
| 922.49 | **AEFVEVTK-**<br>AEFQTGVTK | Query: 79 AEFVEVTK 86<br>AEFVEVTK<br>Sbjct: 249 AEFVEVTK 256 | | |
| 1002.58 | **LVVSTQTALA-**<br>LVVSTQTASP-<br>DPVSNDTALA-<br>DPVSNDTASP | Query: 98 LVVSTQTALA 107<br>LVVSTQTALA<br>Sbjct: 598 LVVSTQTALA 607 | | |
| 1305.71 | **HLVDE**-EDVTM | Query: 142 HLVDE 146<br>HLVDE<br>Sbjct: 402 HLVDE 406 | | |
| 1567.74 | **LGSFLYEYSR-**<br>LGSEMYEYDK | Query: 154 LGSFLYEYSR 163<br>LGSFLYEYSR<br>Sbjct: 350 LGSFLYEYSR 359 | | |
| 1142.71 | **ZQTALVELLK-**<br>ZQTGSGPNLLK | Query: 176 ZQTALVELLK 185<br>+QTALVELLK<br>Sbjct: 548 KQTALVELLK 557 | | |
| 1639.94 | **QVSTPTL-**<br>QVSTACH-<br>QVSTYGSSVV-<br>VGGNDK | Query: 199 QVSTPTL 205<br>QVSTPTL<br>Sbjct: 440 QVSTPTL 446 | | |

fragment marker ions with loss of $H_2O/NH_3/CO$. Table II-1-2 summarizes the identification results and sequence alignments by conventional BLAST program. Table II-1-3 summarizes the identification results and sequence alignments by ungapped MS BLAST program. Comparing these two tables, it is found that the proteins identified by the BLAST searching using sequences which have the highest Sp for each peptides was also identified by ungapped MS BLAST program. All of them are significantly hit. The peptide sequences covered in BLAST searching were also identified by MS BLAST searching. This result indicates the identification of this unknown protein is significant and the proposed *de novo* sequencing method can appropriately provide correct sequence candidates. After finishing the "first pass" analysis, posttranslational modifications will be continued in the "second pass" approach.

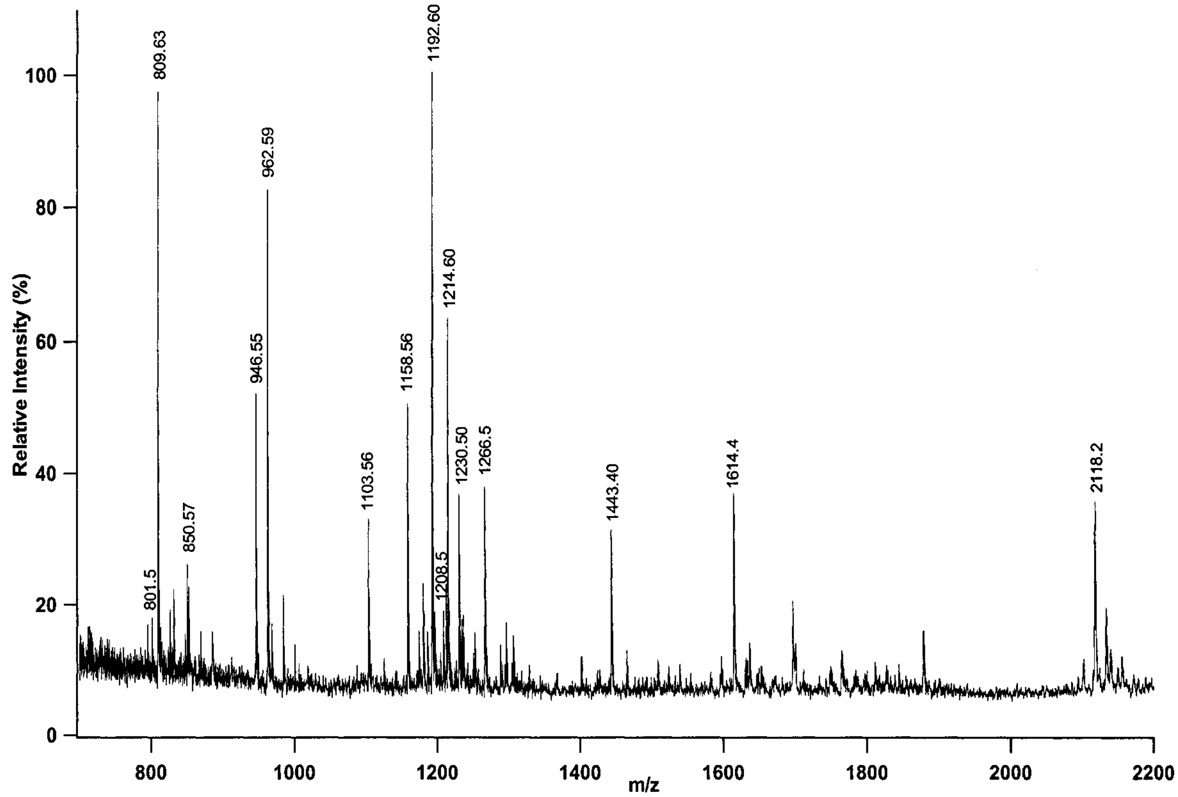


Figure II-1-3. MALDI-TOF peptide mass mapping of one HPLC fraction.

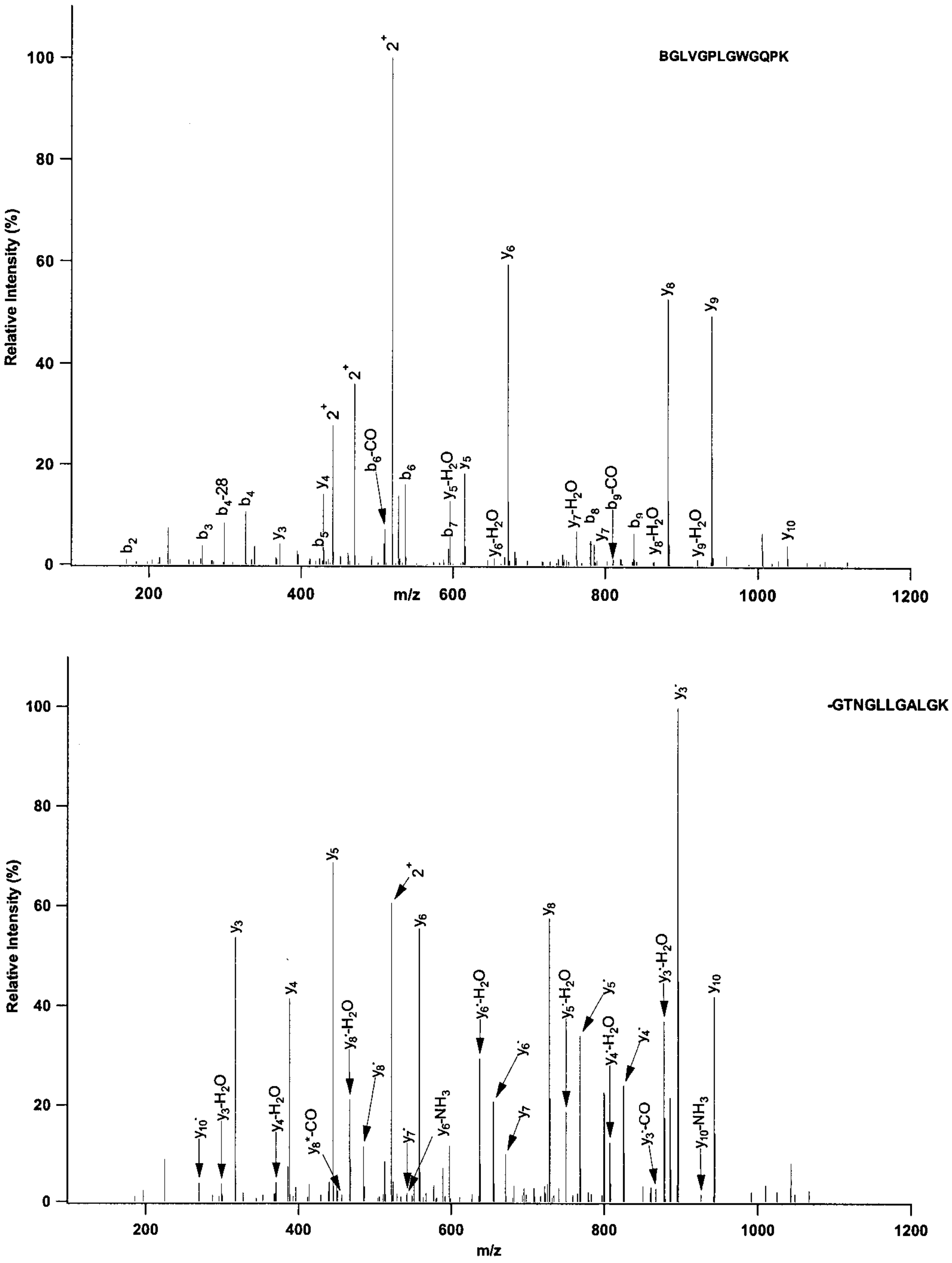


Figure II-1-4. Selected MS/MS spectra and sequences

Table II-1-2. Identification of the unknown protein from one of the HPLC fraction of shark fin by *de novo* interpretation combined with conventional BLAST Searching

| Identification | selected High Score Pairs<br>(BLAST program) | Input unique peptide s | Covered peptides |
|---|---|---|---|
| Q01149<br>CA21-<br>MOUSE<br>P=8E-04 | Query: GPVGAVGPR-AGAAGPAGLR<br>        GPVGAVGPR  G  GP G+R<br>Sbjct: GPVGAVGPR--GPSGPQGIR<br>        994-------------1011<br><br>Query: GFLGAEPWGPK-BGLVGPLGWGQPK-BGPVGAVGPR-AGAAGPAGLR<br>        GF GA P GPK  G  GP G    P    GP G  GPR      G  GL<br>Sbjct: GFPGA-P-GPK--GELGPVG--NP---GPAGPAGPR-----GEVGLP<br>        262----------------------------------------294<br><br>Query: BGPVGAVGPR-AGA       Query: BGVSGLVG<br>        +G VGA GP  AGA                G SG VG<br>Sbjct: RGRVGAPGP--AGA       Sbjct: -GLSGPVG<br>        228--------239               295--301<br><br>Query: GTNGLLGALGK-BGLVGPLQ-V-GAPG   Query: BGLVGPLQVGAPG<br>        GTNGL GA    +G  G L  V GAPG          +G+ G   VGAPG<br>Sbjct: GTNGLTGA----KGATG-LPGVAGAPG   Sbjct: RGIPGA--VGAPG<br>        307---------------------328          681-------691<br><br>Query: AGAAGPAGLR-BGVS----GLVG--GR-BGFVGPEGQR-PLGATGP<br>        AG AGP GLR  G S    GL G  GR  G  GP G R   G TGP<br>Sbjct: AGPAGPPGLR--G-SPGSRGLPGADGR-AGVMGPPGNR---GSTGP<br>        396---------------------------------------434 | 10 | 9 |
| O46392<br>CA21-<br>CANFA<br>P=6E-04 | Query: GPVGAVGPR-AGAAGPAGIR   Query: GFVGPEGQR-PLGATGP<br>        GPVGAVGPR  G  GP GIR          --VGP G   P+G  GP<br>Sbjct: GPVGAVGPR--GPSGPQGIR   Sbjct: --VGPAG---PIGSAGP<br>         988------------------------1005          243-----------254<br><br>Query: QG-PK-GPVGAVGPR   Query: GFLGAEPWGPK<br>        QG P  G VG  GPR          GF GA P GPK<br>Sbjct: QGAP--GSVGPAGPR   Sbjct: GFPGA-P-GPK<br>         1041-------1053          256------274<br><br>Query: VGPEGQR-PLGATGP   Query: GPVGAVGPR-AGAAGPAGI<br>        VGP G   P+G  GP          GPVGA      AGA G  GI<br>Sbjct: VGPTG---PIGSAGP   Sbjct: GPVGA-----AGATGARGI<br>         753---------764           331-------------344<br><br>Query: VGPL-QVGAPG      Query: GVSGLVG      Query: GAAGPAGIR<br>        +GP   VGAPG               GVSG VG               GA GPAG+R<br>Sbjct: IGPVGAVGAPG      Sbjct: GVSGPVG      Sbjct: GATGPAGVR<br>         954-----964                289-295              424---432 | 10 | 8 |

Table II-1-3. Identification of the unknown protein from one of the HPLC fraction of shark fin by *de novo* interpretation combined with ungapped MS BLAST Searching.

| Top identified proteins | High Score Pairs (MS BLAST program) | Input unique peptides | Identified peptides |
|---|---|---|---|
| Q01149/CA2 1-MOUSE (S=293) | Query: GPVGAVGPR<br>GPVGAVGPR<br>Sbjct: 994 GPVGAVGPR 1002<br>Query: AGAAGPAGLR<br>AG AGP GLR<br>Sbjct: 396 AGPAGPPGLR 405<br>Query: GPAGLR<br>GPAG+R<br>Sbjct: 433 GPAGIR 438<br>Query: GTNGLLGA<br>GTNGL GA<br>Sbjct: 307 GTNGLTGA 314<br>Query: BGFVGPEGQR<br>+G VGP+G R<br>Sbjct: 162 RGVVGPQGAR 171<br>Query: VGAPG<br>VGAPG<br>Sbjct: 687 VGAPG 691<br>Query: BGVSGLVG<br>G SG VG<br>Sbjct: 295 GLSG**P**VG 301 | 10 | 7 |
| O46392/CA2 1-CANFA (S=167) | Query: GPVGAVGPR<br>GPVGAVGPR<br>Sbjct: 988 GPVGAVGPR 996<br>Query: GPAGLR<br>GPAG+R<br>Sbjct: 1075 GPAGIR 1080<br>Query: VGAPG<br>VGAPG<br>Sbjct: 960 VGAPG 964<br>Query: GVSG**L**VG<br>GVSG VG<br>Sbjct: 289 GVSG**P**VG 295 | 10 | 4 |

**Posttranslational Modification Analysis.** It is also very interesting to find that in the Table 2 and Table 3, there are two peptides GFLGA and GVSGLVG which contain L (m/z=113) can't match the database-derived sequences GFPGA and GVSGPVG which contain P (m/z=97). When we look back into the highly hit protein collagen, this is easily explained. In collagen, Proline is often posttranslationally modified as Hydroxyproline which has a mass of 113, the same as Leucine. This modification was not input into the sequence generation, the amino acid was reported as Leucine. The presence of this hydroxyproline provides another evidence for the identification of this protein as collagen. It also indicates the potential ability of this *de novo* interpretation protocol to predict posttranslation modifications. For the *de novo* interpretation of the MS/MS spectrum of peptide 1103.56 (m/z), sequences did not match any significant highly hit proteins. It was then considered to be post-translationally modified. Post-translational modification analysis was studied using MALDI MS/MS (Figure II-1-5). The MS/MS spectrum indicates the presence of HexA-Hex-HexA-Hex glycosylation. Based on the literature [43], this kind of glycosylation is possible in the collagen protein family. So the modification study result provides another evidence for the identified proteins which were highly hit in BLAST searching.

### II. 1. 4. Conclusions and Perspectives

This chapter demonstrates that identification of the unknown proteome that does not have complete genome database and also shares very little sequence similarity with other known homologies can be performed by the proposed *de novo* interpretation combined with BLAST searching program. The protocol of the *de novo* interpretation presented here improves the data quality through an extraction step that discards the other

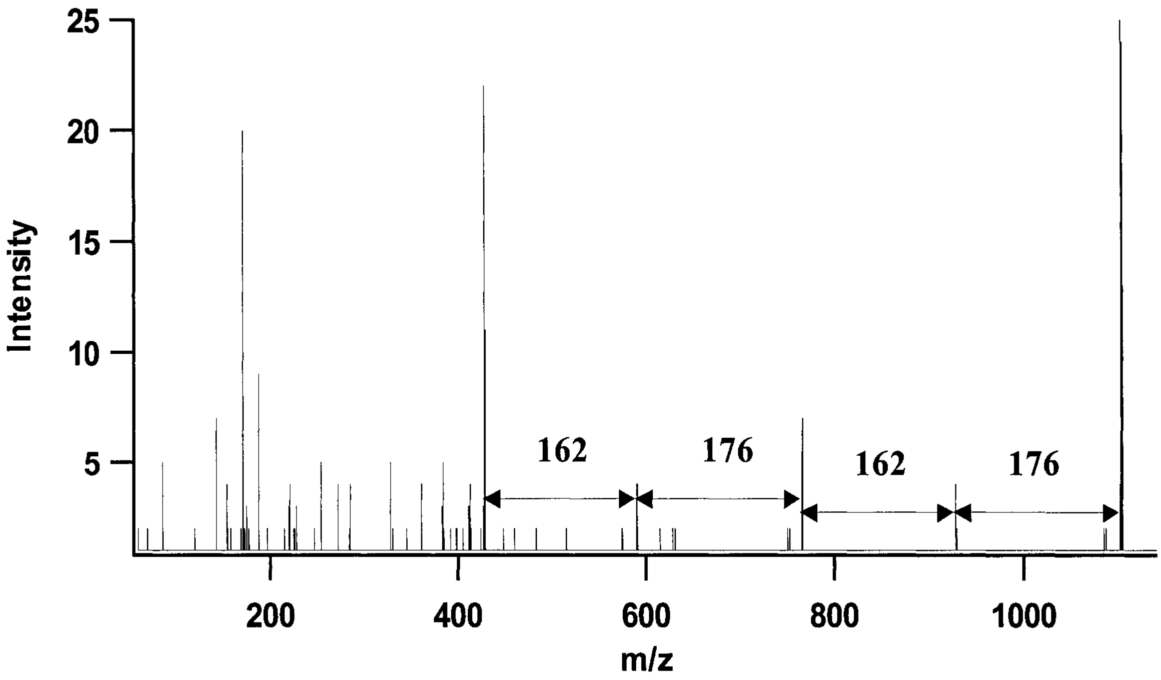


Figure II-1-5. MALDI MS/MS spectrum of peptide 1103.56 (m/z)

unexpected fragment ions. Therefore, not only the calculation step needed to classify the fragment ion types is greatly reduced but also the candidate sequences decrease. The peptide mass is calculated through multiple points average and the random errors resulting from random factors is reduced. using the "two pass" approach, sequences are rapidly, correctly obtained and the potential posttranslational modifications are predicted. Because the calculation was greatly reduced, this method is also suitable for manual evaluation of MS/MS data. The limitation of this protocol is that it largely depends on the quality of MS/MS spectra and is better for small peptides whose mass is less than

1500 (m/z). A way to overcome this problem is to digest proteins using multiple enzymes or combined with chemical cleavage in order to obtain smaller peptide fragments.

### II. 1. 5. Cited Literature.

(1) Fontenele, J. B.; Araujo, G. B.; Alencar, J. W.; Viana, G. S. B. *Biol. Pharm. Bull.* **1997**, *20*, 1151-1154.

(2) Lee, A. K.; Beuzekom, M.V.; Glowacki, J.; Langer, R. *Comp. Biochem. Physiol.* **1984**, *78B*, 609-616

(3) Moses, M.; Langer, R. *Science* **1990**, *248*, 1408-1410.

(4) Fontenele, J. B.; Viana, G. S. B.; Xavier-Filho, J.; Alencar, J. W. *Braz. J. Med. Biol. Res.* **1996**, *29*, 643-646.

(5) Blackstock, W. P.; Weir, M. P. *Trends Biotechnol.* **1999**, *17*, 121-127.

(6) Pandev, A.; Mann, M. *Nature* **2000**, *405*, 837-846.

(7) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871-2882.

(8) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466-469.

(9) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, . *Anal. Chem.* **1996**, *68*, 850-858.

(10) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390-4399.

(11) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290-299.

**(12)** Shevchenko, A.; Keller, P.; Scheiffele, P.; Mann. M.; Simons, K. *Electrophoresis* **1997**, *18*, 2591-2600.

(13) Meri, S.; Baumann, M. *Biomecular Engineering* **2001**, *18*, 213-220.

(14) Jungblut, P. R.; Muller, E. C.; Mattow, J.; Kaufmann, S. H. E. *Infection and Immunity* **2001**, *69,* 5905-5907.

(15) Creighton, T. E. *Proteins: Structures and Molecular Principles*, W. H. Freeman: N. Y. 1984.

(16) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U. S. A.* **1996,** *93*, 14440-14445.

(17) Wilm, M.;Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature*, **1996**, *379*, 466-469.

(18) Bradley, C. V.; Williams, D. H.;Hanley, M. R. *Biochem. Biophys. Res. Commun.* **1982**, *104*, 1223-1230.

(19) Doucette, A.; Li, Liang *Proteomics* **2001**, 7, 157-170.

(20) Chait, B. T.; Wang, R.; Beavis, R. C.; Kent, S. B. *Science* **1993**, *262*, 89-92.

(21) Tarr, G. E. US Patent # 5824556.

(22) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 6233-6237.

(23) Pfeifer, T.; Rucknagel, P.; Kuellertz, G.; Schierhorn, A. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 362-369.

(24) Martin, M.; Manfredo, Q.; Giovanni, M.; Peter, J. *Anal. Chem.* **2000**, *72*, 4047-4057.

(25) Tsugita, A.; Takamoto, K.; Ataka, T.; Sakuhara, T.; Uchida, T. US Patent #6046053.

(26) Schneider, L. V.; Hall, M. P.; Peterson, J. N. US Patent #6379971.

(27) Schnoelzer, M.; Jedrzejewski, P.; Lehmann, W. D. *Electrophoresis* **1996**, 17, 945-953.

(28) Zhang, Z.; McElvain, J. S. *Anal. Chem.* **2000**, 72, 2337-2350.

(29) Sakurai, T.; Matsuo, T.; Matsuda, H.; Katakuse, I. *Biomed. Mass Spectrom.* **1984**, *11*, 396-399.

(30) Biemann, K.; Cone, C.; Webster, B. R.; Arsenault, G. P. *J. Am. Chem. Soc.* **1966**, *88*, 5598-5606.

(31) Ishikawa, K.; Niwa, Y. *Biomed. Environ. Mass. Spectrom.* **1986**, *13*, 373-380.

(32) Siegel, M. M.; Bauman, N. *Biomed. Environ. Mass. Spectrom.* **1988**, *15*, 333-343.

(33) Johnson, R. S.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945-957.

**(34)** Yates, J. R., III; Griffin, P. R.; Hood, L. E. *Computer Aided Interpretation of Low Energy MS/MS Mass Spectra of Peptides*, Academic Press: San Diego, CA, 1991.

**(35)** Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067-1075.

(36) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594-2604.

(37) Bartels, C. *Biomed. Environ. Mass Spectrom.***1990**, *19*, 363-368.

(38) Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 326-336.

(39) Scarberry, R. E.; Zhang, Z.; Knapp, D. R. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 947-961.

(40) Altschul, S. F.; Madden, T. L.; Schaffer, A. A. ; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nuleic Acids Res.* **1997**, *25*, 3389-3402.

(41) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917-1926.

(42) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426-1436.

(43) Piez, K. A.; Reddi, A. H. *Extracellular Matrix Biochemistry,* Elsevier: New York, 1984.

# Chapter 2. Two-dimensional Mass Spectra Generated from the Analysis of $^{15}$N-Labeled and Unlabeled Peptides for Efficient Protein Identification and *de novo* Peptide Sequencing[1]

Protein identification has been greatly facilitated by database searches against protein sequences derived from product ion spectra of peptides. This approach is primarily based on the use of fragment ion mass information contained in a MS/MS spectrum. Unambiguous protein identification from a spectrum with low sequence coverage or poor spectral quality can be a major challenge. A two-dimensional (2D) mass spectrometric method is presented in which the numbers of nitrogen atoms in the molecular ion and the fragment ions are used to provide additional discriminating power for much improved protein identification and *de novo* peptide sequencing. The nitrogen number is determined by analyzing the mass difference of corresponding peak pairs in overlaid spectra of $^{15}$N-labeled and unlabeled peptides. These peptides are produced by enzymatic or chemical cleavage of proteins from cells grown in $^{15}$N-enriched and normal media, respectively. It is demonstrated that, using 2D information, i.e., m/z and its associated nitrogen number, this method can, not only confirm protein identification results generated by MS/MS database searching, but also identify peptides that are not possible to identify by database searching alone. Examples are presented of analyzing *Escherichia coli* K12 extracts that yielded relatively poor MS/MS spectra, presumably from the digests of low abundance proteins, which can still give positive protein

[1] A version of this chapter has been accepted for publication as:
Hongying Zhong, Sandra Marcus and Liang Li, "Two-dimensional Mass Spectra Generated from the Analysis of $^{15}$N-Labeled and Unlabeled Peptides for Efficient Protein Identification and *de novo* Peptide Sequencing" *J. Proteome Res.* **2004,** in press. Dr. Sandra Marcus performed cell culture of *E. coli* K12.

identification using this method. Additionally, this 2D-MS method can facilitate spectral interpretation for *de novo* peptide sequencing and identification of post-translational or other chemical modifications. It is envisioned that this method should be particularly useful for proteome expression profiling of organelles or cells that can be grown in $^{15}$N-enriched media.

### II. 2. 1. Introduction

Mass spectrometry combined with computer algorithms for database searching based on MS and MS/MS data generated from the analysis of proteolytic peptides has been widely used for protein identification [1-4]. The efficiency of this protein identification approach ultimately requires that the acquired mass spectra be matched accurately to protein sequences from the corresponding database entries inferred from known (or predicted) gene sequences using bioinformatics tools [5]. Database entries are expanding rapidly with the availability of a growing number of completely sequenced genomes of many species and organisms. With the increase in database size in general, the possibility of MS/MS spectra being matched with a number of peptides with similar matching scores increases as well. This problem will increase, as there is a growing need in proteomics applications for the analysis of comprehensive proteomes comprised of both high and low abundance proteins. Only one peptide may be detected from a low abundance protein and its product ion spectrum may be of low quality in terms of sequence coverage and signal-to-background ratio. Database searching using low quality spectra often results in poor matching to database entries.

Another important issue related to database searching for protein identification is that many novel proteins resulting from co- or post-translational modifications or gene products undetected by the genomic and bioinformatic tools (e.g., via gene slicing) are not encoded in DNA sequences. Homology searching can only identify the similar parts of the protein sequences. *De novo* peptide sequencing and characterization of chemical modifications are becoming increasingly important, but are very challenging when based on MS/MS spectral interpretation of proteolytic peptides [6].

A strategy for improving protein identification and protein sequencing is to generate additional information on a peptide beyond a mere MS/MS spectrum of the intact peptide. This can be achieved by using chemical derivatization or isotope labeling to produce a peptide having a mass-tag that can be used to facilitate database searching or spectral interpretation. For example, $H_2{}^{18}O$ can be used in protease digestion to introduce the $^{18}O$ tag through the hydrolysis reaction to label all proteolytic peptides uniformly at the C-terminus, allowing the y-type ions to be distinguished from the b-ions [7-9], and providing a convenient means of proteome quantitation based on molecular ion intensities [10-17]. Similarly, N-terminal acetylation or C-terminal methylation can also be used to tag peptides [18], and modification by nicotinyl-n-hydroxysuccinimide at the N-terminus has been used to label proteins [19]. Specific labeling techniques such as isotope-coded affinity tag (ICAT) reagents have been used to isolate and characterize cysteine-containing peptides selectively [20, 21]. Lysine-containing peptides can be characterized by using mass-coded abundance tagging (MCAT) [22]. While these methods require carrying out additional reactions to a protein or peptide, they are effective in improving database searching results and *de novo* sequencing as well as for MS-based quantitation with differential labeling of two samples [23-27].

Isotope labeling can also be done by incorporation of stable isotopes into metabolic products during cell growth. Replacement of $^{13}$C for $^{12}$C, $^{15}$N for $^{14}$N, and $^{2}$H for $^{1}$H in proteins can generate characteristic mass shifts in their isotopic distribution patterns without affecting their chemical or structural properties. Amino acid coded mass tagging has been used for single peptide-based protein identification [28-30]. Uniformly $^{15}$N-labeled proteins have been used for MS-based quantitation from simple organisms such as bacteria [31-33], mammalian cell cultures [33, 34], and multicellular organisms such as *Caenorhabditis elegans* [35], and for improvements in sensitivity and accuracy of molecular mass measurement [36, 37]. Trace labeling of proteins with carbon-13 has been shown to be useful to identify the labeled peptide fragments in complex mixtures containing both labeled and unlabeled peptides [38].

In this chapter, a method will be described for improving protein identification and *de novo* sequencing using whole cell *in vivo* N$^{15}$-labeling combined with tandem MS analysis of labeled and unlabeled protein digests. When cells are grown in normal and N$^{15}$-enriched media, proteins isolated from the two cultures are differentially labeled. Analyzing the digests of the proteins generates two sets of product ion spectra, and by overlaying the two MS/MS spectra obtained from the N$^{15}$-labeled peptide and its corresponding unlabeled peptide, the number of nitrogen atoms present in the molecular ion and fragment ions can be readily determined. The nitrogen number reflects the amino acid composition. Using the analysis of proteins isolated from *E. coli* cell cultures as an example, this method is demonstrated to be very effective for improved protein identification and *de novo* sequencing.

### II. 2. 2. Experimental Section

**Materials.** All chemicals for gel electrophoresis were from BioRad (Hercules, CA) and the other chemicals were from Sigma (St. Louis, MO) and were analytical grade. HPLC grade methanol and acetonitrile were from Fisher (Mississauga, ON).

**Sample Preparation.** *E. coli* K-12 (*E. coli*, ATCC 47076) was from the American Type Culture Collection. A single *E. coli* K12 colony was used to inoculate 10 ml of LB broth (BBL, Becton Dickinson). The culture was incubated overnight with shaking at 37 °C. 1.5 ml of this saturated culture was added to 90 ml of labeled or unlabeled growth medium in a 500 ml baffled Erlenmeyer flask. Bio-Express Cell Growth Media was from Cambridge Isotope Laboratories and came unlabeled and $^{15}$N-labeled (96-99%). Cells were grown as above and harvested after 7 hours. The optical density (at 600 nm) for the unlabeled cells was 3.56 and for $^{15}$N-labeled cells was 3.84.

For the in-gel protein identification experiments, 10 mg of lyophilized cells were suspended in 1 ml of 50 mM Tris-acetate buffer (pH 7.5) containing 1% Triton X100 and 1 mM ethylenediamine tetraacetate (EDTA), and then subjected to sonication (Branson probe sonicator, Branson Ultrasonics Corp., Danbury, CT) for 2 minutes. The cell suspension was centrifuged for 5 minutes at 10,000 g and a sample of the supernatant was subjected to sodium dodecylsulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Protein concentration was determined by using the Coomassie® plus protein assay protocol (Pierce, Rockford, IL) performed on a microplate reader (Thermomax) at 650 nm. For SDS-PAGE, 50 μl of supernatant containing about 400 μg of proteins was mixed with an equal volume of sample buffer {1.25 ml 0.5 M Tris-HCl (pH 6.8), 2.50 ml glycerol, 2.00 ml 10% SDS, 0.20 ml 0.5% (v/v) bromophenol blue, and 0.50 ml β-mercaptoethanol, made up to a total volume of 10.00 ml with 3.55 ml deionized water} and then heated at 95°C for 4 minutes. For gel electrophoresis, 20 μl of the reduced sample containing about 80 μg of proteins was then loaded in each lane of a 10% acrylamide gel using a Mini-PROTEAN® 3 cell system (BioRad). Electrophoresis was carried out at a constant voltage of 200 V for about 1 hour. After electrophoresis, gels were stained with Coomassie Blue R-250 (BioRad).

The bands were excised and in-gel digested according to the protocol developed by Shevchenko and co-workers [39]. In brief, an excised gel piece was placed in a 0.6-ml siliconized vial and 25 μl of acetonitrile was added to the vial to dehydrate the gel piece with vortexing for about 20 minutes until the gel piece became completely white. The gel piece was placed in another vial and dried by Speedvac. The dried gel piece was re-hydrated by vortexing with 25 μl of 45 mM dithiothreitol (DTT) in 100 mM $NH_4HCO_3$ and then incubated at 37°C for 1 hour. After removal of the excess liquid, 25 μl of 100 mM iodoacetamide in 100 mM $NH_4HCO_3$ was added and vortexed for about 5 minutes. The sample was incubated in the dark at room temperature for 45 minutes. The excess liquid in the vial was again removed and 40 μl of 100 mM $NH_4HCO_3$ was added and vortexed for 10 minutes. After removing the liquid, 40 μl of acetonitrile was added and vortexed for 10 minutes. The gel piece was placed in a new vial and dried by Speedvac. The gel was swelled in 30 μl of 0.5 μg/μl of trypsin in 100 mM $NH_4HCO_3$ with 3 mM $CaCl_2$ on ice for 45 minutes and was then incubated at 37°C for 4 hours. The liquid around the gel piece was removed and saved to a new vial. After adding 50 μl of 50% acetonitrile with 0.1% trifluoroacetic acid (TFA), the gel piece was subjected to sonication for 10 minutes. The liquid was removed and pooled with the previously saved liquid. This extraction process was repeated three times. Another extraction was carried out by adding 50 μl of 75% acetonitrile with 0.1% TFA to the gel, followed by sonication for 10 minutes. The extracts were combined and dried to about 20 μl by Speedvac. The extracted tryptic peptides were further concentrated and desalted with a C18 Ziptip (Millipore, Bedford, MA, USA).

For the in-solution proteome analysis approach, proteins were extracted by suspending 10 mg of lyophilized cells in 1 ml of 0.1% TFA solution, followed by sonication for 2 minutes and centrifugation for 5 minutes at 10,000 g. Pre-fractionation by reversed-phase High Performance Liquid Chromatography (HPLC) of the complex protein mixture was carried out with an Agilent 1100 HPLC system (Palo Alto, CA) by injection of 50 μl (about 80 μg of proteins) of the supernatant on a C8 column (4.6 mm ID x 250 mm, 5 μm, 300 A°) (Vydac, Hesperia, CA) at a flow rate of 1 ml/min, at room temperature, with a linear gradient from 5% to 25% B over 10 minutes and then to 85% B over 60 minutes, where mobile phase A was water containing 0.1% (v/v) TFA, and mobile phase B was acetonitrile with 0.1% (v/v) TFA. After 10 minutes of elution one-minute fractions were collected. After concentration by Speedvac to about 5 μl, 15 μl of 100 mM $NH_4HCO_3$ and 5 μl of 45 mM DTT were added to each fraction and then incubated for 20 minutes at 37°C. Then 5 μl of 100 mM iodoacetamide was added and the mixture was allowed to stand at room temperature in the dark for 15 minutes. About 0.5 μl of 0.5 μg/μl trypsin stock solution was added and incubated for 30 minutes at 37°C. The tryptic peptides were then desalted with a C18 Ziptip (Millipore) and were ready for LC MS/MS.

**MS and Database Searching.** MS and MS/MS spectra were acquired on an electropray ionization LCQ DECA (ThermoFinnigan, CA) ion trap mass spectrometer coupled to a microflow HPLC system (ThermoFinnigan). LC separation was done on a capillary C18 column (150 μm i. d. x 150 mm long, 5 μm, 300 A°) (LC Packings, CA). Mobile phase A was water with 0.5% acetic acid and mobile Phase B was acetonitrile with 0.5% acetic acid. A linear gradient was run from 5% to 85% B over 60 minutes at a flow rate of 1 μl/minute.

Database searching was performed without limitations on protein molecular weight, pI or species. Database searching using tandem mass spectra was performed by SEQUEST (ThermoFinnigan) or by MASCOT (http://www.matrixscience.com). The BLAST program of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) was used for sequence validation.

### II. 2. 3. Results and Discussion

**Experimental protocol.** Figure 1 illustrates the experimental and data interpretation workflow of the 2D-MS method for enhancing the performance of database searching for protein identification and *de novo* peptide sequencing. In this method, the number of nitrogen atoms present in the molecular and fragment ions provides the second dimension of information with the m/z values as the first dimension. The nitrogen number of a given ion can be readily determined, by calculating the mass difference between the peak pair shown in the overlaid spectra of the product ions of peptides with and without $^{15}$N labeling. Among the 20 common amino acids, most of them have one nitrogen atom resulting in a 1 Da mass shift except R, which has 4 nitrogen atoms, H has 3 nitrogen atoms, and N, Q, K, and W have 2 nitrogen atoms. Thus the mass shifts between the $^{15}$N-labeled and unlabeled peptides and their corresponding fragment ions reflect their amino acid compositions. This 2D-MS method can be used for protein identification based on both in-gel and in-solution proteome analysis approaches. As Figure 1 illustrates, to apply this method to in-gel proteome analysis, proteins extracted from cells grown in normal and $^{15}$N-enriched media were independently separated by gel electrophoresis. The stained gel bands of interest were excised for in-gel digestion, followed by peptide extraction and MS/MS analysis of the peptides. For in-solution proteome analysis, proteins were digested by an enzyme and the resultant peptides were analyzed by LC

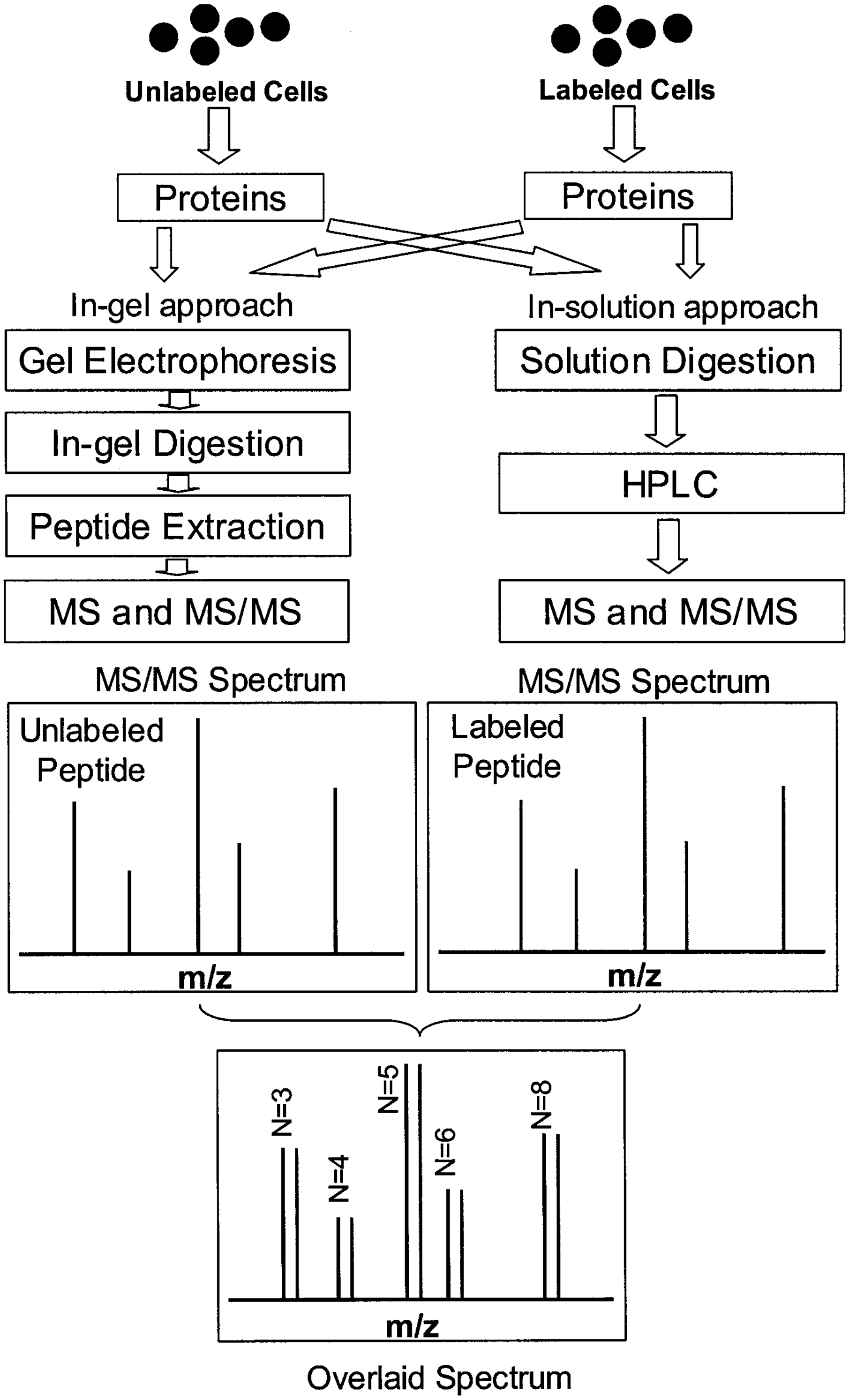


Figure II-2-1. Experimental and data processing workflow for the 2D-MS method.

MS/MS. In both cases, the MS/MS spectra of peptide pairs with and without labeling were individually produced and then overlaid for the analysis of nitrogen numbers. For this work, matching the peptide pair from the individually collected MS/MS spectra was done manually based on the similarity of gel image patterns in the in-gel approach and on retention times of peptides in the in-solution approach. This peptide pairing process should be automatable in the future by modifying the algorithms of existing MS and MS/MS database searching programs (e.g., replacing $^{14}$N with $^{15}$N in database sequences, followed by searching the product ion spectrum of the labeled peptide against this database).

**Confirmation of database searching by 2D-MS method.** To illustrate how the 2D-MS method can be used to improve database search results, Figure II-2-2 presents an example where the identification of a high abundance protein separated by SDS-PAGE of the *E. coli* K12 extract was attempted. In this case, both SEQUEST and MASCOT database searching of product ion spectra of the peptides extracted from the gel band identified the protein as OMPA-ECOLI (P02934). Since this protein was present in high abundance in the sample, its identification was made possible with high score matches of several peptides from the same protein (i.e., seven peptides listed in Figure II-2-2). The nitrogen number information derived from the comparison of MS/MS spectra of labeled and unlabeled peptides confirmed the MS/MS data searching results. This is shown in the overlaid spectra in Figure II-2-2 where the nitrogen numbers match with the amino acid compositions of the fragment ions. A number of other high abundance proteins detected from the gel electrophoresis or LC MS/MS experiments have been examined and in all cases, the nitrogen number of the molecular and fragment ions was able to provide additional information to confirm the protein identification results generated by MS/MS database searching.

OMPA-ECOLI (P02934)

| Scans | Sequence | $MH^+$ | Z | XC | DeltaCn | sp | Rsp | Ions |
|---|---|---|---|---|---|---|---|---|
| 340-342 | IGSDAYNQGLSER | 1410.5 | 2 | 2.867 | 0.579 | 1016.9 | 1 | 18/24 |
| 361-363 | GIKDVVTQPQA | 1156.3 | 2 | 2.552 | 0.535 | 807.6 | 1 | 15/20 |
| 1339-1344 | DGSVVVLGYTDR | 1281.4 | 2 | 3.563 | 0.605 | 1338.9 | 1 | 18/22 |
| **1434-1437** | **AALIDCLAPDR** | **1214.4** | **2** | **2.624** | **0.531** | **799.8** | **1** | **19/33** |
| 1517-1519 | SDVLFNFNK | 1084.2 | 2 | 3.072 | 0.533 | 730.6 | 1 | 13/16 |
| 1680-1682 | AQSVVDYLISK | 1223.4 | 2 | 2.852 | 0.641 | 956.8 | 1 | 17/20 |
| 1846-1848 | LGYPITDDLDIYTR | 1655.8 | 2 | 2.981 | 0.696 | 969.0 | 1 | 19/26 |

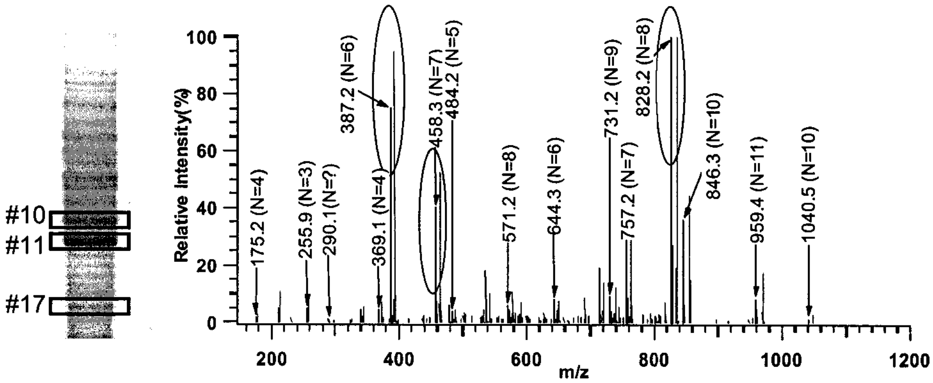


Figure II-2-2. MASCOT database search results from MS/MS spectra of seven peptides detected from in-gel digestion of gel band #11. One of the overlaid MS/MS spectra matched with the sequence AALIDCLAPDR of protein OMPA-ECOLI (P02934) is shown. The peaks are labeled with m/z values and nitrogen numbers (N) determined from the overlaid spectra.

**Eliminating ambiguous identification.** In addition to confirming the database searching results, the 2D-MS method can provide discriminating power to resolve ambiguous peptide assignments. An example of this is shown in Figure II-2-3. Figure II-2-3 shows the product ion spectrum of a peptide from in-gel digestion of band #17 in the gel image shown in Figure II-2-2. This seemingly good quality spectrum was unfortunately matched with several possible peptide candidates from MASCOT database search (see inset of Figure II-2-3A). The 1st and 2nd ranked peptides have the sequences of GFGFITPADGSK (N=13) and YPFLTESLAR (N=13), respectively. Panels B and C of Figure II-2-3 show the fragment ion assignments of these two peptides given by the database search results. It is clear that the product ion spectrum matches well with both peptides by visual inspection. Such ambiguous peptide assignments with similar scores are not uncommon in analyzing complex proteome digests. Figure II-2-3A shows the overlaid spectra of the labeled and unlabeled peptides. The molecular ion masses of the labeled and unlabeled peptides have a mass difference of 13. Thus this peptide contains 13 nitrogen atoms, consistent with the nitrogen number of the top two ranked peptides. However, if we compare the nitrogen numbers of the fragment ions determined from the overlaid spectra shown in Figure II-2-3A with those predicted from fragment ion assignments shown in Figure II-2-3B or 3C, it can be seen that only the sequence assignment shown in Figure II-2-3B is correct (the nitrogen-number-matched peaks are indicated by dots). This example illustrates that with the aid of the nitrogen number, unambiguous assignment can be attained in situations where two or more peptides can be assigned to the same MS/MS spectrum with similar matching scores.

**Identification of proteins with incomplete MS/MS spectra.** With the 2D-MS method unambiguous peptide identification can be achieved using a smaller number of fragment ions found in the product ion spectrum compared to that used by MS/MS database searching alone. This can be illustrated in a positive control experiment where MS/MS spectra of peptides from a previously identified protein are examined. Taking the

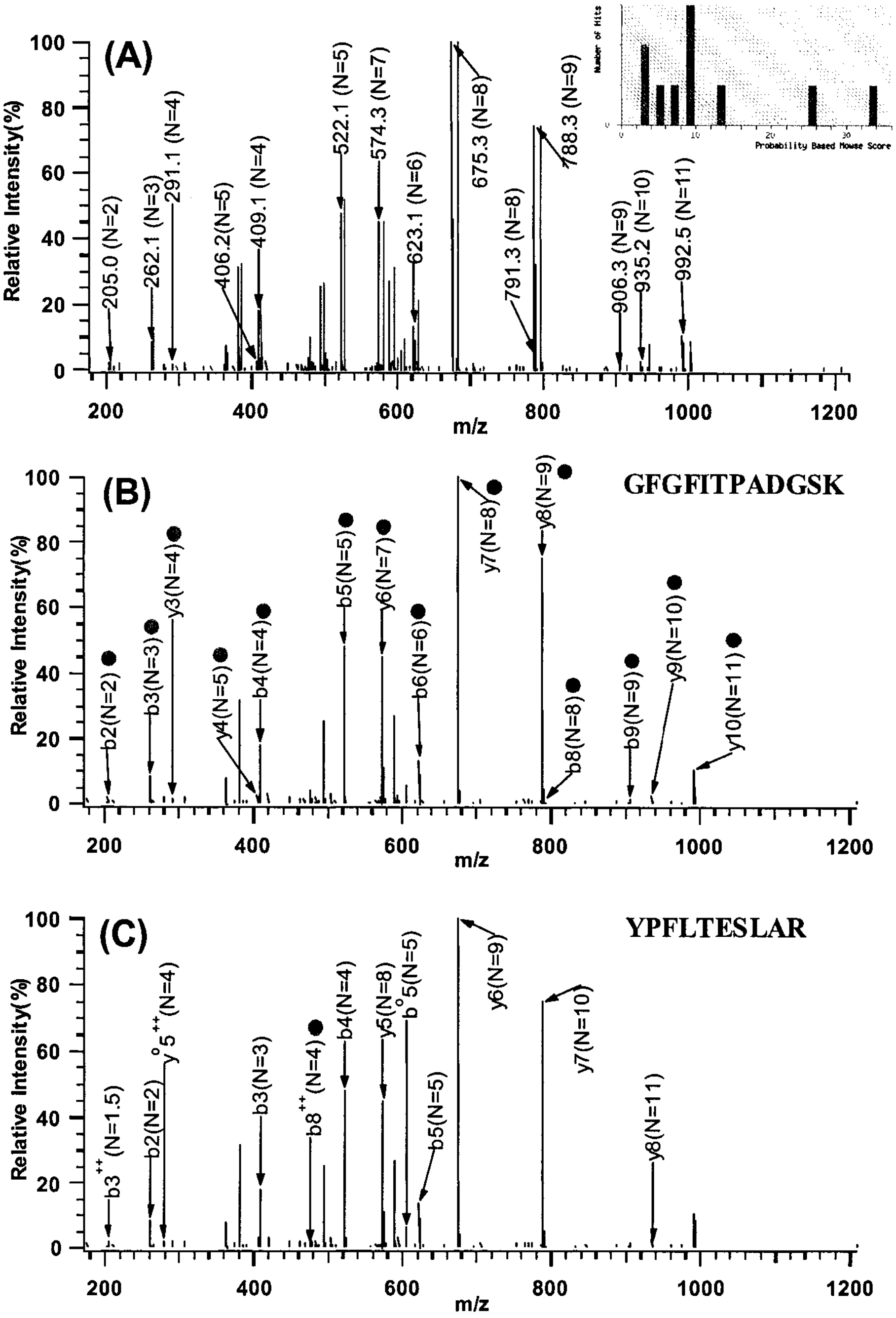


Figure II-2-3. (A) Overlaid MS/MS spectra of a peptide from gel band #17 along with the summary of the MASCOT search results shown in the inset. (B) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the top ranked sequence GFGFITPADGSK. The calculated nitrogen number for each fragment ion is shown in parentheses. The dot indicates that the nitrogen number of the peak matches with the experimentally determined number from the overlaid spectra. (C) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the $2^{nd}$ ranked sequence YPFLTESLAR.

spectrum shown in Figure II-2-2 as an example to demonstrate this, we deliberately deleted the spectral information from this product ion spectrum. It is assumed that only a single peptide (highlighted with bold font in the peptide list shown in Figure II-2-2) was detected and only the three most intense fragment ions of this peptide (indicated by circles in the spectrum) were observed in the MS/MS spectrum. Database searching using these three fragment ions produced matches with several peptides. Table II-2-1 lists the top 7 matched peptides by the MASCOT program. The correct peptide has a matching score ranked at #5. As Table II-2-1 shows, the molecular ion masses of all seven peptides match well with the peptide mass determined in the experiment. The fragment ions match with b or y ions or a combination of the two types for the listed peptides. However, according to the molecular ion mass shift of the labeled and unlabelled peptides, the total nitrogen number of this peptide should be 14. Out of the seven peptides, only the 5th ranked peptide has a sequence containing a total nitrogen number of 14. Further discrimination of these seven peptides is evident in the matched nitrogen numbers of the fragment ions (see last column of Table II-2-1). Thus the combination of only three fragment ions and the nitrogen numbers of the intact peptide and fragment ions allow for unique identification of this peptide as AALIDCLAPDR. The results from this control experiment illustrate that it should be possible to use the 2D-MS approach to identify peptides using relatively low quality MS/MS spectra. This is significant for real world applications where proteins are present in varying abundances. In situations where the amount of a protein is low, a single peptide may be detected from the protein and the quality of the MS/MS spectrum in terms of the number of fragments observed and their intensities may not be sufficiently high for positive protein identification using MS/MS database searching. Two examples of identifying proteins from relatively poor product ion spectra are shown in Figures II-2-4 and II-2-5.

Table II-2-1. Top 7 matched peptides from MASCOT database search using only the three most intense fragment ions shown in Figure II-2-2.

| No. | Sequence | Score | Mass error (%) | Nitrogen number in molecular ion | | Nitrogen number in fragment ion | |
|---|---|---|---|---|---|---|---|
| | | | | Calc. | Match? | Calc. | Match? |
| 1 | EEKAALADNKK | 38 | 0.01 | 15 | no | b3: N=4 | no |
| | | | | | | b4: N=5 | no |
| | | | | | | b8: N=9 | no |
| 2 | VESAAGSKKPIK | 30 | 0.06 | 15 | no | b4: N=4 | no |
| | | | | | | b5: N=5 | no |
| | | | | | | y8: N=11 | no |
| 3 | AMALLKERLGL | 27 | 0.06 | 15 | no | b4: N=4 | no |
| | | | | | | y4: N=7 | yes |
| | | | | | | y7: N=11 | no |
| 4 | IICAVLDGIIK | 27 | 0.06 | 12 | no | b3: N=3 | no |
| | | | | | | b4: N=4 | no |
| | | | | | | y8: N=9 | no |
| 5 | **AALIDCLAPDR** | 27 | 0.06 | 14 | yes | y3: N=6 | yes |
| | | | | | | y4: N=7 | yes |
| | | | | | | b8: N=8 | yes |
| 6 | LSDALEKDKPV | 26 | 0.06 | 13 | no | b3: N=3 | no |
| | | | | | | y4: N=5 | no |
| | | | | | | y7: N=9 | no |
| 7 | GRLVDTDAVIR | 26 | 0.06 | 17 | no | y3: N=6 | yes |
| | | | | | | y4: N=7 | yes |
| | | | | | | b8: N=11 | no |

From the gel band #11 shown in Figure II-2-2, four proteins were unambiguously identified through MS/MS database searching. But there were also other several peptides from the same band that could not be identified with high confidence using the product ion spectral information alone. Figure II-2-4 shows one of the MS/MS spectra. Very few

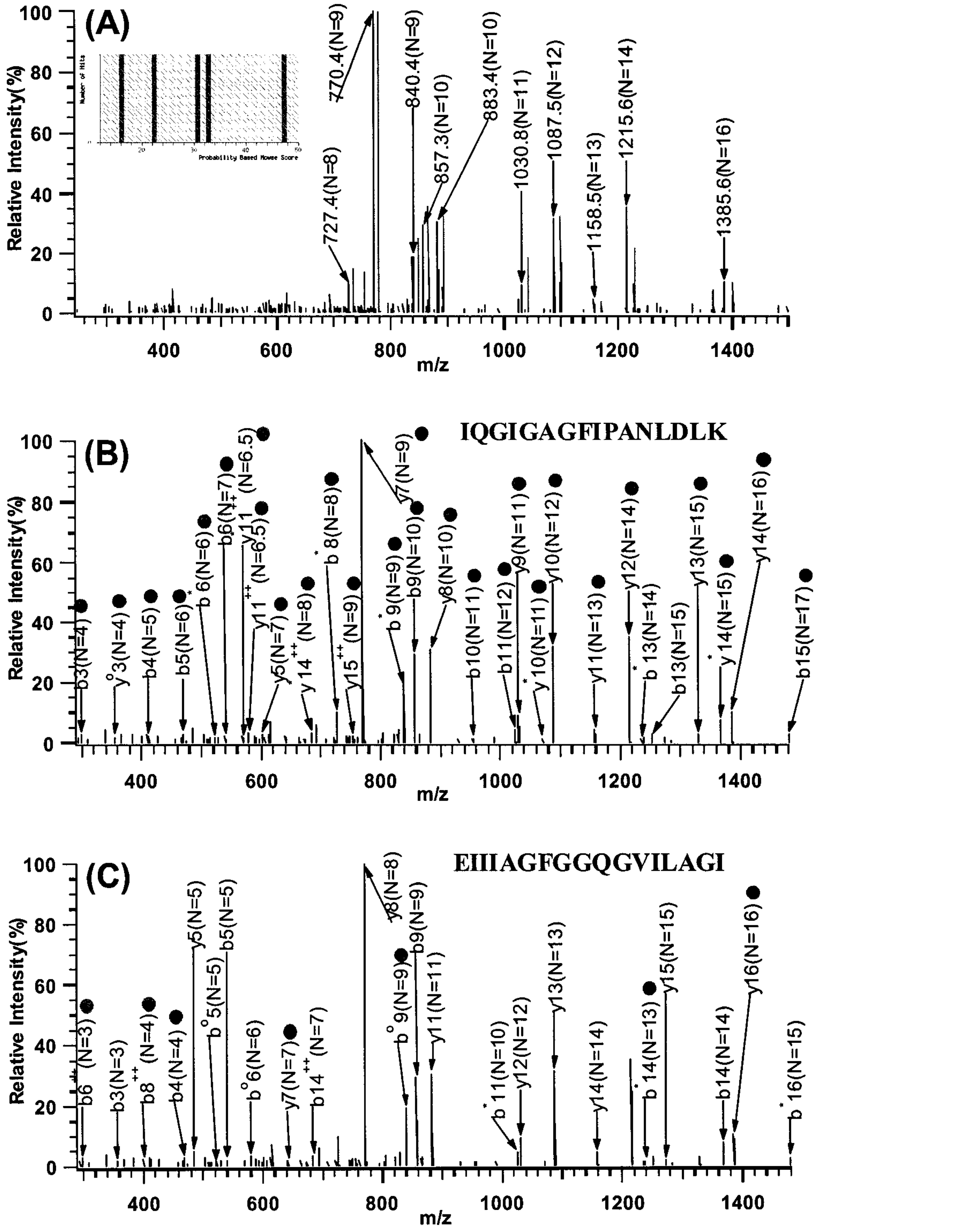


Figure II-2-4. (A) Overlaid MS/MS spectra of a peptide from gel band #11 along with the summary of the MASCOT search results shown in the inset. (B) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the top ranked sequence IQGIGAGFIPANLDLK. (C) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the 2nd ranked sequence EIIIAGFGGQGVILAGI.

fragment ions could be discerned from the background signals in the low mass region of the spectrum. The top two peptides that have good match with this spectrum as labeled are IQGIGAGFIPANLDLK (N=19) and EIIIAGFGGQGVILAGI (N=18). As shown in panels B and C of Figure II-2-4, the fragment ions of these two peptides match with the MS/MS spectrum of the unlabeled peptide very well. The fragment ion peaks can be assigned to b and y ions. The calculated nitrogen numbers based on the amino acid compositions of the fragment ions are also shown along with the peak assignments. Figure II-2-4A shows the overlaid MS/MS spectra of the labeled and unlabeled peptides. The nitrogen number determined from the overlaid spectra is shown for each pair of peaks. Comparing the nitrogen numbers of the fragment ions shown in Figure II-2-4A or II-2-4B with those corresponding numbers in Figure II-2-3C indicates that IQGIGAGFIPANLDLK (N=19) is the correct peptide. The sequence match of EIIIAGFGGQGVILAGI (N=18) can be eliminated.

Another example from gel band #11 is shown in Figure II-2-5. In this case, the MS/MS spectrum has poor signal-to-noise ratios and displays very few discernible fragment ion peaks (i.e., m/z 1150.7, 829.9, and 359.1). Database searching did not yield significant matches, as shown in the inset of Figure II-2-5A. The top two matched peptides with very low scores are listed in the figure. The peak assignments based on the sequences of these peptides are shown in panels B and C, respectively, of Figure II-2-5. The overlaid spectra of unlabeled and labeled peptides are shown in Figure II-2-5A. The 2$^{nd}$ ranked sequence LRELAENNPLGDYLR has its molecular ion and fragment ion masses matched with their corresponding nitrogen numbers (labeled with black dots), while

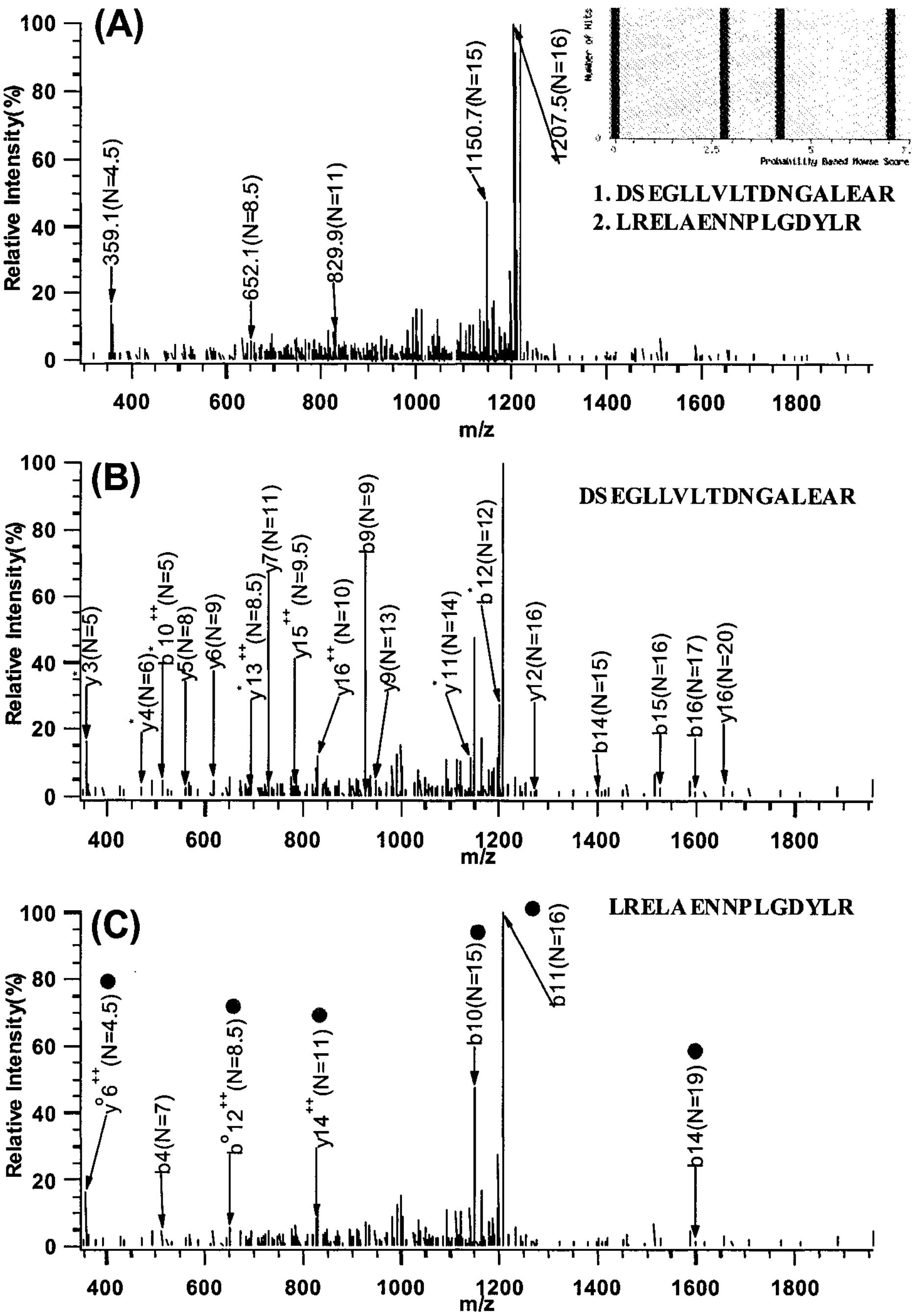


Figure II-2-5. (A) Overlaid MS/MS spectra of a peptide from gel band #11 along with the summary of the MASCOT search results shown in the inset. (B) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the top ranked sequence DSEGLLVLTDNGALEAR. (C) MS/MS spectrum of the unlabeled peptide with fragment ion peaks assigned to the 2nd ranked sequence LRELAEDNPLGDYLR.

the other sequence does not match. As with the positive control experiment described earlier, assignment of LRELAENNPLGDYLR as the correct peptide can be made based on the fragment ion and nitrogen number matching of three relatively intense peaks (plus two less intense ones) with the database sequence.

**Facilitating of *de novo* peptide sequencing.** An additional feature of the 2D-MS method is that it can greatly facilitate the *de novo* peptide sequencing process. *De novo* sequencing is useful for the identification of unknown proteins isolated from species with no genome database, as well as for the identification of mutations or errors of genome sequences. For *de novo* sequencing using tandem MS spectra, one major challenge is to distinguish b and y ions for unambiguous assignment of peptide sequence. Chemical derivatization is one option, but it often requires a relatively large amount of sample. For characterizing unknown proteins from species with no genome database, homology searching can assist in performing *de novo* sequencing, but can only identify the similar part of the protein and determination of a stretch of unknown sequence of unmatched parts is still needed. Using the 2D-MS method, *de novo* interpretation of the MS/MS spectrum can be done more easily because it determines the sequence not only by residual amino acid masses, but also by their corresponding nitrogen numbers.

The process of the 2D-MS method for *de novo* sequencing can be illustrated using one of the MS/MS spectra obtained from an ESI-ion trap tandem mass spectrometric analysis of the digest from gel band #11. Figure II-2-6A shows the overlaid MS/MS spectra of the labeled and unlabeled peptides. With the overlaid spectra, the peaks of the peptide fragment ions (i.e., present as peak pairs with mass difference corresponding to the nitrogen number difference) can be readily discerned from the background peaks (i.e., no peak pairing), which eliminates false assignment of possible sequences to background peaks. Based on the complementary relationship of b ion and y ion, seven pairs of data can be extracted from the fragment ion masses for sequencing. The nitrogen number of each fragment ions is labeled in Figure II-2-6A. For each pair of complementary b and y ions, the sum of their nitrogen numbers is equal to the total nitrogen number of the peptide. In this case, the total nitrogen number of the peptide is 14.

The y-series can be read from the observed m/z 175.2 (N=4) peak. The peak at m/z 290.2 (N=5) is the $y_2$ ion with a sequence of DR. Using the incremental change of nitrogen number as a guide, we can readily assign the next peak at m/z 387.2 (N=6) as $y_3$ (PDR), and the next y-series peak at m/z 458.3 (N=7) as $y_4$ (APDR). Similarly, other y-series peaks can be assigned.

The sequence from the peak with the smallest nitrogen number, N=3, at m/z 255.9 can be read. In the same series of peaks (i.e., b or y series), the nitrogen number should increase as the sequence length increases. The logical choice to extend the series of the N=3 peak is to examine the fragment ion with N=4 having a higher m/z value, i.e., m/z 369.1 (N=4). The mass difference between these two fragment ions is 113 Da and the nitrogen number increment is 1. So the sequence gap can be identified as amino acid L or I. Next to the m/z 369.1 (N=4) peak, there is one peak with N=5 and two peaks with N=6. It is

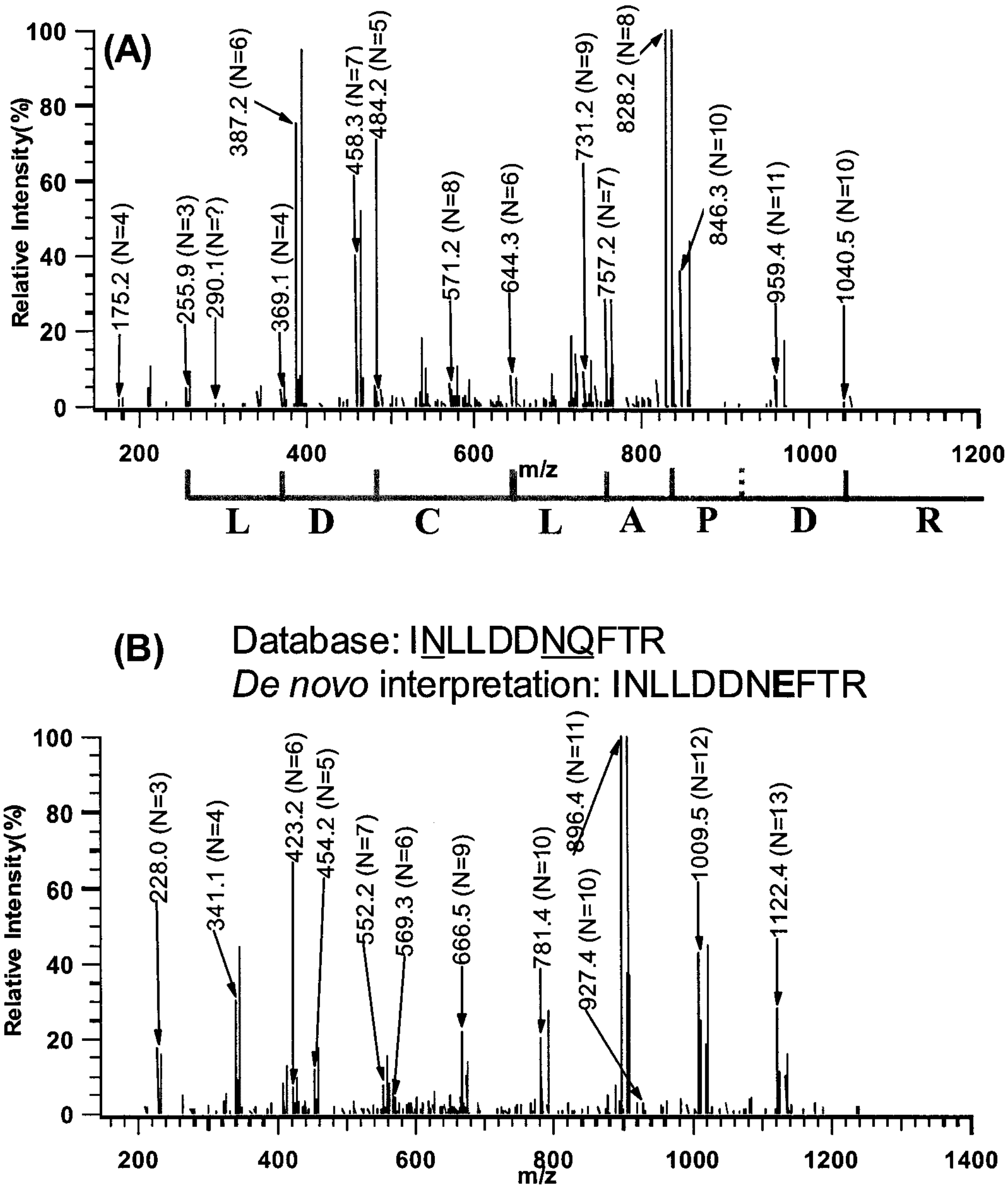


Figure II-2-6. (A) Overlaid MS/MS spectra of a peptide from gel band #11 along with the amino acid sequence reading generated from the spectral interpretation. (B) Overlaid MS/MS spectra of a peptide from gel band #10. Spectral interpretation allows the detection of deamidation of Q in the matched database sequence.

obvious that the peak at m/z 387.2 (N=6) cannot be assigned to the same series, because the mass difference between this peak and m/z 369.1 peak is very small and does not match with any residual amino acid masses. Because the peak at m/z 484.2 (N=5) has a smaller nitrogen number, yet greater mass than the peak at m/z 369.1 (N=6), it must be in the same series as the peak at m/z 369.1 (N=4). The mass and nitrogen number differences between the peak at m/z 369.1 (N=4) and the peak at m/z 484.2 (N=5) correspond to a sequence gap of amino acid D. The next peak in the same series should be the peak at m/z 644.3 (N=6) having a mass increase of 160 Da which matches with the mass of GC, CG, or carboamidomethylated C. However, the nitrogen number difference between these two peaks is only one. Thus, the only possible assignment is the modified C. Using the same logic, the next two peaks (m/z 757.2 and 828.2) in the same series can be assigned as an addition of L (or I) and then A. From the peak at m/z 828.2 (N=8) to the peak at m/z 1040.5 (N=10), there is a large mass gap. The mass difference is 212 Da with a nitrogen number increase of 2, corresponding to PD, DP, VL/I, or L/IV. The mass difference between the molecular ion peak (N=14) and the m/z 1040.5 (N=10) is 175 Da, which can be assigned to the residual mass of R with a corresponding increase of nitrogen number by 4. This is also consistent with trypsin digestion characteristics. Since the sequences of DR and PDR had been determined from the y-series ions, the mass difference between the m/z 828.2 peak and m/z 1040.5 peak should be PD.

The combination of the y- and b-series ion assignments gives the sequence reading of L(or I)DCL(or I)APDR from the MS/MS spectrum shown in FigureII-2-6A. Some amino acids are still missing due to the incomplete series of peaks in high and low mass regions. When the sequence was entered into the BLAST program for sequence matching, the top two matched sequences were AALIDCLAPDR (*E. coli*) with nitrogen number of 14 and KALIACLAPDR (*Haemophilus influenzae)* with nitrogen number of 15. Clearly, only the sequence from *E. coli* matches the nitrogen number, while the sequence from *H. influenzae* has one more nitrogen number than the experimental mass shift. Thus, an unambiguous sequence assignment for this peptide can be made to be AALIDCLAPDR (*E. coli*).

The 2D-MS method is also well suited to characterization of post-translational modifications (PTMs) or other chemical modifications. Difficulties encountered with MS/MS database searching for unambiguous protein identification are often associated with chemical modifications that are not predicted in the proteome database established from genome-translated protein sequences. Possible modifications (such as oxidized M and reduction of disulfide bonds) can be entered into the database search engine. However, as the number of entries for possible modifications increases, the possibility of false positive matching also increases as many combinations of amino acid sequences with modifications can match with a MS/MS spectrum. The nitrogen number information provided in the 2D-MS method can facilitate the identification of chemical modifications.

An example is shown in Figure II-2-6B for the identification of the deamidation of asparagines. The non-enzymatic deamidation of asparagine is one of the most common PTMs. The role of deamidation is not known with certainty. It has been postulated that it may provide a signal of protein degradation thereby regulating intracellular levels. As shown in Figure II-2-6B, there are three possible deamidation sites in the peptide sequence. Spectral interpretation indicates that only Q was deamidated and changed to E.

Table II-2-2. Fragment ion sequences and their corresponding masses and nitrogen numbers from peptide INLLDDNQFTR.

| Sequence | m/z | Nitrogen number |
|---|---|---|
| IN | 228.0 | 3 |
| INL | 341.1 | 4 |
| INLL | 454.2 | 5 |
| INLLD | 569.3 | 6 |
| FTR | 423.2 | 6 |
| EFTR | 522.2 | 7 |
| NEFTR | 666.5 | 9 |
| DNEFTR | 781.4 | 10 |
| DDNEFTR | 896.4 | 11 |
| LDDNEFTR | 1009.5 | 12 |
| LLDDNEFTR | 1122.4 | 13 |

In the overlaid spectra shown in Figure II-2-6B, the nitrogen number of each fragment ions was calculated from the corresponding mass shift of the labeled peptide in the spectra. The total nitrogen number in this peptide was determined to be 17, one nitrogen more than that predicted from the peptide identified by the database searching approach.

Table II-2 summarizes the fragment ion masses and their corresponding nitrogen numbers. The m/z 228.0 fragment ion has 3 nitrogen atoms. Both the mass and nitrogen number indicate that this is IN where N is not deamidated. The fragment ions INL, INLL, INLLD have 4, 5, 6 nitrogen atoms, respectively, again suggesting that N was not deamidated. From the C-terminus, FTR has the same mass (m/z=423.2) and nitrogen number (N=6) as predicted from the database sequence. The mass difference between the fragment ion at m/z 423.2 (N=6) and at m/z 522.2 (N=7) is 129.0 Da, corresponds to E, instead of Q. Also, the nitrogen number difference is only 1, corresponding to E, not Q. For the next fragment ion at m/z 666.5 (N=9), the mass difference is 114.3 Da and the nitrogen number difference is 2. Thus the next amino acid is un-deamidated N, not D. Other fragment ions at m/z 781.4 (N=10), 896.4 (N=11), 1009.5 (N=12) and 1122.4 (N=13) further confirm the above interpretation. In summary, the one nitrogen atom difference between the database sequence and the experimental finding results from the deamidation of Q in the peptide. This example illustrates that interpretation of the MS/MS spectrum from a modified peptide can be done readily with the 2D-MS method by reading sequences and following the nitrogen number changes.

### II. 2. 4. Conclusions

A method of combining m/z information from a product ion spectrum and nitrogen number information of the molecular and fragment ions determined from mass shifts of $^{15}N$ labeled and unlabeled peptides has been demonstrated to be very useful for protein identification. Peptide or protein labeling can be achieved from an *in vivo* whole cell isotope labeling experiment where cells are grown in $^{15}$N-enriched media. Product ion spectra of the unlabeled peptide isolated from a protein digest and its corresponding labeled peptide are individually produced and then superimposed to generate overlaid spectra from which the number of nitrogens in the molecular ion as well as fragment ions can be determined. This two-dimensional MS (i.e., m/z and nitrogen number) method provides several advantages. It can be used to confirm the protein identification results generated from MS/MS database searching for relatively high abundance proteins where searching scores are high. In cases where the MS/MS spectral quality is good, but matches with several peptides with similar scores are obtained, the 2D-MS method can be used to provide additional discriminating power to arrive at a unique peptide identification. It is further shown that the 2D-MS method can identify a peptide using a smaller number of fragment ions than that used by the MS/MS database search alone. In one example, peptide identification was achieved using only three fragment ions. Using the overlaid spectra, fragment ion peaks can be discerned from the background peaks by examining the presence and absence of peak pairs with mass difference corresponding to the nitrogen number difference. It is thus possible to identify peptides using low quality spectra where the number of fragment ions detected is small and the signal-to-background ratios are relatively poor. Finally, it is demonstrated that the 2D-MS method can facilitate the *de novo* peptide sequencing process and the identification of PTMs or other chemical modifications of proteins.

While the interpretation of the 2D-MS data was manually done for this work, automation of the process should be readily achievable with modifications to current MS/MS database search programs. It is envisioned that this 2D-MS method will be particularly useful for proteome expression profiling of organelles or cells that can be grown in $^{15}$N-enriched media. Compared to the current approach of using MS/MS database searching for protein identification, the 2D-MS method should provide more comprehensive proteome coverage with high confidence of protein identification and PTM and other chemical modification assignments.

### II. 2. 5. Cited Literature

(1) Pandev, A.; Mann, M. *Nature* **2000**, *405*, 837-846.

(2) Yates, J. R., III. *J. Mass Spectrom.* **1998**, *33*, 1-19.

(3) Jungblut, P.; Thiede, B. *Mass Spectrom. Rev.* **1997**, *16*, 145-162.

(4) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.

(5) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 14440-14445.

(6) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466-469.

(7) Schnolzer, M.; Jedrzejewski, P.; Lehmann, W. D. *Electrophoresis* **1996**, *17*, 945-953.

(8) Shevchenko, A.; Chernushevich, I.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015-1024.

(9) Uttenweiler-Joseph, S.; Neubauer, G.; Christoforidis, S.; Zerial, M.; Wilm, M. *Proteomics* **2001**, *1*, 668-682.

(10) Mirgorodskaya, O. A.; Kozmin, Y. P.; Titov, M. I.; Komer, R.; Sonksen, C. P.; Roepstorff, P. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1226-1232.

(11) Kosaka, T.; Takazawa, T.; Nakamura, T. *Anal. Chem.* **2000,** *72*, 1179-1185.

(12) Stewart, I. I.; Thomson, T.; Figeys, D. *Rapid Commun. Mass Spectrom.* **2001,** *15*, 2456-2465.

(13) Wang, Y. K.; Ma, Z.; Quinn, D. F.; Fu, E. W. *Anal. Chem.* **2001**, *73*, 3742-3750.

(14) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001,** *73*, 2836-2842.

(15) Yao, X.; Afonso, C.; Fenselau, C. *J. Proteome Res.* **2003**, *2*, 147-152.

(16) Heller, M.; Mattou, H.; Menzel, C.; Yao, X. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 704-718.

(17) Johnson, K. L.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 437-445.

(18) Geng, M.; Ji, J.; Regnier, F. E. *J. Chromatogr.,* A **2000,** *870*, 295-313.

(19) Münchbach, M.; Quadroni, M.; Miotto, G.; James, P. *Anal. Chem.* **2000,** *72*, 4047-4057.

(20) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999,** *17*, 994-999.

(21) Gygi, S. P.; Rist, B.; Griffin, T. J.; Eng, J.; Aebersold, R. *J. Proteome Res.* **2002,** *1*, 47-54.

(22) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002,** *20*, 163-170.

(23) Goshe, M.B.; Smith, R.D. *Curr. Opin. Biotechnol.* **2003**, *37*, 133-145.

(24) Goodlett, D.R.; Yi, E.C. *Trends in Anal. Chem.* **2003**, *22*, 282-290.

(25) Regnier, F.E.; Riggs, L.; Zhang, R.; Xiong, L.; Liu, P.; Chakraborty, A.; Seeley, E.; Sioma, C.; Thompson, R.A. *J. Mass Spectrom.* **2002**, *37*, 133-145.

(26) Liu, P.; Regnier, F. E. *J. Proteome Res.* **2002**, *1*, 443-450.

(27) Shi, Y.; Xiang, R.; Crawford, J. K.; Colangelo, C. M.; Horvath, C.; Wilkins, J. *J. Proteome Res.* **2004**, *3*, 104-111.

(28) Chen, X.; Smith, L. M.; Bradbury, E. M. *Anal. Chem.* **2000,** *72*, 1134-1143.

(29) Veenstra, T. D.; Martinovi, S.; Anderson, G. A.; Pasa-Tolic, L.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 78-82.

(30) Pan, S.; Gu, S.; Bradbury, E. M.; Chen, X. *Anal. Chem.* **2003,** *75*, 1316-1324.

(31) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** *96*, 6591-6596.

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

(33) Condrads, T. P.; Alving, K.; Veenstra, T. D.; Belov, M. E.; Anderson, G. A.; Anderson, D. J.; Lipton, M. S.; Pasa-Tolic, L.; Udseth, H. R.; Chrisler, W. B.; Thrall, B. D.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 2132-2139.

(34) Ong, S.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376-386.

(35) Krijgsveld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H. A.; Heck, A. J. R. *Nat. Biotechnol.* **2003**, *21*, 927-931.

(36) Marshall, A. G.; Senko, M. W.; Li, W.; Li, M.; Dillon, S.; Guan, S.; Logan, T. M. *J. Amer. Chem. Soc.* **1997,** *119,* 433-434.

(37) Jensen, P. K.; Pasa Tolic, L.; Anderson, G. A.; Horner, J. A.; Lipton, M. S.; Bruce, J. E.; Smith, R. D. *Anal. Chem.* **1999,** *71,* 2076-2084.

(38) Zou, J.; Turner, A.N.; Phelps, R.G. *Anal. Chem.* **2004**, *76*, 1445-1452.

(39) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850-858.

## Chapter 3. Chemical Derivatization for *de novo* Peptide Sequencing

Enhanced *de novo* peptide sequencing can be achieved through either an *in vivo* or *in vitro* isotope labeling technique. In some cases, the sequence tag can only be obtained by chemical derivatization *in vitro.* Microwave irradiation has been increasingly used for chemical reactions. Increased reaction rates have been observed. The most common use of microwave in proteomics is the analysis of amino acid compositions. In this chapter, microwave was applied to chemical derivatization of the peptide terminus. Unambiguous identification of peptide sequences can be obtained by comparison of MS/MS spectra of the derivatizated and underivatizated peptides. This is especially useful in identifying low abundant peptides based on the MS/MS spectrum of a single peptide or a peptide with incomplete MS/MS spectra.

### II. 3. 1. Introduction

Chemical derivatization of a peptide terminus often results in decreased signal intensity and is often found to be very time consuming and labor intensive. However, in many cases of real world analysis, it has to be chosen in order to achieve unambiguous identification. In this work, a sample from Korea containing a peptide with m/z value of 2195 Da was required for sequence analysis. For this kind of small peptide, enzymatically introducing a sequence tag such as $^{18}O$ is not practical because there may be no suitable cleavage sites. Additionally, this is a single peptide based analysis and a very good tandem MS/MS spectrum is needed to achieve unambiguous identification even when the peptide has a very specific sequence. Unfortunately, the molecular weight of the peptide is beyond the optimal mass range of tandem MS/MS analysis and a high quality spectrum cannot be expected. Accordingly, C-terminal methylation was chosen to tag the peptide. This can not only locates the acidic group but also removes the ambiguity between Glu, Gln and Lys. The reported procedure for this reaction involves three steps for methylation. The first step is to make the methanolic HCl. Add 800 µl acetyl chloride drop-wise, slowly to 5 ml of dry methanol while stirring and let this mixture stand at room temperature for 5 min. Secondly, make the methyl esters. Add 50 µl of methanolic HCl reagent to 1 nmol of dry protein or peptide. Let stand for 2 hrs at ambient temperature. Thirdly, lyophilize the sample for storage. Concerns associated with this reaction are the yield, toxicity and time consumption [1-3]. Microwave assisted reactions have been used in organic synthesis and analytical chemistry with improved reaction rates and/or yields have been reported [4-9]. In this work, a much simplified methylation reaction of the C-terminus of a peptide was developed and has been applied to a real analysis where the peptide sequence was required.

### II. 3. 2. Experimental Section

**Materials.** All chemicals were from Sigma (St. Louis, MO) and were analytical grade. HPLC grade methanol and acetonitrile were from Fisher (Mississauga, ON). The sample was from Korea University.

**Sample Preparation.** The sample was cleaned using C18 Ziptip (Millipore, Bedford, MA, USA). It was eluted with 50% acetonitrile. The organic solvent was removed by vacuum centrifuge. TFA and methanol were added to the solution so that the final concentration was 0.1%TFA and 50% methanol in water. The sample was then put into a microwave oven with rotation for 2 minutes of irradiation (900W, 2450MHz). A cup of water was put beside the sample to absorb the extra energy. The water cup was not covered in order to avoid the potential for explosion. Caution should be taken during microwave heating [10]. After irradiation, the sample was cooled down and opened in a fumehood. The sample was dried using a vacuum centrifuge and prepared for further MALDI MS and MALDI MS/MS experiments.

**MALDI MS and MALDI MS/MS experiments.** For peptide mass mapping by MALDI time-of-flight (TOF) MS, a two-layer sample/matrix preparation method was used [10] with α-cyano-4-hydroxy-cinnamic acid (HCCA) as matrix. A sample of 0.7 μl of the first-layer matrix solution containing 12 mg/ml of HCCA in 20% methanol/acetone was deposited on the MALDI target and air-dried. The dried sample was re-dissolved in the second layer matrix solution (50% acetonitrile/water saturated with HCCA) and then 0.5 μl of the sample was deposited onto the first layer. The MALDI-TOF mass spectra were obtained on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany).

For MALDI MS/MS analysis of the required peptide, product ion spectra of peptides were obtained in a QSTAR MALDI Qq-TOF mass spectrometer (MDS Sciex, Aurora, Ontario) on the same spot of the peptide sample as that used for MALDI peptide mass mapping.

**Data Interpretation and Database Searching.** Database searching using MS/MS spectra was performed by MASCOT (http://www.matrixscience.com). All the database searching was done against SwissProt using no specification of enzyme type. Methionine oxidization, deamidation of asparagine and glutamine were set as variable modifications.

### II. 3. 3. Results and Discussion

**MALDI MS experiments.** Figure II-3-1. (A) is the MALDI-TOF mass spectrum of the sample. There are only two peptides and one of them with an m/z value of 2195Da is required for sequence analysis. Salt adducts from the separation were observed in the spectrum. C18 Ziptip was used for desalting and the resultant peptide mass mapping is shown in Figure II-3-1 (B). For protein identification based on single peptide with huge m/z value, additional information is needed to confirm the database searching results in order to achieve the un-ambiguity.

**MALDI MS/MS experiment.** The peptide was then methylated. The derivatized and un-derivatized peptides were subjected to the MALDI MS/MS experiments. Figure II-3-2 (A) is the MALDI MS/MS spectrum of the un-derivatized peptide. Figure II-3-2 (B) is the MALDI MS/MS spectrum of the derivatized peptide. Using the MS/MS data of the underivatized peptide for database searching, the peptide was identified as

**SQRFPKADFTEISKIVTDL.** The ions matching the database are labeled in the spectrum. In this sequence, basic residues R and K are close to the N-terminus so the dominant fragment ions in the MS/MS spectrum are the b ions. The corresponding derivatized peptide has a mass shift of 14 Da. However, the resultant observed fragments $b_2$, $b_3$, $b_4$, $b_6$, $b_7$, $b_8$ and $b_{11}$ do not have a mass shift, while $b_{18}$ has a mass shift of 14 Da. These characteristics are consistent with the identified amino acid sequence. The mass shift of $b_{18}$ indicates that the C-terminal D rather than L was derivatized because D has an acidic residue and is easier to be methylated than the C-terminal L.

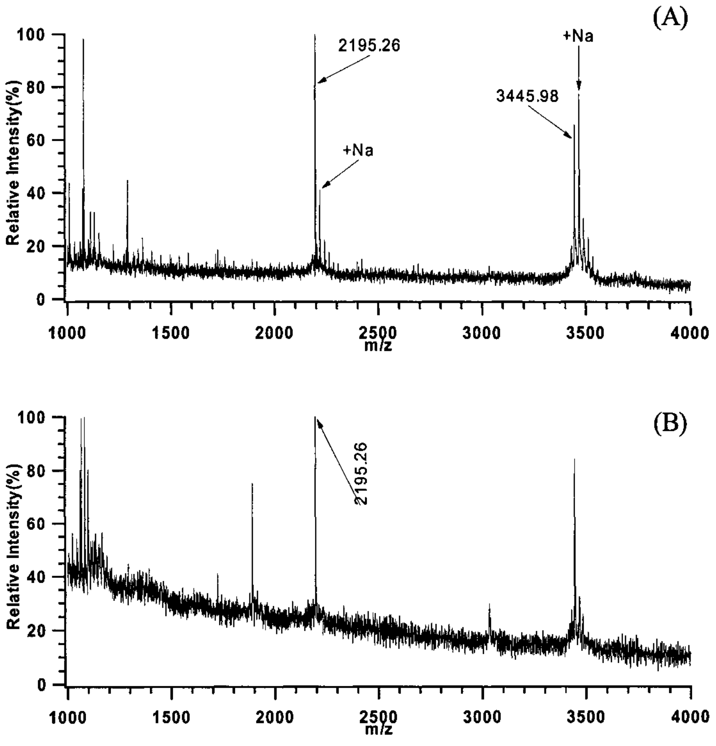


Figure II-3-1. MALDI-TOF mass mapping of (A) the original sample and (B) Ziptip cleaned sample.

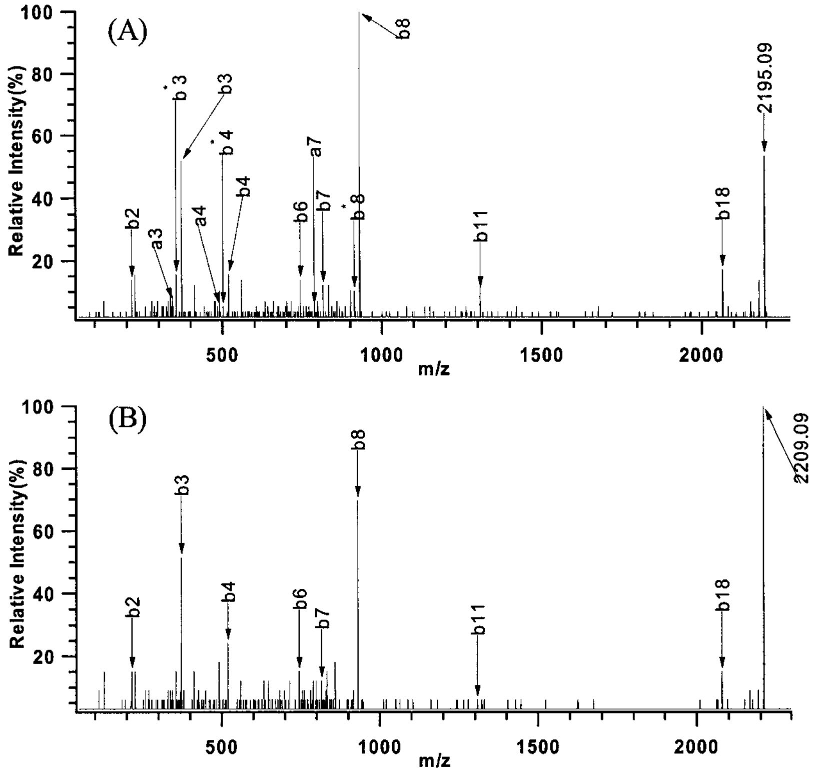


Figure II-3-2. MALDI MS/MS spectra of (A) un-derivatized peptide. (B) derivatized peptide.

## II. 3. 4. Conclusions

Chemical derivatization is still an indispensable tool to enhance protein or peptide identification. It provides additional information that helps eliminate false positive or confirm true negative identification. The application presented herein unambiguously identifies a peptide sequence based on a single peptide MS/MS spectrum by using C-terminal methylation. Microwave assisted reaction was found to provide much improved reaction rates and is very convenient for application. It only needs a few minutes while the conventional method needs a few hours.

### II. 3. 5. Cited literature

(1) Falick, A. M.; Matlby, D. A. *Anal. Biochem.* **1989,** *182,* 165-169.

(2) Goodlett, D. R. et al. *Rapid Commun. Mass Spectrom.* **2001,** *15,* 1214-1221.

(3) Shevchenko, A. et al. *Rapid Commun. Mass Spectrom.* **1997,** *11,* 1015-1024.

(4) Kidwai, M. *Pure Appl. Chem.* **2001,** *73,* 147-151.

(5) Varma, R. S. **2001,** *73,* 193-198.

(6) Lew, A.; Krutzik, P. O.; Hart, M. E.; Chamberlin, A. R. *J. Combinational Chem.* **2001,** *4*, 95-105.

(7) Luque-Garcia, J. L.; Luque de Castro, M. D. *Anal. Chem.* **2001,** *73,* 5903-5908.

(8) Holler, U.; Wolter, D.; Hofmann, P.; Spitzer, V. *J. Agric. Food Chem.* **2003,** *51,* 1539-1542.

(9) Miedel, M. C.; Hulmes, J. D.; Pan, Y. C. E. *J. Biochem. Biophys. Methods* **1989,** *18,* 37-52.

(10) Kingston, H.M.; Haswell, S.J. (Editors), Microwave-Enhanced Chemistry: Fundamentals, Sample Preparation, and Applications. ACS: Washington, D.C., 1997.

# Part III. Intact Protein *de novo* Sequencing for Protein Identification and Detection of Posttranslational Modifications

# Chapter 1. Methodology Development of Intact Protein *de novo* Sequencing by Microwave-Assisted Ladder Generation[1]

A new technique for sequencing proteins and determining modifications with high speed, sensitivity and specificity was developed. Intact protein fragmentation by microwave assisted acid hydrolysis was investigated. It was found that proteins mixed with strong acid could be hydrolyzed with microwave irradiation within a minute to form two series of polypeptide ladders with each containing either the N- or C-terminal amino acids of the protein. Mass spectrometric analysis of the hydrolysate produced a simple mass spectrum consisting of peaks from these polypeptide ladders, allowing direct reading of the amino acid sequence and modifications of the protein. As examples, we applied this technique to the determination of protein phosphorylation sites, covalently bonded heme group, methylation and acetylation at N-terminus of a protein and for determining the oxidization sites and disulfide bonds in a protein. Proteins with different pI values have been examed using the presented method. This technique can potentially be automated for large-scale protein annotation.

[1] A version of this chapter has been accepted for publication as:
Hongying Zhong, Ying Zhang, Zhihui Wen and Liang Li, "Protein sequencing by mass analysis of polypeptide-ladders after controlled protein hydrolysis", *Nature Biotechnology,* **2004,** October, in press.
Dr. Ying Zhang and Mr. Zhihui Wen purified the *E. coli* K12 proteins using strong cation ion exchange and C 8 reversed phase chromatography.

### III. 1. 1. Introduction

Characterization of protein modification is essential and urgently needed for the study of protein function using functional genomic and proteomic approaches [1-3]. However, current techniques are not efficient for determination of protein modifications. Protein modifications, such as posttranslational modifications (PTMs), can only be characterized by examining the protein directly. At present, reading the amino acid sequence of a protein is performed by the Edman degradation method and increasingly by tandem mass spectrometry (MS/MS) [4-7]. Other MS-based techniques, such as ladder sequencing [8-15], in-source fragmentation [16, 17], and chemical derivatization [18, 19], have been reported for sequencing peptides with varying degrees of success. Compared to the Edman method, the MS approach has the advantages of high sensitivity and generating structural information on modifications. However, to map the sequence of an entire protein for examining all possible modifications, one often needs to produce, detect, and sequence many short, partially overlapping peptides, which can be very difficult. The technique developed here is specifically cleaved at the amide bonds of intact proteins and produces sequence specific ladders containing N- and C-terminal amino acids. The resultant MALDI-TOF spectra provide a clear sequence readout.

### III. 1. 2. Experimental Section

**Materials.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany.

**Generation of sequence ladders.** For the MAP (Mass Analysis of Polypeptide Ladders) sequencing experiment, a microliter of protein sample was mixed with an equal volume of 6M HCl in a 0.6-mL polypropylene vial. The vial was capped and then placed inside a household microwave oven with 900W output at 2450 MHz. A water container containing 100 ml of water was placed beside the sample vial so that the extra microwave energy was absorbed mainly by the water. The microwave oven was turned on for, typically, 60 sec. After the microwave irradiation for less than 2 minutes, the bottom of the vial was found to be slightly warm. The temperature of the solution inside the vial was unknown; but no visible boiling or depletion of the solution was noted. After microwave irradiation, the sample vial was removed from the microwave oven and the solution in the vial was dried under vacuum centrifugation.

**MALDI-TOF detection of sequence ladders.** The dried sample was re-dissolved in a matrix solution of α-cyano-4-hydroxycinnamic acid (HCCA). The mixture was then deposited on a sample target using a two-layer sample preparation method[34] for matrix-assisted laser desorption ionization (MALDI) analysis. MALDI MS experiments were carried out on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany) using a linear mode of operation. Ionization was performed with a 337-nm pulsed nitrogen laser.

### III. 1. 3. Results and Discussion

**Principle of the developed technique.** Figure III-1-1 shows the schematic for the technique. It illustrates our protein sequencing technique that is based on the mass analysis of polypeptide ladders (MAP) of a protein after it has been subjected to brief hydrolysis in HCl with the assistance of microwave irradiation (MI). This technique allows direct sequencing of peptides and even proteins with high sensitivity and speed. This technique has been validated using a variety of proteins with different pI value and proteins with acid labile peptide bonds such as the D-P bond. These results are summarized in Table III-1-1. Cytochrome C (Horse heart) was used to demonstrate the time course of the sequence development. Figure III-1-2 shows the MALDI spectra of 1 pmol cytochrome c hydrolyzed under different conditions. Without MI, a small number of polypeptides in the low mass region were observed after the protein and HCl mixture was kept at room temperature for 5 min (FigureIII-1-2A). These polypeptides were found to be exclusively N- and C-terminal peptides. If the mixture was kept for longer periods, e.g., 15 hrs, more intense terminal peptide peaks were observed (FigureIII-1-2B). The speed of hydrolysis could be dramatically accelerated using MI, possibly due to microwave-induced rapid heating of the solution and conformational/structural changes of the proteins along the peptide bonds. When the microwave radiation was applied for only a short duration (10 s), a number of peaks were detected at m/z below the molecular ion region and they were distributed across a broad mass range (Fig. III-1-2C). As the irradiation time was increased, the intensities of these terminal peptide peaks also increased (Fig. III-1-2D-F).

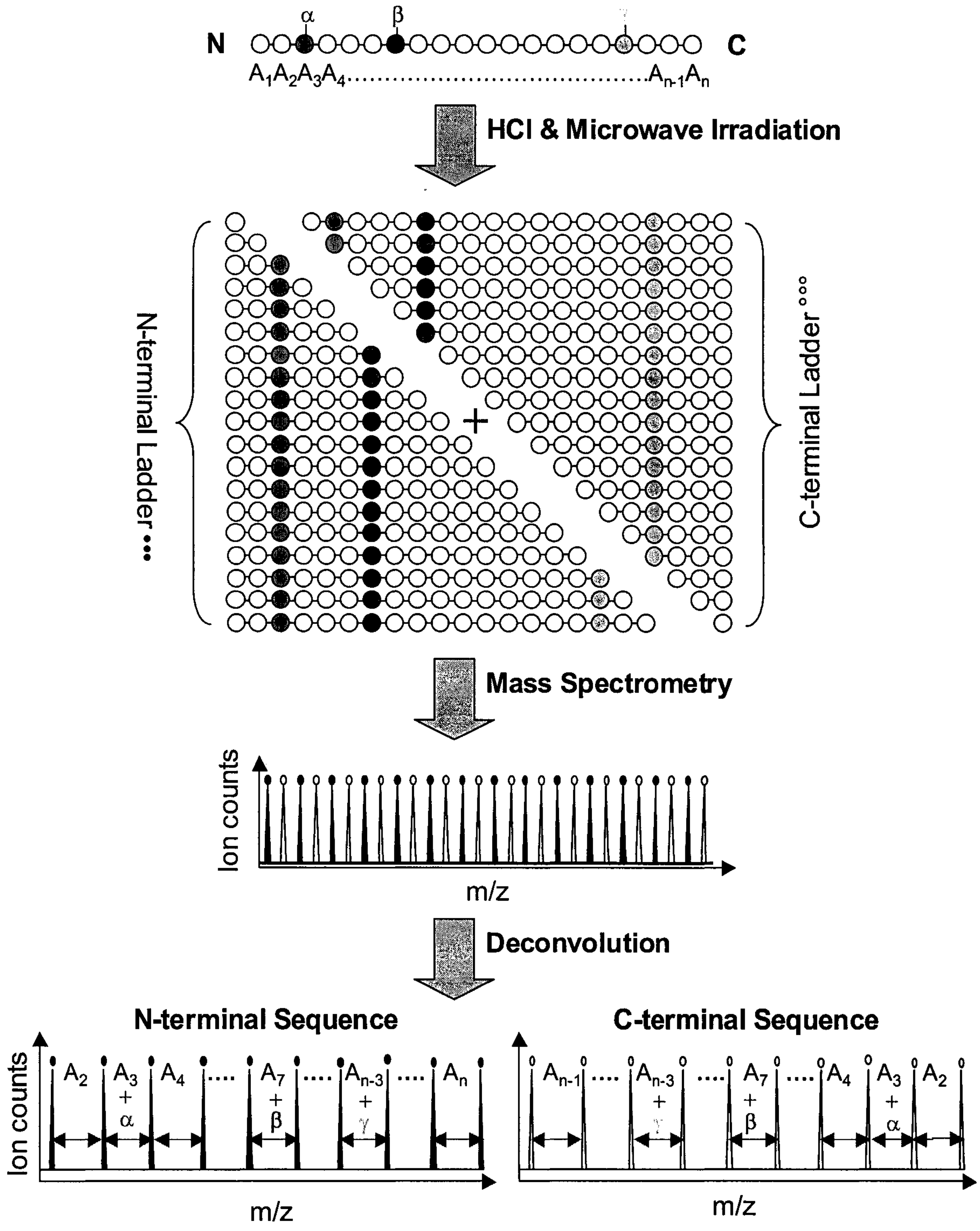


Figure III-1-1. Schematic of the mass analysis of polypeptide-ladders (MAP) sequencing technique. The mass spectrum produced consisted of peaks exclusively from the N- and C-terminal polypeptides of the protein. Amino acid sequences and any modifications were read from the mass differences of adjacent peaks within the same series of the polypeptide ladder.

Mass spectra displaying peaks over a broad mass range with good signal-to-noise ratios could be generated after MI for 30 to 90 s. As the irradiation time was increased further, many internal peptides started to appear and the high mass peptide peaks, along with the molecular ion peak, disappeared (Fig. III-1-2G-I).

Table III-1-1. Standard proteins that have been used for MAPs examination.

| proteins | ID (Swiss-Prot) | Mr(calc.) | pI | PTMs |
|---|---|---|---|---|
| Ubiquitin-human | P02248 | 8564.84Da | 6.56 | no |
| Cytochrome-horse | P00004 | 11701.55Da | 9.59 | Acetylation and heme group |
| Lysozyme-chicken | P00698 | 14313.14Da | 9.32 | Signal peptide cleavage; 4 disulfide bonds |
| Apomyoglobin-horse | P02188 | 16951.48Da | 7.36 | No heme group |
| α-casein-bovine | S1: P02662 | 24529.92Da | 4.91 | Signal peptide cleavage; 9 phosphorylation sites |
| Hypothetic protein-*E. coli* K12 | P56614 | 5884.22 | 9.87 | Cleavage of M start codon |
| 50S ribosomal protein- *E. coli* K12 | P02436 | 6372.58 | 10.25 | Cleavage of M start codon; mpnpmethylation of A at 2 |

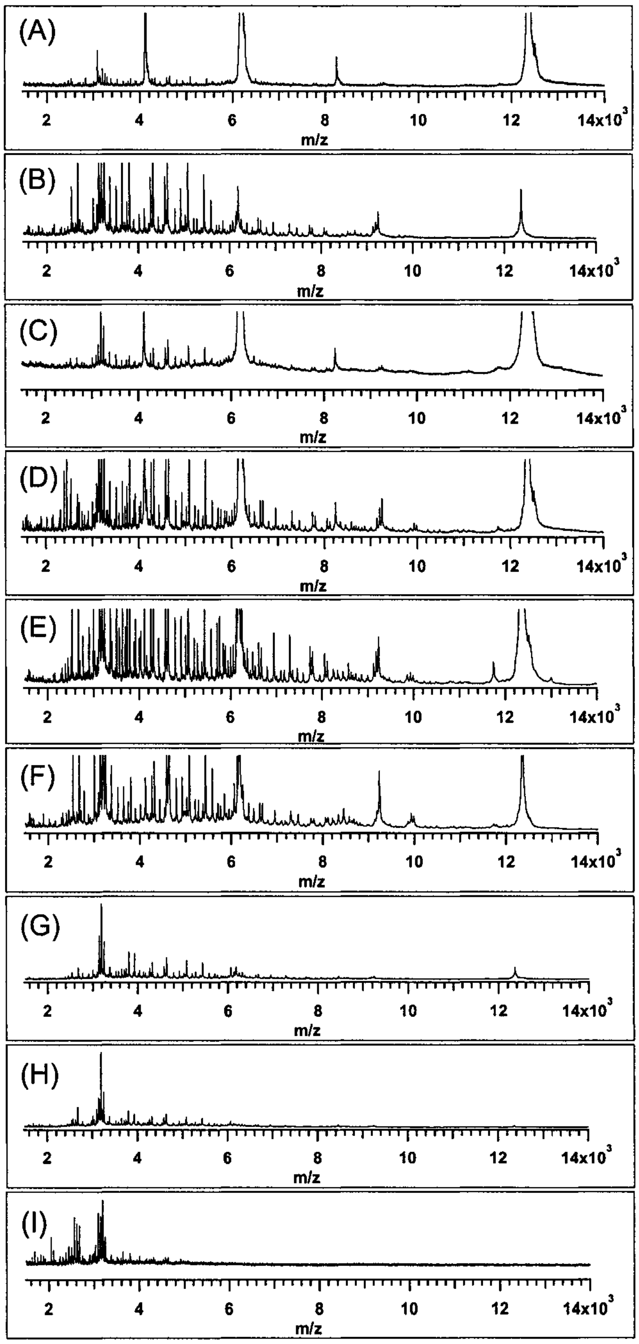


Figure III-1-2. MALDI spectra of 1 pmol of cytochrome c after (A) mixing with 6M HCl for 5 min without MI, (B) mixing with 6M HCl for 15 hr without MI, and mixing with 6M HCl applying MI for (C) 10 sec, (D) 30 sec, (E) 1 min, (F) 2 min, (G) 3 min, (H) 4 min, and (I) 5 min.

Similar time dependencies to those shown in Figure III-1-2 were observed for other proteins, such as human ubiquitin (MW 8,565 Da) and horse apomyoglobin (MW 16,951 Da) and other proteins. It appears therefore that when the protein hydrolysis process is properly controlled, e.g., using 6M HCl with MI for 1 min, mass spectrometric analysis of the resulting hydrolysate generates a spectrum consisting of peaks from exclusively terminal peptides with no internal peptides, which makes the reading of protein sequence very easy. This somewhat surprising observation may be explained by considering the reaction rates of peptide bond breakage and the concentration differences among the intact protein, terminal peptides produced from the first hydrolysis process, and internal peptides produced from the follow-up hydrolyses of the terminal peptides. As Fig. III-1-2C shows, with 10 sec MI, a number of terminal peptides have already been generated; but their signal intensities are much lower than that of the intact protein, indicating that the relative abundances of these terminal peptides to the intact protein are very low. As the irradiation increases from 10 sec to 60 sec, more intense terminal peptide signals are generated, but their relative intensities to that of the intact protein are still low. Furthermore, it appears that all peptide bonds of the intact protein are cleaved once and only small variations in relative intensity of adjacent terminal peptides are found. This suggests that the reaction constant for peptide bond breakage is similar for all peptide bonds. Therefore, in the first hydrolysis process, an intact protein consisting of many peptide bonds is most likely to undergo parallel reactions by breaking any one of the peptide bonds once, to form, collectively, many terminal peptides. If only the first hydrolysis of the intact protein took place, the amount of an individual terminal peptide would equal the amount of the intact protein hydrolyzed divided by the number of peptide bonds broken in the protein and, at the early stage of hydrolysis where the intact protein concentration is high, the terminal peptide concentration would be expected to be very low. In reality, the terminal peptides formed early on in the hydrolysis process, e.g., those terminal peptides giving arise the peaks shown in Fig. III-1-2C after 10 sec MI, could further hydrolyze to form shortened peptides including internal peptides (i.e., follow-up hydrolysis). However, as indicated earlier and shown in Fig. III-1-2C and 2D, the amount or concentration of an individual terminal peptide generated early on in the hydrolysis process is relatively low compared to that of the intact protein. Since the hydrolysis reagents (i.e., water and acid) are in excess, the acid hydrolysis rate of the intact protein or a terminal peptide is pseudo-first order in protein or peptide concentration, i.e., $v = k[C]$, where v is the reaction rate, k is the reaction constant, and [C] is protein or peptide concentration. Thus, at the early stage of hydrolysis, a terminal peptide is generated very quickly from the intact protein and only a small portion of it is further hydrolyzed. The net result of this dynamic equilibrium is that an excess amount of the terminal peptide is accumulated until its concentration is comparable to that of the intact protein. Beyond this time the terminal peptide will not be replenished from the hydrolysis of the intact protein as quickly and the concentrations of the internal peptides generated from the follow-up hydrolysis of the terminal peptide will build up. As shown in Figure III-1-3G, after 3 min MI, terminal peptides generate similar peak intensities as that of cytochrome c and many internal peptides are observed. Once all intact protein molecules are consumed and shortly thereafter, the internal peptide peaks become the dominating feature of the spectra (Figure III-1-3H, I). Eventually all peptides will be hydrolyzed to form amino acids.

From the above discussion, we can conclude that so long as the hydrolysis process is controlled to the extent that there is a small amount of intact protein remaining in the hydrolysate, the hydrolysis rate of a terminal peptide to form internal peptides is always much smaller than that of the intact protein. As a result, the terminal peptides will be the dominant components in the final solution along with the intact protein. Due to a limited detection dynamic range of the mass spectrometer and the ion signal suppression effect in MALDI, direct MALDI analysis of the hydrolysate only allows the detection of the intact protein and the terminal peptides. Signals from the low abundance internal peptides are suppressed and not seen in the MALDI spectrum. As a result, peptide peaks detected in the spectrum are exclusively from the terminal peptides. We note that, to account more accurately for the acid hydrolysis process, we are currently in the process of determining hydrolysis reaction constants of proteins under various conditions with the goal of developing a quantitative description of the hydrolysis dynamics involved in the MAP sequencing technique.

Using the properly controlled hydrolysis conditions as shown in Fig. III-1-2, the MAP sequencing technique is found to be generally applicable to a wide range of proteins, including proteins containing internal disulfide bonds and proteins containing acid labile bonds such as D-P. Since the success of the MAP sequencing technique depends on near uniform hydrolysis of all peptide bonds in a protein, it would appear to be difficult to apply this technique for sequencing proteins consisting of acid labile bonds such as D-P. It turns out that this technique can be readily applied for sequencing these proteins. This

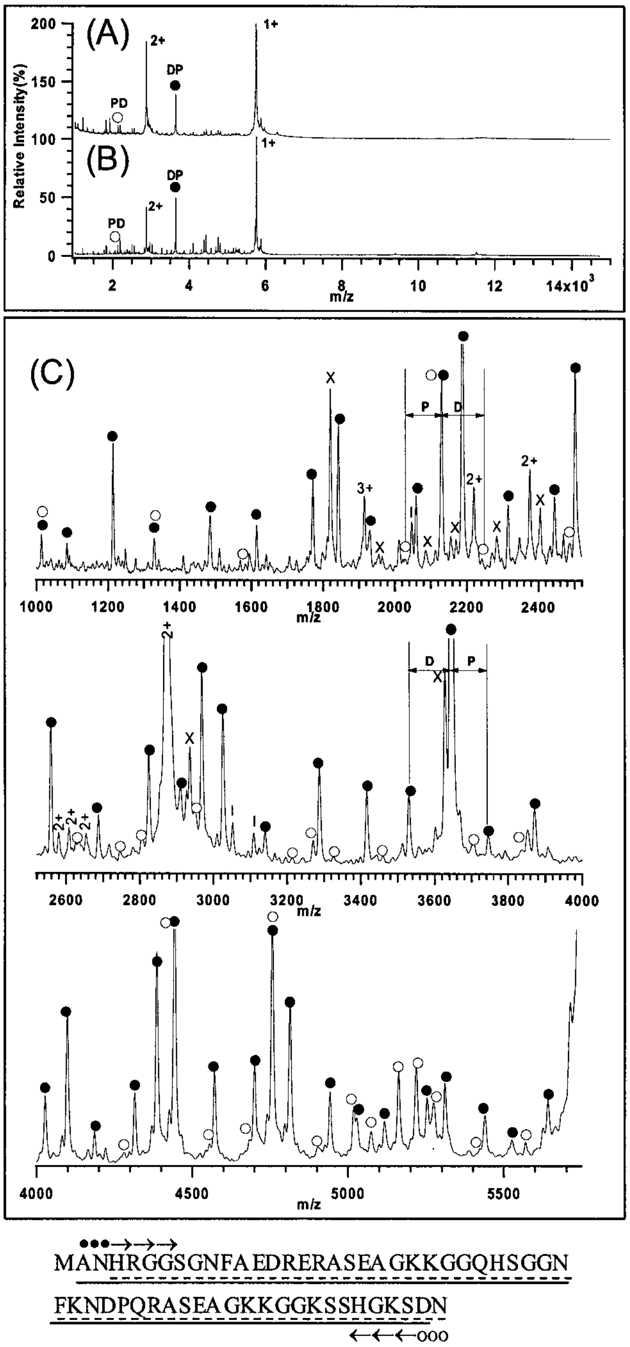
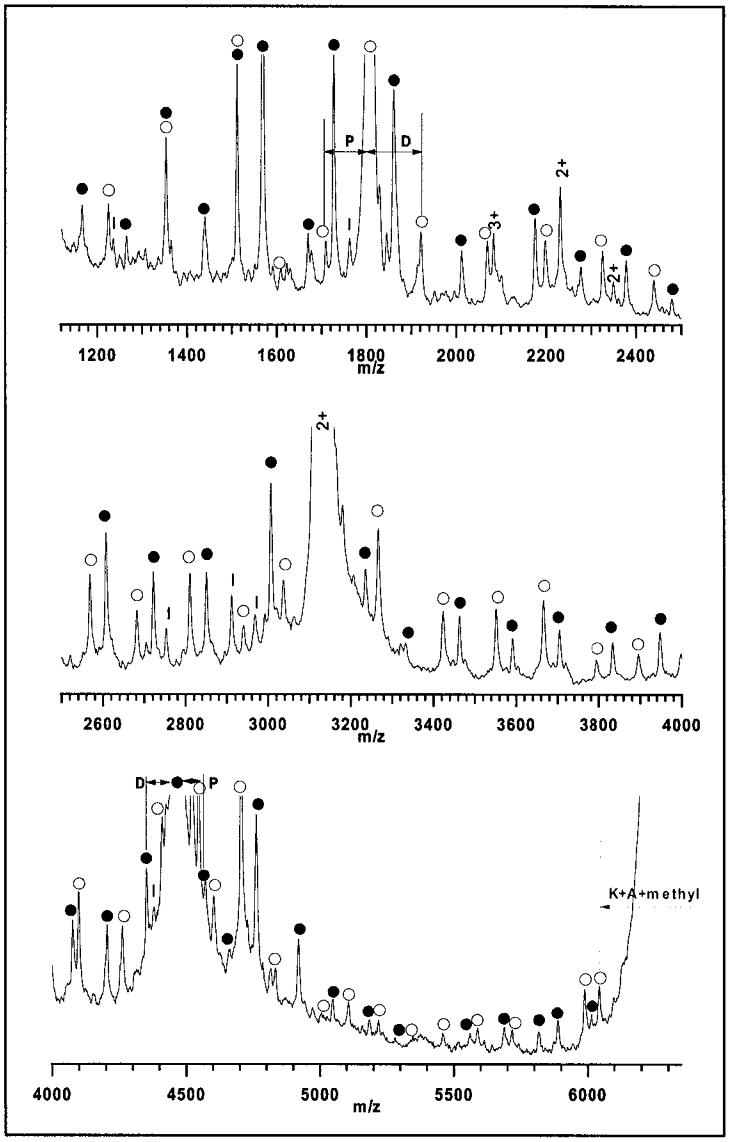


Figure III-1-3. MALDI spectra of a HPLC-fractionated sample from *E. coli* K12 extract after mixing with 6M HCl and applying microwave irradiation for (A) 30 s and (B) 1 min. (C) Expanded MALDI spectra of (B). The protein was identified as YMDF-ECOLI (P56614) and determined to have the Met start codon cleaved. Bottom panel: sequences determined from the N-terminal ladder (•) with a solid underline and from the C-terminal ladder (○) with a dashed underline. Peaks labeled with "I" are possibly from the internal fragments of the two terminal peptides generated from the D-P bond breakage. Peaks labeled with "X" are from other impurities.

is shown in Figure III-1-3 where a protein containing a D-P bond isolated from an *E. coli* extract was subjected to MAP sequencing. Panels A and B of Figure III-1-3 show the MALDI spectra obtained from the sample after acid hydrolysis with MI for 30 and 60 sec, respectively. A number of polypeptide peaks are observed, including one intense peak arising from the N-terminal peptide generated from the breakage of the D-P bond of the protein identified as YMDF-ECOLI (P56614). The corresponding C-terminal peptide is also observed. With the increase in irradiation time from 30 sec to 60 sec, the signal-to-noise ratios of the polypeptide peaks are improved. However, the relative intensities calculated from the peak areas between other polypeptides and the two peptides from the D-P bond breakage remain similar. The expanded spectrum of Figure III-1-3B is shown in Figure III-1-3C where the entire sequence of the protein can be read from the N- and C-terminal ladders. Three peaks labeled with "I" are the possible internal fragments from the follow-up hydrolysis of the two terminal peptides generated from the D-P bond breakage. Several other peaks labeled as "X" in Figure III-1-3C are from the impurities present in the sample. These impurities have much lower concentrations than that of the main protein in the original sample. Thus they did not hydrolyze extensively. More importantly, the presence of the impurity peaks did not interfere with the assignment of the terminal peptide ladders of the main protein. Another example of MAP sequencing of proteins containing D-P bonds [i.e., RL33-ECOLI (P02436) isolated from *E. coli*] is given Figure III-1-4. It is clear that the presence of acid labile bonds, such as D-P, in a protein does not prevent the generation of a complete set of terminal ladders.



●●●→→→
AKGIREKIKLVSSAGTGHFYTTTKNKRTKPEKLELKK

FDPVVRQHVIYKEAKIK
←←←○○○

Figure III-1-4. Expanded MALDI spectra of an HPLC-fractionated sample from an *E. coli* K12 extract after mixing with 6M HCl and applying microwave irradiation for 1 min. The protein was identified as RL33-ECOLI (P02436) and determined to have M start codon cleaved and monomethylation of A at residue 2. Bottom panel: sequences determined from the N-terminal ladder (•) with a solid underline and from the C-terminal ladder (○) with a dashed underline. Peaks labeled with "I" are possibly from the internal fragments of the two terminal peptides generated from the D-P bond breakage of the protein.

**Detection Sensitivity.** The detection sensitivity of the MAP technique was also examined. For small proteins with molecular masses of up to about 14,000 Da, the sample required for the experiment is generally less than 1 pmol. Lower amounts of sample can be used, but as the amount decreases, the polypeptide peaks in the high mass region start to decrease. Human ubiquitin was used to demonstrate the detection limit as shown in Fig. III-1-5. This level of detection is consistent with the current practice of using a microliter sample deposition method where small proteins and large peptides can be detected at 1 fmol using MALDI-TOF [20]. With nanoliter sample deposition, MALDI sensitivity can be improved by 100-fold or more [21-23]. Thus, future work on miniaturizing the hydrolysis process followed by nanoliter sample deposition should significantly improve the overall sensitivity of the MAP technique.

**Location of Modifications.** With the generation of polypeptide ladders, information on sequence and PTMs can be deduced from the mass spectra. As Fig. III-1-1 illustrates, the mass difference between adjacent peaks of the same series of polypeptides corresponds to the mass of an amino acid residue, which forms the basis for its identification as well as its modification, if any. Acid labile modifications may be destroyed during the hydrolysis; but they are rare [24]. Among the twenty common amino acids, Leu and Ile have the same mass and, therefore, cannot be distinguished by this method. Gln and Lys have similar masses and cannot be readily distinguished within the molecular mass measurement accuracy of time-of-flight MS; but it should be entirely possible with MALDI FT-ICR MS [25]. However, distinguishing these pairs of amino acids is only required for *de novo* sequencing of an unknown protein from a species with no or little

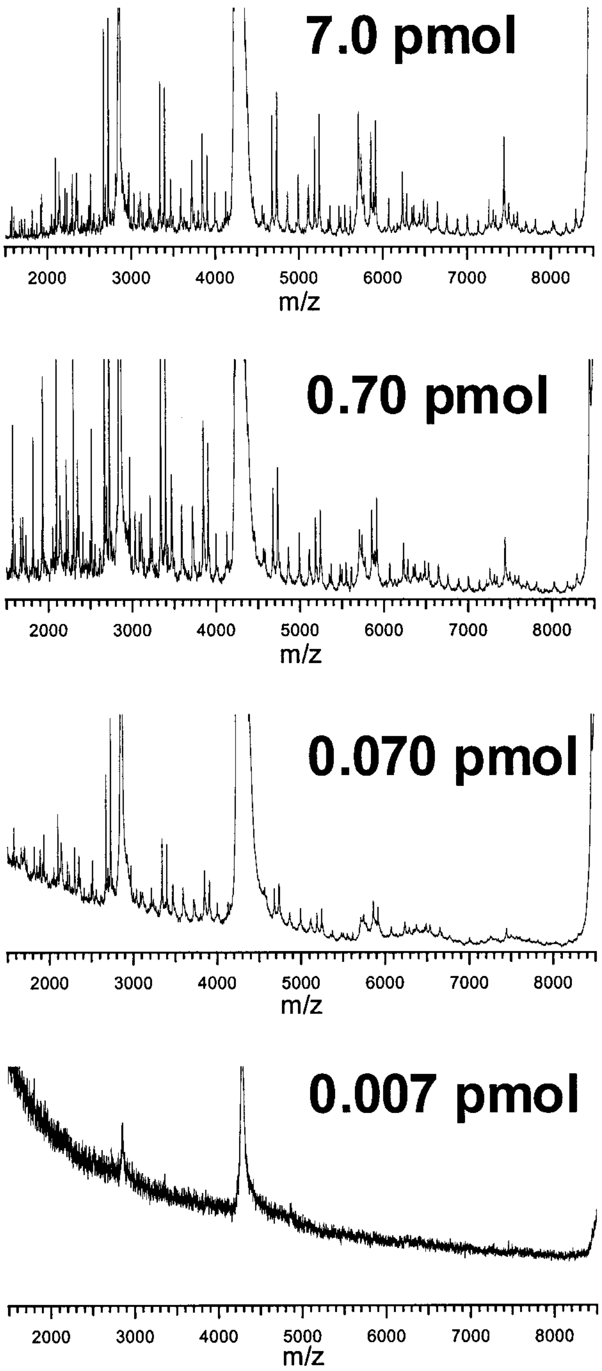
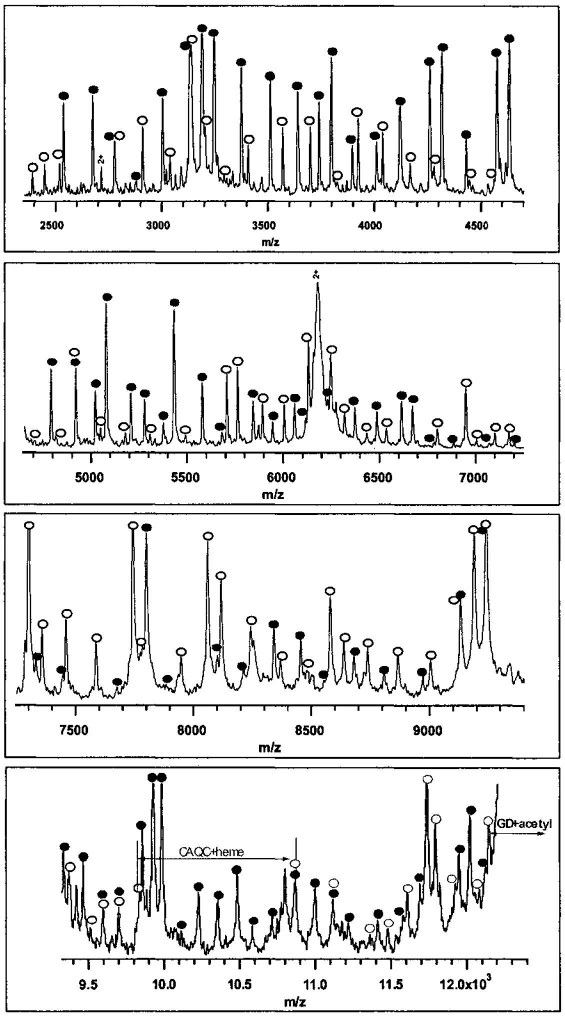


Figure III-1-5. MAP sequencing of human ubiquitin using different amounts of starting materials ranging from 7 fmol to 7 pmol.



● ● ●→→→

**GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKN**

**KGITWKEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE**

←←← o o o

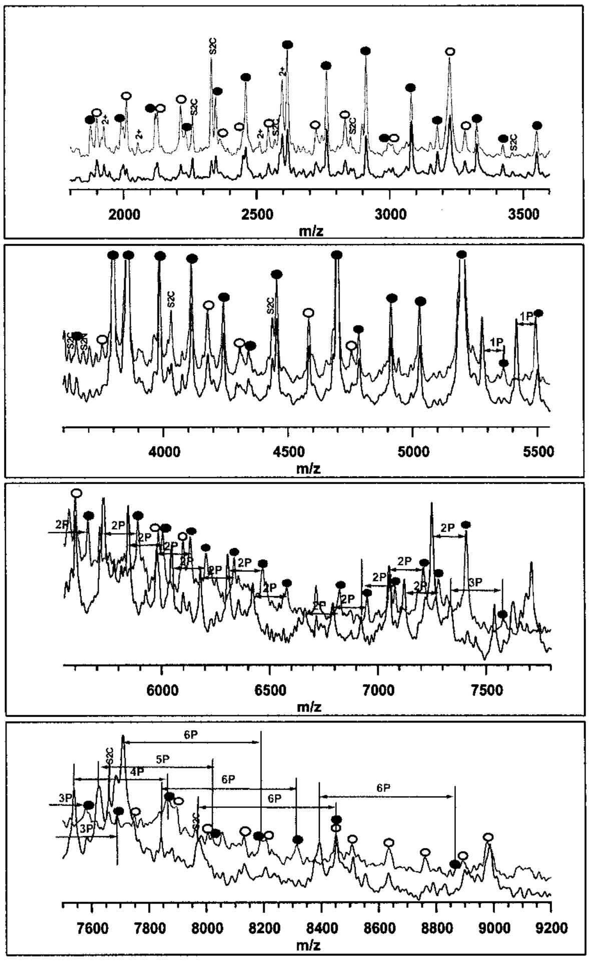
Figure III-1-6. Expanded MALDI spectra from Fig. III-1-2 (E) except that the high mass region (9,200-12,400 Da) was generated from the same sample, but using another setting with a low-mass cutoff that favored the detection of high mass ions. Bottom panel: sequences determined from the N-terminal ladder (•) with a solid underline and from the C-terminal ladder (○) with a dashed underline.

genome or proteome database. Fortunately, with the rapid expansion of genome databases as well as the possibility of performing cross-species database searching, *de novo* sequencing is becoming an increasingly rare practice. Thus the MAP sequencing technique should be very powerful for re-examining protein sequences translated from their genome and determining their modifications.

As an example, the expanded MALDI spectra of cytochrome c after 60-sec MI in HCl are shown in Figure III-1-6. A long stretch of sequence can be deduced from a ladder and the sequences read from two ladders have an overlap. The combination of the two allows the determination of the entire sequence of cytochrome c with information on possible modifications. For example, the mass difference between the molecular ion peak and the peak at m/z 12147 is 214 Da, which matches with the mass of acetylated glycine linked to D (i.e., Ac-GD). Another modification found in cytochrome c is the heme group covalently bonded to cysteines 14 and 17. In this case, polypeptide ladders from 1-17 to 1-104 all have a positive mass shift corresponding to the mass of heme. Polypeptides 1-13, 1-12, etc., do not have the mass shift. Furthermore, because of the heme attachment to cysteines 14 and 17 that apparently hinders the hydrolysis of internal peptide bonds, polypeptide peaks generated from the hydrolysis of peptide bonds between these two residues were not detected. Taken together, the evidence confirms the heme modification and its location on cysteines 14 and 17 in cytochrome C.

One of the anticipated major applications of the MAP sequencing technique is the determination the phosphorylation sites of a phosphorylated protein. An example is illustrated in Figure III-1-7. In this case, an α-S1-casein sample and a de-phosphorylated α-S1-casein sample were subjected to MAP sequencing. Because both samples contain about 15% of α-S2-casein, there were a number of low intensity peaks (some are labeled in the figure) in addition to the more intense polypeptide peaks generated from α-S1-casein. Despite the presence of these low intensity peaks, the major peaks could be readily assigned to the sequence of α-S1-casein. In addition, mass shifts corresponding to one or more phosphate groups in 80 Da increments could be identified, which in turn provided the information for identification of the phosphorylation sites. Determining the modification site is easy and unambiguous, because all polypeptides in a ladder containing one modification will have the 80-Da mass shift until they encounter another modification which will further increase the mass by another 80 Da. In this case, the polypeptide ladders from N-terminal 16-31 to 16-84 and C-terminal 137-214 to 198-214 were detected, resulting in the survey of 114 amino acids. In addition, mass analysis of terminal peptides 16-31 and 198-214 indicated that there were no modifications on these peptides. Thus a total of 145 amino acids were examined. In the protein sequence covered by these amino acids, there are six known phosphorylation sites. The MAP sequencing technique as shown in Figure III-1-7 detected all six sites in one experiment. The internal sequence from residue 85 to 136 was not covered in this experiment due to the suppression of high mass peaks (>9,000 Da) in MALDI. One obvious approach to map this portion of the sequence would use chemical or limited enzyme digestion of the protein to generate large peptides, followed by HPLC separation and MAP sequencing. However, we believe that technical advances in the near future such as using more powerful mass spectrometers [25-27] and better sample preparation [28, 29] should allow direct sequencing of a protein of this size (see below).

While the MAP sequencing technique requires a relatively pure protein for unambiguous amino acid sequencing and determination of chemical modifications, the above example also illustrates that sequence and modification information can still be obtained even when the sample contains about 15% of other components. The applicability of the technique for sequencing a protein containing impurities clearly depends on the type of information to be generated and the nature of the impurities, i.e., whether the impurities will interfere with the assignment of the sequence ladders from the protein of interest. In the example given in Figure III-1-3, the impurity peaks shown did not interfere with the reading of the terminal ladders, allowing complete sequencing of the protein.



•••→→→

MKLLILTCLVAVALA**RPKHPIKHQGLPQE**

**VLNENLLRFFVAPFPEVFGKEKVNELSKD**

**IGSESTEDQAMEDIKQMEAESISSSEEIVPN**

SVEQKHIQKEDVPSERYLGYLEQLLRLKKY

KVPQLEIVPNSAEERL**HSMKEGIHAQQKEP**

**MIGVNQELAYFYPELFRQFYQLDAYPSGA**

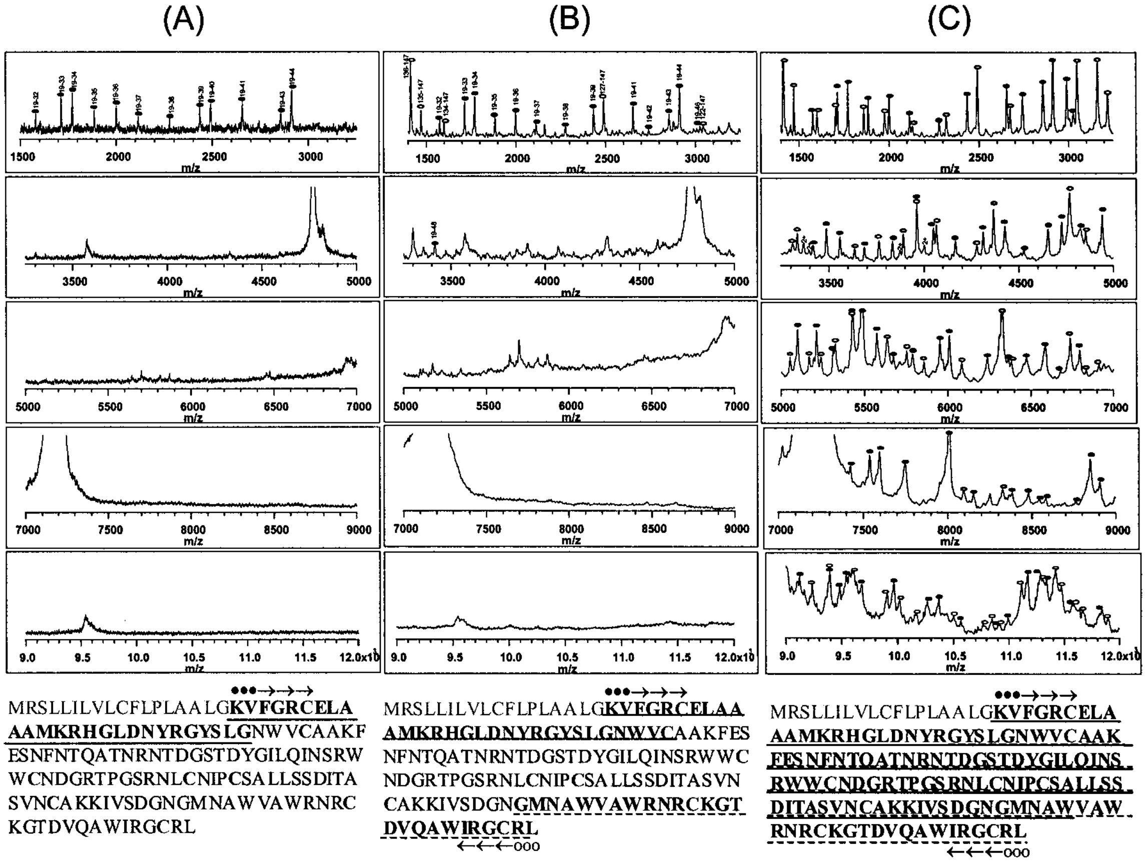
**WYYVPLGTQYTDAPSFSDIPNPIGSENSEK**

**TTMPLW**

←←←ooo

Figure III-1-7. MALDI spectrum (top) of a sample containing 85% α-S1-casein and MALDI spectrum (bottom) of a sample containing 85% de-phosphorylated α-S1-casein. Peaks labeled with "•" and "o" are those from α-S1-casein terminal polypeptides.

The disulfide bond is another important PTM that provides intra- and intermolecular crosslinking within proteins and determines the protein stability. Figure III-1-8 shows the MALDI-spectra when lysozyme was reduced to different extents and then subjected to MAP sequencing. Without reduction, only a few peaks from the N-terminal peptides could be observed (Fig. III-1-8A), indicating that cysteines near both terminals were involved in the formation of disulfide bonds. While the genome sequence predicts protein sequence of lysozyme as shown in Fig. III-1-8, the polypeptide peaks shown in Fig. III-1-8A all correspond to the starting amino acid residue 19. Thus the signal peptide 1-18 was cleaved in the mature form of this protein. Cleavage of a signal peptide from an expressed protein is a common PTM. The MAP sequencing technique provides a rapid means of identifying this type of PTM. As Figure III-1-8B illustrates, after 2 hr reduction in 45 mM DTT at 37 °C, followed by MAP, N-terminal ladders as well as a few C-terminal ladders were detected. Several other peaks were observed in the mass range from about 3,000 Da to 6,000 Da, but the signals were still low indicating that the protein was not completely unfolded. After 15 hr reduction, additional and more intense polypeptide ladders were observed (Figure III-1-8C) with 100% sequence coverage, indicating the opening of the protein that makes it more susceptible to hydrolysis. Mass information on the protein itself and polypeptides indicated that there were four disulfide bonds; but the exact locations of the disulfide bonds could not be determined. Nevertheless, this work illustrates that disulfide bond reduction can be monitored by MAP sequencing. Future development of MAP sequencing, in combination with H/D exchange or chemical derivatization to study structure-dependence of disulfide bond reduction rates, may allow for the determination of complicated disulfide bond linkages.



(A)
●●●→→→
MRSLLILVLCFLPLAALG**KVFGRCELA**
**AAMKRHGLDNYRGYSLG**NWVCAAKF
ESNFNTQATNRNTDGSTDYGILQINSRW
WCNDGRTPGSRNLCNIPCSALLSSDITA
SVNCAKKIVSDGNGMNAWVAWRNRC
KGTDVQAWIRGCRL

(B)
●●●→→→
MRSLLILVLCFLPLAALG**KVFGRCELAA**
**AMKRHGLDNYRGYSLGNWVC**AAKFES
NFNTQATNRNTDGSTDYGILQINSRWWC
NDGRTPGSRNLCNIPCSALLSSDITASVN
CAKKIVSDGN**GMNAWVAWRNRCKGT**
**DVQAWIRGCRL**
←←←ooo

(C)
●●●→→→
MRSLLILVLCFLPLAALG**KVFGRCELA**
**AAMKRHGLDNYRGYSLGNWVCAAK**
**FESNFNTQATNRNTDGSTDYGILQINS**
**RWWCNDGRTPGSRNLCNIPCSALLSS**
**DITASVNCAKKIVSDGNGMNAWVAW**
**RNRCKGTDVQAWIRGCRL**
←←←ooo

Figure III-1-8. MALDI spectra of 1 pmol of lysozyme after (A) no reduction, (B) 2 hr reduction in 45 mM DTT at 37 °C, and (C) 15 hr reduction in 45 mM DTT at 37 °C, followed by directly mixing with 6M HCl and applying microwave irradiation for 1 min.

**Effect of surfactants.** In applying the MAP sequencing technique to real world samples, another important consideration is whether this technique can tolerate common additives, such as surfactants. SDS is perhaps the most difficult surfactant to deal with in the mass spectrometric analysis of proteins[30]. However, recent advances in sample preparation methods, such as the use of a two-layer method, allow direct analysis of protein and peptide samples containing a small amount of SDS, albeit with reduced sensitivity [31, 32]. We have investigated the effect of SDS on acid hydrolysis with MI and the effect of SDS on MALDI-TOF detection (see Figure III-1-9 (A) and (B)). Figure III-1-9 (A) shows the effect of SDS on MI assisted acid hydrolysis and (B) shows the effect of SDS on MALDI-TOF detection. (A) MALDI spectra of the hydrolysates generated from solutions containing 1 μg/μl of cytochrome c, 3M HCl, and various percentages of SDS after applying microwave irradiation for 1 min. (B) MALDI spectra of the cytochrome c hydrolysates mixed with various percentages of SDS. Note that, in (A), the number of peaks detected and the signal-to-noise ratios of the spectra obtained from the samples containing up to 0.1% SDS are similar. However, when the sample containing 0.3% SDS was analyzed, the spectral quality deteriorated. With 0.5% SDS, a poor spectrum was obtained. With SDS concentrations greater than 0.1%, protein was observed to precipitate out. As a control, hydrolysates of cytochrome c were produced in the absence of SDS. This was followed by addition of various different amounts of SDS. The mixtures were then analyzed by MALDI. The results of this control experiment are shown in (B). Similar concentration dependences of spectral quality are obtained, indicating that decreasing spectral quality as the SDS concentration increases is mainly due to the interference of SDS with MALDI detection. It is found that the effect of SDS on acid hydrolysis is small for samples containing less than 0.1% SDS. Thus the technique appears to provide moderate tolerance to SDS, which should prove to be important in situations where the presence of SDS in a sample is unavoidable, as in the case of membrane proteins.

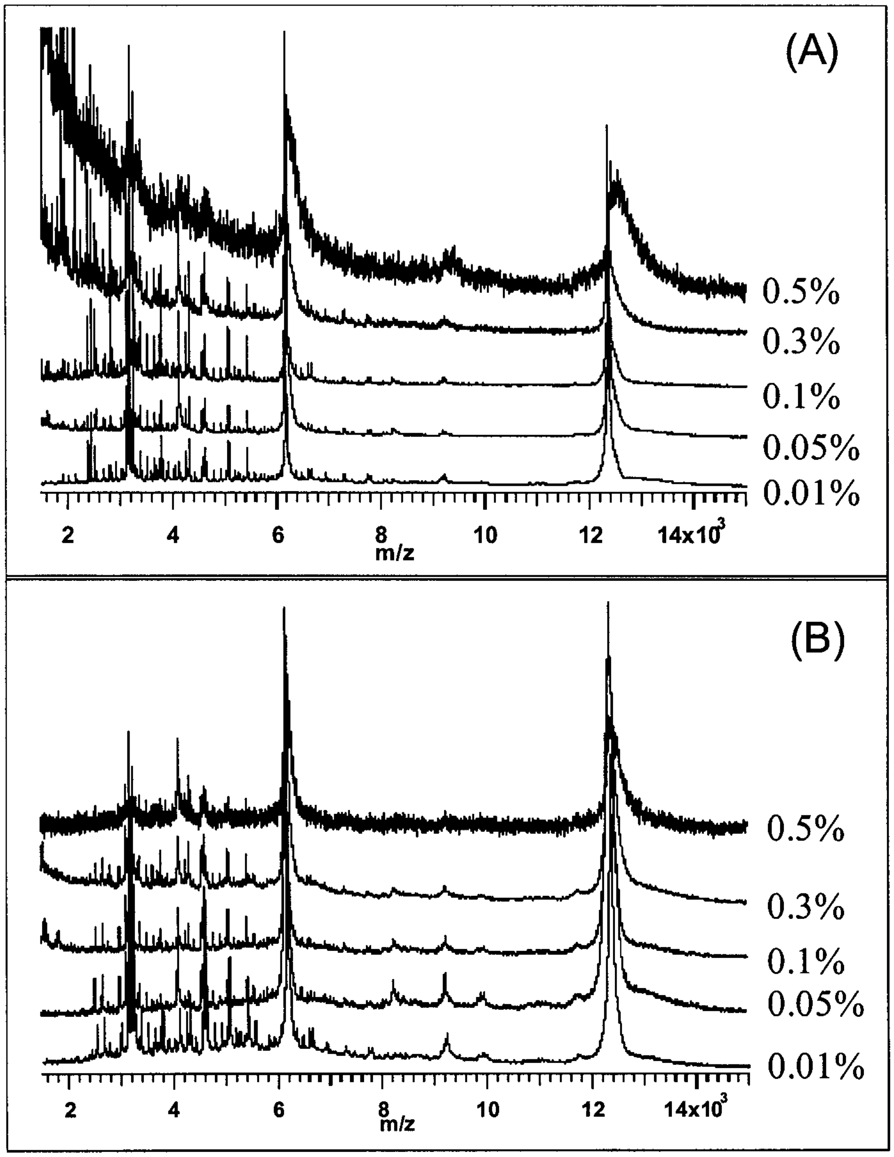


Figure III-1-9. Effect of SDS on acid hydrolysis with MI. (A) MALDI spectra of the hydrolysates generated from solutions containing 1 μg/μl of cytochrome c, 3M HCl, and various percentages of SDS after applying microwave irradiation for 1 min. (B) MALDI spectra of the cytochrome c hydrolysates mixed with various percentages of SDS.

### III. 1. 4. Conclusions

In conclusion, it should be noted that future improvements in mass spectrometric instrumentation and sample handling methodology should expand the applicability of the MAP technique to large proteins. The examples given here show that, for the current MAP technique, there is an upper mass limit for detecting polypeptide peaks. The useful mass region is generally limited to below 14,000 Da. This limit is most likely due to the problem associated with peptide detection, not the hydrolysis process itself. As Figure III-1-2D-F shows, the peak intensity decreases as the polypeptide mass increases, which is expected in MALDI-TOF where the ionization efficiency drops as the analyte mass increases [20]. Detector saturation also plays a role [29]. We note that analyzing polypeptide ladders is analogous to the analysis of an industrial polymer such as polystyrene with a broad oligomer distribution where high mass oligomers are usually not detected [28, 29]. Analyzing a polydisperse polymer can be carried out using size-exclusion chromatography to pre-fractionate the sample into several narrow polydisperse polymers, followed by MALDI analysis of individual fractions [33]. Similarly we believe that the upper mass limit of the current MAP technique can be extended by using chromatography to mass-fractionate polypeptides after hydrolysis, and then to analyze the individual fractions by MALDI. Ultimately the mass limit is likely to be imposed by the mass resolution required to resolve adjacent peptides and mass measurement accuracy. Resolution requirements may be relaxed by designing experiments to fractionate N- and C-terminal peptides into two groups (e.g., using an affinity tag at the terminus combined with affinity purification), followed by MAP sequencing of the two fractions. The use of orthogonal MALDI-TOF [27, 28] which provides better resolution and accuracy than the conventional MALDI-TOF used in this work, and MALDI FT-ICR [25] with superior resolution and accuracy than TOF may greatly extend the useful mass range of the MAP sequencing technique, allowing for sequencing large proteins.

In summary, we discovered that proteins could be readily hydrolyzed after a brief exposure to microwave irradiation to form predominately two series of polypeptide ladders: one containing the N-terminal amino acid and another one containing the C-terminal amino acid. MALDI analysis of the hydrolysate produced a simple mass spectrum consisting of peaks from the N- and C-terminal peptide ladders exclusively. Mass analysis of the polypeptide ladders allowed rapid determination of protein sequences and modifications.

### III. 1. 5. Cited literature

(1) Yates, J. R., III. *Trends Genet.* **2000,** *16*, 5-8.

(2) Pandey, A. and Mann, M. *Nature* **2000,** *405*, 837-846.

(3) Aebersold, R. H. and Goodlett, D. R. *Chem. Rev.* **2001,** *101*, 269-295.

(4) Hunt, D. F. *et al. Science* **1992,** *255*, 1261-1263 .

(5) Kinter, M. & Sherman, N. E. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, Desiderio, D. M.; Nibbering, N. M. M., Ed.; John Wiley and Son's Ltd.: New York, 2000.

(6) Kelleher, N. L. *et al. J. Am. Chem. Soc.* **1999,** *121*, 806-812.

(7) Sze, S. K., Ge, Y., Oh, H., & McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*, 1774-1779.

(8) Patterson, D. H., Tarr, G. E., Regnier, F. E., & Martin, S. A. *Anal. Chem.* **1995,** *67*, 3971-3978.

(9) Chait, B. T., Wang, R., Beavis, R. C., & Kent, S. B. H. *Science* **1993,** *262*, 8992-8996.

(10) Tsugita, A., Takamoto, K., Kamo, M., & Iwadate, H. *Eur. J. Biochem.* **1992,** *206*, 691-696.

(11) Vorm, O. & Roepstorff, P. *Bio. Mass Spectrom.* **1994,** *23*, 734-740.

(12) Zubarev, R. A., Chivanov, V. D., Hakansson, P. & Sundqvist, B. U. R. *Rapid Commun. Mass Spectrom.* **1994,** *8*, 906-912.

(13) Gobom, J., Mirgorodskaya, E., Nordhoff, E., Hojrup, P., & Roepstorff, P. *Anal. Chem.* **1999,** *71*, 919-927.

(14) Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., & Standing K. G. *Anal. Chem.* **2000,** *72*, 2132-2141.

(15) Lin, S. H., Tornatore, P. & Weinberger, S. R. *Eur. J. Mass Spectrom.* **2001,** *7*, 131-141.

(16) Reiber, D. C., Brown, R. S., Weinberger, S., Kenny, J. & Bailey, J. *Anal. Chem.* **1998,** *70*, 1214-1222.

(17) Lennon, J. J. & Walsh, K. A. *Protein Science* **1997,** *6*, 2446-2453.

(18) Keough, T., Youngquist, R. S. & Lacey, M. P. *Natl. Acad. Sci. U.S.A.* **1999,** *96*, 7131 -7136.

(19) Shevchenko, A. *et al. Rapid Commun. Mass Spectrom.* **1997,** *11*, 1015-1024.

(20) Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. *Anal. Chem.* **1991,** *63*, 1193-1199.

(21) Li, L., Golding, R. E. & Whittal, R. M. *J. Am. Chem. Soc.* **1996,** *118*, 11662-11663.

(22) Whittal, R. M., Keller, B. O. & Li, L. *Anal Chem.* **1998,** *70*, 5344-5347.

(23) Solouki, T., Marto, J. A., White, F. M., Guan, S. & Marshall, A. G. *Anal. Chem.* **1995,** *67*, 4139-4144.

(24) Graves, D. J., Martin, B. L., Wang, J. H. *Co- and Post-translational Modification of Proteins*, Oxford University Press: New York, 1994.

(25) Jones, J. J., Stump, M. J., Fleming, R. C., Lay, J. O. & Wilkins, C. L. *Anal. Chem.* **2003,** *75*, 1340-1347.

(26) Krutchinsky, A. N. *et al. Rapid Commun. Mass Spectrom.* **1998,** *12*, 508-518.

(27) Loboda, A. V., Ackloo, S. & Chernushevich, I. V. *Rapid Commun. Mass Spectrom.* **2003,** *17*, 2508-2516.

(28) Schriemer, D. C. & Li, L. *Anal. Chem.***1997,** *69*, 4169-4175.

(29) Schriemer, D. C. & Li, L. *Anal. Chem.* **1997,** *69*, 4176-4183.

(30) Vorm, O., Chait, B. T., Roepstorff, P. *Proc. 41st ASMS Conf. Mass Spectrometry and Allied Topics*, San Francisco, CA, May 31-June 4, 1993, pp 621.

(31) Zhang, N., Doucette, A. & Li, L. *Anal. Chem.*, **2001,** *73*, 2968-2975.

(32) Zhang, N., Li, L. *Rapid Commun. Mass Spectrom.*, **2004,** *18*, 889-896.

(33) Nielen, M. W. F. *Mass Spectrom. Reviews* **1999,** *18*, 309-344.

(34) Dai, Y., Whittal, R. M. & Li, L. *Anal. Chem.* **1999,** *71*, 1087-1091.

# Chapter 2. Kinetic Studies of of Intact Protein *de novo* Sequencing by Microwave-Assisted Ladder Generation[1]

The existing techniques for intact protein fragmentation include traditional CAD (Collision Activated Dissociation), IRMPD (Infrared Multiphoton Dissociation) and the recently developed ECD (electron capture dissociation). Instead of gas phase ion fragmentation, the technique of sequence ladder generation using microwave irradiation presented in this work is based on liquid phase ion fragmentation. The specificity of this technique to cleave amide bonds results from the specificity of acid hydrolysis on amide bonds. Therefore, dominant sequence ladders from N- and C-terminus of the protein can be observed in the MALDI-TOF spectrum. With the reduction of disulfide bonds using DTT, strong acid and microwave irradiation denature the 3D structure of the protein and near uniform cleavage on all amide bonds is achieved, resulting in high coverage of amino acid residues. Detection by MALDI-TOF makes it possible for much improved sensitivity in comparison with ESI. Most importantly, there are far fewer multiple charged ion clusters resulting in significantly simplified spectra that are easier to interpret. In this chapter, the kinetic process, reaction rate constant, steric effect, side reactions and ion suppression in MALDI-TOF detection will be discussed in detail.

[1] A version of this chapter will be submitted for publication as:
Hongying Zhong and Liang Li, " Kinetic studies of microwave-assisted sequence ladder generation from intact proteins ".

### III. 2. 1. Introduction

Microwave is a form of electromagnetic radiation. It is between the infrared and radiofrequencies in energy. The most commonly used microwave ovens, operate at 12.2 cm (2450MHz). At this wavelength, oscillations occur $4.9\times10^9$ times per second [1]. Microwave has been used in organic synthesis and has been found to dramatically accelerate reactions and improve yields [2-8]. In the area of protein biochemistry, the use of microwave-assisted reactions is very limited. It has been most commonly used for the determination of amino acid compositions [9, 10]. More recently, microwave induced protein denaturation has been reported [11]. It is expected that tightly folded proteins will give greatly enhanced proteolysis rates under microwave irradiation than those with conventional heating.

This work applies microwave irradiation for acid hydrolysis to generate sequence ladders that contain either the N-terminal amino acid or the C-terminal amino acid. Protein molecules are seriously denatured because the hydrogen bonds are destroyed under the extremely acidic conditions. Therefore all the amide bonds are exposed to hydrolysis. The originally generated polypeptides could undergo further hydrolysis that may result in internal peptides. However, by controlling the reaction extent, these secondary cleavages could be limited and will not affect the sequence reading. The kinetic process will be theoretically simulated. Quantitative determination of the reaction rate will be discussed in this chapter as well. The challenge is for intact protein quantitation by mass spectrometry. Though isotope dilution has already been successfully demonstrated for comparison of peptide intensities [12-22], it is almost impossible to apply this technique to intact protein quantitation. This is because in the linear mode in which proteins are detected by MALDI-TOF, the resolution is not high enough to differentiate the isotope clusters. Additionally, chemical derivatization is also difficult to apply to intact proteins. Alternative reactions on many different amino acid residues produces poor yield. A new concept for intact protein quantitation is therefore studied in this chapter.

### III. 2. 2. Experimental Section

**Materials.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany.

**Generation of sequence ladders.** For the MAP sequencing experiment, a microliter of protein sample was mixed with an equal volume of 6M HCl in a 0.6-mL polypropylene vial. The vial was capped and then placed inside a household microwave oven with 900W output at 2450 MHz. A water container containing 100 mL of water was placed besides the sample vial so that the extra microwave energy was absorbed mainly by the water. The microwave oven was turned on for, typically, 60 s. After the microwave irradiation for less than 2 minutes, the bottom of the vial was found to be slightly warm. The temperature of the solution inside the vial was unknown; but no visible boiling or depletion of the solution was noted. After microwave irradiation, the sample vial was taken out of the microwave oven and the solution in the vial was dried under vacuum centrifugation.

**MALDI-TOF detection of sequence ladders.** The dried sample was re-dissolved in a matrix solution of α-cyano-4-hydroxycinnamic acid (HCCA). The mixture was then deposited on a sample target using a two-layer sample preparation method[34] for matrix-assisted laser desorption ionization (MALDI) analysis. MALDI MS experiments were carried out on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany) using a linear mode of operation. Ionization was performed with a 337-nm pulsed nitrogen laser.

**SDS-PAGE Displaying of the Resultant Mixtures of C- and N-terminal Ladders.** For gel electrophoresis, the hydrolysate of about 10 μg of the protein was put into a speedVac to remove the acid and then re-dissolved in 20 μl of the sample buffer. Sample containing about 10 μg of protein hydrolysate was then loaded in each lane of a 10% acrylamide gel using a Mini-PROTEAN® 3 cell system (BioRad). Electrophoresis was carried out at a constant voltage of 200 V for about 1 hour. After electrophoresis, gels were stained with Coomassie Blue R-250 (BioRad) first and then silver stained on the top.

### III. 2. 3. Results and Discussion

**Mechanism of Protein Hydrolysis Using Microwave Irradiation in Strong Acid Solution.** Figure III-2-1 shows a possible mechanism for hydrolysis [23]. Under acidic conditions, protonation of the amidecarbonyl makes the amide more electrophilic. The carbonyl group reacts with the hydronium ion to form a highly resonance stabilized carbocation. Water as a weak nucleophile then can attack the carbocation. This acid-base reaction is strongly favored. The amino group then becomes a good leaving group. An electron pair can be donated from either of the alcohol oxygen atoms to the carbon atom with the protonated amino group acting as a leaving group. The amine then acts as a base to remove the proton from the protonated carboxylic acid. It is expected that this step is favored by a large reaction constant and is not reversible.

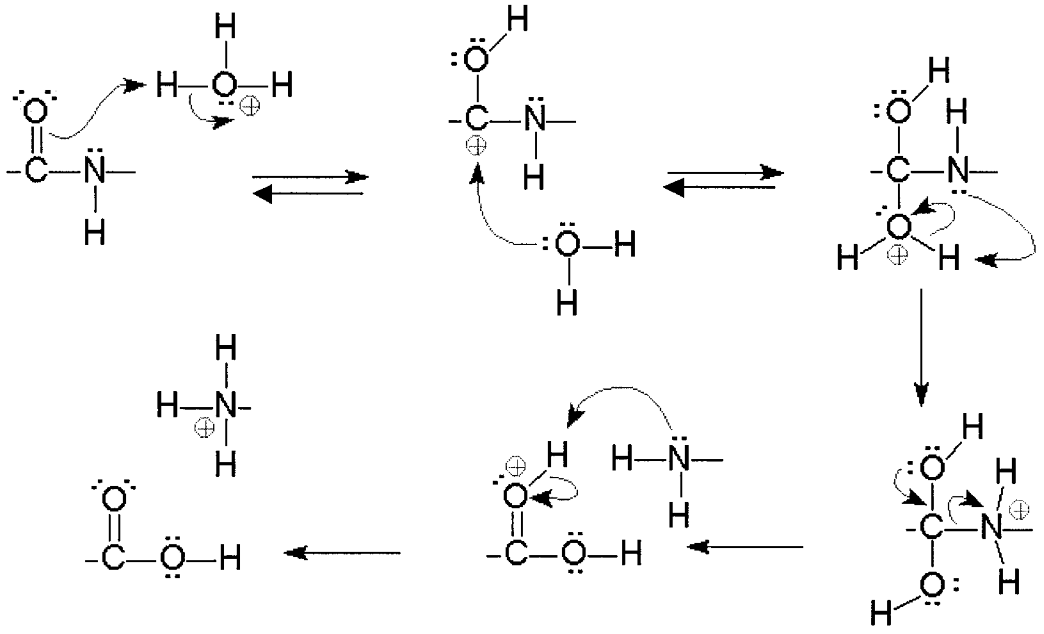


Figure III-2-1. Possible mechanism of protein hydrolysis using microwave irradiation in strong acid solution.

**Theoretical Simulation of the Kinetic Process.** Figure III-2-2 shows the schematic diagram of the hydrolysis. A is the original protein that has n amide bonds. To simplify the calculation, only the C-terminal ladder is discussed in the following. B shows the peptides from the C-terminal ladders. C shows the product resulting from the following-up hydrolysis of the original terminal peptides. An excess of water and acid were remained in the solution so that the concentration is constant during the course of reaction. A *pseudo first order* reaction with respect to the protein concentration was considered for the following calculations. K is the total reaction rate constant and $k_i$ is the reaction rate constant of each hydrolysis reaction on each amide bond. The initial concentration of the protein A is $[A]_0$ and the initial concentration of both B and C is 0.

At the time of t=0, $[A]=[A]_0$, $[B]=[C]=0$

$$[A]_0=[A]+\Sigma[B]_i+\Sigma[C]_i$$

$$\frac{dA}{dt}+\Sigma\frac{dB}{dt}+\Sigma\frac{dC}{dt}=0$$

$$-\frac{dA}{dt}=k_1[A]+k_2[A]+\cdots\cdots k_n[A]=K[A]$$

To decrease the complexity, assuming $k_1=k_2=\cdots\cdots k_n=k$, $K=nk$

$$-\int\frac{dA}{A}=\int Kdt$$

$$[A]=[A]_0e^{-Kt}$$

$$K=\frac{\ln\frac{[A]}{[A]_0}}{-t} \qquad (1)$$

$$\frac{d[B]_i}{dt}=k_i[A]-\kappa_i[B]_i=K\frac{[A]_0}{n}e^{-Kt}-\kappa_i[B]_i$$

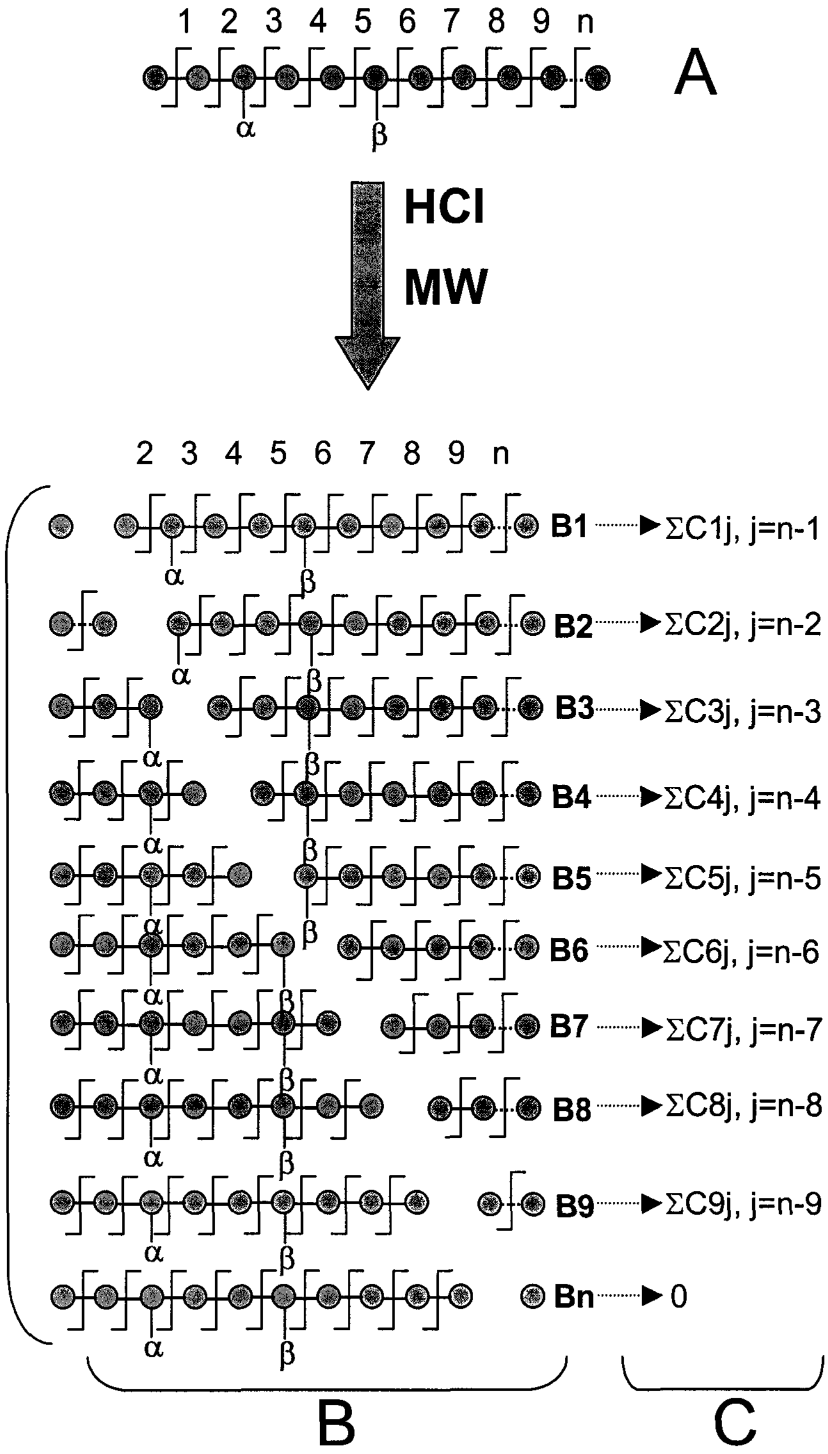


Figure III-2-2. Kinetic process of intact protein fragmentation by microwave assisted acid hydrolysis. A is the original protein. B is the C-terminal ladder and C is the products of the following-up hydrolysis of the C-terminal ladder.

so $[B]_i = \frac{K\frac{[A]_0}{n}}{\kappa_i - K}(e^{-Kt} - e^{-\kappa t})$, $\kappa_i = \sum_{i+1}^{n} k_i$

$$\frac{[B]_i}{[A]_0} = \frac{K}{n(\kappa_i - K)}(e^{-Kt} - e^{-\kappa t}) \qquad (2)$$

To reach the maximum value of $[B]_i$, $\frac{d[B]_i}{dt} = 0$

so $K\frac{[A]_0}{n} e^{-Kt} - \kappa_i [B]_i = 0$

$$t_{max} = \frac{\ln\frac{\kappa_i}{K}}{\kappa_i - K}$$

According to the assuming that $K = nk$ and $\kappa_i = \sum_{i+1}^{n} k = (n-i)k$

$$t_{max} = \frac{\ln\frac{i-n}{n}}{ik}$$

$$t_{1max} = \frac{n\ln(1-\frac{1}{n})}{-K} \quad (i=1)$$

$$[B]_{imax} = \frac{[A]_0}{n}\left(\frac{nk}{(n-i)k}\right)^{\frac{(n-i)k}{-k}} = \frac{[A]_0}{n}\left(\frac{n}{n-i}\right)^{i-n}$$

$$[B]_{1max} = \frac{[A]_0}{n}\left(\frac{n}{n-1}\right)^{1-n}$$

At the time of $t_{1max}$,

$[B]_i = \frac{K\frac{[A]_0}{n}}{\kappa_i - K}(e^{-Kt} - e^{-\kappa t})$, $\kappa_i = \sum_{i+1}^{n} k$

It can be calculated, $[B]_1 < [B]_2 < [B]_3 ....... < [B]_n$ while masses $[M]_1 > [M]_2 > [M]_3 ...... > [M]_n$

So in the following calculation, only $[B]_1$ was considered for the estimation of the interference from follow-up hydrolysis of terminal peptides.

At a specific time t, the total intensity of internal peptides $\Sigma\Sigma[C]$ resulting from follow-up hydrolysis of terminal peptides can be estimated as:

$\Sigma\Sigma[C]=[A]_0-[A]-\Sigma[B]=[A]_0(1-e^{-Kt}-n[B]_1)=[A]_0(1-e^{-Kt}+n(e^{-Kt}-e^{Kt(1/n-1)})$

Assuming $[C]_{11}=[C]_{22}=[C]_{33}=\ldots\ldots\ldots=[C]_{ij}$

$\Sigma\Sigma[C]=\frac{n^2}{2}[C]_{ij}$

then the ratio of the internal peptides and the terminal peptides ( I/S ) can be estimated as:

$$\text{I/S}=\frac{[C]_{ij}}{[B]_i}=\frac{n^2[1-e^{-Kt}+n(e^{-Kt}-e^{Kt(1/n-1)})]}{2(e^{-Kt}(e^{Kt/n}-1)} \qquad (3)$$

Figure III-2-2 shows the simulation result. n is the number of polypeptide bonds. A is the intensity from protein. B is the intensity of expected terminal peptide signal and C is the expected internal peptide signal.

For n=100, at the point of Kt equal to 0.2, the ratio of the internal peptides and terminal peptides is 0.0024. If [A], the original concentration of the analyzed protein is 1 pmol then the concentration of $B_1$ at this point is 1.6 fmol; At the point of Kt equal to 1, the ratio of the internal peptides and terminal peptides is 0.014 and the concentration of $B_1$ at this point is 3.7 fmol. Lower ratio of the internal peptides and the terminal peptides can be found with the decreasing of the polypeptide size. Considering the detection limit for a large fragment of polypeptide, the experimental condition was set to favor the forming of

$B_l$. Between Kt 0.2 and 1.0, $B_l$ has optimal concentration and also other polypeptides have high enough signals for detecting as shown in Figure III-2-3.

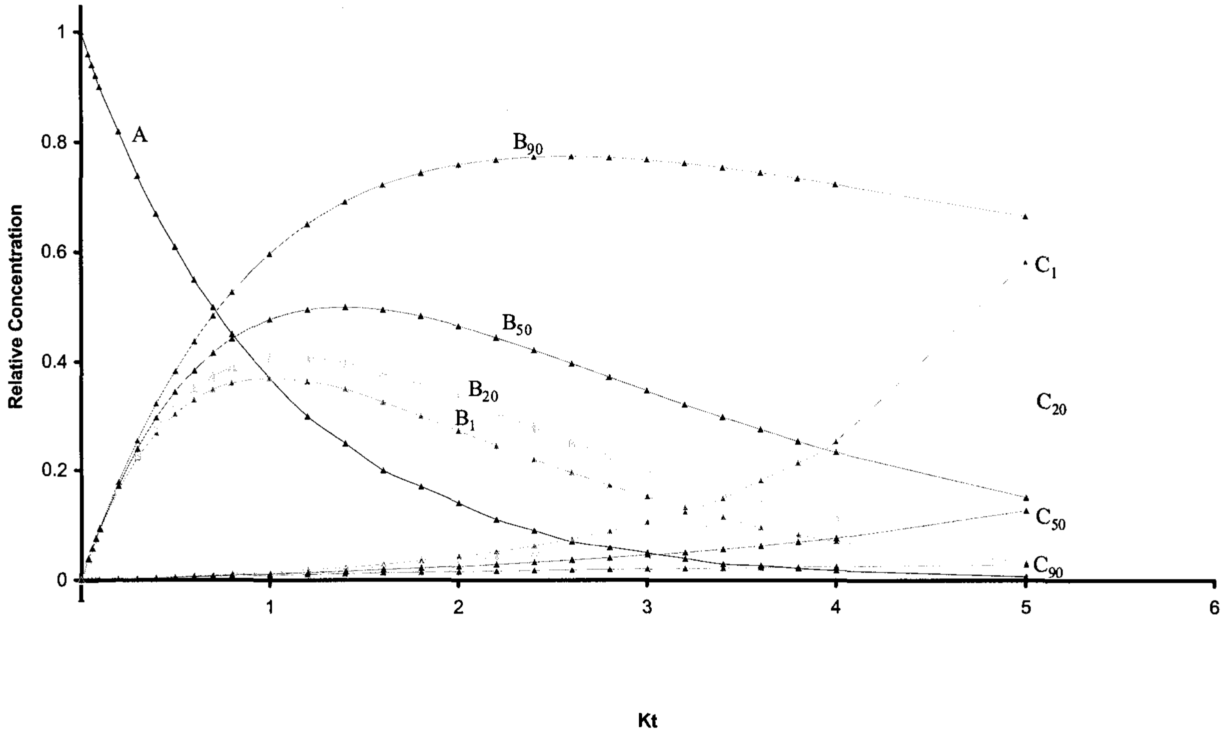


Figure III-2-3. Theoretical simulation of the intact protein fragmentation by microwave assisted hydrolysis (n=100).

**Determination of the Reaction Rate Constant K of Cytochrome C (horse heart) and $\mathbf{t_{max}}$.** The reaction rate can be calculated from equation (1) K=$\frac{\ln\frac{[A]}{[A]_0}}{-t}$. Cytochrome c from bovine heart was chosen as an internal calibration standard for the quantitative determination of concentration change of cytochrome c (horse heart) with the time for the reaction. These two proteins have the same modifications, such as covalently bound heme group and acetylation at the N-terminus. They have almost the same sequences except for a few amino acids as shown in Figure III-2-4. The differing amino acids are in bold font. It can be considered that the two types of cytochrome c will have similar ionization efficiencies and cytochrome c (bovine heart) can be used as internal calibration standard. Figure III-2-5 (A) to (E) shows MALDI-TOF spectra of the mixtures of these two proteins at different ratios. It can be seen that there is good linear relationship between the concentration ratio and the peak intensity ratio. Cytochrome c from horse heart was hydrolyzed under microwave irradiation (900W, 2450MHz) for various time periods and traditional heating at various temperatures for various times. Then cytochrome c (bovine heart) at an equal amount to the original cytochrome c (horse heart) was added to the solution. It was found that 60 s microwave irradiation had a similar reaction rate as traditional heating at 90°C for 60 s as shown in Figure III-2-6.

**Eliminating steric preference.** From Figure III-2-2, it can be found that uniform cleavage from A to B is critical for high sequence coverage and low noise level. In this study, different hydrolysis conditions were investigated. Fig. III-2-7 shows the MALDI-TOF results for cytochrome c (horse heart) hydrolysis under various conditions. All the

**GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGF TYTDANKNKGITWKEETLMEYLENPKKYIPGTKMIFAGIKKKTERED LIAYLKKATNE-Cytochrome C (horse heart)**

**GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGF SYTDANKNKGITWGEETLMEYLENPKKYIPGTKMIFAGIKKKGERED LIAYLKKATNE-Cytochrome C (bovine heart)**

Figure III-2-4. Amino acid sequences of cytochrome C from horse heart and bovine heart

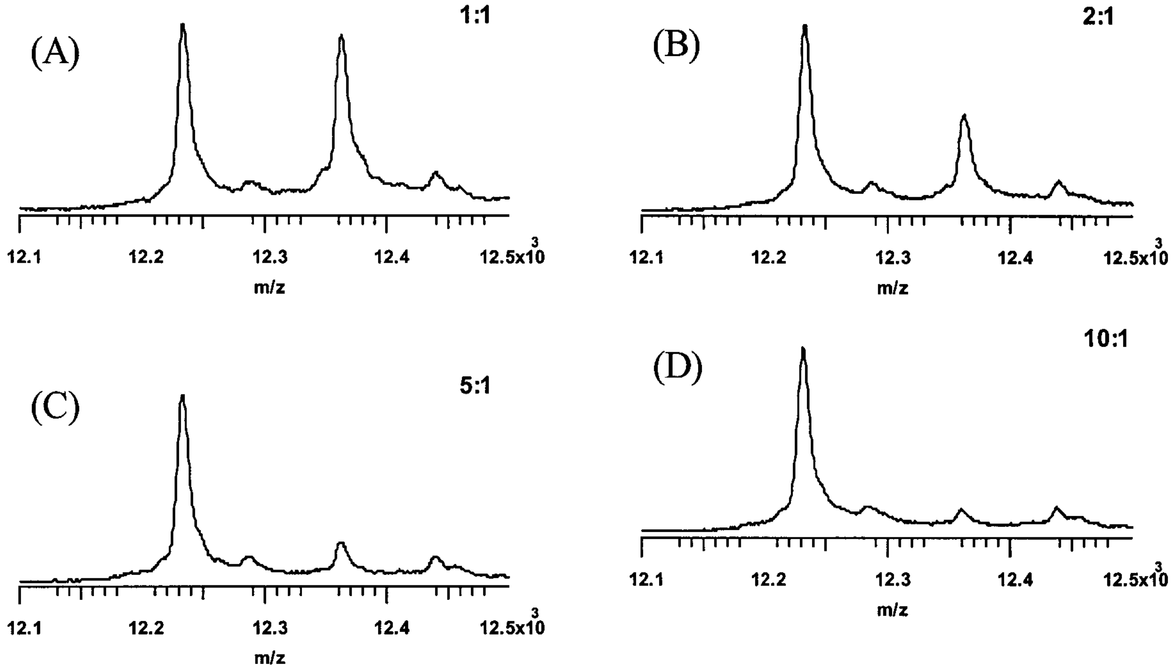


Figure III-2-5 (A) to (E) MALDI-TOF spectra of CYC-HORSE and CYC-BOVINE with different ratio.

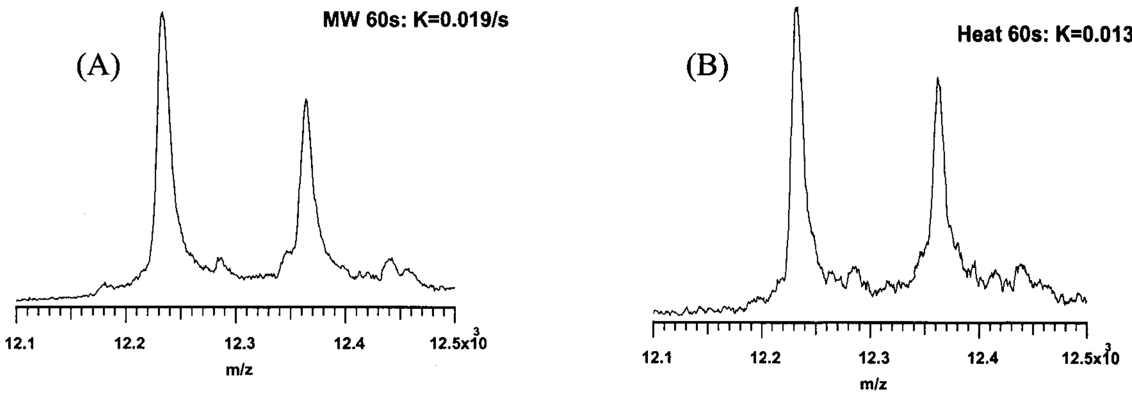
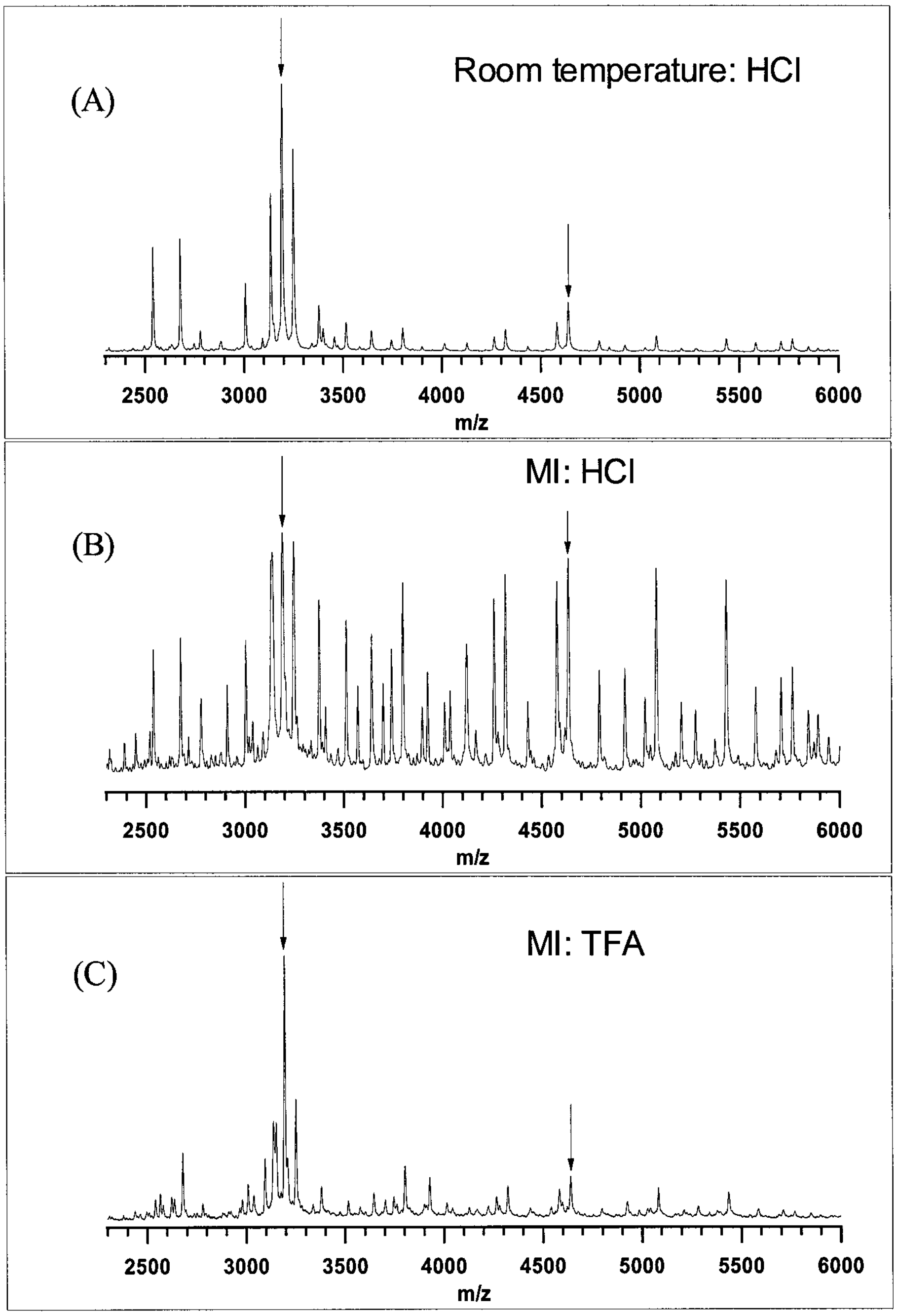


Figure III-2-6. Kinetic comparison of protein hydrolysis under microwave irradiation conditions and conventional heating conditions.

spectra were acquired at the $t_{max}$ under the specific conditions. Under microwave irradiation using HCl as catalyst, the spectrum within similar mass range displays similar intensity. The two peaks indicated by arrows correspond to two different cleavage sites: one is from the site with two small amino acid residues, G, and another one is from the site with one small amino acid residue, G, and one big amino acid residue R. Irradiated by microwave, these two peaks show similar reaction rates. It is probably because of the heating and molecular alignment resulting from microwave irradiation. There is less steric preference. All the polypeptide bonds are exposed to acid cleavage and provide more sequence specific ladders. Using the same HCl at room temperature with no microwave irradiation, these two peaks show a significant difference. It is obvious that the traditional reaction preferentially cleaves the small amino acid residue. With traditional heating, fewer peaks are observed, and some polypeptide bonds are not cleaved. Using the same microwave irradiation but with 5% TFA as catalyst, these two peaks show significant differences again. Compared with HCl, TFA is a bigger molecule and it prefers small amino acid residues. So both the microwave and the acid used in the experiments affect the steric preference of the reaction. Besides the steric effect, the 3D structure of proteins also affects the reaction. Strong acid used in the experiments denatures the protein 3D structure and the microwave irradiation further destroys the protein folding. All of these factors work together to cleave the intact protein into a complete sequence ladder.

**Competition of side reactions.** Oxidization and deamidation are two major side reactions of this ladder generation technique. Oxidization of methionine and tryptophan was avoided by adding DTT to the protein solution or by purging with nitrogen gas. The main concern here is the deamidation of N and Q to D and E, respectively, that affects the mass accuracy and mass resolution. Because there is only a 1 Da mass shift for deamidation, in this chapter, high resolution reflectron mode of MALDI-TOF was used to monitor the deamidation reaction results in one Da mass difference. Figure III-2-8 shows that within one minute there is less deamidation reaction for both the small protein substance P (Figure III-2-8 (A)) and the bigger protein cytochrome c (Figure III-2-8 (B)). However, serious deamidation occurs with the increasing reaction time. Comparing with Figure III-2-2, the time period where there is less deamidation side reaction is just the optimal time to achieve high signal intensity and low noise level. At more than 1 minute, even though there is still good signal intensity and acceptable S/N ration, the deamidation will result in bad resolution and bad mass accuracy, finally resulting in the failture of sequence reading.



→GDVEKGKKIFVQKCAQCHTVEKG(G)
→GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFG(R)

Figure III-2-7. Steric preference under different hydrolysis conditions. (A) CYC digested in HCl at room temperature. (B) CYC digested in HCl under microwave irradiation. (C) CYC digested in TFA under microwave irradiation.

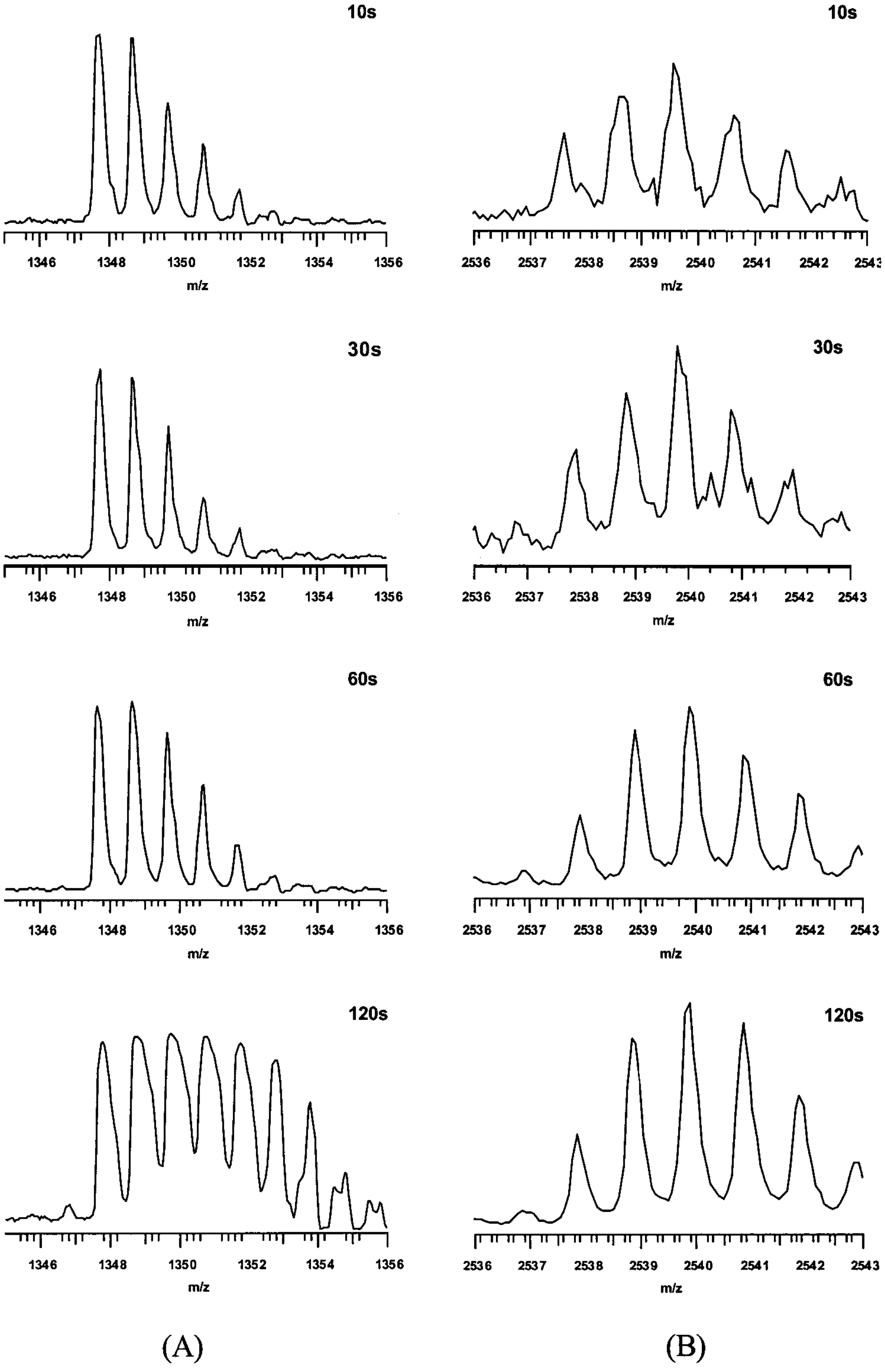


Figure III-2-8. Time course study of the competition of deamidation side reaction. (A) substance P. (B) A peptide from the hydrolysate of cytochrome c.

**Ion suppression in MALDI-TOF detection.** Ion suppression in MAPs experiments includes two different types. One is the suppression of the signal to the noise. From the theoretical calculation described earlier, it was found that within the optimal reaction time the ratio of the internal peptides and the terminal peptides can be 0.014 for cytochrome c. Ion suppression experiments demonstrated that this interference can be ignored and it will not affect the reading of sequences from the MALDI-TOF spectrum. Figure III-2-5 shows the peak intensity ratio of mixtures with different concentration ratios. Even at the ratio of 10 to 1, the weaker peak has already been suppressed. So it can be expected that with the ratio of the internal peptides and the terminal peptides as 0.014, the interference will be mainly suppressed by coexisting other ions with much higher concentrations.

Another mode of suppression is the suppression of the high molecular ions to the low molecular ions. The MALDI-TOF spectrum of the resultant mixtures of polypeptides produced by microwave-assisted protein hydrolysis always displays very low intensity in high mass region. In order to demonstrate that this phenomena results from the ion suppression in MALDI-TOF detection instead of the reaction of microwave-assisted hydrolysis, additional SDS-PAGE experiments were done to display the polypeptides with high masses that are not shown in MALDI-TOF spectrum. Figure III-2-9 is the MALDI-TOF spectrum of polypeptide mixtures produced by myoglobin hydrolysis and the inset is the corresponding image of the SDS-PAGE separation. The MALDI-TOF spectrum does not show the existence of polypeptides with high masses. However, the SDS-PAGE image demonstrates that in the high mass region, there are continuous dark bands. Very small polypeptides were not stained. The analysis of the SDS PAGE of the myoglobin hydrolysate complementarily explains the corresponding MALDI-TOF spectrum and confirms the theoretical expectation again.

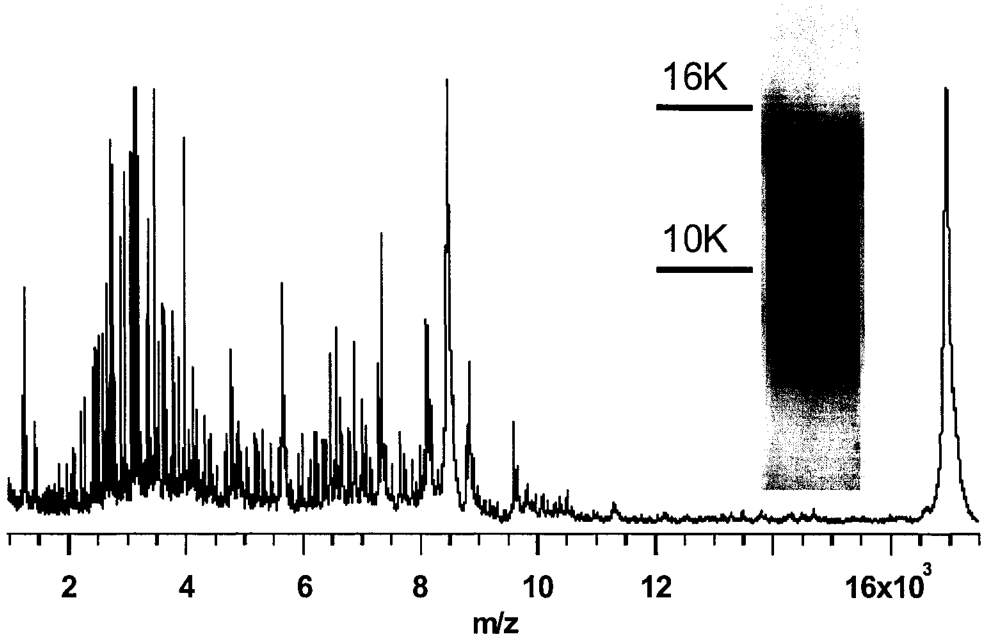


Figure III-2-9. Ion suppression in MALDI-TOF. MALDI-TOF spectrum of myoglobin hydrolysate and the inset is the corresponding SDS PAGE image.

### III. 2. 4. Conclusions.

The kinetic process of MAPs technique was experimentally and theoretically demonstrated. Strong acid is not only just a catalyst for hydrolysis reaction but also protonates the protein and destroys the hydrogen bonds. It is possible while the protein molecules align and realign with the electric field produced by microwave irradiation, the 3D structure is further destroyed and all the amide bonds are near uniformly exposed to acid cleavage, resulting in the generation of C- and N- terminal sequence ladders. Low temperature and acids with larger volume prefer to cleave at sites that have lower activation energy and therefore internal peptides and incomplete sequence ladders can be observed. Side reactions can be avoided by chemicals or by controlling the extent of the reaction. Though ion suppression in MALDI-TOF detection is effective at reducing noise from follow up hydrolysis, it limits the application of MAPs technique to proteins with higher masses. The optimum of the present MAPs technique is 12, 000 Da as discussed in chapter 1.

### III. 2. 5. Cited literature

(1) Halliday, D.; Resnick, R.; Walker, J. *Fundamentals of Physics,* 5th ed.; John Wiley and Sons: New York, 1997, pp570-pp615.

(2) Kidwai, M. *Pure Appl. Chem.* **2001,** *73,* 147-151.

(3) Varma, R. S. *Pure Appl. Chem.* **2001,** *73,* 193-198.

(4) Lew, A.; Krutzik, P. O.; Hart, M. E.; Chamberlin, A. R. *J Combinational Chem.* **2001,** *4,* 95-105.

(5) Luque-Garcia, J. L.; Luque de Castro, M. D. *Anal. Chem.* **2001,** *73,* 5903-5908.

(6) Holler, U.; Wolter, D.; Hofmann, P.; Spitzer, V. *J Agric. Food Chem.* **2003,** *51,* 1539-1542.

(7) Miedel, M. C.; Hulmes, J. D.; Pan, Y. C. E. *J Biochem. Biophys. Methods* **1989,** *18,* 37-52.

(8) Olofsson, K.; Kim, S. Y.; Larhed, M.; Curran, D. P.; Hallberg, A. *J. Org. Chem.* **1999,** *64,* 4539-4541.

(9) Weiss, M.; Manneberg, M.; Juranville, J. F.; Lahm, H. W.; Fountoulakis, M. *J. Chromatogr A.* **1998,** *795,* 263-275.

(10) Yu, H. M.; Chen, S. T.; Chiou, S. H.; Wang, K. T. *J. Chromatogr. A* **1988,** *456,* 357-362.

(11) Pramanik, B. N. et al. *Protein Sci.* **2002,** *11,* 2676-2687.

(12) Dancik, D.; Addona, T.; Clauser, K.; Vath, J.; Pevzner, P. *J. Comput. Biol.* **1999,** *6,* 327-342.

(13) Kosaka, T.; Takazawa, T.; Nakamura, T. *Anal. Chem.* **2000,** *72,* 1179-1185.

(14) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001,** *73,* 2836-2842.

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

(16) Kuyama, H. et al. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 1642-1650.

(17) Gygi, S. P. et al. *Nat. Biotechnol.* **1999,** *17,* 994-999.

(18) Shen, M. et al. *Mol. Cell. Proteomics* **2003,** *2,* 315-324.

(19) Hale, J. E.; Butler, J. P.; Knierman, M. D.; Becker, G. W. *Anal. Biochem.* **2000,** *287,* 110-117.

(20) Chelius, D.; Shaler, T. A. *Bioconjugate Chem.* **2003,** *14,* 205-211.

(21) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002,** *20*, 163-170.

(22) Chen, X.; Smith, L. M.; Bradbury, E. M. *Anal. Chem.* **2000,** *72*, 1134-1143.

(23) Roberts, J. D.; Caserio, M. C. *Basic Principles of Organic Chemistry*, W. A. Benjamin, Inc.: New York, 1964.

# Chapter 3. Sequential Digestion for the Analysis of Proteins with Large Masses by Microwave-Assisted Ladder Generation

Due to the ion suppression, the intensity of polypeptides in high mass region is much lower than that of low mass region. MAPs technique is limited to about 12,000 Da. An option is to chemically cleave the large proteins into several polypeptides and then subject these to MAPs. A rapid, non-toxic method was developed to cleave large proteins into large peptide segments and then further to generate sequence ladders.

## III. 3. 1. Introduction

Using chemical methods, large proteins can be chemically cleaved into several segments [1]. The possible cleavage sites include methiony-X [2-11], tryptophanyl-X [12-16], aspartyl-X [17-19], Cysteinyl-X [20-22] and asparaginyl-glycyl bonds [23-26]. However, the reported methods are either toxicity, low efficiency or induce the formation of adducts. D-P bond is an acid labile bond. It is reported that 70% formic acid can generate protein segments. Unfortunately, formylation at the N-terminus of the protein is often observed upon treatment with formic acid. Additionally, complete cleavage and high yield of the protein segments are very difficult to obtain. As discussed in the previous chapters, microwave-assisted hydrolysis of proteins does show cleavage preference when acids such as TFA are used. In this chapter, further discussion will be presented using carbonic anhydrase as an example to show how to achieve the goal of generating large peptides from large proteins using TFA.

### III. 3. 2. Experimental Section

**Chemicals.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For HPLC separation, MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany.

**Cleavage of the large protein into segments.** 10 µl (1 µg/µl) of the protein solution with 1% TFA was placed in a 1.5 ml polypropylene centrifuge vial, capped and sealed with Teflon tape. The vial was placed in a domestic 900W (2450 MHz) microwave oven. 100 ml of water in a loosely covered container was placed besides the sample vial to absorb excess microwave energy. 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 were 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 vacuum centrifuge to remove the acid. The protein digest was re-suspended in 100 µl of 0.1% TFA aqueous solution and centrifuged at 16000 g for 5 minutes to remove any possible residual particles. 90 µl of the solution was injected into the HPLC system for separation and fractionation.

Care was taken when handling the concentrated acids as well as other precautions that were followed in the use of the microwave oven (see Safety Considerations).

**HPLC Separation.** Reversed-phase HPLC separations of the peptides were made using a Vydac C18 column (1 mm i. d. × 150 mm, 5 μm, 300 A°; Vydac, Hesperia, CA) on an Agilent 1100 capillary HPLC system. At a flow rate of 50 μl/min, a linear gradient from 5% B to 85% B over 70 minutes was used, where mobile phase A was water containing 0.1% (v/v) TFA, and mobile phase B was acetonitrile with 0.1% (v/v) TFA. A UV detector was used at 214 nm. During the separation period from 10 to 50 minutes, the eluate was fractionated by peaks.

**MAPs of the protein segments.** For the MAP sequencing experiment, microliters of protein sample was mixed with an equal volume of 6M HCl in a 0.6-mL polypropylene vial. The vial was capped and then placed inside a household microwave oven with 900W output at 2450 MHz. A water container containing 100 ml of water was placed besides the sample vial so that the extra microwave energy was absorbed mainly by the water. The microwave oven was turned on for, typically, 60 s. After microwave irradiation, the sample vial was taken out of the microwave oven and the solution in the vial was dried under vacuum centrifugation.

**MALDI MS and MS/MS.** For peptide mass mapping by MALDI time-of-flight (TOF) MS, a two-layer sample/matrix preparation method was used with α-cyano-4-hydroxy-cinnamic acid (HCCA) as matrix. 0.7 μl of the first-layer matrix solution containing 12 mg/ml of HCCA in 20% methanol/acetone was deposited on the MALDI target and air-dried. 0.5 μl of sample was mixed with 0.5 μl of matrix (50% acetonitrile/water saturated with HCCA) and then deposited onto the first layer. The MALDI-TOF mass spectra were obtained on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany).

### III. 3. 3. Results and Discussion

**Principle of the sequential digestion.** The acid labile D-P bond is not common in proteins so it can be used to generate large segments from intact proteins. The resultant segments can then be separated and fractionated by HPLC. Figure III-3-1 shows the methodology of the sequential digestion process. Each fraction with optimal size (less than 15 KDa and more than 2 KDa) then can be analyzed using MAPs technique.

**Cleavage of carbonic anhydrase into segments.** As shown in Figure III-3-2, carbonic anhydrase (Bovine) has two D-P bonds indicated by the squares. Theoretically, 3 segments should be obtained as indicated by the dashed line, solid line and square dotted line, with masses of 4633.99, 15206.90 and 9222.73 Da respectively. These masses are in the optimal mass range for MAPs technique.

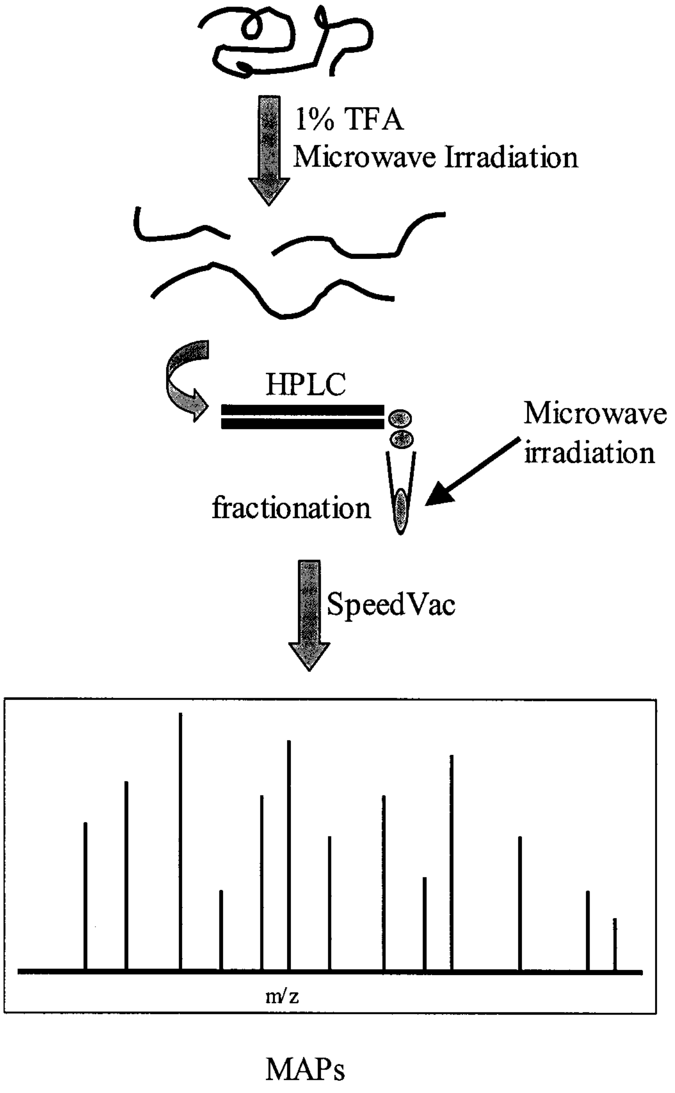
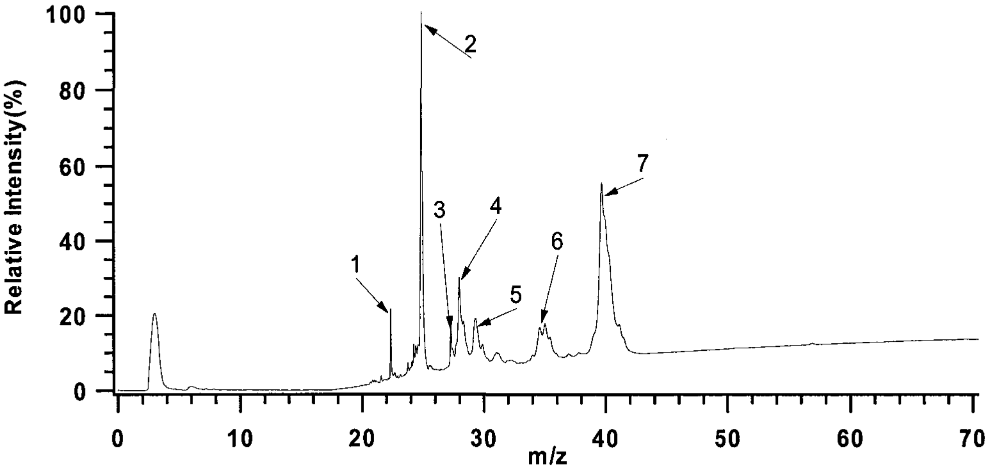


Figure III-3-1. Methodology of the sequential digestion for the analysis of large size proteins by microwave-assisted ladder generation.

SHHWGYGKHNGPEHWHKDFPIANGERQSPVDIDTKAVVQDPALK
PLALVYGEATSRRMVNNGHSFNVEYDDSQDKAVLKDGPLTGTYR
LVQFHFHWGSSDDQGSEHTVDRKKYAAELHLVHWNTKYGDFGT
AAQQPDGLAVVGVFLKVGDANPALQKVLDALDSIKTKGKSTDFP
NFDPGSLLPNVLDYWTYPGSLTTPPLLESVTWIVLKEPISVSSQQM
LKFRTLNFNAEGEPELLMLANWRPAQPLKNRQVRGFPK

Figure III-3-2. Sequence of carbonic anhydrase (bovine). D-P bonds are indicated by squares. The various underlines indicate the different segments.

**HPLC separation and fractionation of the resultant segments.** Carbonic anhydrase was hydrolyzed in 1% TFA by microwave irradiation for 15 minutes. The resultant hydrolyzate was separated and fractionated by HPLC, as shown in Figure III-3-3. Each fraction was detected by MALDI-TOF. The three theoretical segments are all observed (peak 2, peak 6, peak 7). Additionally there are 3 other segments from the N- or C-terminus (peak 1, peak 3, peak 5) and one segment from an internal peptide (peak 4). The corresponding molecular weight of each peak is listed in the spectrum that is in the range from 2 KDa to 15 KDa. Compared with the CNBr method, this method is fast, non-toxic and has high yield. Intact protein peak was not detected in either the liquid chromatograph or the MALDI-TOF, meaning that all the protein had been converted into segments.



| Peak# | MH⁺ (Da) | Peak# | MH⁺ (Da) |
|---|---|---|---|
| 1 | 2250.50 | 5 | 9066.27 |
| 2 | 4634.91 | 6 | 15207.29 |
| 3 | 4452.07 | 7 | 9223.27 |
| 4 | 3465.49 | | |

Figure III-3-3. Liquid chromatography of the resultant segments from carbonic anhydrase (bovine). Molecular mass detected by MALDI-TOF was listed in the spectrum.

**MAPs of the segments.** Sequence analysis by MAPs technique was applied to the eluant. Peak 6 has weak signal but higher molecular weight so it was chosen as an example to display the MAPs. Figure III-3-4 is the MALDI-TOF spectrum of the sequence ladders for peak 6. Even though the Signal/Noise of this spectrum is not very good due to the low amount, the sequence can still be clearly seen. Figure III-3-4 (A) is the MALDI-TOF spectrum of the intact protein. Multiple charged states are labelled in the spectrum. The other unlabeled peaks are from the co-eluted impurities. Figure III-3-

4 (B) is the corresponding MALDI-TOF spectrum of sequence ladders generated by microwave irradiation for 1 minute in 3M HCl. Peaks indicated by "•" are from N-terminus and peaks indicated by "o" are from C-terminus. Unlabeled peaks are from co-eluted impurities. So it is obvious that sequential digestion is a potential powerful option to apply MAPs technique to deal with proteins with high masses. Combined with multidimensional HPLC separation, further improved results can be expected.

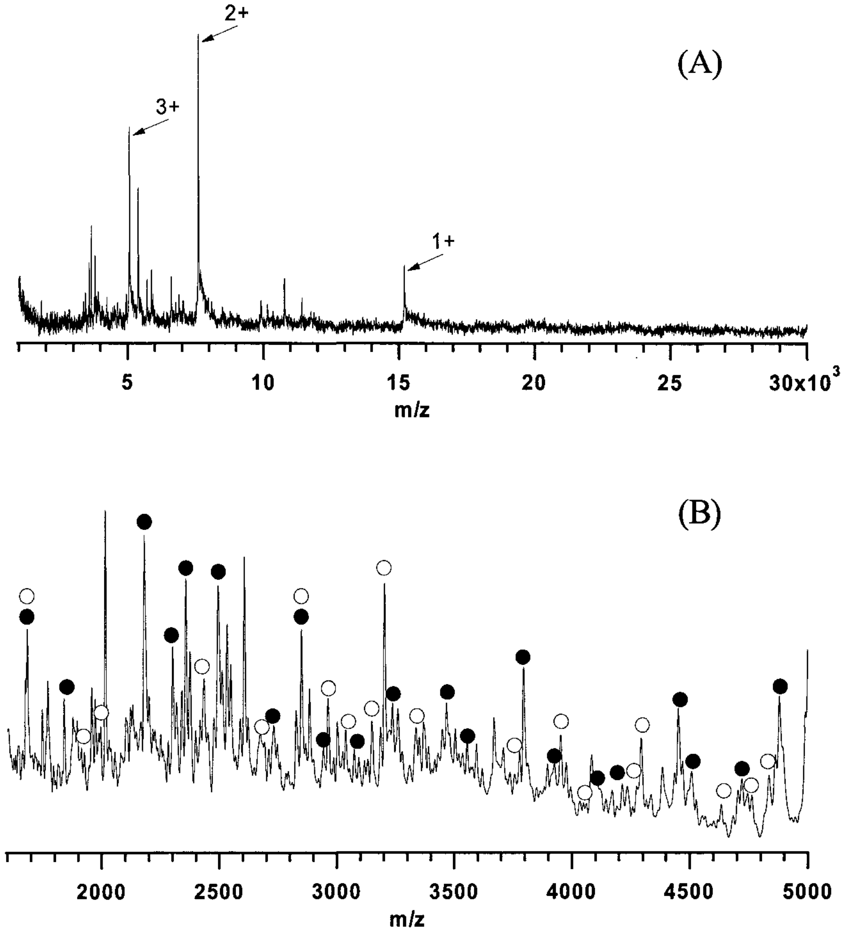


Figure III-3-4. Sequential digestion for analysis proteins with high molecular masses. (A) MALDI-TOF spectrum of a segment from carbonic anhydrase (bovine). (B) MALDI-TOF spectrum of the corresponding sequence ladder generated by microwave irradiation for 1 minute in 3 M HCl. Peaks indicated by "•" are from the N-terminus and peaks indicated by "o" are from the C-terminus.

### III. 3. 4. Conclusions

The dynamic range of MALDI-TOF limits the application of the MAPs technique to the application to proteins with low molecular masses. Larger proteins can be made amenable to MAPs by chemically cleaving the large protein into a few segments, based on acid hydrolysis of the rare D-P bond. Microwave assisted sequential acid digestion developed in this work is an efficient and rapid technique to deal with this situation. Combined with separation tools, it is possible to cover the whole sequence of large proteins using the MAPs technique. Not only does this method provide a significantly improved reaction rate, but also it avoids the commonly used toxic CNBr and efficiently produces protein segments.

### III. 3. 5. Cited literature

(1) Walker, J. M. *The Protein Protocols Handbook;* Humana Press: New Jersey, 2002.

(2) Yuan, G.; Bin, J. C.; McKay, D. J.; Synder, F. F. *J. Biol. Chem.* **1999,** *274,* 8175-8180.

(3) Malouf, N. N.; McMahon, D.; Oakeley, A. E.; Anderson, P. A. W. *J. Biol. Chem.* **1992,** *267,* 9269-9274.

(4) Dong, M.; Ding, X. Q.; Pinon, D. I.; Hadac, E. M.; Oda, R. P.; Landers, J. P.; Miller, L. J. *J. Biol. Chem.* **1999,** *274,* 4778-4785.

(5) Wallace, D. S.; Hofsteenge, J.; Store, S. R. *Eur. J. Biochem.* **1990,** *188,* 61-66.

(6) Fontana, A.; Gross, E. *Fragmentation of polypeptides by chemical methods in practical protein chemistry A handbook,* Darbre, A., Ed.; John Wiley and Son's Ltd.: Chichester, 1986, pp67-120.

(7) Morrison, J. R.; Fidge, N. H.; Greo, B. *Anal. Biochem.* **1990,** *186,* 145-152.

(8) Kaiser, R.; Metzka, L. *Anal. Biochem.* **1999,** *266,* 1-8.

(9) Beavis, R. C.; Chait, B. T. *Proc. Natl. Acad. Sci. USA* **1990,** *87,* 6873-6877.

(10) Caprioli, R. M.; Whaley, B.; Mock, K. K.; Cottrell, J. S. *Techniques in Protein Chemistry II,* Angeletti, R. M., Ed.; Academic Press: San Diego, 1991.

(11) Murphy, C. M.; Fenselau, C. Wilson, K. J.; Fischer, S.; Yuau, P. M. *Anal. Chem,* **1995,** *67,* 1644-1645.

(12) Rahali, V.; Gueguen, J. *J. Prot. Chem.* **1999,** *18,* 1-12.

(13) Vestling, M. M.; Kelly, M. A.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1994,** *8,* 786-790.

(14) Fontana, A.; Dalzoppo, D.; Grandi, C.; Zambonin, M. *Meth. Enzymol.* **1983,** *91,* 311-318.

(15) Lischwe, M. A.; Sung, M. T. *J. Biochem.* **1997,** *252,* 4976-4980.

(16) Savige, W. E.; Fontana, A. *Meth. Enzymol.* **1977,***47,* 459-469.

(17) Ingris, A. S. *Meth. Enzymol.* **1983,** *91,* 324-332.

(18) Landon, M. *Meth. Enzymol.* **1977,** *47,* 132-145.

(19) Yu, W.; Vath, J. E.; Huberty, M. C.; Martin, S. A. *Anal. Chem.* **1993,** *65,* 3015-3023.

(20) Wilson, K. J.; Fischer, S.; Yuau, P. M. *Methods in protein sequence analysis,* Wittman,-Liebold, B. Ed.; Springer-Verlag: Berlin, 1989, pp310-314.

(21) Aitken, A. *Methods in Molecular Biology,* Vol 32: *Basic protein and peptide protocols,* Walker, J. M., Ed.; Humana Press: New Jersey, 1994, pp351-360.

(22) Stark, G. R. *Meth. Enzymol.* **1977,** *47,* 129-132.

(23) Bornstein, P.; Balian, G. *Meth. Enzymol.* **1977,** *47,* 132-145.

(24) Kwong, M. Y.; Harris, R. J. *Protein Sci.* **1994,** *3,* 147-149.

(25) Blodgett, J. K.; Londin, G. M.; Collins, K. D. *J Am. Chem. Soc.* **1985,** *107,* 4305-4313.

(26) Saris, C. J. M.; Van Eenbergen, J.; Jenks, B. G.; Bloemers, H. P. J. *Anal. Biochem.* **1983,** *132,* 54-67.

# Chapter 4. Identification of Amino Acid Mutation in Proteins by Microwave-Assisted Ladder Generation[1]

Genetic variation influences disease susceptibility and diagnosis as well as the response to drug treatments. Variations in promoter regions may lead to over- or under-expression of a particular gene and potentially to abnormal levels of translated proteins. It has been discovered that protein and gene expression may not correlate well, and thus identification of amino acid mutations at the protein level is urgently needed. Several problems exist for the analysis or identification of amino acid mutation using current techniques. In the following chapter, the application of the MAPs technique to the identification of amino acid mutation will be discussed.

## III. 4. 1. Introduction

As more and more genome projects are completed, interest is shifting to the role of genetic variation in phenotype [1, 2]. The most common type of human genetic variation is the single-nucleotide polymorphism (SNP) that is a base position at which two alternative bases occur at appreciable frequency (>1%) in the population. The complete human genome will release at least 1 million SNPs in the non-repetitive coding regions of genes that are expected to contribute significantly to genetic risk for common

[1] A version of this chapter will be submitted for publication as:
Hongying Zhong and Liang Li, "Identification of amino acid mutations by microwave-assisted sequence ladder generation".

disease [3]. Common SNPs occurring in the coding regions (cSNPs) of genes have already been linked to diseases and referred to as mutations [4].

Links between anemias and hemoglobin variations are well-established [5-9]. It is estimated that more than 150 million people carry Hb variations. Routine Hb diagnosis is increasing for clinical screening of newborns for abnormal hemoglobin diseases. The routine method of screening abnormal hemoglobin is performed by comparing several pieces of information present in the database. Electrophoretic mobilities, chromatographic retention time, functional properties, ethnic distribution and clinical presentation are taken into account. Finally, mass spectrometry methods are used to confirm the screening results. Hereditary hemochromatosis (HH) is another cSNPs related disease [10-12] that has an estimated carrier frequency of 1 in 8 or 1 in 10 individuals of northern`` European descent. It is a common autosomal recessive disease associated with the loss of regulation of dietary iron absorption and excessive iron deposition in major organs of the body. These diseases all result from the amino acid variation.

Mass spectrometry has been used to identify SNPs in short DNA molecules (<100 bases) based on the detection of m/z value [13-17]. However, protein and gene expression may not correlate well and sequence variation in the coding regions can affect the activity proteins more directly. In order to identify amino acid variations at the protein level, several problems existing in the current techniques need to be overcome. One is the isolation of peptides carrying amino acids with a mutation. The conventional bottom-up approach is not efficient and very difficult to cover the whole sequence [18, 19]. A top-down approach using high resolution and accuracy FT-ICR circumvents this problem to some extent. However, the dynamic range and sensitivity are low [20-23]. Also, it is difficult to apply to proteins with high molecular masses due to the limitations of the quadrupole used for ion transfer and focusing. Additionally, the complicated spectra from the top-down approach are difficult to interpretate resulting from the existence of multiple charge states from ESI and presence of other fragment ions. Another difficulty with the identification of amino acid variation is with the databases. Tandem mass spectrometry can provide sequence information but the final identification of proteins depends on the accurate match of experimental data with data in the databases predicted from DNA sequences [24-26]. Unfortunately, most amino acid mutations cannot be predicted from DNA sequences. Software to automatically match tandem mass spectra to sequences has been developed but it is not effective for unanticipated sequence variations and does not specifically indicate the type and site of the amino acid variation.

In this work, the application of the MAPs technique for the efficient and rapid identification of amino acid mutations has been achieved and is demonstrated on a model protein cytochrome C with different amino acids.

### III. 4. 2. Experimental Section

**Chemicals.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For HPLC separation, MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany.

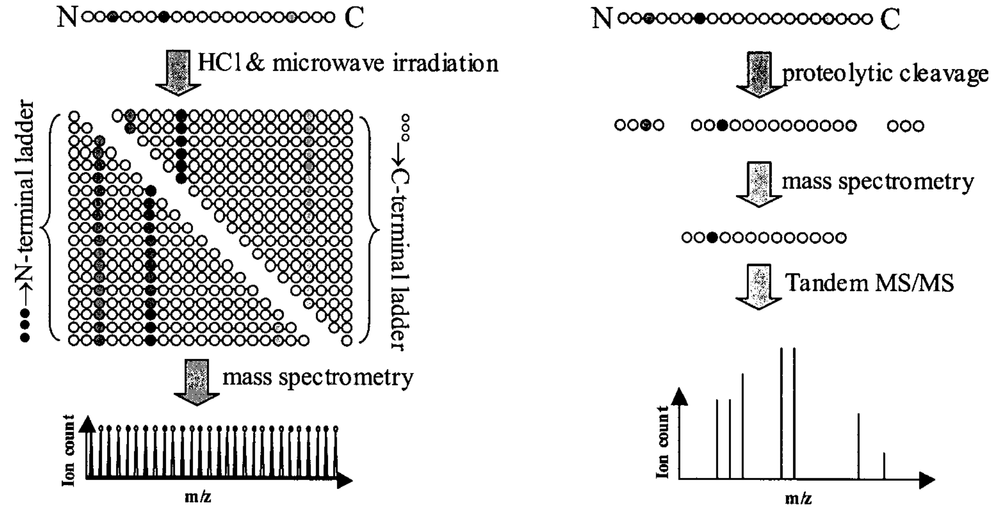
**Generation of sequence ladders.** For the MAP sequencing experiment, a microliter of protein sample was mixed with an equal volume of 6M HCl in a 0.6-mL polypropylene vial. The vial was capped and then placed inside a household microwave oven with 900W output at 2450 MHz. A water container containing 100 ml of water was placed besides the sample vial so that the excess microwave energy was absorbed mainly by the water. The microwave oven was turned on for, typically, 60 s. After the microwave irradiation for less than 2 minutes, the bottom of the vial was found to be slightly warm. The temperature of the solution inside the vial was unknown; but no visible boiling or depletion of the solution was noted. After microwave irradiation, the sample vial was taken out of the microwave oven and the solution in the vial was dried under vacuum centrifugation.

**Generation of tryptic peptides.** For the enzymatic digestion, 5 μl cytochrome C (1μg/μl) was mixed with 10 μl $NH_4HCO_3$ (100mM) and 0.3 μl trypsin (0.5 μg/μl) was added to the mixture and incubate at 37°C for 2 hours. The resultant peptides were acidified with TFA and concentrated/desalted by C18 ZIPTIP(Millipore). 50% acetonitrile was used to eluate proteolytic peptides. Organic solvent in the eluted solution was removed by SpeedVac drying for MALDI-TOF analysis.

**MALDI-TOF detection of sequence ladders.** The dried sample was re-dissolved in a matrix solution of α-cyano-4-hydroxycinnamic acid (HCCA). The mixture was then deposited on a sample target using a two-layer sample preparation method for matrix-assisted laser desorption ionization (MALDI) analysis. MALDI MS experiments were carried out on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany) using a linear mode of operation. Ionization was performed with a 337-nm pulsed nitrogen laser.

### III. 4. 3. Results and Discussion

**Principle of MAPS technique for the identification of amino acid variations.** Figure III-4-1 is the diagram of the MAPs technique versus conventional tandem mass spectrometry technique. In conventional tandem mass spectrometry, intact proteins are always enzymatically cleaved into proteolytic peptides first so the peptides without optimal size may not be detected in this sample preparation. Further, MS/MS spectra of the peptides with amino acid variations may not be identified by database searching due to the uncertainty. However, in the MAPs technique, one amino acid difference will result in a series mass shift of terminal peptides and makes the identification clear, confident and fast.



Sequence ladder-based detection of variation

Proteolytic peptide-based detection of variation

Figure III-4-1. Diagram of sequence ladder based technique versus proteolytic peptide based mass spectrometry

**Cytochrome C as a model protein for the identification of amino acid variation.** Cytochrome C from different species differ by just a few different amino acids. However, they experience the same posttranslational modifications, making them optimal as a model protein to demonstrate the MAPs technique for the identification of amino acid variations. Figure III-4-2 is the MALDI-TOF spectra of tryptic peptides of cytochrome C from horse heart and bovine heart. The amino acid sequences are listed in the spectra. These two MALDI-TOF peptide mass mapping display very similar patterns. Thus it is hard to find the differences due to the interference from background peaks. Additionally, the tryptic peptide did not cover the whole sequence. Though the two different amino acids are covered in this case, the signal intensity of the peptides is very low.

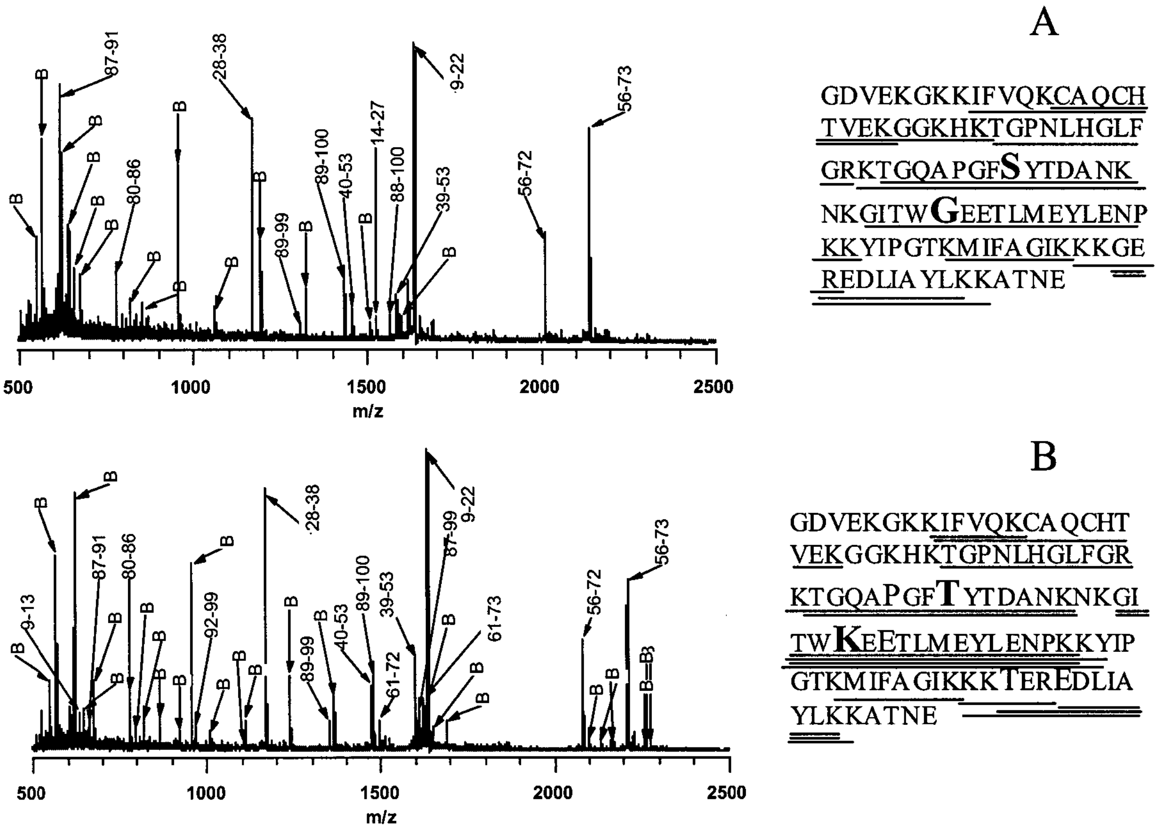


Figure III-4-2. MALDI-TOF tryptic peptide mass mapping and sequences of (A) cytochrome C(horse heart) (B) cytochrome C (bovine heart). Bold amino acids are different. Underlines indicate the sequence coverage by tryptic peptides.

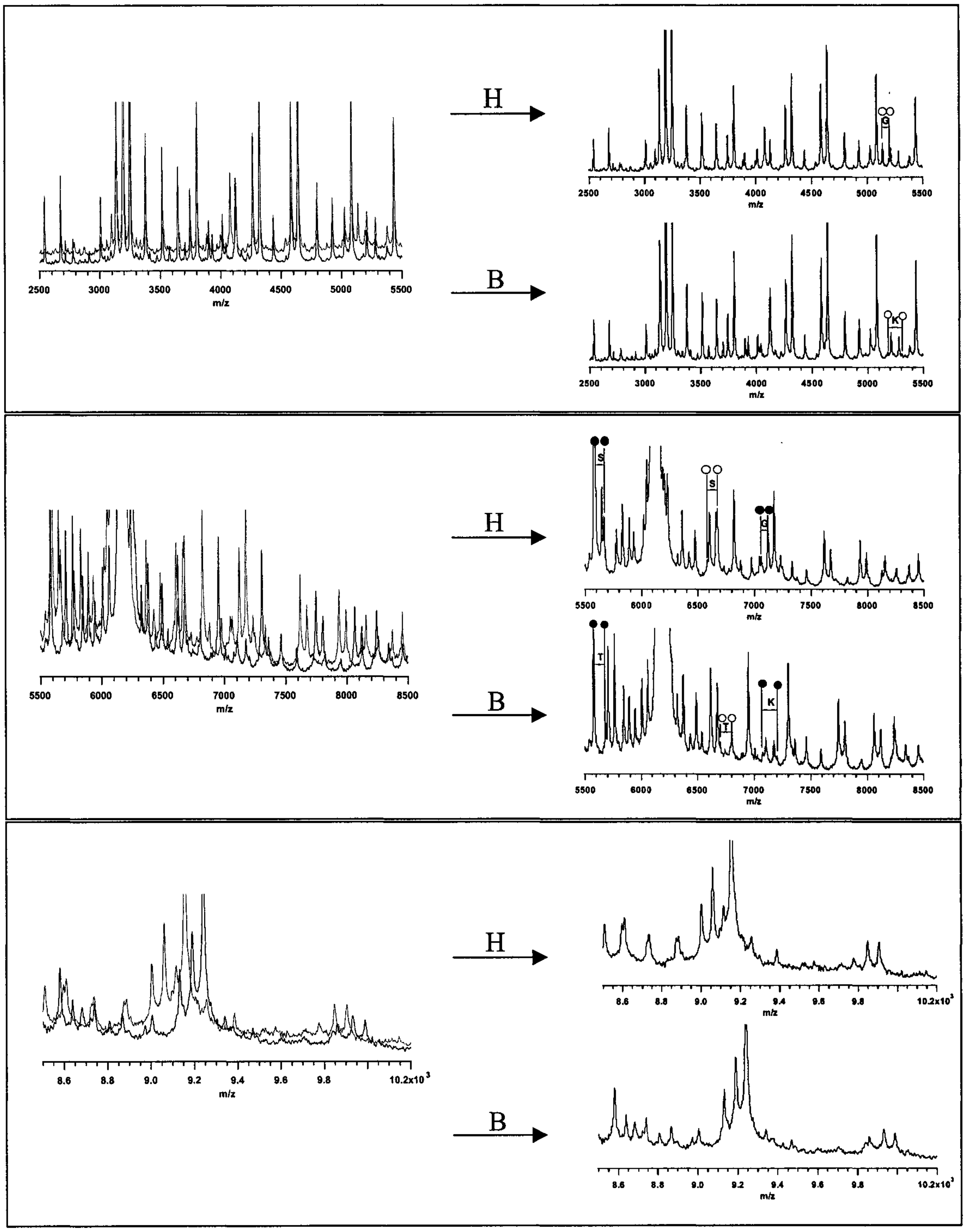


Figure III-4-3. Maps of cytochrome C from horse heart and bovine heart. Left panel is the overlaid spectra. The top spectra are from horse heart and bottom spectra are from bovine heart. Right panel is the separated spectra. The different amino acids are labeled in the spectra. "•" labeled peaks are from N-terminus and "o" labeled peaks are from C-terminus.

**Identification of amino acid variation by MAPs technique.** The intact protein of cytochrome C from horse heart and bovine heart was subjected to MAPs analysis. MALDI-TOF spectra of the resultant terminal ladders display significant differences that differentiate and locate the amino acid variation. Cytochrome C from horse heart was used as control. Figure III-4-3 is the MALDI-TOF spectra of sequence ladders generated by microwave irradiation. H indicates horse heart and B indicates bovine heart. The different amino acids at 47 and 60 produce a series of mass shifts from 5000 Da in both C-terminal ladders and N-terminal ladders. Compared with enzymatic digestion, the MAPs technique has no interference from background peaks resulting from autolysis of trypsin and the spectra clearly display masses shift and thus amino acid mutations. In the spectra of the tryptic peptides, there are only a few different peptides buried in other background peaks in addition to the incomplete sequence coverageWith MAPs technique, the two coexisting terminal ladders further complement each other and provides high sequence coverage.

### III. 4. 4. Conclusions

Currently the dominant view of the molecular basis of human disease is the concept that the accumulation of multiple mutations within genes of a single cell drives neoplastic transformation and ultimately leads to tomorigenesis. Detection of the mutations remains an important topic in functional proteomics. The MAPs technique presented in this work should be general and applicable to the rapid identification of amino acid mutations in many different kinds of proteins. This method provides a strategy to identify amino acid mutations at the protein level and it is useful for the high level of interest in establishing the phenotypic impact of polymorphisms and for the studies of the effect of nonsilent coding polymorphisms on function, expression level and turnover rate of the corresponding protein.

### III. 4. 5. Cited literature

(1) Wang, D. G. et al. *Science* **1998,** *280,* 1077-1082.

(2) Rieder, M. J.; Taylor, S. L.; Clark, A. G.; Nickerson, D. A. *Nat. Genet.* **1999,** *22,* 59-62.

(3) Collins, A.; Lonjou, C.; Morton, N. E. *Proc. Natl. Acad. Sci. USA.* **1999,** *96,* 15173-15177.

(4) Cargill, M. et al. *Nat. Genet.* **1999,** *22,* 231-238.

(5) Lacombe, C.; Riou, J.; Godard, C.; Rosa, J.; Galacteros, F. *Acta Haematol.* **1986,** *78,* 119-135.

(6) Riou, J. et al. *J. Clin. Chem.* **1997,** *43,* 34-37.

(7) Schneider, R. G.; Barwick, R. C. *Hemoglobin* **1982**, *6,* 199-208.

(8) Witkowska, H. E.; Bitsch, F.; Shackleton, C. H. *Hemoglobin* **1993,** *17,* 227-242.

(9) Center, I. H. I. *Hemoglobin* **1997,** *21,* 507-602.

(10) Feder, J. N. et al. *Proc. Natl. Acad. Sci. USA.* **1998,** *95,* 1472-1477.

(11) Jouanolle, A. M. et al. *Human Genet.* **1997,** *100,* 544-547.

(12) Feder, J. N. *Ann. Intern. Med.* **1999,** *130,* 953-962.

(13) Murray, K. K. *J. Mass Spectrom.* **1996,** *31,* 1203-1215.

(14) Koster, H. et al. *Nature Biotechnol.* **1996,** *14,* 1123-1128.

(15) Fu, D. J. et al. *Nature Biotechnol* **1998**, *16*, 381-384.

(16) Little, D. P.; Braun, A.; O' Donnell, M. J.; Koster, H. *Nat. Med.* **1997,** *3,* 1413-1416.

(17) Haff, L. A.; Smirnov, I. P. *Genome Res.* **1997,** *7,* 378-388.

(18) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001,** *73,* 5683-5690.

(19) Mann, M.; Jensen, O. N. *Nature Biotechnol.* **2003,** *21,* 255-261.

(20) Kelleher, N. L. *et al. J. Am. Chem. Soc.* **1999,** *121*, 806-812.

(21) Sze, S. K., Ge, Y., Oh, H., & McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*, 1774-1779.

(22) Sze, S. K.; Ge, Y.; Oh, H.; McLafferty, F. W. *Anal. Chem.* **2003,** *75,* 1599-1603.

(23) Fridriksson, E. K. et al. *Biochem.* **2000,** *39,* 3369-3376.

(24) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976-989.

(25) Mann, M.; Wilm, M. *Anal. Chem.* **1994,** *66,* 4390-4399.

(26) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995,** *67,* 1426-1436.

# Chapter 5. Analysis of Human Breast Cancer Cell Line MCF7 by Microwave-Assisted Ladder Generation[1]

In order to demonstrate that the MAPs technique is general and applicable to a broad range of proteins isolated from real world samples, proteins isolated from a human breast cancer cell line were separated and fractionated by two-dimensional HPLC and then subjected to MAPs analysis. Several modifications were discovered.

## III. 5. 1. Introduction

Despite tremendous advances in the understanding of the molecular basis of diseases such as cancer, huge gaps still exist in our understanding of disease pathogenesis and in the development of effective strategies for earlier diagnosis and earlier treatment [1]. Breast cancer is the most common malignancy among women, with a lifetime risk of more than 10% [2-8]. Discovery of disease-related posttranslational modification and inherited or environmental induced amino acid mutations are important for women's health worldwide.

[1] A version of this chapter will be submitted for publication as:
Hongying Zhong and Liang Li, "A new tool for disease proteomics ".

### III. 5. 2. Experimental Section

**Chemicals.** α-Cyano-4-hydroxycinnamic acid (HCCA), sinapic acid (SA), 2,5-dihydroxybenzoic acid (DHB), dithiothreitol (DTT) and HPLC grade acetone, methanol, acetonitrile, acetic acid, formic acid and trifluoroacetic acid (TFA) were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). HCCA was recrystallized from ethanol (95%) at 50°C prior to use. Analytical grade HCl was from Caledon Laboratories (Edmonton, Alberta, Canada). Water used in all experiments was from a NANOpure water system (Barnstead/Thermolyne). Phenylmethyl sulfonyl fluoride (PMSF) was purchased from Bioshop Canada Inc. (Burlington, ON). All protein standards were from Sigma (St. Louis, Mo).

**Fractionation of cytosolic proteins from human cancer cell line MCF7.** Human breast cancer cell lines, MCF7 cells, were grown in 15 cm diameter plates in ATCC media at 37°C for 2 weeks. The plates were aspirated with media to form a monolayer and placed on a cold metal tray. The plates were washed 3 times with 10 mL PBS$^{++}$ (0.9 mM $CaCl_2$ and 0.5 mM $MgCl_2$). 2.5 ml saponin lysis buffer (0.2% saponin in 50 mM pH=7.5 Tris-Cl with 1 mM PMSF) was added to each dish and left to incubate on ice for 5 min with constant rocking. The cells were scraped and the buffer solution was collected. The mixture was centrifuged at 12000 rpm for 15 min at 4°C. The supernatant contained the cytosolic proteins. DTT was added to the supernatant so that the final concentration of DTT was 20 mM. The solution was incubated for 1 hr at 37°C. Iodoacetamide was added to the final concentration of 20 mM and left to stand for 1 h at room temperature in the dark. All disulfide linkages should have been reduced and carbamidomethylated.

**2D HPLC separation.** Protein mixtures were first separated by cation ion exchange. The protein mixture was acidified by 0.1%TFA prior to cation ion exchange separation (Vydac 400VHP 81, 1mm i. d. ×150mm, 5 μm, 300 A°). A linear gradient over 60 minutes (A: 20% acetonitrile in 0.1% water. B: 1M NaCl in A) was used. Each fraction was further separated by reversed-phase HPLC (C8, Vydac, i. d. 1mm, 150 mm, 5 μm, 300 A°). A linear gradient over 60 minutes (A: 0.1% TFA in water. B: 0.1%TFA in acetonitrile) was used.

**Sequence ladder generation by controlled acid digestion.** The proteins of interest in HPLC fractions were concentrated by a vacuum centrifugation. 1 μl of 1.8M DTT was added and incubated at 37°C overnight. The sample volume was reduced down to about 0.2 μl by a vacuum centrifugation. An equal volume of 6 M HCl was added so that the final concentration of HCl was 3 M. The volume was kept as small as possible so that the concentration of proteins of interest would be high enough to be analyzed. The sample containing 3 M HCl was either irradiated by 900 W microwave for 1 minutes or heated at 95°C for 1 minute. Extra HCl was removed from the resultant polypeptides by vacuum centrifugation.

**Sample preparation for MALDI MS analysis.** The two-layer sample deposition method with 4-HCCA as matrix was used in the MALDI MS analysis. The first layer solution containing 12 mg/ml 4-HCCA in 20% methanol/acetone was deposited on a stainless metal target. The dried polypeptides were re-dissolved in 0.2 μl the 2nd layer HCCA solution (3mg recrystallized HCCA was dissolved in 125 μl DD water and 125 μl acetonitrile). The 2nd HCCA layer solution was pipetted a few times so that all the dried polypeptides would be re-dissolved into the 2nd layer solution of HCCA. The mixture with 4-HCCA and polypeptides was deposited just onto the first layer and air-dried. The spot was washed twice with DD water and air-dried.

**MALDI-TOF mass spectrometry.** MALDI-TOF MS was performed on a Bruker Reflex III Matrix-Assisted Laser Desorption Ionization time of flight mass spectrometer (Bremen/Leipzig, Germany) equipped with a SCOUT 384 multiprobe inlet and a 337nm nitrogen laser operated with a 3ns pulse in positive ion mode with delayed extraction using linear mode. The spectra were acquired by averaging 100-500 individual laser shots and processed with the Bruker supporting software. The spectra were semi-internal calibrated by spotting a standard protein very close to the sample spot and performing calibration. Each spectrum was reprocessed using the Igor Pro software package (WaveMetrics, Lake Oswego, Oregon, USA). All the spectra were normalized to the most intense signal in the displayed mass range.

**Data interpretation and database searching.** The resultant polypeptides are sequence ladders either from the C-terminal or N-terminal. Because the sum of each set of the data is equal to a constant (MH+18), the data can be pre-extracted and all the other fragment ions resulting from impurities or other internal fragmentation can be removed from sequence reading. Polypeptides with molecular mass less than 1500 Da were not taken into consideration because many fragment ions from the matrix are in this mass region. Sequences were developed from the extracted data and put into BLAST searching program in NCBI (The National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/BLAST/) for protein identification and modification detection. The mass calculated from the identified protein was then compared with the apparent mass in the MALDI-TOF mass spectrum to confirm the identification results.

### III. 5. 3. Results and Discussion

**Identification of oxidization sites in ubiquitin.** Figure III-5-1 shows the finding of ubiquitin modifications that play an important role in cancer development. Ubiquitin tagging of proteins for destruction that is an important stage in the body's defence against cancer [9, 10]. In order to make a comparison, another cell line from human cancer HT29 was prepared using the same method as that of MCF7. Protein extractions are first separated by a cation ion exchange chromatography. The same fraction was further separated by RP-HPLC. Figure III-5-1 (A) is the RP chromatography of the fraction from HT29 and Figure III-5-1 (B) is the RP chromatography from MCF7. MALDI-TOF spectra of peak α (Figure III-5-1 (C)) and peak β (Figure III-5-1 (D)) show that the intact protein mass of peaks α and β are 8581.93 and 8565.72 Da respectively and there is 16 Da mass differences. Based on the chromatographic the peak area, 25% of ubiquitin has been modified. Further MAPs analysis (Figure III-5-2) indicates that these two proteins have the same sequence ladder from the C-terminus (from 4-76 to 62-76). The N-terminus has the predicted amino acid sequence but all polypeptide ladders from 1-21 to 1-72 of α peak has 16Da mass shift. Thus this 16Da mass shift is associated with the first amino acid M. M was oxidized. Oxidatized forms of proteins have been demonstrated to accumulate during aging, oxidative stress and in some pathological conditions [11-15]. The huge amount of oxidized ubiquitin found in MCF7 is interesting and needs further study.

**Identification of acetylation at the N-terminus of proteins.** Figure III-5-3 is the MAPs analysis of a thymosin beta-4 containing hematopoietic system regulatory peptide. Acetylation was found on the first amino acid at N-terminus. Loss of acetylation at the N-terminus is clearly displayed in the spectra. Figure III-5-3 (A) is the MALDI-TOF spectrum of the intact protein and (B) is the MALDI-TOF spectra of sequence ladders generated by microwave irradiation. It has been reported that N-terminal acetylation plays an important role in protein stability, protection of N-terminus and regulation of protein-DNA interactions (histones) [16].

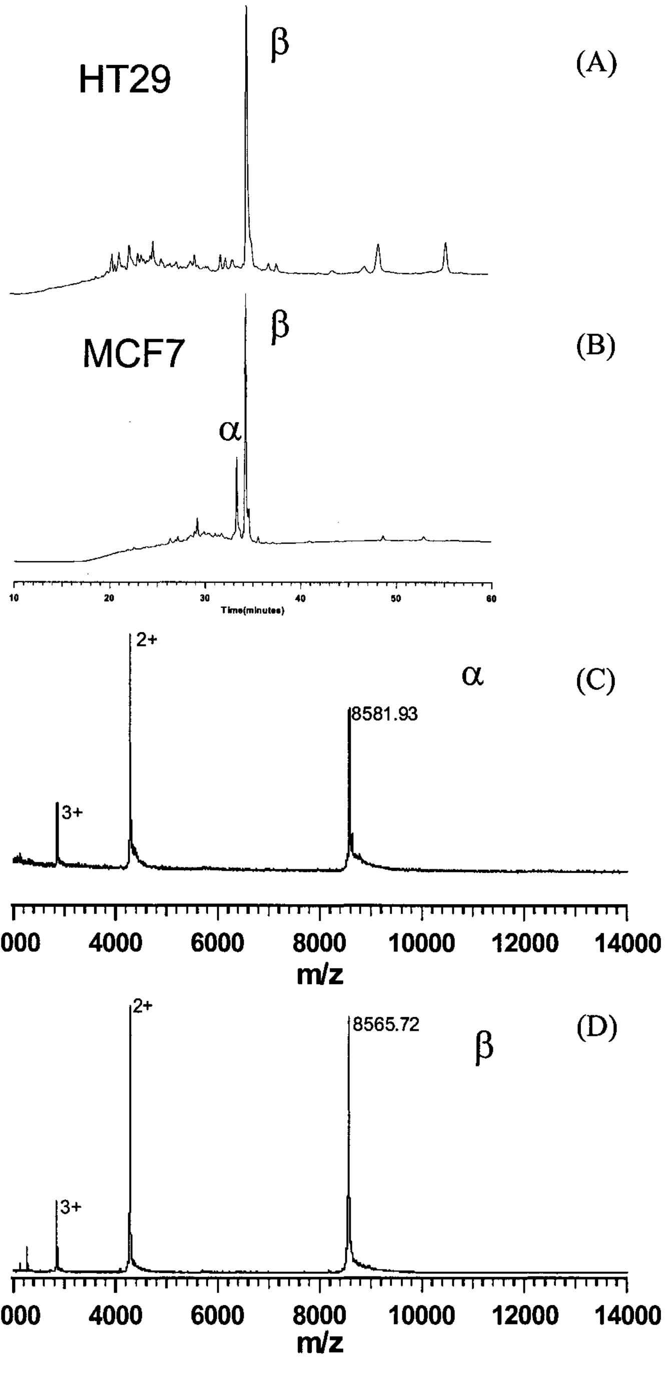
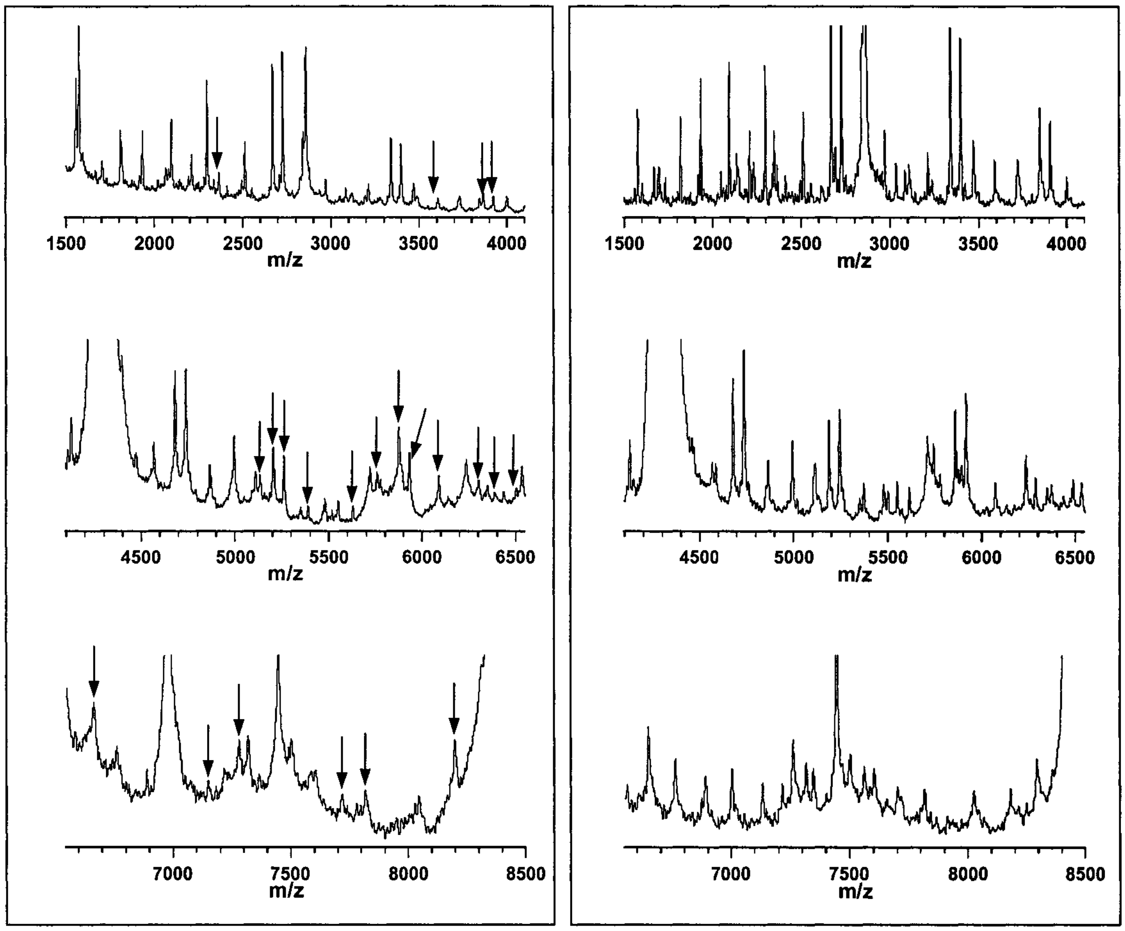


Figure III-5-1. RP chromatography and and MALDI-TOF spectra of cancer cell fractions. (A) RP chromatography of a cation ion exchange fraction of HT29 cell. (B) RP chromatography of the same cation ion exchange fraction of MCF7 cell. (C) MALDI-TOF spectrum of α peak. (D) MALDI-TOF spectrum of β peak.



MAPs of α peak

MAPs of β peak

Figure III-5-2. MAPs of the isolated peaks of MCF7. Left panel ARE the MALDI-TOF spectra of α peak of MCF7. Right panel are the MALDI-TOF spectra of β peak of MCF7. The peaks that have a 16 Da mass shift are labeled by arrows in the left panel spectra.

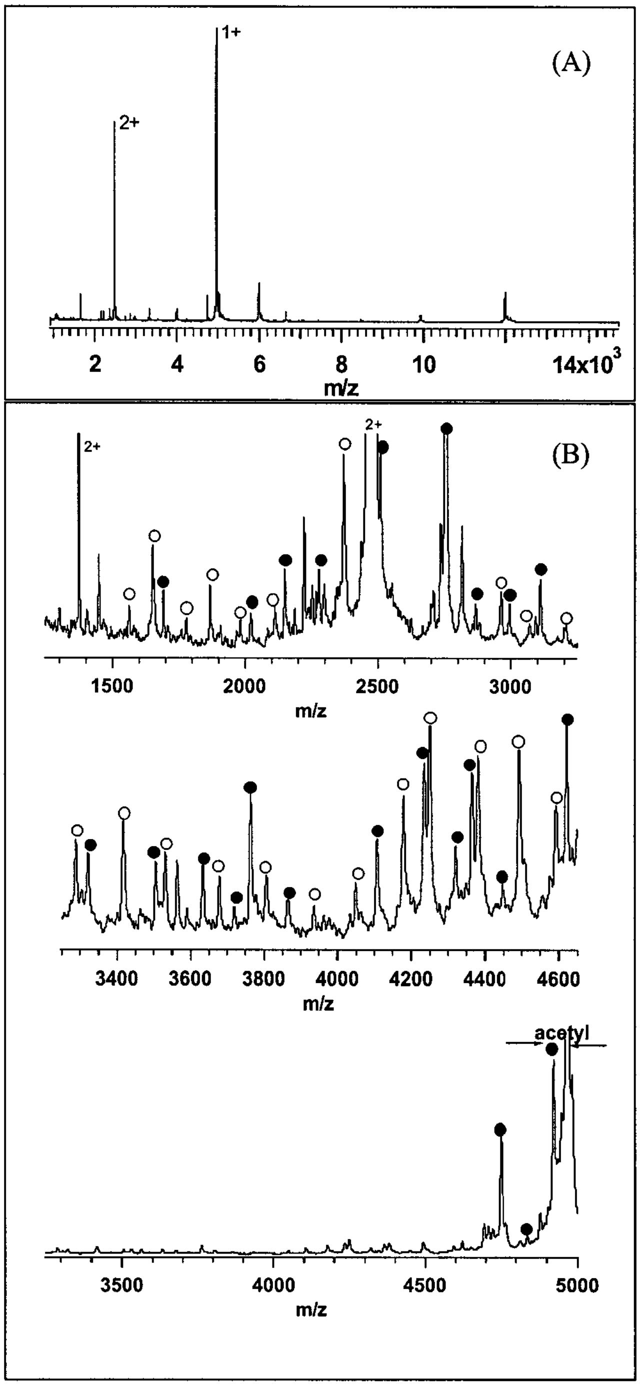


Figure III-5-3. MAPs analysis of N-terminus acetylated protein from MCF7 (A) is the MALDI-TOF spectrum of the intact protein. (B) is the MALDI-TOF spectra of sequence ladders generated by microwave irradiation.

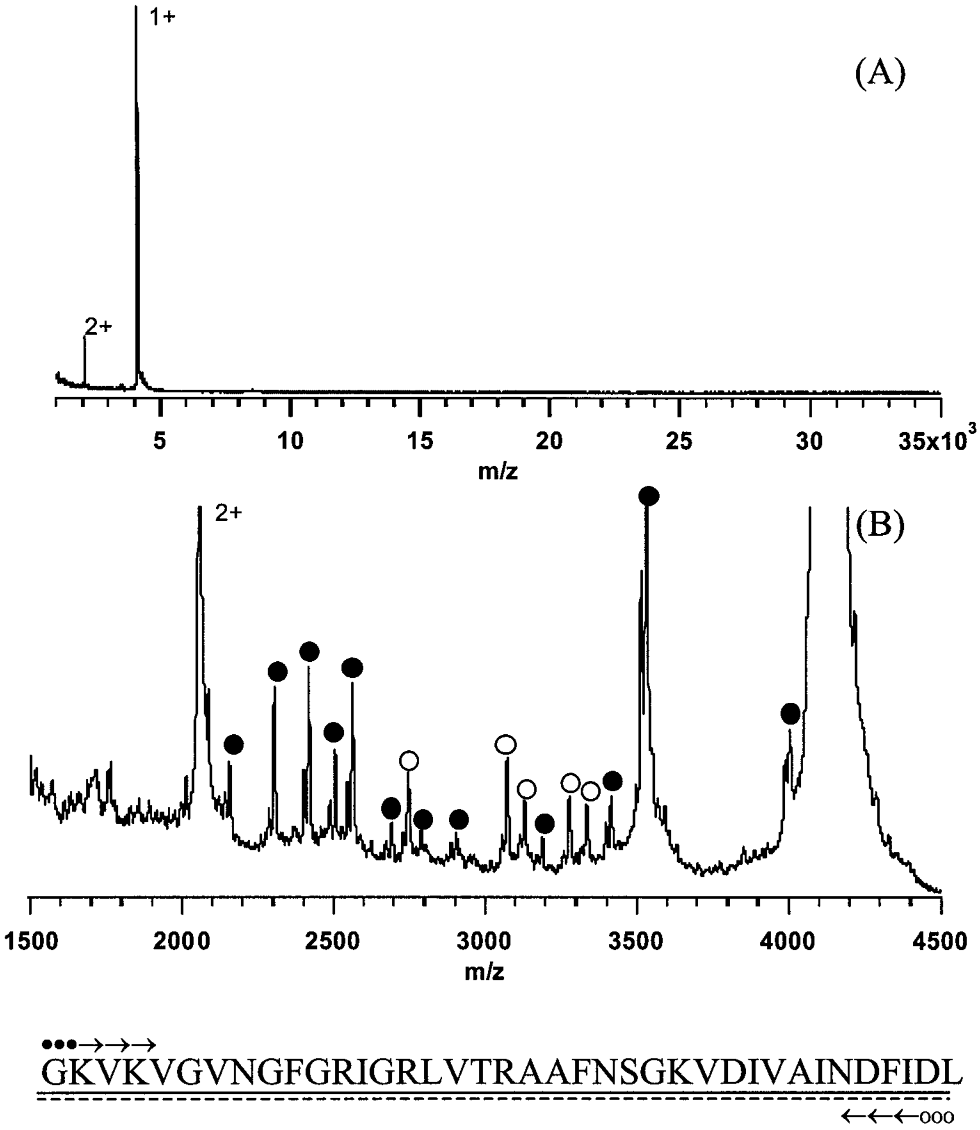


Figure III-5-4. MAPs analysis of a fragment from G3P2-HUMAN isolated from breast cancer cell line MCF7. (A) MALDI-TOF spectrum of the intact protein. (B) MALDI-TOF spectrum of the sequence ladder generated by microwave irradiation. Solid underline indicates the sequence coverage from the N-terminus and dashed underline indicates the sequence from the C-terminus. "•" labeled peaks are from the N-terminus and "o" labeled peaks are from the C-terminus.

**Identification of protein fragments from large proteins.** Figure III-5-4 is the MAPs analysis of a fragment from the large protein G3P2-HUMAN (PP04406). The identification results indicate that this small peptide is from the N-terminus (1-39) of G3P2-HUMAN that has a molecular weight of 36 kDa. It maybe a degradation product from G3P2 [17-18] but further detailed biological studies is needed to find out the reason.

### III. 5. 4. Conclusions

The MAPs technique is efficient at identifying posttranslational modifications in cancer research. Combined with chromatography, it is possible to isolate and identify both the identity of posttranslational modifications and the quantitative level of posttranslational modifications. Once post-translational modifications (PTMs) analysis can routinely be done at the proteomics level, the involvement of PTMs in disease can be studied much more systematically than has thus far been possible. Although many examples of PTMs in disease are known, it is very likely that these examples are just the top. Proteomic PTM analysis by MAPs technique will thus contribute to our understanding of disease etiology and deliver many new targets for research against disease.

### III. 5. 5. Cited literature

(1) Wulfkuhle, J. D.; Liotta, L. A.; Petricoin, E. F. *Nature Rev. Cancer* **2003,** *3,* 267-275.

(2) Narod, S. A. *Nature Rev. Cancer* **2002,** *2,* 113-121.

(3) King, M. C.; Marks, J. H.; Mandell, J. B. *Science* **2003,** *302,* 643-646.

(4) Mitra, K.; Marquis, J. C.; Hillier, S. M.; Rye, P. T.; Zayas, B.; Lee, A. S.; Essigmann, J. M.; Croy, R. G. *J. Am. Chem. Soc.* **2002,** *124,* 1862-1863.

(5) Brown, K. J.; Fenselau, C. *J. Proteome Res.* **2004,** *3,* 455-462.

(6) Woodbury, R. L.; Varnum, S. M.; Zangar, R. C. *J. Proteome Res* **2002,** *1,* 233-237.

(7) Zhu, K.; Kim, J.; Yoo, C.; Miller, F. R.; Lubman, D. M. *Anal. Chem.* **2003,** *75,* 6209-6217.

(8) Lee, K. J. *et al. J. Am. Chem. Soc.* **2002,** *124,* 12439-12446.

(9) Ghosh, M.; Huang, K.; Berberich, S. J. *Biochem.* **2003,** *42,* 2291-2299.

(10) Kamura, T. *et al. Science* **1999,** *284,* 657-661.

(11) Dean, R. T.; Fu, S.; Stocker, R.; Davies, M. J. *Biochem.* **1997,** *324,* 1-18.

(12) Barlow, J. N.; Zhang, Z.; John, P.; Baldwin, J. E.; Schofield, C. J. *Biochem.* **1997,** *36,* 3563-3569.

(13) Berlett, B. S.; Stadtman, E. R. *J. Bio. Chem.* **1997,** 272, 20313-20316.

(14) Ishikawa, Y.; Yamamoto, Y.; Otsubo, M.; Theg, S. M.; Tamura, N. *Biochem.* **2002,** *41,* 1972-1980.

(15) Levine, R. L.; Mosoni, L.; Berlett, B. S.; Stadtman, E. R. *Proc. Natl. Acad. Sci. USA.* **1996,** *93,* 15036-15040.

(16) Mann, M.; Jensen, O. N. *Nature Biotechnol.* **2003,** *21,* 255-261.

(17) Kim, M. S.; Yoo, K. J.; Kang, I.; Chung, H. M.; Baek, K. H. *Int. J. Oncol.* **2004**, *25,* 373-379.

(18) Wasinger, V. C.; Smith, I. H. *FEMS Microbio. Letters* **1998,** *169,* 375-382.

# Part IV. Analysis of Membrane Protein Sequences and Detection of Posttranslational Modifications

# Chapter 1. Membrane Proteome Analysis by Microwave-Assisted Acid Hydrolysis of Proteins and Liquid Chromatography MALDI-MS/MS[1]

Simple and efficient digestion of proteins, particularly hydrophobic membrane proteins, is of significance for comprehensive proteome analysis using the bottom-up approach. A microwave-assisted acid hydrolysis (MAAH) method for rapid protein degradation for peptide mass mapping and tandem mass spectrometric analysis of peptides for protein identification is reported in this chapter. It uses 25% trifluoroacetic acid (TFA) aqueous solution to suspend membrane proteins, followed by microwave irradiation for 10 minutes. This detergent-free method generates peptide mixtures that can be directly analyzed by liquid chromatography (LC) matrix-assisted laser desorption ionization (MALDI) mass spectrometry (MS) without the need of extensive sample cleanup. LC-MALDI MS/MS analysis of the hydrolysate from 5 μg of a model transmembrane protein, bacteriorhodopsin, resulted in almost complete sequence coverage by the peptides detected, including the identification of two posttranslational modification sites. Cleavage of peptide bonds inside all 7 transmembrane domains took place, generating peptides of sizes amenable to MS/MS to determine possible sequence errors or modifications within these domains. Cleavage specificity, such as glycine residue cleavage, was observed. Terminal peptides were found to be present in relatively high abundance in the hydrolysate, particularly when low concentrations of proteins were used

[1] A version of this chapter has been submitted for publication as:
Hongying Zhong, Sandra Marcus and Liang Li, "Membrane Proteome Analysis by Microwave-Assisted Acid Hydrolysis of Proteins and Liquid Chromatography MALDI-MS/MS"· Dr. Sandra Marcus performed the cell culture and membrane protein fractionation.

for MAAH. It was shown that these peptides could still be detected from MAAH of bacteriorhodopsin at a protein concentration of 1 ng/μl or 37 fmol/μl.

### IV. 1. 1. Introduction

Membrane proteins play an important role in many biological processes. However, the intrinsic difficulty of solubilization makes them difficult to be analyzed by proteomic methods developed for soluble proteins [1, 2]. Insoluble membrane protein aggregates cannot be readily degraded by proteases. Therefore buffers containing concentrated urea, detergents and other salts are often chosen for membrane protein solubilization. In addition to the interference encountered with mass spectrometric detection, concentrated urea or detergents denature many proteases and decrease their activity to cleave proteins.

There are a number of reports describing technical advances to enable membrane protein analysis using the shotgun proteomics approach in which proteins are dissolved in a buffer containing surfactants or in an organic solvent, followed by protein digestion and liquid chromatography (LC) tandem mass spectrometry (MS/MS) analysis of the peptides [3-11]. For example, Han *et al.* used 0.5% SDS to solubilize a membrane-enriched microsomal fraction and then did trypsin digestion in diluted SDS solution [3]. While trypsin is commonly used for protein digestion, the limited tryptic cleavage sites in hydrophobic domains make it difficult to generate peptides with optimal size to cover all of the transmembrane domains. Cyanogen bromide (CNBr)-mediated enzymatic digestion [4-7] has been used to circumvent this problem, although CNBr is highly toxic.

Norris *et al.* utilized a novel cleavable detergent, PPS (3-[3-(1,1-bisalkyloxyethyl)pyridin-1-yl]propane-1-sulfonate), to analyze membrane proteins and cell lysates [8] and found that, following cleavage, detection sensitivity and spectral quality from matrix-assisted laser desorption ionization (MALDI) MS are significantly improved for dilute solutions of even moderately soluble proteins. Washburn *et al.* developed a surfactant-free method that used 90% formic acid to solubilize proteins in the presence of CNBr, with further enzymatic digestion of the CNBr-cleaved protein fragments by LysC and trypsin [6]. Blonder *et al.* reported a method of using 60% methanol to extract, solubilize and digest membrane proteins with trypsin [9-11]. More recently, Wu *et al.* reported a method for comprehensive membrane protein analysis using non-specific Proteinase K for digestion and subsequent analysis by LC electrospray ionization (ESI) MS/MS [12]. Using this method, not only extensive sequence coverage of membrane proteins could be obtained but also the protein topology in cells could be determined [12].

Recently, acid-catalyzed hydrolysis has been developed as an alternative technique to degrade the proteins into small peptides for MS analysis. The analytical application of acid hydrolysis for generating peptides can be dated back to the classic protein sequencing work of Sanger [13]. Prior to the use of MS as a major tool for protein identification, limited acid hydrolysis had been used as a means of generating peptides from proteins separated by SDS-PAGE for peptide mapping and Edman microsequencing [14, 15]. With recent advances in MS for proteome analysis, several groups have explored the use of limited acid hydrolysis to generate peptides for peptide mass mapping or tandem MS analysis [16-24]. A variety of experimental conditions including the use of acids of different strengths, organic and inorganic acids, and liquid and vapor phases have been studied for optimal degradation of proteins in solution, in gel, or surface-bound. For example, Anqun Li *et al.* demonstrated that using formic acid, cleavage at aspartyl residues of proteins was efficient and specific for both soluble and insoluble proteins [21]. Shevchenko *et al.* reported the use of 6 M HCl for in-gel digestion of bacteriorhodopsin (BR) [22] and demonstrated that acid hydrolysis was superior to that of trypsin digestion for handling this very hydrophobic membrane protein. MALDI hybrid quadrupole time-of-flight (Qq-TOF) mass spectrometry was used to generate product ion spectra of peptides extracted from the in-gel acid hydrolysis method.

In this work, a method for membrane proteome analysis was discussed. It is based on microwave-assisted acid hydrolysis using trifluoroacetic acid (TFA) for protein degradation, followed by LC-MALDI MS/MS of the resultant peptides. The analytical performance of this method is demonstrated using the integral membrane protein, BR, as a model system. The experiments were performed on a LC-MALDI MS/MS system with a heated droplet interface [25].

### IV. 1. 2. Experimental Section

**Materials and Reagents.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For HPLC separation, MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used

(Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany. Human breast cancer cell line, MCF7 cells (ATCC HTB-22), was purchased from the American Type Culture Collection (Manassas, VA).

**Acid Hydrolysis.** In the experiments on method development and evaluation of protein sequence coverage by the peptides generated from BR using acid hydrolysis, a stock solution of BR was prepared by suspending 1 mg of BR in 1000 µl of an acid such as 25% TFA with 20 mM DTT added in a 1.5 ml vial. The reason for adding DTT was to avoid oxidization of methionine, tryptophan and other amino acids. 10 µl of the protein suspension was placed in a 1.5 ml polypropylene centrifuge vial, capped and sealed with Teflon tape. The vial was placed in a domestic 900W (2450 MHz) microwave oven. 100 ml of water in a loosely covered container was placed besides the sample vial to absorb extra microwave energy. 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 vacuum centrifuge to remove the acid. The BR digest was re-suspended in 100 µl of 0.1% TFA aqueous solution and centrifuged at 16000 g for 5 minutes to remove any possible residual particles. 50 µl of the solution was injected into the HPLC system for analysis.

For the sensitivity and acid hydrolysis efficiency experiments, the 1 mg/ml stock solution of BR was diluted to various concentrations, from which 10 μl samples were irradiated for 10 minutes.

Care was taken when handling the concentrated acids as well as other precautions that were followed in the use of the microwave oven (see Safety Considerations).

**HPLC Separation**. Reversed-phase HPLC separations of the peptides were made using a Vydac C18 column (1 mm i. d. × 150 mm, 5 μm, 300 A°; Vydac, Hesperia, CA) on an Agilent 1100 capillary HPLC system. At a flow rate of 50 μl/min, a linear gradient from 5% B to 85% B over 70 minutes was used, where mobile phase A was water containing 0.1% (v/v) TFA, and mobile phase B was acetonitrile with 0.1% (v/v) TFA. During the separation period from 10 to 50 minutes, the eluate was fractionated at 30 s intervals and directly deposited to the MALDI target (Applied Biosystems, Boston, MA) using the heated droplet interface [25].

**MALDI MS and MS/MS**. For peptide mass mapping by MALDI time-of-flight (TOF) MS, a two-layer sample/matrix preparation method was used [26] with α-cyano-4-hydroxy-cinnamic acid (HCCA) as matrix. 0.7 μl of the first-layer matrix solution containing 12 mg/ml of HCCA in 20% methanol/acetone was deposited on the MALDI target and air-dried. 0.5 μl of sample was mixed with 0.5 μl of matrix (50% acetonitrile/water saturated with HCCA) and then deposited onto the first layer. The

MALDI-TOF mass spectra were obtained on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany).

For MALDI MS/MS analysis of the HPLC fractions, the peptide samples were prepared using a dried droplet method, in which 2,5-dihydroxybenzoic acid (DHB) was used as the matrix. 0.3 μl of matrix solution in 50% acetonitrile/water saturated with DHB was put into each well and then pipetted several times before spotting on the target. Product ion spectra of peptides were obtained in a QSTAR MALDI Qq-TOF mass spectrometer (MDS Sciex, Ontario).

**Data Interpretation and Database Searching.** Database searching using MS/MS spectra was performed by MASCOT (http://www.matrixscience.com). All the database searching was done against SwissProt using no specification of enzyme type. Methionine oxidization, carbamidomethylation of cysteine, and deamidation of asparagine and glutamine were set as variable modifications. Potential protein candidates with the highest MOWSE scores were determined from database search. The MS/MS spectra of the matched peptides were examined manually to see if the major peaks observed were matched with the expected fragmentation patterns. If they agreed well, the identification was considered to be positive.

**Hydropathy Calculations.** Proteins identified were examined using the ProtParam program available at the EXPASY web site (http://us.expasy.org/tools/protparam.html) that allows for calculation of the grand average of hydrophobicity (GRAVY). Positive values are considered as hydrophobic and negative values are considered as hydrophilic.

**Safety Considerations.** Microwave ovens have been used extensively for protein hydrolysis to generate amino acids for amino acid composition analysis [27]. Procedures and cautions in handling liquid samples for microwave experiments can be found in the literature [27]. Although the sample volume used in this work was generally small (10 μl), care was still exercised. The microwave oven was turned off and unplugged from the power after each use. In case of microwave oven malfunction that may result in continued radiation generation even after it is switched off, unplugging from the power should prevent any potential accident from such failure [27]. Care was also taken in handling acids and possible acid vapors during and after microwave hydrolysis. After microwave irradiation of the sample vial it was allowed to cool inside the microwave oven. The cooled sample vial was opened in a well-vented bench area, such as inside a fume-hood.

### IV. 1. 3. Results and Discussion

**Microwave-assisted acid hydrolysis vs. other methods.** Figure VI-1-1 shows the workflow of the proteome analysis method. Membrane proteins are first extracted from cell lysates using a proper solvent system usually containing surfactants. The sample is then reduced and alkylated, followed by acetone-precipitation. The resulting powder is washed to remove salts, buffers, and soluble proteins. Instead of using surfactants or organic solvents to re-dissolve the proteins, in our method, the protein powder is suspended in an acid such as TFA and then subjected to microwave irradiation. The hydrolysate generated from this microwave-assisted acid hydrolysis (MAAH) process is directly analyzed by LC MS and MS/MS. In this work, LC-MALDI TOF and LC-MALDI Qq-TOF are used for peptide mass mapping and protein identification by MS/MS database searching, respectively.

The ability to hydrolyze proteins from a protein suspension eliminates the need to use strong surfactants, facilitating the downstream mass spectrometric analysis. Strong surfactants are known to cause interferences in MS analysis and must be carefully removed. For example, in the work of Hixson *et al.* [28], 500 µg of BR was mixed with 10 µl of 260 mM SDS for initial solubilization. This was followed by addition of 100 µl of buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, and 5 mM DTT for further solubilization. The dissolved BR was digested by trypsin for LC-ESI MS/MS. However, prior to MS analysis, concentrated urea and SDS had to be removed by an overnight dialysis process. Nevertheless, using this approach, they detected an extensive list of peptides that provided complete sequence coverage of BR [28].

MAAH can significantly increase the speed of protein degradation for proteome analysis using database searching of product ion spectra of peptides. As demonstrated below, MAAH of proteins are generally completed within 10 min, as opposed to overnight enzymatic digestion. An even longer digestion time is required for enzymatic digestion involving the use of surfactants or organic solvents to solubilize the proteins, due to the reduction of enzyme activities. While conventional acid hydrolysis can shorten the protein degradation time, for membrane protein analysis, MAAH offers a distinct advantage over other acid hydrolysis methods, i.e., it provides more efficient cleavage inside transmembrane domains. For example, organic acids such as 90% formic acid have been used for membrane protein solubilization, followed by hydrolysis with or without heating of the protein/acid mixture. Protein hydrolysis using formic acid yields peptides resulting from the preferential cleavage of hydrophilic aspartyl residues, especially the D-P bond. The method appears to work well with hydrophilic proteins and some membrane proteins [21] and it is found to be very useful for generating large peptides for mapping posttranslational or chemical modifications of proteins (as described in PART III, Chapter3). However, the method provides limited cleavage sites at the hydrophobic transmembrane domains where D-P bonds are not expected to be present. In addition, N-terminal formylation of peptides is extensively observed. This is shown in Figure IV-1-2A where BR was mixed with 70% formic acid [29] and heated at 110°C for 4 hours. The MALDI spectrum of the hydrolysate (from about 200 ng of BR) displays a number of modified peptides and the overall signal-to-noise ratio of the spectrum is not good. Decreasing the formic acid concentration can decrease the extent of N-terminal formylation, but solubilization of membrane proteins decreases as well, resulting in poor acid cleavage efficiency. For example, no peptide peaks were detected when the mixture of 5% formic acid and BR, after heating at 110°C for 4 hours, was analyzed.

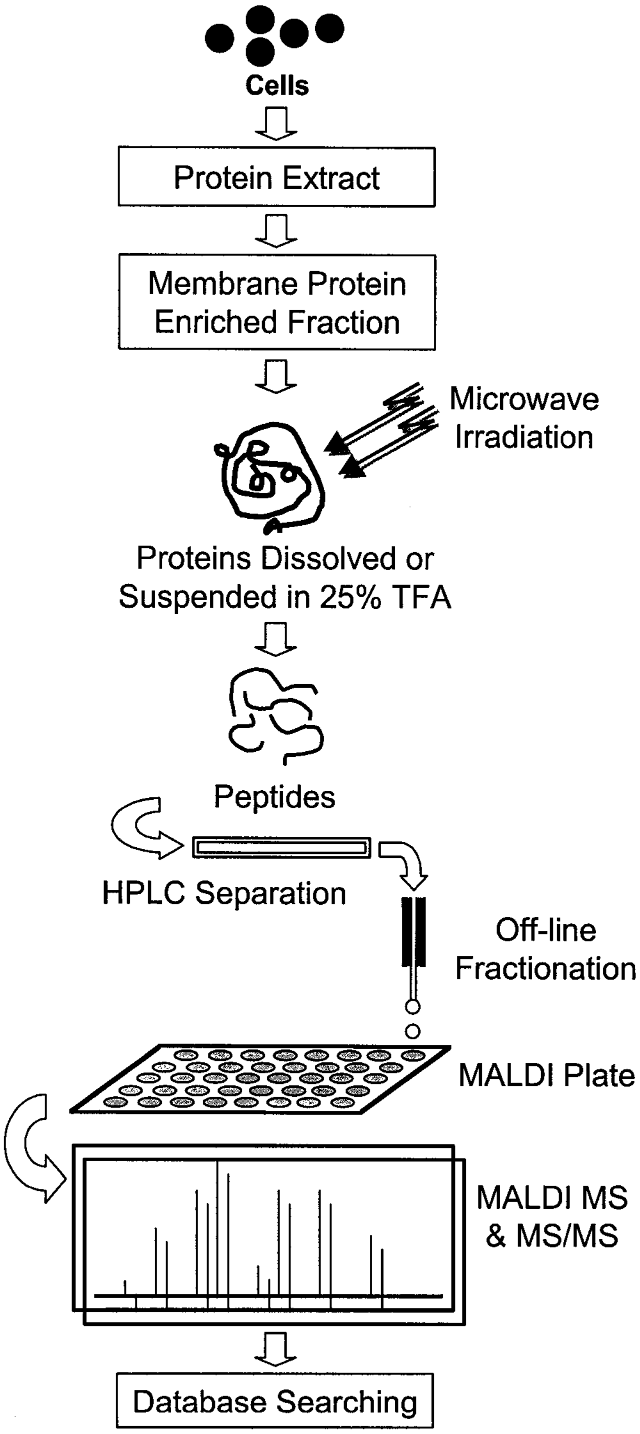


Figure IV-1-1. Workflow for the analysis of enriched membrane proteins by microwave-assisted acid hydrolysis for protein degradation, HPLC separation of peptides, and MALDI MS/MS with database search for protein identification.

MAAH can overcome the solubility problem associated with conventional acid hydrolysis. This is best illustrated in acid hydrolysis of BR using TFA. Figure IV-1-2B shows the MALDI spectrum of BR hydrolyzed in 25% TFA by conventional heating at 110°C for 4 hours. Only one peptide peak from the N-terminus was observed. It was found that upon heating, BR aggregated, which might have prevented further hydrolysis. In contrast, when the same mixture of 25% TFA and BR was subjected to MAAH, no protein aggregates were observed after 10 minutes of microwave irradiation. Figure IV-1-2C shows the MALDI spectrum of the BR hydrolysate from MAAH. A number of peptide peaks with their masses assignable to BR fragments were observed with good signal-to-noise ratios. One plausible explanation for this striking difference between MAAH and acid hydrolysis with conventional heating is that microwave irradiation can assist in disrupting the protein aggregates in cases where proteins are merely suspended in an acid, preventing the solubilized proteins from aggregation, and denaturing and unfolding the proteins to expose both hydrophobic and hydrophilic residues to acid hydrolysis.

**Acid Type and Concentration.** To optimize the performance of MAAH, we examined several different acids at varying concentrations to hydrolyze BR. A proper acid for MAAH should not produce any interfering side reactions and be easily removed by vacuum centrifugation. $HNO_3$ is oxidizing acids so it was not chosen for the experiments. $H_3PO_4$ and $H_2SO_4$ are not volatile and thus cannot be removed by vacuum centrifugation. Formic acid (shown above) and acetic acid (data not shown) were found

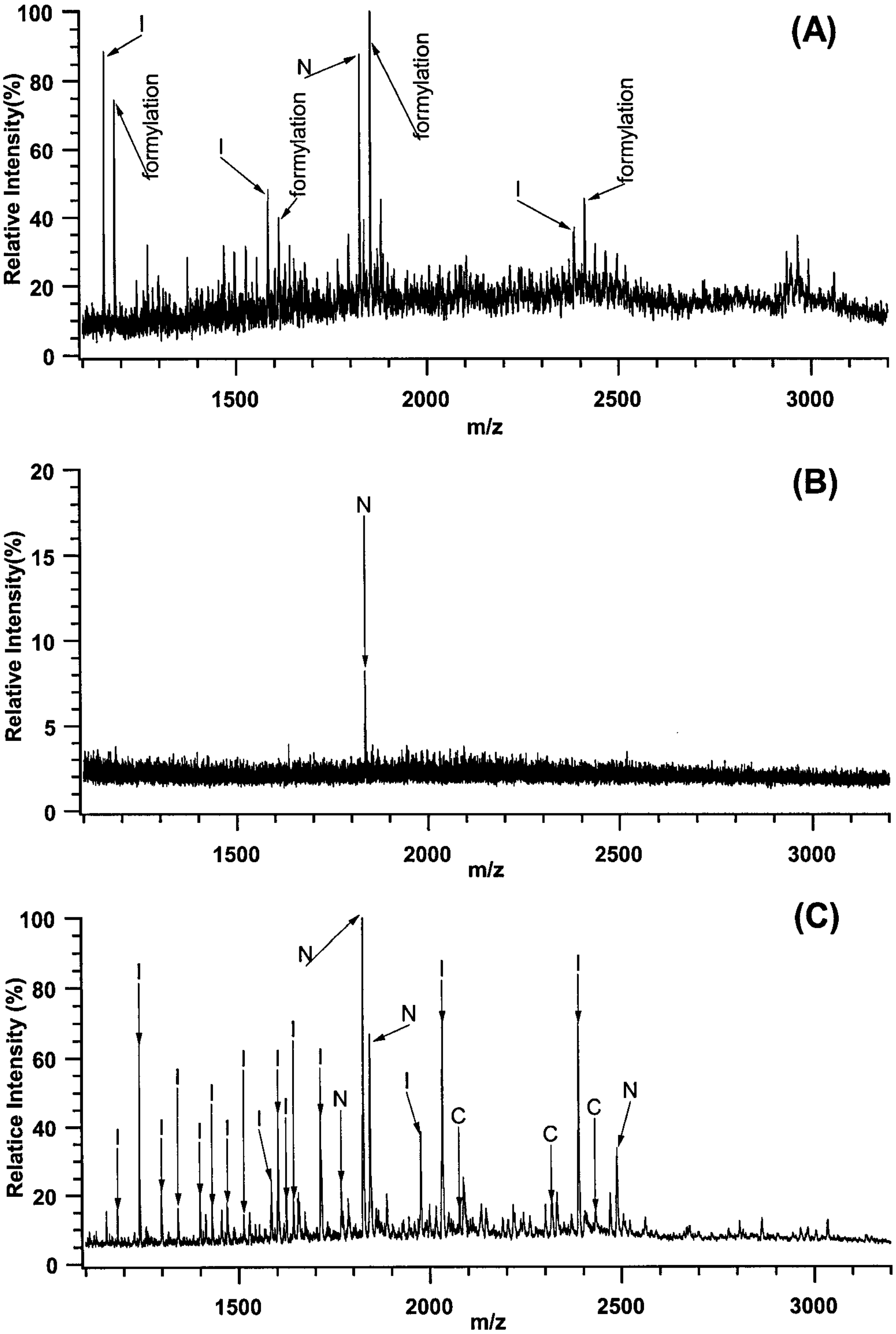


Figure IV-1-2. MALDI MS spectra of the peptides generated from the hydrolysis of 200 ng of bacteriorhodopsin under different conditions: (A) 70% formic acid heated at 110°C for 4 hours, (B) 25% TFA heated at 110°C for 4 hours, and (C) 25% TFA digested by microwave irradiation for 10 minutes. "I" indicates peptides resulting from internal fragmentation. "C" and "N" indicate C- and N-terminal peptides, respectively.

to modify the N-terminus of peptides so they were not studied further. In the end, HCl and TFA were selected for in-depth studies to determine their suitability for MAAH.

It was found that acid type, acid concentration and microwave irradiation time all affect the number and type of peptides produced from MAAH of BR. Table IV-1-1 summarizes the conditions of two level experiments. Table IV-1-2 summarizes the experiment design and the corresponding results from MALDI-TOF peptide mass mapping. Low acid concentration (e.g., 0.1 M HCl or 0.3 M TFA) and short irradiation time (e.g., 2 minutes) resulted in fragments containing the C- and/or N-terminus. As increasing acid concentration and irradiation time, more internal fragment ions were observed, along with the C- and N-terminal fragment ions. At a higher concentration (e.g., 1.5 M HCl or TFA) and a longer irradiation time (e.g., 10 minutes), HCl resulted in more non-specific cleavages (see next section). In addition, 1.5 M TFA gave a better signal-to-noise ratio than that obtained with 1.5 M HCl.

Table IV-1-1. Experiments of two levels for each of the three factors.

| Factors | Value of (-) | Value of (+) |
|---|---|---|
| Acid (A) | TFA | HCl |
| Concentration (C) | 0.3 M | 1.5 M |
| Irradiation time (I) | 2 minutes | 10 minutes |

Table 2. Experimental design and corresponding results.

| Run | A | C | I | Interaction | A | C | I | MALDI-TOF spectra |
|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | -1 | 0 | 0 | 0 | no signal |
| 2 | + | - | - | A | 1 | 0 | 0 | C-terminal fragmentation |
| 3 | - | + | - | C | 0 | 1 | 0 | C-and N-terminal fragmentation |
| 4 | + | + | - | AC | 1 | 1 | 0 | C-and N-terminal fragmentation+ internal fragmentation |
| 5 | - | - | + | I | 0 | 0 | 1 | C-and N-terminal fragmentation + internal fragmentation |
| 6 | + | - | + | AI | 1 | 0 | 1 | C-terminal fragmentation + internal fragmentation |
| 7 | - | + | + | CI | 0 | 1 | 1 | C-and N-terminal fragmentation + internal fragmentation |
| 8 | + | + | + | ACI | 1 | 1 | 1 | Internal fragmentation + non-specific fragmentation |

From these results, we concluded that TFA was a better choice for generating relatively more specific internal cleavage products and producing optimal peptides for tandem MS analysis. We then carried out a detailed optimization of acid concentration and digestion time studies for TFA. It was found that 3 M TFA (25% v/v) digestion for 10 min gave the optimal results in terms of the number of peptides generated and sequence information covered for BR.

It is worth noting that some of the observations found during the acid optimization experiments seem to suggest that protein denaturation may play an important role in hydrolysis. Protein denaturation is likely facilitated by heating resulting from microwave irradiation while hydrolysis of polypeptide bonds takes place. For example, 0.3 M TFA digestion for 2 minutes did not yield any signal while 1.5 M TFA digestion for the same irradiation time yielded peptides from both the C- and N-terminus. However, 0.3 M HCl digestion for 2 minutes of irradiation yielded peptides from the C-terminus, while 1.5 M

HCl digestion for the same irradiation time yielded peptides from both the C- and N-terminus as well as peptides resulting from internal fragmentation. These observations can be attributed to 0.3 M TFA and heating for 2 minutes irradiation being insufficient to denature the protein. So membrane proteins under these "mild" conditions were not exposed to the cleavage of polypeptide bonds. Increasing acid concentration or increasing irradiation time yielded hydrolytic peptides from N- and C-terminus. Further increases in acid concentration or irradiation time not only generated hydrolytic peptides from the C- and N-terminus but also yielded hydrolytic peptides from internal fragmentations.

**Acid Cleavage Specificity**. It was found that in using TFA for MAAH there was acid-cleavage specificity to produce the peptides. This is consistent with those reported by Gobom *et al.* [20] using direct MALDI analysis of proteins digested by acid vapors. In our experiment, BR was digested by 25% TFA with microwave irradiation for 10 minutes. To examine the acid-cleavage specificity and whether the very hydrophobic domains of a membrane protein could be cleaved by the acid, the hydrolysate of BR was separated by C18 reversed-phase chromatography, followed by direct deposition of the fractions to the MALDI target. MALDI MS and MS/MS spectra of peptides on each fraction were then obtained and the peptides were identified by MS/MS database searching. Figure IV-1-3 shows some of the representative MALDI MS spectra.

One interesting finding from the BR acid hydrolysis experiment is the readiness of the cleavage on both sides of glycine using 25% TFA for digestion. TFA is an ion-pairing reagent commonly used as a modifier for HPLC separation of peptides and proteins. It is likely that the extent of the formation of TFA/protein complex plays some role in hydrolysis. Glycine has the least steric effect on TFA/proton attachment to the peptide bond which is presumed to initiate and catalyze the hydrolysis process. In the MALDI MS spectra, peptide pairs with 57 Da mass differences produced by the cleavage on both sides of glycine are always abundant. Even after HPLC separation, as shown in Figure 3A as an example, pairs of peptides are still observed in the same fraction, suggesting peptides with and without a terminal glycine have similar chromatographic retentions. This glycine cleavage specificity can be used to aid in protein identification. This can be illustrated using the MALDI MS spectrum of fraction 34 shown in Figure 3A. The monoisotopic mass of the peak labeled as 85-97 is 1525.66 Da. For peptide mass mapping, this peak matches 6 peptides from BR with 200 ppm error tolerance: AVEGVSQAQITGRPE (8-22; 1525.69 Da), GVSQAQITGRPEWI (11-24; 1525.73 Da), GGEQNPIYWARYA (85-97; 1525.66 Da), LLLLDLALLVDADQ (105-118; 1525.82 Da), VWWAISTAAMLYI (149-161; 1525.85 Da), and VWLIGSEGAGIVPLN (201-215; 1525.78 Da). However, the two adjacent peaks have a mass difference of 57 Da, which indicates the presence of two adjacent glycines. Therefore the sequence for the peptide with a mass of 1525.66 Da should be GGEQNPIYWARYA (85-97). The sequence for the adjacent peak with a mass of 1468.78 Da should be GEQNPIYWARYA (86-97). And the sequence for the peptide with a mass of 1411.72 Da should be EQNPIYWARYA (87-97). Subsequent MALDI-MS/MS experiments have confirmed the results (data not shown). For an unknown sample, co-elution of the peptide pair can be very useful for protein identification and confirmation of the database search results.

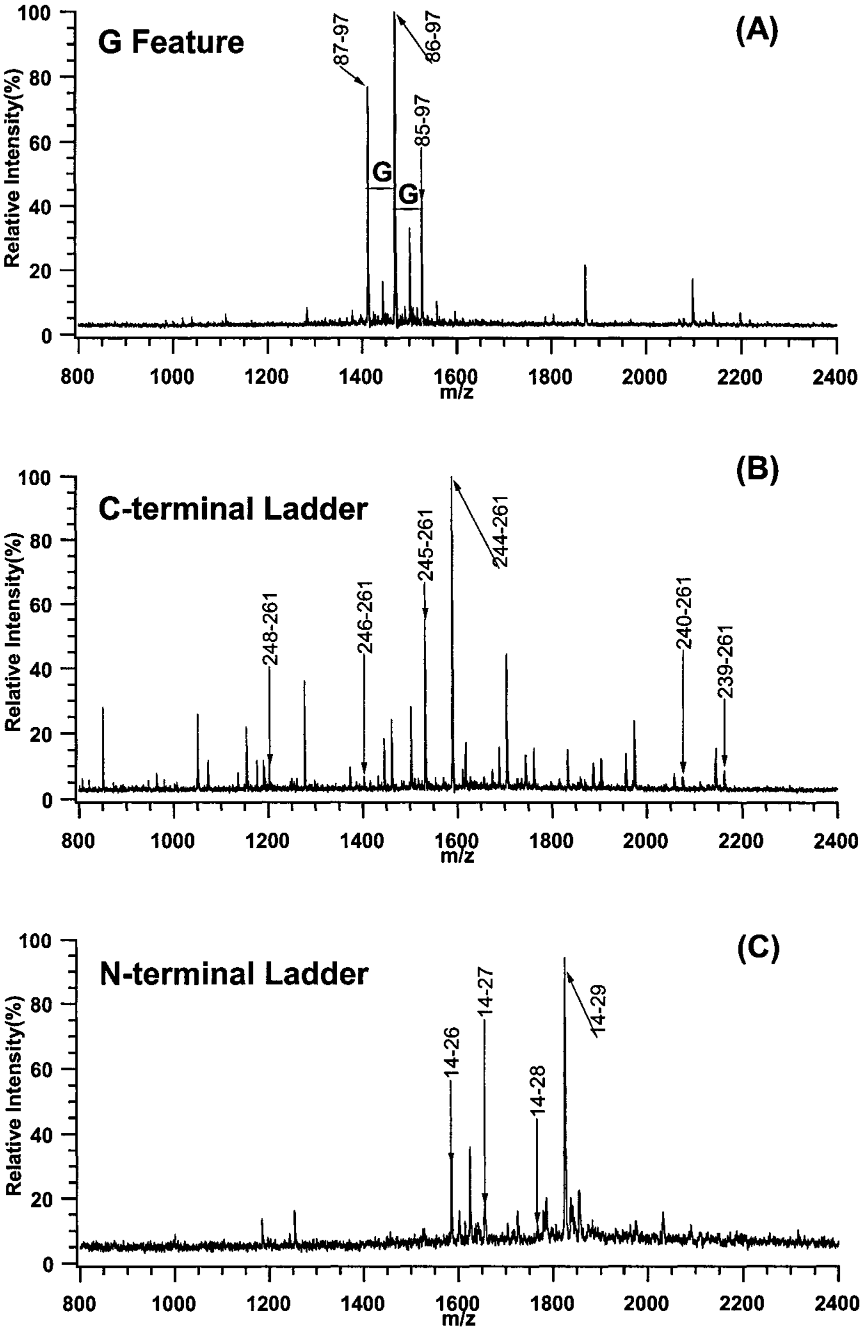


Figure IV-1-3. MALDI MS spectra of LC fractionated bacteriorhodopsin peptides: (A) fraction 34, (B) fraction 25, and (C) fraction 64. A reversed-phase HPLC separation was performed by injecting 5 μg of the bacteriorhodopsin hydrolysate. The peptides were directly fractionated onto a MALDI plate using a heated droplet LC-MALDI interface.

MAAH can also produce useful information in the form of sequence ladders generated from the C- or N-terminus that, not only provide peptide masses, but also give some sequence information. For BR, peptide ladders could be observed in some fractions. For example, Figure IV-1-3B shows the MALDI MS spectrum of fraction 25. C-terminal fragmentation is observed in this spectrum and is labeled as 248-261, 246-261, 245-261, 244-261, 240-261 and 239-261. Figure IV-1-3C shows the MALDI MS spectrum of fraction 64. N-terminal fragmentation is observed in this spectrum and is labeled as 14-26, 14-27, 14-28, and 14-29.

MALDI MS/MS analysis of the LC fractionated peptides from the BR hydrolysate, combined with database searching, reveals effective coverage of both hydrophilic and hydrophobic parts of the protein. Figure IV-1-4 shows the MALDI MS/MS spectra of two peptides from the most hydrophobic transmembrane domains of BR. The underlined sequences indicate the amino acid residues inside the membrane. The MALDI MS/MS spectra of peptides covering all 7 transmembrane domains of BR are given in the supplementary information as Figure IV-1-5. The peptides detected and sequence coverage of BR from this LC-MALDI MS/MS experiment are given in the Table IV-1-3 and Figure IV-1-6, respectively. As Figure IV-1-6 shows, the peptides detected covered the entire sequence except for three internal amino acids.

The ability to generate high sequence coverage using MAAH can be very useful for determining posttranslational modification (PTM) of a protein. Many proteins experience PTMs, such as peptide cleavage from the C- or N-terminus. This method yields peptide ladders from the C- and N-terminus and it helps determine whether this type of PTM has occurred. In the case of BR, signal peptide cleavage from the N-terminus (see Figure IV-1-3C) and final cleavage of C-terminal D (see Figure IV-1-7A) were observed. In the spectrum shown in Figure IV-1-7 (B), N-terminal transformation of glutamine to pyroglutamic acid was also found for this protein.

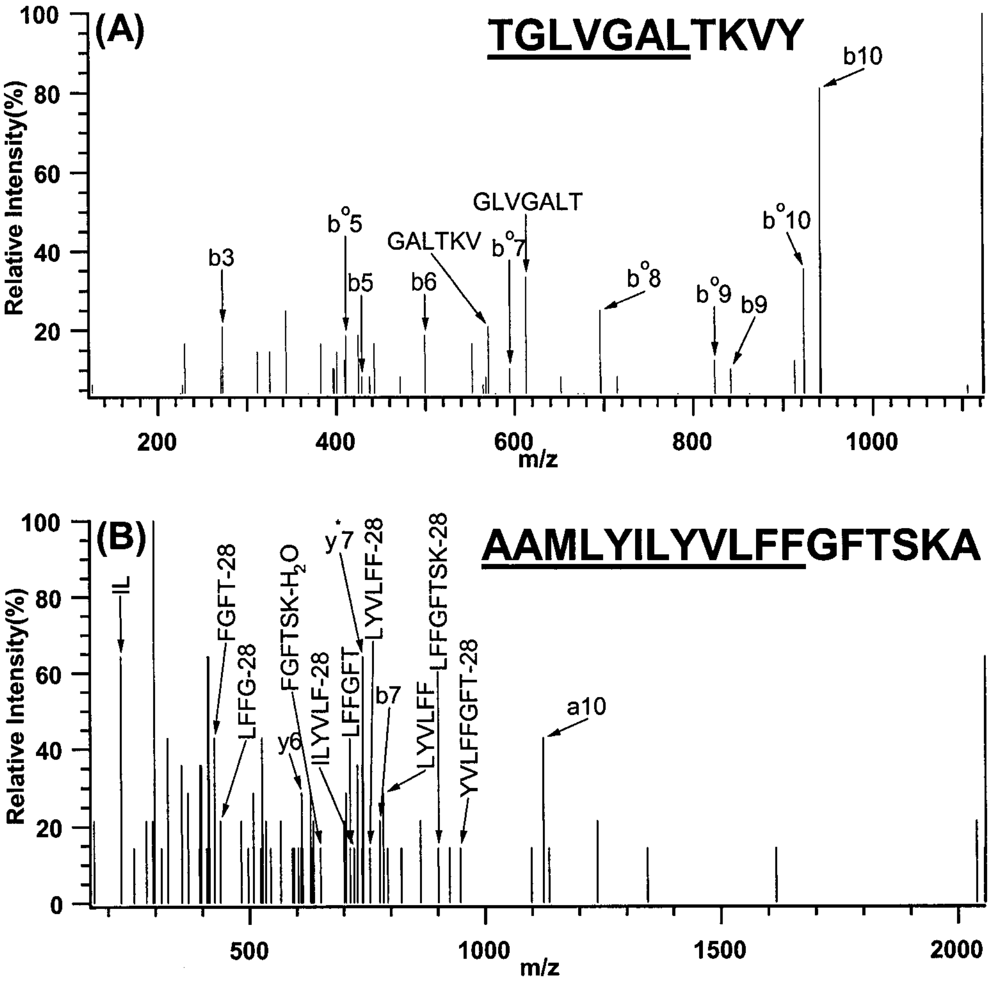


Figure IV-1-4. MALDI MS/MS spectra of peptides from the most hydrophobic transmembrane domains of bacteriorhodopsin: (A) a peptide from transmembrane domain D with a GRAVY value of 1.910, and (B) a peptide from transmembrane domain E with a GRAVY value of 1.930.

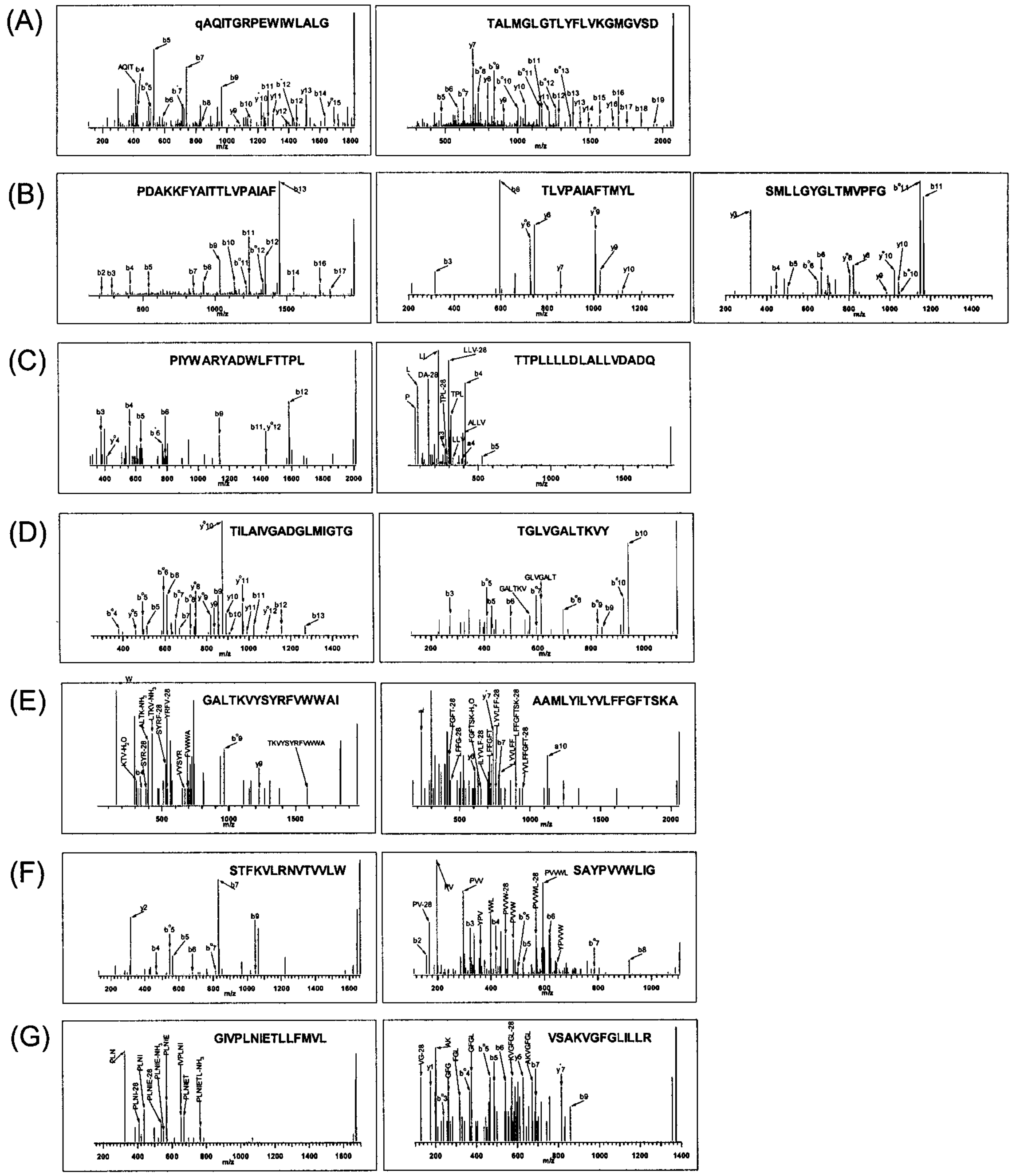


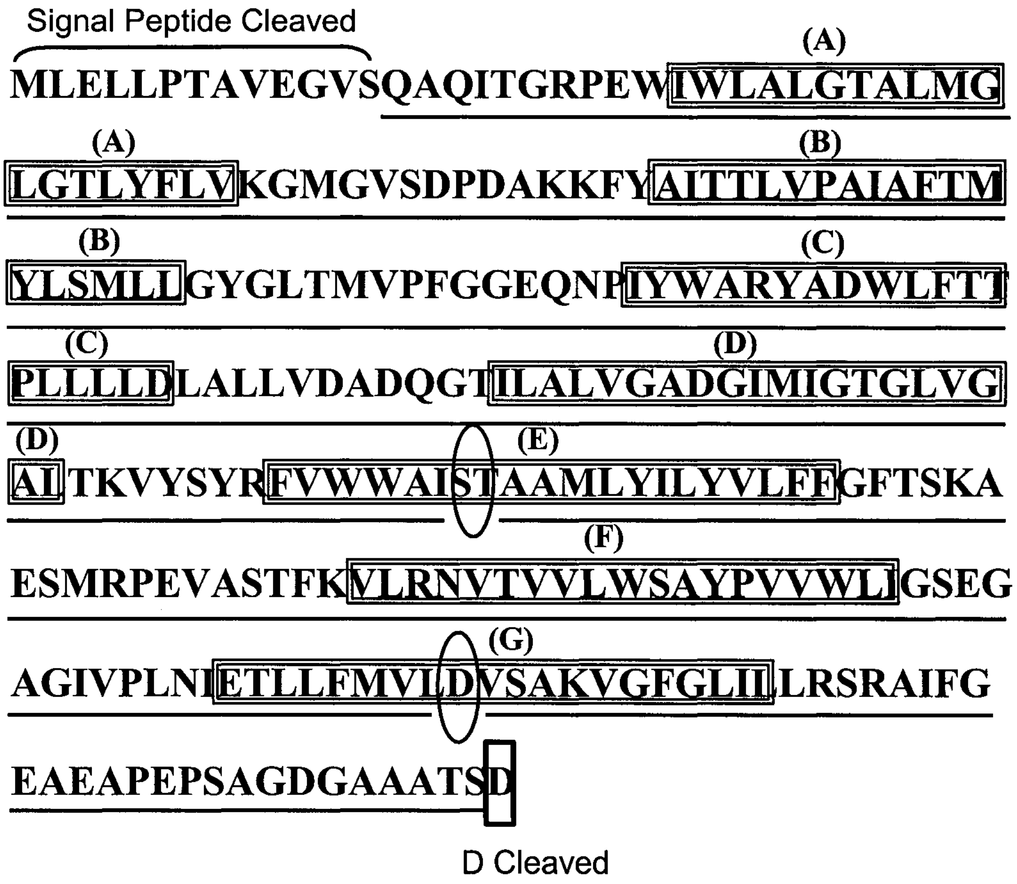
Figure IV-1-5. MALDI MS/MS spectra of peptides from the 7 transmembrane domains of bacteriorhodopsin. (A) Peptides from transmembrane domain A with a GRAVY value of 1.842. (B) Peptides from transmembrane domain B with a GRAVY value of 1.821. (C) Peptides from transmembrane domain C with a GRAVY value of 0.579. (D) Peptides from transmembrane domain D with a GRAVY value of 1.910. (E) Peptides from transmembrane domain E with a GRAVY value of 1.930. (F) Peptides from transmembrane domain F with a GRAVY value of 1.511. (G) Peptides from transmembrane domain G with a GRAVY value of 1.610.

Table IV-1-3. Peptides detected by LC-MALDI MS/MS analysis of the hydrolysate generated by MAAH of 5 μg of bacteriorhodopsin.

| Position | observed | Mr(expt) | Mr(calc) | Delta | sequence |
|---|---|---|---|---|---|
| 14 - 21 | 854.40 | 853.39 | 853.43 | -0.04 | QAQITGRP |
| 14 - 22 | 983.40 | 982.39 | 982.47 | -0.08 | QAQITGRPE |
| 14 - 26 | 1581.70 | 1580.69 | 1580.80 | -0.11 | QAQITGRPEWIWL |
| 14 - 27 | 1652.90 | 1651.89 | 1651.91 | -0.02 | QAQITGRPEWIWLA |
| 14 - 28 | 1766.20 | 1765.19 | 1764.92 | 0.27 | QAQITGRPEWIWLAL |
| 14 - 29 | 1823.10 | 1822.09 | 1821.94 | 0.15 | QAQITGRPEWIWLALG |
| 17 - 29 | 1511.80 | 1510.79 | 1510.83 | -0.04 | ITGRPEWIWLALG |
| 18 - 29 | 1398.70 | 1397.69 | 1397.75 | -0.06 | TGRPEWIWLALG |
| 20 - 28 | 1183.60 | 1182.59 | 1182.65 | -0.06 | RPEWIWLAL |
| 20 - 29 | 1240.90 | 1239.89 | 1239.68 | 0.21 | RPEWIWLALG |
| 20 - 34 | 1713.90 | 1712.89 | 1712.91 | -0.02 | RPEWIWLALGTALMG |
| 20 - 36 | 1884.20 | 1883.19 | 1883.01 | 0.18 | RPEWIWLALGTALMGLG |
| 30 - 44 | 1583.90 | 1582.89 | 1582.88 | 0.01 | TALMGLGTLYFLVKG |
| 30 - 45 | 1715.00 | 1713.99 | 1713.92 | 0.07 | TALMGLGTLYFLVKGM |
| 30 - 46 | 1772.10 | 1771.09 | 1770.94 | 0.15 | TALMGLGTLYFLVKGMG |
| 30 - 48 | 1958.20 | 1957.19 | 1957.04 | 0.15 | TALMGLGTLYFLVKGMGVS |
| 30 - 49 | 2073.20 | 2072.19 | 2072.07 | 0.12 | TALMGLGTLYFLVKGMGVSD |
| 32 - 49 | 1917.10 | 1916.09 | 1915.98 | 0.11 | LMGLGTLYFLVKGMGVSD |
| 34 - 49 | 1656.90 | 1655.89 | 1655.86 | 0.03 | GLGTLYFLVKGMGVSD |
| 35 - 43 | 1053.60 | 1052.59 | 1052.63 | -0.03 | LGTLYFLVK |
| 35 - 44 | 1110.70 | 1109.69 | 1109.65 | 0.04 | LGTLYFLVKG |
| 35 - 49 | 1599.80 | 1598.79 | 1598.84 | -0.05 | LGTLYFLVKGMGVSD |
| 35 - 49 | 1615.80 | 1614.79 | 1614.83 | -0.04 | LGTLYFLVKGMGVSD |
| 36 - 45 | 1128.60 | 1127.59 | 1127.60 | -0.01 | GTLYFLVKGM |
| 36 - 48 | 1371.70 | 1370.69 | 1370.73 | -0.03 | GTLYFLVKGMGVS |
| 36 - 49 | 1486.80 | 1485.79 | 1485.75 | 0.04 | GTLYFLVKGMGVSD |
| 37 - 45 | 1071.60 | 1070.59 | 1070.58 | 0.01 | TLYFLVKGM |
| 37 - 45 | 1087.60 | 1086.59 | 1086.58 | 0.01 | TLYFLVKGM |
| 37 - 46 | 1128.60 | 1127.59 | 1127.60 | -0.01 | TLYFLVKGMG |
| 37 - 46 | 1144.80 | 1143.79 | 1143.60 | 0.19 | TLYFLVKGMG |
| 37 - 48 | 1314.70 | 1313.69 | 1313.71 | -0.01 | TLYFLVKGMGVS |
| 37 - 48 | 1330.80 | 1329.79 | 1329.70 | 0.09 | TLYFLVKGMGVS |
| 37 - 49 | 1429.70 | 1428.69 | 1428.73 | -0.04 | TLYFLVKGMGVSD |
| 37 - 49 | 1445.90 | 1444.89 | 1444.73 | 0.16 | TLYFLVKGMGVSD |
| 39 - 49 | 1215.80 | 1214.79 | 1214.60 | 0.19 | YFLVKGMGVSD |
| 40 - 49 | 1052.70 | 1051.69 | 1051.54 | 0.15 | FLVKGMGVSD |
| 45 - 59 | 1643.00 | 1641.99 | 1641.81 | 0.18 | MGVSDPDAKKFYAIT |
| 47 - 59 | 1454.90 | 1453.89 | 1453.75 | 0.15 | VSDPDAKKFYAIT |
| 50 - 56 | 868.40 | 867.39 | 867.45 | -0.06 | PDAKKFY |
| 50 - 58 | 1052.70 | 1051.69 | 1051.57 | 0.12 | PDAKKFYAI |
| 50 - 59 | 1153.80 | 1152.79 | 1152.62 | 0.17 | PDAKKFYAIT |
| 50 - 61 | 1367.90 | 1366.89 | 1366.75 | 0.14 | PDAKKFYAITTL |
| 50 - 62 | 1466.80 | 1465.79 | 1465.82 | -0.03 | PDAKKFYAITTLV |
| 50 - 64 | 1635.10 | 1634.09 | 1633.91 | 0.18 | PDAKKFYAITTLVPA |
| 50 - 66 | 1819.10 | 1818.09 | 1818.03 | 0.06 | PDAKKFYAITTLVPAIA |
| 50 - 67 | 1966.20 | 1965.19 | 1965.10 | 0.09 | PDAKKFYAITTLVPAIAF |
| 60 - 71 | 1339.90 | 1338.89 | 1338.73 | 0.16 | TLVPAIAFTMYL |
| 72 - 85 | 1485.90 | 1484.89 | 1484.74 | 0.15 | SMLLGYGLTMVPFG |
| 77 - 85 | 984.50 | 983.49 | 983.48 | 0.01 | YGLTMVPFG |

| | | | | | |
|---|---|---|---|---|---|
| 79 - 97 | 2215.10 | 2214.09 | 2214.05 | 0.04 | LTMVPFGGEQNPIYWARYA |
| 80 - 89 | 1081.60 | 1080.59 | 1080.44 | 0.15 | TMVPFGGEQN |
| 85 - 95 | 1291.80 | 1290.79 | 1290.60 | 0.19 | GGEQNPIYWAR |
| 85 - 96 | 1454.80 | 1453.79 | 1453.66 | 0.13 | GGEQNPIYWARY |
| 85 - 97 | 1525.90 | 1524.89 | 1524.70 | 0.19 | GGEQNPIYWARYA |
| 85 - 98 | 1642.00 | 1640.99 | 1640.71 | 0.28 | GGEQNPIYWARYAD |
| 86 - 95 | 1234.70 | 1233.69 | 1233.58 | 0.11 | GEQNPIYWAR |
| 86 - 96 | 1397.80 | 1396.79 | 1396.64 | 0.15 | GEQNPIYWARY |
| 86 - 97 | 1467.90 | 1466.89 | 1466.69 | 0.20 | GEQNPIYWARYA |
| 86 - 98 | 1583.90 | 1582.89 | 1582.71 | 0.19 | GEQNPIYWARYAD |
| 86 - 101 | 2029.00 | 2027.99 | 2027.95 | 0.04 | GEQNPIYWARYADWLF |
| 87 - 95 | 1177.70 | 1176.69 | 1176.56 | 0.14 | EQNPIYWAR |
| 87 - 96 | 1340.80 | 1339.79 | 1339.62 | 0.17 | EQNPIYWARY |
| 87 - 97 | 1410.80 | 1409.79 | 1409.67 | 0.12 | EQNPIYWARYA |
| 87 - 98 | 1526.90 | 1525.89 | 1525.68 | 0.21 | EQNPIYWARYAD |
| 88 - 96 | 1193.70 | 1192.69 | 1192.57 | 0.13 | QNPIYWARY |
| 88 - 97 | 1282.80 | 1281.79 | 1281.61 | 0.18 | QNPIYWARYA |
| 88 - 98 | 1397.80 | 1396.79 | 1396.64 | 0.15 | QNPIYWARYAD |
| 89 - 96 | 1082.70 | 1081.69 | 1081.53 | 0.16 | NPIYWARY |
| 89 - 97 | 1153.70 | 1152.69 | 1152.57 | 0.12 | NPIYWARYA |
| 90 - 95 | 805.60 | 804.59 | 804.43 | 0.16 | PIYWAR |
| 90 - 96 | 968.60 | 967.59 | 967.49 | 0.10 | PIYWARY |
| 90 - 97 | 1039.70 | 1038.69 | 1038.53 | 0.16 | PIYWARYA |
| 90 - 98 | 1154.70 | 1153.69 | 1153.56 | 0.14 | PIYWARYAD |
| 90 - 101 | 1600.90 | 1599.89 | 1599.79 | 0.10 | PIYWARYADWLF |
| 90 - 105 | 2013.10 | 2012.09 | 2012.02 | 0.07 | PIYWARYADWLFTTPL |
| 94 - 101 | 1041.60 | 1040.59 | 1040.51 | 0.08 | ARYADWLF |
| 95 - 101 | 970.60 | 969.59 | 969.47 | 0.12 | RYADWLF |
| 102 - 112 | 1182.90 | 1181.89 | 1181.73 | 0.16 | TTPLLLLDLAL |
| 102 - 118 | 1824.10 | 1823.09 | 1823.03 | 0.06 | TTPLLLLDLALLVDADQ |
| 119 - 135 | 1558.90 | 1557.89 | 1557.84 | 0.05 | GTILALVGADGIMIGTG |
| 120 - 135 | 1501.80 | 1500.79 | 1500.82 | -0.03 | TILALVGADGIMIGTG |
| 134 - 144 | 1121.80 | 1120.79 | 1120.65 | 0.14 | TGLVGALTKVY |
| 134 - 147 | 1527.90 | 1526.89 | 1526.85 | 0.04 | TGLVGALTKVYSYR |
| 136 - 144 | 963.70 | 962.69 | 962.58 | 0.11 | LVGALTKVY |
| 138 - 153 | 1960.10 | 1959.09 | 1959.04 | 0.05 | GALTKVYSYRFVWWAI |
| 145 - 154 | 1314.80 | 1313.79 | 1313.66 | 0.13 | SYRFVWWAIS |
| 156 - 173 | 2055.20 | 2054.19 | 2054.09 | 0.10 | AAMLYILYVLFFGFTSKA |
| 167 - 181 | 1657.00 | 1655.99 | 1655.80 | 0.19 | FGFTSKAESMRPEVA |
| 168 - 177 | 1113.50 | 1112.49 | 1112.53 | -0.04 | GFTSKAESMR |
| 168 - 181 | 1509.70 | 1508.69 | 1508.73 | -0.04 | GFTSKAESMRPEVA |
| 169 - 181 | 1452.70 | 1451.69 | 1451.71 | -0.02 | FTSKAESMRPEVA |
| 170 - 181 | 1305.60 | 1304.59 | 1304.64 | -0.05 | TSKAESMRPEVA |
| 175 - 182 | 876.50 | 875.49 | 875.42 | 0.07 | SMRPEVAS |
| 175 - 188 | 1621.10 | 1620.09 | 1619.88 | 0.21 | SMRPEVASTFKVLR |
| 177 - 188 | 1402.90 | 1401.89 | 1401.81 | 0.08 | RPEVASTFKVLR |
| 182 - 189 | 964.50 | 963.49 | 963.55 | -0.06 | STFKVLRN |
| 182 - 195 | 1662.90 | 1661.89 | 1661.95 | -0.06 | STFKVLRNVTVVLW |
| 196 - 205 | 1104.70 | 1103.69 | 1103.60 | 0.09 | SAYPVVWLIG |
| 206 - 217 | 1198.70 | 1197.69 | 1197.62 | 0.07 | SEGAGIVPLNIE |
| 210 - 224 | 1672.10 | 1671.09 | 1670.97 | 0.10 | GIVPLNIETLLFMVL |
| 225 - 244 | 2119.20 | 2118.19 | 2118.23 | -0.04 | DVSAKVGFGLILLRSRAIFG |
| 226 - 238 | 1372.80 | 1371.79 | 1371.86 | -0.07 | VSAKVGFGLILLR |
| 226 - 243 | 1947.10 | 1946.09 | 1946.18 | -0.09 | VSAKVGFGLILLRSRAIF |
| 226 - 244 | 2004.10 | 2003.09 | 2003.20 | -0.11 | VSAKVGFGLILLRSRAIFG |
| 229 - 238 | 1115.80 | 1114.79 | 1114.72 | 0.07 | KVGFGLILLR |
| 229 - 244 | 1747.20 | 1746.19 | 1746.07 | 0.12 | KVGFGLILLRSRAIFG |

| | | | | | |
|---|---|---|---|---|---|
| 229 - 246 | 1947.20 | 1946.19 | 1946.15 | 0.04 | KVGFGLILLRSRAIFGEA |
| 231 - 244 | 1520.10 | 1519.09 | 1518.90 | 0.19 | GFGLILLRSRAIFG |
| 232 - 239 | 918.60 | 917.59 | 917.57 | 0.02 | FGLILLRS |
| 232 - 240 | 1074.70 | 1073.69 | 1073.67 | 0.02 | FGLILLRSR |
| 232 - 241 | 1145.70 | 1144.69 | 1144.71 | -0.02 | FGLILLRSRA |
| 232 - 244 | 1462.90 | 1461.89 | 1461.88 | 0.01 | FGLILLRSRAIFG |
| 232 - 246 | 1663.10 | 1662.09 | 1661.96 | 0.13 | FGLILLRSRAIFGEA |
| 232 - 248 | 1863.10 | 1862.09 | 1862.04 | 0.05 | FGLILLRSRAIFGEAEA |
| 233 - 244 | 1315.80 | 1314.79 | 1314.81 | -0.02 | GLILLRSRAIFG |
| 233 - 246 | 1515.90 | 1514.89 | 1514.89 | -0.00 | GLILLRSRAIFGEA |
| 233 - 247 | 1644.90 | 1643.89 | 1643.94 | -0.04 | GLILLRSRAIFGEAE |
| 233 - 248 | 1715.90 | 1714.89 | 1714.97 | -0.08 | GLILLRSRAIFGEAEA |
| 233 - 253 | 2197.20 | 2196.19 | 2196.19 | 0.00 | GLILLRSRAIFGEAEAPEPSA |
| 233 - 254 | 2254.20 | 2253.19 | 2253.21 | -0.02 | GLILLRSRAIFGEAEAPEPSAG |
| 234 - 243 | 1201.80 | 1200.79 | 1200.77 | 0.02 | LILLRSRAIF |
| 234 - 244 | 1258.80 | 1257.79 | 1257.79 | 0.00 | LILLRSRAIFG |
| 234 - 248 | 1658.90 | 1657.89 | 1657.95 | -0.06 | LILLRSRAIFGEAEA |
| 237 - 254 | 1858.20 | 1857.19 | 1856.94 | 0.25 | LRSRAIFGEAEAPEPSAG |
| 239 - 246 | 850.60 | 849.59 | 849.43 | 0.16 | SRAIFGEA |
| 239 - 248 | 1050.70 | 1049.69 | 1049.51 | 0.18 | SRAIFGEAEA |
| 239 - 249 | 1147.70 | 1146.69 | 1146.57 | 0.13 | SRAIFGEAEAP |
| 239 - 250 | 1276.80 | 1275.79 | 1275.61 | 0.18 | SRAIFGEAEAPE |
| 239 - 251 | 1373.80 | 1372.79 | 1372.66 | 0.13 | SRAIFGEAEAPEP |
| 239 - 252 | 1460.80 | 1459.79 | 1459.69 | 0.10 | SRAIFGEAEAPEPS |
| 239 - 254 | 1588.90 | 1587.89 | 1587.75 | 0.14 | SRAIFGEAEAPEPSAG |
| 239 - 255 | 1704.00 | 1702.99 | 1702.78 | 0.21 | SRAIFGEAEAPEPSAGD |
| 239 - 257 | 1832.10 | 1831.09 | 1830.84 | 0.25 | SRAIFGEAEAPEPSAGDGA |
| 239 - 259 | 1974.10 | 1973.09 | 1972.91 | 0.18 | SRAIFGEAEAPEPSAGDGAAA |
| 239 - 261 | 2162.10 | 2161.09 | 2160.99 | 0.10 | SRAIFGEAEAPEPSAGDGAAATS |
| 240 - 249 | 1060.70 | 1059.69 | 1059.53 | 0.16 | RAIFGEAEAP |
| 240 - 250 | 1189.70 | 1188.69 | 1188.58 | 0.11 | RAIFGEAEAPE |
| 240 - 251 | 1286.80 | 1285.79 | 1285.63 | 0.16 | RAIFGEAEAPEP |
| 240 - 253 | 1444.90 | 1443.89 | 1443.70 | 0.19 | RAIFGEAEAPEPSA |
| 240 - 254 | 1501.90 | 1500.89 | 1500.72 | 0.17 | RAIFGEAEAPEPSAG |
| 240 - 261 | 2075.05 | 2074.04 | 2073.96 | 0.08 | RAIFGEAEAPEPSAGDGAAATS |
| 242 - 254 | 1274.60 | 1273.59 | 1273.58 | 0.01 | IFGEAEAPEPSAG |
| 242 - 255 | 1389.60 | 1388.59 | 1388.61 | -0.02 | IFGEAEAPEPSAGD |
| 244 - 258 | 1328.50 | 1327.49 | 1327.55 | -0.06 | GEAEAPEPSAGDGAA |
| 244 - 259 | 1399.60 | 1398.59 | 1398.59 | 0.00 | GEAEAPEPSAGDGAAA |
| 244 - 261 | 1587.60 | 1586.59 | 1586.67 | -0.08 | GEAEAPEPSAGDGAAATS |
| 245 - 259 | 1342.50 | 1341.49 | 1341.57 | -0.08 | EAEAPEPSAGDGAAA |
| 245 - 261 | 1530.60 | 1529.59 | 1529.65 | -0.06 | EAEAPEPSAGDGAAATS |
| 246 - 261 | 1401.70 | 1400.69 | 1400.60 | 0.06 | AEAPEPSAGDGAAATS |
| 248 - 261 | 1201.60 | 1200.59 | 1200.52 | 0.07 | APEPSAGDGAAATS |
| 249 - 261 | 1130.50 | 1129.49 | 1129.49 | 0.00 | PEPSAGDGAAATS |



Signal Peptide Cleaved
(A)
MLELLPTAVEGVSQAQITGRPEWIWLALGTALMG
(A)
(B)
LGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTM
(B)
(C)
YLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTT
(C)
(D)
PLLLLDLALLVDADQGTILALVGADGIMIGTGLVG
(D)
(E)
AITKVYSYRFVWWAISTAAMLYILYVLFFGFTSKA
(F)
ESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEG
(G)
AGIVPLNIETLLFMVLDVSAKVGFGLIILLRSRAIFG
EAEAPEPSAGDGAAATSD
D Cleaved

Figure IV-1-6. Sequence coverage of bacteriorhodopsin peptides generated using 25% TFA and 10 min microwave digestion and detected by LC-MALDI MS and MS/MS. 5 µg of the bacteriorhodopsin hydrolysate was injected for LC-MALDI. Double-line rectangles indicate the seven transmembrane domains and underlinings indicate the peptides identified by MALDI MS/MS and MASCOT search. Three circled amino acid residues were not covered by the peptides detected.

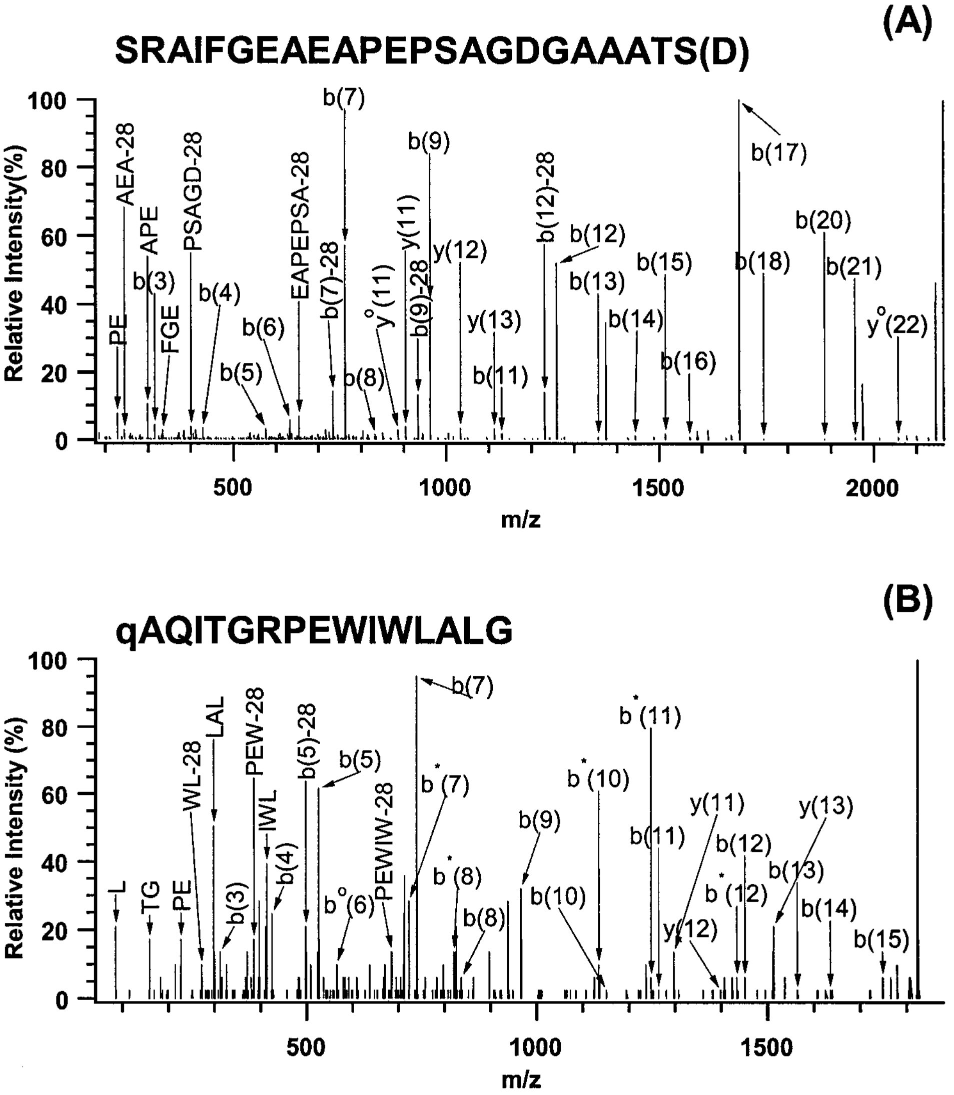


Figure IV-1-7. MALDI MS/MS spectra of bacteriorhodopsin peptides containing posttranslational modifications: (A) a peptide showing cleavage of C-terminal aspartic acid, and (B) a peptide showing N-terminal transformation of glutamine to pyroglutamic acid.

**Detection Sensitivity.** As has been demonstrated, the MAAH method can provide more cleavage sites, resulting in extensive sequence coverage. On the other hand, multiple cleavages on a protein can potentially result in many peptides that may complicate the downstream analysis by LC-MS. However, reducing the protein concentration decreases significantly the number of peptides detected by MALDI MS. As illustrated in Figure 6, different concentrations of BR were hydrolyzed by 25% TFA with microwave irradiation for 10 minutes. As the BR concentration was decreased from 100 ng/μl to 1 ng/μl, the signal-to-noise ratio and the number of peptide peaks decreased. Interestingly, the terminal sequence ladders can still be observed in the lowest concentration spectrum, while a number of peptides from internal fragmentation disappear. The results shown in Figure IV-1-8 demonstrate that the MAAH method can still generate several peptides from a protein concentration of as low as to 1 ng/μl. Unlike enzyme digestion, where ion suppression from protease autolysis peptides can occur when a low concentration of protein is used for digestion, MAAH does not introduce background peaks, as shown in Figure IV-1-8C. It is clear that MAAH is an effective method for digesting low concentrations of proteins and the spectra generated consist of a few peaks that should be readily handled by conventional LC separation strategies, such as 1D- or 2D-LC methods.

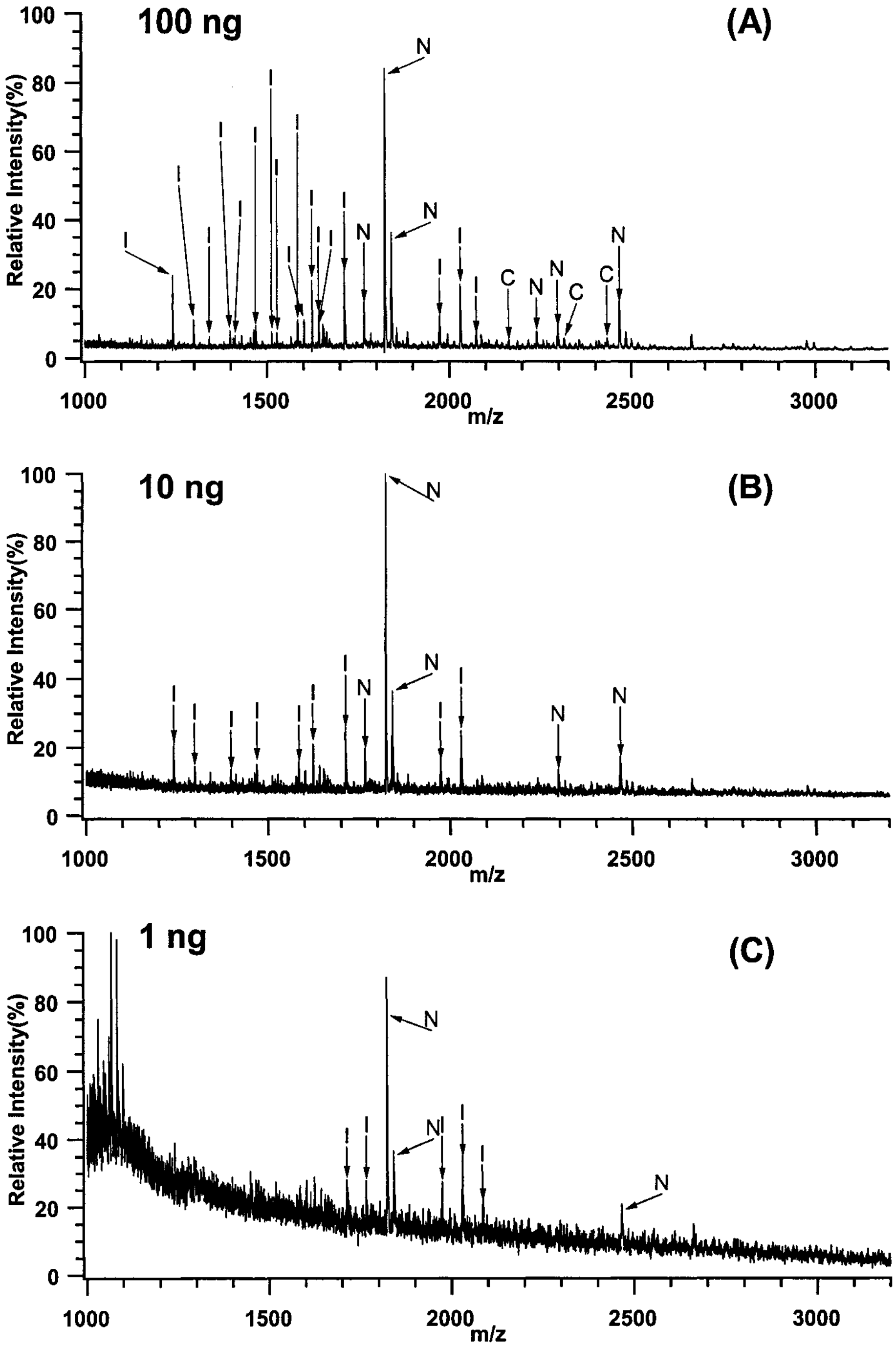


Figure IV-1-8. MALDI MS spectra of hydrolysates generated from different concentrations of bacteriorhodopsin: (A) 100 ng/μl, (B) 10 ng/μl, and (C) 1 ng/μl. In each case, 10 μl of protein solution in 25% TFA was subjected to microwave irradiation for 10 minutes. The hydrolysate was then dried by Speedvac and re-dissolved in 5 μl of the second-layer HCCA matrix solution. 0.5 μl of the matrix/analyte solution was deposited to the first layer for MALDI analysis.

### IV. 1. 4. Conclusions

A method of using microwave-assisted acid hydrolysis to degrade membrane proteins into peptides for MS characterization has been described in this chapter. It works for proteins merely suspended in 25% TFA aqueous solution and does not require the use of strong surfactants to solubilize proteins. The presented method expands the range of proteins that can be quickly analyzed by MS and circumvents the reliance on specific proteases. It improves throughput by reducing the time required for protein digestion. This method also offers good sequence coverage including transmembrane domains. The unique cleavage specificity such as C- or N-terminal fragmentation and the feature of glycine cleavage provides useful information for positive protein identification.

Compared to enzymatic digestion, this MAAH method is fast and detergent-free. It involves a simple sample handling process and there are no background peptides, such as those from protease autolysis in enzyme digestion, introduced in MAAH. This method also provides some advantages over other acid hydrolysis techniques. Protein aggregates, which prevent degradation, are observed in acid hydrolysis of membrane proteins with conventional heating. Under microwave irradiation, not only faster degradation can be obtained, but also efficient cleavage in transmembrane domains can be achieved, resulting in more sequence coverage.

### IV. 1. 5. Cited Literature

(1) Wu, C. C.; Yates J. R., III. *Nat. Biotechnol.* **2003**, *21*, 262-267.

(2) Rabilloud, T. *Nat. Biotechnol.* **2003**, *21*, 508-510.

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

(4) Ball, L. E.; Oatis, J. E.; Dharmasiri, K.; Busman, M.; Wang, J.; Cowden, L. B.;

(5) Galijatovic, A.; Chen, L.; Crouch, R. K.; Knapp, D. R. *Prot. Sci.* **1998,** *7,* 758-764. Ablonczy, Z.; Kono, M.; Crouch, R. K.; Knapp, D. R. *Anal. Chem.* **2001,** *73,* 4774-4779.

(6) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242-247.

(7) Quach, T. T. T.; Li, N.; Richard, D. P.; Zheng, J.; Keller, B. O.; Li, L. *J. Prot. Res.* **2003,** *2,* 543-552.

(8) Norris, J. L.; Porter, N.A.; Caprioli, R. M., *Anal.Chem.* **2003,** *75,* 6642-6647.

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

(10) Goshe, M. B.; Blonder, J.; Smith, R. D., *J. Prot. Res.* **2003,** *2,* 153-161.

(11) Blonder, J.; Conrads, T. P.; Yu, L.; Terunuma, A.; Janini, G. M.; Issaq, H. J.; Vogel, J. C.; Veenstra, T. D., *Proteomics* **2004,** *4,* 31-45.

(12) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., III. *Nat. Biotechnol.* **2003,** *21*, 532-538.

(13) Sanger, F.; Thompson, E.O.P. *Biochem. J.* **1953**, *53*, 353-366.

(14) Sonderegger, P.R.; Jaussi, H.; Gehring, K.; Brunschweiler, K.; Christen, P. *Anal. Biochem.* **1982**, *122*, 298-301.

(15) Vanfleteren, J.R.; Raymackers, J.G.; van Bun, S.M.; Meheus, L.A. *BioTechnique* **1992**, *12*, 550-557.

(16) Tsugita, A.; Takamoto, K.; Kamo, M.; Iwadate, H. *Eur. J. Biochem*, **1992**, *206*, 691-696.

(17) Tsugita, A.; Kamo, M.; Miyazaki, K.; Takayama, M.; Kawakami, T.; Shen, R.; Nozawa, T. *Electrophoresis* **1998**, *19*, 928-938.

(18) Vorm, O.; Roepstorff, P. *Biol. Mass Spectrom.* **1994**, *23*, 734-740.

(19) Zubarev, R.A.; Chivanov, V.D.; Hakansson, P.; Sundkvist, B.U.R. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 906-912.

(20) Gobom, J.; Mirgorodskaya, E.; Nordhoff, E.; Hojrup, P.; Roepstorff, P. *Anal. Chem.* **1999**, *71*, 919-927.

(21) Li, A.; Sowder, R.C. II; Henderson, L.E.; Moore, S.P.; Garfinfel, D.J.; Fisher, R.J. *Anal. Chem.* **2001**, *73*, 5395-5402.

(22) Shevchenko, A.; Loboda, A.; Shevchenko, A.; Ens, W.; Standing, K.G. *Anal. Chem.* **2000**, *72*, 2132-2141.

(23) Lin, S.H.; Tornatore, P.; Weinberger, S.R.; King, D.; Orlando, R. *Eur. J. Mass Spectrom.* **2001**, *7*, 131-141.

(24) Ocana, M.F.; Neubert, H.; Przyborowska, A.; Parker, R.; Bramley, P.; Halket, J.; Patel, R. *Analyst* **2004**, *129*, 111-115.

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

(26) Dai, Y. Q.; Whittal, R. M.; Li, L. *Anal. Chem.*, **1999**, *71*, 1087-1091.

(27) Kingston, H.M.; Haswell, S.J. (Editors), *Microwave-Enhanced Chemistry: Fundamentals, Sample Preparation, and Applications.* ACS: Washington, D.C., 1997.

(28) Hixson, K. K., Rodriguez, N., Camp II, D.G., Lipton, M. S., Smith, R. D.

*Electrophoresis* **2002,** *23*, 3224-3232.

(29) Smith, B. J. in *The Protein Protocols Handbook*, 2nd Ed.; Walker, J. M., Ed.; Humana Press: Totowa, NJ, 2002; 485-491.

# Chapter 2. Identification of Membrane Proteins Isolated from Human Breast Cancer Cell Line by Microwave-assisted Acid Hydrolysis and HPLC MALDI MS/MS[1]

Membrane proteins play an important role in many fundamental biological processes such as cell-cell interactions, signal transduction, and material transport. They have been extensively targeted for drug design and account for about 70% of all known drug targets. Identification of membrane proteins is important for understanding complex biological structure and function, also critical for earlier diagnosis and earlier treatment of diseases. In this chapter, the microwave-assisted acid hydrolysis was applied to identify proteins in a membrane protein fraction isolated from a human breast cancer cell line.

## IV. 2. 1. Introduction

Membrane proteins are important with regard to human diseases. However, the percentage of fully characterized membrane proteins is still low in datasets although global genomic analyses [1] predict that 20-30% of all open reading frames encode integral membrane proteins. The under-representation of membrane proteins in proteomic analyses is greatly due to the intrinsic difficulties of solubilization, low abundance and extensive posttranslational modifications [2-4]. General mass spectrometric methods that analyze peptides generated by enzymatic proteolysis developed for analysis of soluble

[1] A version of this chapter has been submitted for publication as:
Hongying Zhong, Sandra Marcus and Liang Li, "Membrane Proteome Analysis by Microwave-Assisted Acid Hydrolysis of Proteins and Liquid Chromatography MALDI-MS/MS". Dr. Sandra Marcus performed cell culture and membrane protein fractionation.

proteins are not amenable to membrane proteins [5-7]. Insoluble membrane protein aggregates cannot be accessed by proteases. Therefore, buffers containing concentrated urea, detergents and other salts are often chosen for membrane protein solubilization. In addition to their interference with mass spectrometric detection, concentrated urea or detergents needed for membrane protein solubilization denature many proteases and therefore decrease their activity to cleave proteins.

Microwave-assisted acid hydrolysis has been successfully used as a detergent-free method to identify the model integral membrane protein. The unique cleavage characterization such as G-feature, C- or N-terminal sequence ladders provide additional information for efficient protein identification. Combined with the MALDI-TOF that has high mass accuracy, high resolution and high sensitivity, it is possible to identify many hydrophobic membrane proteins. In this chapter, it was applied to identify membrane protein fractions isolated from a human breast cancer cell line MCF7. With one-dimensional LC-MALDI MS/MS, a total of 119 proteins, including 41 membrane-associated or membrane proteins containing 1 to 12 transmembrane domains, were identified by MS/MS database searching based on matches of at least 2 peptides to a protein.

### IV. 2. 2. Experimental Section

**Materials and Reagents.** Unless otherwise noted, all chemicals were purchased from Sigma (St. Louis, MO) and were of analytical grade. For HPLC separation, MS analysis and preparation of digestions, HPLC grade water, methanol and acetonitrile were used (Fisher Scientific, Mississauga, ON). 37% HCl (ACS grade) was from Merck KGaA, Darmstadt, Germany. Human breast cancer cell line, MCF7 cells (ATCC HTB-22), was purchased from the American Type Culture Collection (Manassas, VA).

**Membrane Protein Fractionation**. MCF7 cells were grown in 15 cm diameter plates in ATCC medium at 37°C for 2 weeks. The growth medium was aspirated to leave a monolayer of cells on the plates, which were then placed on an ice-cold metal tray. The plates were washed 3 times with ice-cold 10 ml PBS$^{++}$ buffer (0.9 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.7 mM $KH_2PO_4$, 8 mM $NaHPO_4$, 1 mM KCl, and 0.1 M NaCl) and then 2.5 ml saponin lysis buffer [0.2% saponin in 50 mM Tris-HCl pH 7.5 with 1 mM phenylmethanesulfonyl fluoride (PMSF)] was added to each dish. Following incubation on ice for 5 min with constant rocking, the cells were scraped from the plates and collected with the buffer solution. The suspension was centrifuged at 17400g for 15 min at 4°C. The supernatant contained cytosolic proteins. The cell pellets were washed by resuspending in a total volume of 10 ml wash buffer (50 mM Tris-HCl pH 7.5 with 1 mM PMSF). The pellets were combined into one centrifuge tube and centrifuged at 17400g for 15 min at 4°C. The pellet was resuspended in 12 ml Triton X-100 buffer (1% Triton X-100 in 50 mM Tris-HCl pH 7.5 with 1 mM PMSF) and incubated on ice for 15 min with vortexing at 5 min intervals to release membrane proteins. The preparation was then centrifuged (as above) to remove insoluble materials. Dithiothreitol (DTT) was added to the supernatant to a final concentration of 20 mM. The solution was incubated for 1 hour at 37°C. Iodoacetamide was added to a final concentration of 20 mM and the solution was left to stand for 1 hour at room temperature in the dark. All disulfide linkages should have been reduced and carbamidomethylated. The supernatant was then mixed with 3 volumes of cold acetone (-20°C) to precipitate the proteins and remove detergents and other chemicals. It was then centrifuged as described above and the protein precipitate was lyophilized and stored at -70°C.

**Acid Hydrolysis.** In the analysis of the membrane protein fraction from MCF7 cells, 1 mg of the lyophilized sample was suspended in 100 μl of 25% TFA aqueous solution. In both cases, 10 μl of the protein suspension was placed in a 1.5 ml polypropylene centrifuge vial, capped and sealed with Teflon tape. The vial was placed in a domestic 900W (2450 MHz) microwave oven. 100 ml of water in a loosely covered container was placed besides the sample vial to absorb extra microwave energy. 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 vacuum centrifuge to remove the acid. The dried sample was re-suspended in 100 μl of 0.1% TFA aqueous solution and centrifuged at 16000 g for 5 minutes. 50 μl of the solution was injected into the HPLC for analysis.

Care was taken when handling the concentrated acids as well as other precautions that were followed in the use of the microwave oven (see Safety Considerations).

**HPLC Separation**. Reversed-phase HPLC separations of the peptides were made using a Vydac C18 column (1 mm ID × 150 mm; Vydac, Hesperia, CA) on an Agilent 1100 capillary HPLC system. At a flow rate of 50 μl/min, a linear gradient from 5% B to 85% B over 70 minutes was used, where mobile phase A was water containing 0.1% (v/v) TFA, and mobile phase B was acetonitrile with 0.1% (v/v) TFA. During the separation period from 10 to 50 minutes, the eluate was fractionated at 30 s intervals and directly deposited to the MALDI target (Applied Biosystems, Boston, MA) using the heated droplet interface [8] .

**MALDI MS and MS/MS**. For peptide mass mapping by MALDI time-of-flight (TOF) MS, a two-layer sample/matrix preparation method was used [9] with α-cyano-4-hydroxy-cinnamic acid (HCCA) as matrix. 0.7 μl of the first-layer matrix solution containing 12 mg/ml of HCCA in 20% methanol/acetone was deposited on the MALDI target and air-dried. 0.5 μl of sample was mixed with 0.5 μl of matrix (50% acetonitrile/water saturated with HCCA) and then deposited onto the first layer. The MALDI-TOF mass spectra were obtained on a Bruker Reflex III time-of-flight mass spectrometer (Bremen/Leipzig, Germany).

For MALDI MS/MS analysis of the HPLC fractions, the peptide samples were prepared using a dried droplet method, in which 2,5-dihydroxybenzoic acid (DHB) was used as the matrix. 0.3 μl of matrix solution in 50% acetonitrile/water saturated with DHB was put into each well and then pipetted several times before spotting on the target. Product ion spectra of peptides were obtained in a QSTAR MALDI Qq-TOF mass spectrometer (MDS Sciex, Ontario).

**Data Interpretation and Database Searching.** Database searching using MS/MS spectra was performed by MASCOT (http://www.matrixscience.com). All the database searching was done against SwissProt using no specification of enzyme type. Methionine oxidization, carbamidomethylation of cysteine, and deamidation of asparagine and glutamine were set as variable modifications. Potential protein candidates with the highest MOWSE scores were determined from database search. The MS/MS spectra of the matched peptides were examined manually to see if the major peaks observed were matched with the expected fragmentation patterns. If they agreed well, the identification was considered to be positive.

**Hydropathy Calculations.** Proteins identified were examined using the ProtParam program available at the EXPASY web site (http://us.expasy.org/tools/protparam.html) that allows for calculation of the grand average of hydrophobicity (GRAVY). Positive values are considered as hydrophobic and negative values are considered as hydrophilic.

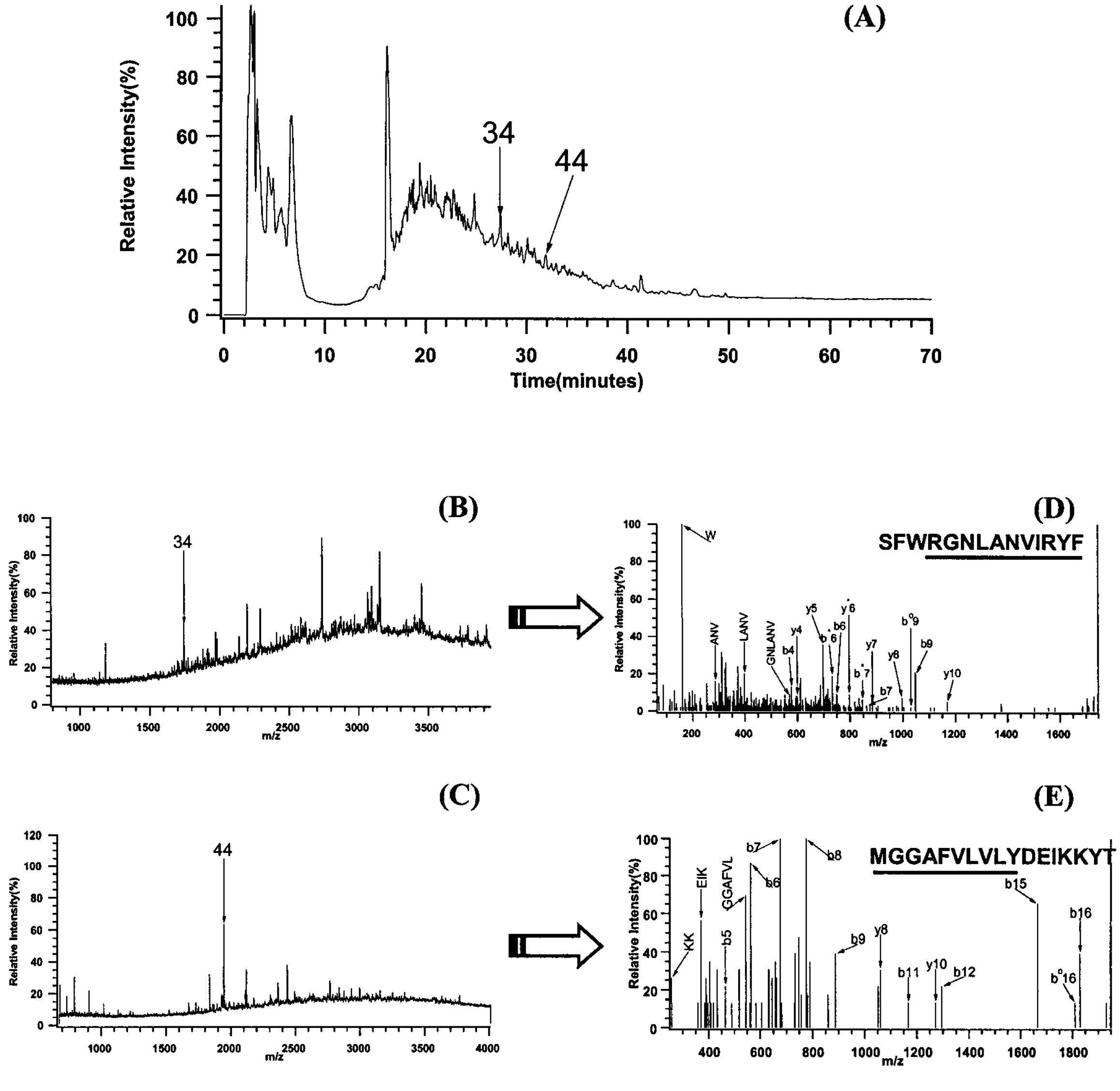


Figure IV-2-1. (A) UV chromatogram from a reserved-phase HPLC separation of the hydrolysate generated from microwave-assisted acid hydrolysis of a membrane protein enriched fraction from human breast cell line MCF7. About 50 μg of the protein hydrolysate was injected. MALDI MS spectra of (B) fraction 34 and (C) fraction 44. MS/MS spectra of (D) a peptide from fraction 34 (GRAVY 0.400) and (E) a peptide from fraction 44 (GRAVY 1.037). These two product ion spectra were matched with the peptide sequences corresponding to parts of the transmembrane domains of the integral membrane protein ATP carrier (P05141) fractionated from human breast cancer cell line MCF7.

### IV. 2. 4. Results and Discussion

The heated droplet LC-MALDI interface reported previously was used for on-line peptide fractionation [8]. The fractions were directly collected into a 100-well MALDI plate. Figure IV-2-1 (A) shows the UV chromatogram from the C18 column separation of the MAAH hydrolysate. No late eluting peaks (i.e., retention time > 50 minutes) are observed, indicating that most proteins, if not all, were hydrolyzed. Panels B and C of Figure IV-2-1 show the MALDI MS spectra of two fractions. The MS/MS spectra of two peptides from the transmembrane domains of an integral membrane protein ATP carrier (P05141) are shown in panels D and E of Figure IV-2-1. The underlined sequences indicate the amino acids residing inside the transmembrane domains. For this apparently high abundance protein, a total of 11 peptides were detected, providing sequence coverage of 24%, including 10 peptides containing parts of the transmembrane domains.

Using MALDI MS/MS with one-dimensional LC separation, a total of 119 proteins were identified by MS/MS database searching based on matches of at least 2 peptides to a protein. Among them, 41 proteins are membrane proteins with GRAVY scores ranging from -0.860 to 0.780. As described earlier, sequence ladders containing the N- or C-terminus of BR were present in the spectrum even from the hydrolysate generated using a small amount of protein. In the analysis of this real complex mixture, many proteins were identified from terminal peptides (the peptide sequences used to identify each protein are given in Table IV-2-1). For example, two peptides identified for the soluble protein P19338 were from its C-terminus, i.e., DHKPQGKKTKFE and

GGDHKPQGKKTKFE. Three peptides identified for integral membrane protein Q9NZ01 were from its C-terminus, i.e., PPLRMPIIPFLL, PLRMPIIPFLL, and RMPIIPFLL. The sequence ladders provided additional information to confirm the protein identification results, and distinguish similar sequences from different proteins generated by database searching. For example, in the list of the proteins identified, there are three ATP carrier protein isoforms with differences in the C-terminal amino acid. Sequence ladders from the C-terminus clearly show the terminal amino acid difference, allowing the differentiation of these isoforms: the fibroblast isoform having T as the C-terminus, the T1 isoform having V as the C-terminus, and the T2 isoform having I as the C-terminus.

Table IV-2-1. Identified proteins from membrane protein fraction of human breast cancer cell line MCF7.

| ID | Description | Subcellular location | Pep[a)] | TMD[b)] | GRAVY[c)] |
|---|---|---|---|---|---|
| P21817 | Ryanodine receptor 1 | Integral membrane | MLPIGLNMCAPTDQDLITLA<br>RSNCALFSTNL | 12 | -0.309 |
| P98198 | Phospholipid-transporting ATPase ID | Integral membrane | HSNGLFDMFPN<br>NNRQSQVLINGILQQEQWMN | 10 | -0.057 |
| O43861 | Potential phospholipid-transporting ATPase IIB | Integral membrane | ALILTELLMVA<br>CLTGVEDQLQADV | 10 | 0.040 |
| Q9Y2W3 | Proton-associated sugar transporter A | Integral membrane | FSSLVSFL<br>FSSLVSFLG | 8 | 0.154 |
| Q04656 | Copper-transporter ATPase 1 | Integral membrane | AYSLIILLL<br>AYSLIILL<br>AYSLIIL | 8 | 0.092 |
| Q14973 | Sodium/bile acid cotransporter | Integral membrane | MIIILLCS<br>METGCQNV<br>GGMIIILLCS | 8 | -0.593 |
| Q8NGI8 | Olfactory receptor 5AN1 | Integral membrane | MPQLLILSC<br>QLLILSC | 7 | 0.763 |
| P04035 | 3-hydroxy-3-methylglutary-coenzyme A reductase | Integral membrane | VTLTICMMSMNMFTGNNKIC<br>VTLTICMMSMNMFTGNNKICG | 7 | 0.083 |
| Q8NGY3 | Olfactory receptor 6K3 | Integral membrane | MNNAIKKLFCLQKVL<br>KLFCLQKVLNKP | 7 | 0.710 |
| P07550 | Beta-2 adrenergic receptor | Integral membrane | VHVIQDNL<br>SNGNTGEQSGYHVE | 7 | 0.160 |
| P05141 | ATP carrier protein, fibroblast isoform | Integral membrane | MGGAFVLVLYDEIKKYT<br>GAFVLVLYDEIKKYT<br>GGAFVLVLYDEIKKYT<br>GMGGAFVLVLYDEIKKY<br>GMGGAFVLVLYDEIKKYT<br>TDAAVSFAKDFLAG<br>AFVLVLYDEIKKYT<br>TADKQYKGIIDCVVRIPKEQ | 6 | 0.024 |

| | | | | | |
|---|---|---|---|---|---|
| | | | VLRGMGGAFVLVLYDEIKKYT<br>SNVLRGMGGAFVLVLYDEIKKYT<br>SFWRGNLANVIRYF | | |
| P12235 | ATP carrier protein, isoform T1 | Integral membrane | GMGGAFVLVLYDEIKKY<br>SFWRGNLANVIRYF<br>MGGAFVLVLYDEIKKYV<br>AFVLVLYDEIKKYV<br>GAFVLVLYDEIKKYV<br>GGAFVLVLYDEIKKYV<br>SNVLRGMGGAFVLVLYDEIKKYV | 6 | 0.059 |
| P12236 | ATP carrier protein T2 | Integral membrane | AFVLVLYDELKKVI<br>GAFVLVLYDELKKVI<br>GGAFVLVLYDELKKVI<br>SFWRGNLANVIRYF | 6 | 0.075 |
| P54852 | Epithelial membrane protein-3 | Integral membrane | ALHILILIL<br>ALHILILI<br>ALHILIL | 4 | 0.780 |
| P51572 | B-cell receptor-associated protein 31 | Integral membrane | KGAAVDGGKLDVG<br>AESASEAAKKYME | 3 | -0.166 |
| Q9NZ01 | Synaptic glycoprotein SC2 | Integral membrane | PPLRMPIIPFLL<br>PPLRMPIIPFL<br>PLRMPIIPFLL<br>RMPIIPFLL<br>YPPLRMPIIPFL | 3 | 0.037 |
| P18850 | Cyclic-AMP-dependent transcription factor ATF-6 alpha | Type II membrane | QGSNSQLMAVQ<br>ALEQGSNSQLMAVQYT<br>ENVINGQDYEVMMQIDCQVM | 1 | -0.507 |
| P27824 | calnexin | Type I membrane | PDFFEDLEPFRMTPF<br>DWDERPKIP<br>PSYQGIWKPRKIPN | 1 | -0.874 |
| Q14517 | Cadherin-related tumor suppressor homolog precursor | Type I membrane | HLALLLLLL<br>HLALLLLL<br>HLALLLL | 1 | -0.295 |
| P28827 | Receptor-type protein-tyrosine phosphatase | Type I membrane | AHLLLPQL<br>AHLLLPQ | 1 | -0.381 |
| P07099 | Epoxide hydrolase 1 | Membrane-bound | SFYEFYKIIPLLTD<br>SFYEFYKIIPLL<br>PGSFYEFYKIIPLLTD | 1 | -0.261 |
| P78357 | Contactin associated protein 1 precursor | Type I membrane | QTAFHGCMELLKVDG<br>DGYVQRFILN | 1 | -0.315 |
| Q13873 | Bone morphogenetic protein receptor type II | Type I membrane | QQNLPKRPTSLPLNTKN<br>QQNLPKRPTSLPLNTK | 1 | -0.543 |
| P05026 | Sodium/potassium-transporting ATPase beta-1 chain | Type II membrane | NPNVLPVQCT<br>LLQPKYLQPL | 1 | -0.578 |
| Q13349 | Integrin alpha-D precursor | Type I membrane | KQLQEKIYAVEGTQS<br>LIGAPHYYEQ<br>DLILIGAPHYYEQ | 1 | -0.078 |
| P55283 | Cadherin-4 precursor | Type I membrane | YVMTITANDADDS<br>IITVTDVNDNP<br>VAQTSSP | 1 | -0.320 |
| Q9Y5H1 | Protocadherin gamma A2 precursor | Type I membrane | QDGPGLLTRAKVIVTV<br>KDGGNPSLSTDAHILLQ | 1 | -0.261 |
| P98155 | Very low-density lipoprotein receptor precursor | Type I membrane | QDDCSDGSDELDCAP<br>GATGTGRKAKCEPSQFQC | 1 | -0.418 |
| P05556 | Integrin beta-1 precursor | Type I membrane | YSVNGNNE<br>IPKSAVGTLSANS<br>QIQPQQLVLRLRSGEP | 1 | -0.407 |
| Q15011 | Methyl methanesulfonate (MMF)-inducible protein 1 | Integral membrane | PAENQPANQNAAPQVVVNPGAN<br>QNAAPQVVVNPGAN | 1 | -0.555 |
| Q9UKJ8 | ADAM 21 precursor | Type I membrane | LGTYIILI<br>LGTYIIL | 1 | -0.111 |
| P13224 | Platelet glycoprotein Ib beta chain precursor (GP-Ib beta) | Type I membrane | GLGLLHALLL<br>LLHALLL | 1 | 0.311 |
| O75396 | Vesicle trafficking protein SEC22b | Type IV membrane | VLLTMIARVADGLPLAA<br>VLLTMIARVAD | 1 | -0.182 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| P56181 | NADH-ubiquinone oxidoreductase | Mitochondrial inner membrane | KGQPQNSKKQSP<br>NSKKQSPPKKP | 0 | -0.860 |
| O00483 | NADH-ubiquinone oxidoreductase MLRQ subunit | Mitochondrial inner membrane | KKHPSLIPLFVFIG<br>AKKHPSLIPLFVFIG<br>QAKKHPSLIPLFVFIG<br>KHPSLIPLFVFIG<br>MLRQIIGQAKKHPSLIPLFVFIG<br>HPSLIPLFVFIG<br>PSLIPLFVFIG | 0 | -0.380 |
| P25705 | ATP synthase alpha chain | Mitochondrial inner membrane | KISEQSDAKLKEIVTNFLAGFEA<br>GKISEQSDAKLKEIVTNFLAGFEA<br>KISEQSDAKLKEIVTNFLAGFE<br>VPVGEELLGRVVDALG<br>GKISEQSDAKLKEIVTNFLAGFE<br>VPVGEELLGRVVDAL<br>AIVDVPVGEELLGRVVDALG | 0 | -0.067 |
| Q08378 | Golgi autoantigen, golgi subfamily A member 3 | peripheral membrane | KTLLQQNQQLKLDLR<br>ALQSQLQQVQLER<br>DQQLEALQQEHLDLMKQL<br>IALTSQELE | 0 | -0.835 |
| Q9Y2B2 | N-acetylglucosaminyl-phosphatidylinositol de-N-acetylase | ER membrane | INLVVTFDAGGVSGHSNHIA<br>QDVLFVLNSKEVAQAKK | 0 | 0.167 |
| P00387 | NADH-cytochrome b5 reductase | Membrane-bound | TGITPMLQVIRAIMKD<br>ITPMLQVIRAIMKD | 0 | -0.187 |
| P14060 | 3 beta-hydroxysteroid dehydrogenase | Membrane-bound | NGILSSVGKFSTVNPVYVGN<br>APYPHSKKLAE | 0 | -0.186 |
| O14939 | Phospholipase D2 | Membrane-associated | DFFQLWQDMAESNANI<br>DTETEPSLIDGAEYQAG | 0 | -0.298 |
| P10809 | 60KDa heat shock protein | mitochondrial | GKGDKAQIEKRIQEIIEQLD<br>GKGDKAQIEKRIQEIIEQL<br>DKAQIEKRIQEIIEQL<br>GDKAQIEKRIQEIIEQL<br>PVEIRRGVMLAVDAVIAELKKQ<br>PVEIRRGVMLAVDAVIAELKK<br>NPVEIRRGVMLAVDAVIAELKK<br>GANPVEIRRGVMLAVDAVIAELK<br>GKGDKAQIEKRIQEIIEQL<br>ANPVEIRRGVMLAVDAVIAELKKQ<br>ADARALMLQGVDLLADAVAVTM<br>GANPVEIRRGVMLAVDAVIAELKKQ | 0 | -0.076 |
| P31930 | Ubiquinol-cytochrome C reductase complex core protein I | mitochondrial | GMFWLRF<br>MFWLRF | 0 | -0.135 |
| P18859 | ATP synthase coupling factor 6 | mitochondrial | SSEYQQELERELFKLKQMF<br>SEYQQELERELFKLKQMF | 0 | -0.478 |
| Q04984 | Heat shock protein | mitochondrial | AGQAFRKFLPLFDRVLVE<br>AGQAFRKFLPLFDRVL | 0 | -0.060 |
| P30084 | Enoyl-CoA hydratase | mitochondrial | GLIDELNQALKIFEED<br>GLIDELNQALKIFE | 0 | -0.047 |
| P06576 | ATP synthase beta chain | mitochondrial | PAPATTFAHLDATTVLSRAIAELG<br>PAPATTFAHLDATTVLSRAIAEL<br>SEVSALLGRIPSAVG<br>GQDVLLFIDNIFRFTQAG<br>QDVLLFIDNIFRFTQAG | 0 | 0.018 |
| P49411 | Elongation factor Tu | mitochondrial | MVKPGSIKPHQKVEA<br>GRHKPFVSHFMPVMF | 0 | -0.122 |
| P19338 | Nucleolin (protein C23) | nuclear | DHKPQGKKTKFE<br>GGDHKPQGKKTKFE | 0 | -1.135 |
| Q14686 | Nuclear receptor coactivator 6 | nuclear | QIMTNQMQGN<br>QGPVNNSPSQVMGIQGQ<br>QPQLPQQQQPPPPSQPQSQ<br>QQGPPSQLM<br>GPPQNQMQVSHGPPNMMQPSLMG<br>QMSCGQNP<br>QNSTVSVAAVGGVVEDNKE<br>QPVSSPGRNPMVQQGNVPP | 0 | -0.699 |

| | | | | | |
|---|---|---|---|---|---|
| Q02446 | Transcription factor Sp4 | nuclear | QNAQDQSNSLQQVQIVG<br>QNAQDQSNSLQQVQIVGQ<br>QPQQLELVTTQLAGNAWQL | 0 | -0.439 |
| Q14498 | RNA-binding region containing protein 2 | nuclear | SSQLQPNGMQN<br>LQLMARLAEGTGLQIPPAA<br>AQQALQMSGSLAFGAVA | 0 | -0.649 |
| Q09472 | E1A-associated protein p300 | nuclear | QPLNMAPQPGLGQVGISPLKP<br>VNQMPTQP<br>NNKKTSKNKSSLSRGNKKK<br>ILHANPQLL | 0 | -0.728 |
| Q13838 | Probable ATP-dependent RNA helicase p47 | nuclear | GQQGVHSNPAMQNMNPMQAG<br>EEVLKKNCPHIVVGT<br>QLEPVTGQVSVLVMCH | 0 | -0.274 |
| P80217 | Interferon-induced 35KDa protein | nuclear | EIHFQKPTRGGGGRGPDS<br>AGSALITFDDPKVAEQVLQQK | 0 | -0.170 |
| Q9Y483 | Metal-response element-binding transcription factor 2 | nuclear | ASKPISDSREVSNGIEKKGKK<br>NTEILNNLADQELQLNHLKNSI | 0 | -0.644 |
| Q02241 | Kinesin-like protein KIF23 | nuclear | PLDADGDNVLQEKEQI<br>LPRCLDMIFNS | 0 | -0.817 |
| Q9UJ98 | Cohesin subunit SA-3 | nuclear | RHSRKQSE<br>MMNALFR | 0 | -0.38.3 |
| Q9Y4X4 | Krueppel-like factor 12 | nuclear | KLSHVHRIPV<br>NKLSHVHRIPV | 0 | -0.604 |
| Q9NZ71 | Helicase-like protein NHL | nuclear | GGAGGQFLSGQEWYRQQA<br>STAAAQQLDPQEHLN | 0 | -0.386 |
| Q8TEQ6 | Gem-associated protein 5 | nuclear | AFQKLQNIKYPSATNNT<br>RHSRKQSE | 0 | -0.398 |
| Q96ST3 | Paired amphipathic helix protein Sin3a | nuclear | SVRNDHGGTVKKPQLNNKPQR<br>PQHPSQPSAQSA | 0 | -0.686 |
| O75444 | Transcription factor Maf | nuclear | ALISNSHQLQGGFDGYARGA<br>AMSNSDLPTSPLAME | 0 | -0.495 |
| Q00444 | Homobox protein Hox-C5 | nuclear | PAGLSQPPAPPQIYP<br>KEEQAQTGQPAGLSQPPAPPQIY | 0 | -0.893 |
| Q03164 | Zinc finger protein HRX | nuclear | ESIPSRSSPEGPDPPVLTEVSKQ<br>DDCGNILPSDIM | 0 | -0.745 |
| O14686 | Myeloid/lymphoid or mixed-lineage leukemia protein 2 | nuclear | QIMTNQMQGN<br>QGPVNNSPSQVMGIQGQ<br>QPQLPQQQQPPPPSQPQSQ<br>QQGPPSQLM<br>GPPQNQMQVSHGPPNMMQPSLMG<br>QMSCGQNP<br>QNSTVSVAAVGGVVEDNKE | 0 | -0.635 |
| P78364 | Polyhomeotic-like protein 1 | nuclear | QPVSSPGRNPMVQQGNVPP<br>QSKPPVAPIKPPQLGAAKM<br>QVNRTPGSNVPLASQLILM | 0 | -0.483 |
| P55060 | Importin-alpha re-exporter | nuclear | ILFSSLILI<br>FSSLILI | 0 | -0.031 |
| Q14692 | Ribosome biogenesis protein BMS1 homolog | nuclear | MGPPKVGK<br>QGQKERRNQKSSLKGAEGQL<br>KGEMQKQAQLNRAEFEDQDDEA | 0 | -0.776 |
| Q9Y534 | RNA-binding protein PIPPin | nuclear,<br>cytoplasm | KFQAVEVVLTQLAP<br>KRTRTYSATARAS | 0 | -0.528 |
| P55786 | Puromycin-sensitive aminopeptidase | Cytoplasm,<br>nuclear | LSEEVRPQDT<br>KPIAAVMNTWT<br>LKILMDKPEMNVVLKN | 0 | -0.199 |
| Q96MG8 | Hypothetic protein | cytoplasm | KLESFIKNSD<br>QLIPQPL | 0 | -0.569 |
| Q9Y613 | FH1/FH2 domains-containing protein | cytoplasm | SDEIMDLLVQSVTKS<br>MPTEEERQKIEEAQL | 0 | -0.294 |
| O00505 | Importin alpha-3 subunit | cytoplasm | QVQAVIDAGLIPMIIHQ<br>KDQVEYLVQQNVIPPF | 0 | -0.119 |
| P11586 | C-1-tetrahydrofolate synthase | cytoplasm | LNEDSTVHGFLVQL<br>LKNQVTQLKEQVPGFTPRL<br>SELDLISRLSREH | 0 | -0.097 |
| Q9UHI6 | Probable ATP-dependent RNA helicase DDX20 | cytoplasm | SGNMNQNQRLDAMAKL<br>QTVNPQNGFVRNKVIEQRVPV | 0 | 0.497 |
| P00352 | Aldehyde dehydrogenase | cytoplasm | MDIDKVAF | 0 | -0.157 |

| | | | | | |
|---|---|---|---|---|---|
| | 1A1 | | QPTVFSNVTDEMR | | |
| Q92871 | Phosphomannomutase 1 | cytoplasm | ETSPGGNDFEIF<br>HLGEELLQDLINFCLS, | 0 | -0.369 |
| P55786 | Puromycin-sensitive aminopeptidase | cytoplasm | LSEEVRPQDT<br>KPIAAVMNTWT<br>LKILMDKPEMNVVLKN | 0 | -0.199 |
| P05787 | Keratin | cytoskeleton | PNIQAVRTQEKEQIK<br>LSPLVLEVD | 0 | -0.602 |
| P11021 | Glucose-regulated protein precursor | ER | STRIPKIQQLVKEFF<br>STRIPKIQQLVKEFFN<br>STRIPKIQQLVKEFFNG<br>TRIPKIQQLVKEFFN<br>GSTRIPKIQQLVKEFF<br>PKIQQLVKEFFNGKEPS<br>TNGDTHLGGEDFDQRV | 0 | -0.487 |
| P19224 | UDP-glucuronosyltransferase 1-6 precursor, microsomal | microsomal | TEYRNNMIVIGLY<br>RPVEPLDL | 0 | -0.028 |
| Q9Y4G6 | Talin 2 | Focal adhesion plaques | QAAAMQLSQCAKNLATSLA<br>QAQAEDLSAQLALI<br>LVRAAQKAAF | 0 | -0.216 |
| P15924 | Desmoplakin (DP) | desmosomal plaque. | LEAQIATGGIIDP<br>EIELKQVMQQ | 0 | -0.824 |
| P25391 | Laminin alpha-1 chain precursor | extracellular | GAGRITPAYEPKTAT<br>NPQTPGGSCQKCDCNP | 0 | -0.339 |
| P20382 | Pro-MCH precursor | | ENKVSKNTGSKHN<br>QGILLSASKSIRNL | 0 | -0.512 |
| Q92870 | Amyloid beta A4 precursor | | AEEKSQPVQGQASTIIGNGDLLLQ<br>QGQQDPNKNLSPTA | 0 | -0.627 |
| O15078 | Hypothetical protein KIAA0373 | | DNKQSLIEELQRKVKKLEN<br>DNKQSLIEELQRKVKKLE | 0 | -0.941 |
| P20929 | nebulin | | LATKERPHHHAGNQTT<br>LAKNMMQIQS<br>PVLLQAQVN | 0 | -0.847 |
| Q9UPQ9 | Protein KIAA 1093 | | QPNSWNKQHQQQQPPQ<br>QIPQFQLACQL | 0 | -0.982 |
| O60309 | Hypothetical protein KIAA 0563 | | LEDIQSSSLQQEAPAQLPQL<br>LEDIQSSSLQQ | 0 | -0.703 |
| P16499 | Rod cGMP-specific 3',5'-cyclic phosphodiesterase alpha | | YLHNCETRRGQILL<br>NVMKKLCFLLQADRM | 0 | -0.366 |
| P08206 | Calpactin I light chain | | PSQMEHAMETMMFTFHKFAG<br>PSQMEHAMETMMFTFHKFA | 0 | -0.384 |
| P11142 | Heat shock cognate 71Kda protein | | STRIPKIQKLLQDFF<br>STRIPKIQKLLQDFFN<br>STRIPKIQKLLQDFFNG<br>TRIPKIQKLLQDFFN<br>GSTRIPKIQKLLQDFF | 0 | -0.456 |
| P09525 | Annexin A4 | | SFEDALLAIVKCMRNK<br>SFEDALLAIVKCMRN<br>SFEDALLAIVKCMR | 0 | -0.447 |
| P21359 | neurofibromin | | CQDPNLLNPIHGIVQS<br>NNFNAVFSRISTR | 0 | -0.137 |
| P54652 | Heat shock-related 70Kda protein 2 | | STRIPKIQKLLQDFFNG<br>STRIPKIQKLLQDFF | 0 | -0.486 |
| Q96A49 | Synapse associated protein 1 | | NFASAATKKITESV<br>SLIKQSAQLTALAAQQQAAG | 0 | -0.750 |
| Q8TEY7 | Ubiquitin carboxyl-terminal hydrolase 33 | | LNIMEPSLLQFYISR<br>KAGYIEDLVLMLPQN | 0 | -0.487 |
| P08758 | Annexin A5 | | NLEQLLLAVVKSIR<br>SIRSIPAYLAETLYYAMKG<br>SIRSIPAYLAETLYYAMKGAG<br>SIRSIPAYLAETLYYAMKGA<br>SGNLEQLLLAVVKSIR<br>TSGNLEQLLLAVVKSIR | 0 | -0.337 |
| Q92520 | Protein FAM3C precursor | | DVAPFIEFLKAIQDG<br>GDVAPFIEFLKAIQDG | 0 | -0.088 |
| P08237 | 6-phosphofructokinase | | AEGAIDKNGKPITSEDIKNG | 0 | -0.174 |

| | | | | |
|---|---|---|---|---|
| P53804 | Tetratricopeptide repeat protein 3 | EGKGIFDSRKNVLGHM<br>KVPPRPILKQKC<br>KEHQVLQDQLQEVY | 0 | -0.589 |
| O95870 | Protein BAT 5 | QSSENKRQL<br>NQSSENKRQL | 0 | -0.243 |
| O94991 | Hypothetical proteinKIAA 0918 | GGASSVPLSVLIL<br>FPCSPAAYT | 0 | -0.331 |
| O43847 | Nardilysin precursor | MCENMQLYPLQDILT<br>VENMCENMQLYPLQ | 0 | -0.449 |
| Q8IWN7 | Retinitis pigmentosa 1-like 1 protein | CGSTGSSHQST<br>KNMDPRLQQT | 0 | -0.940 |
| P98082 | Disabled homolog 2 | QASFSPENAFSANL<br>QASFSPENAFSANLN<br>NQLLNKINEPPK | 0 | -0.631 |
| Q9NYL4 | FK506 binding protein 11 precursor | KQVIPGLEQSL<br>AIIPSHLAYG | 0 | 0.059 |
| Q14008 | CH-TOG protein, | LNILQQLAVAMGPNI<br>VLPPTCIQL<br>LDNKNPKIIVACIE<br>NFTTPRDEYIEQL | 0 | -0.363 |
| Q16825 | Protein tyrosine phosphatase | NSLNNPQPYLQPSPMSS<br>TNSLNNPQPYL | 0 | -0.575 |
| O15078 | Hypothetical protein KIAA0373 | DNKQSLIEELQRKVKKLEN<br>DNKQSLIEELQRKVKKLE | 0 | -0.941 |
| P05387 | 60S acidic ribosomal protein P2 | MRYVASYLLAALGGN<br>MRYVASYLLAALGG<br>MRYVASYLLAALG<br>MRYVASYLLAAL | 0 | -0.237 |
| P03897 | NADH-ubiquinone oxidoreductase chain 3 | MSSLLLIII<br>MSSLLLII | 0 | 0.992 |
| O60610 | Diaphanous protein homolog 1 | ILMATMLNGAAVMD<br>AKKEMASLSAAAITV | 0 | -0.555 |
| P41091 | Eukaryotic translation initiation factor 2 subunit 3 | ILMATMLNGAAVMD<br>MDAALLLIAGNESCPQPQT | 0 | -0.018 |
| O75962 | Triple functional domain protein | QVQQKAEAMLQANH<br>ITASSLQEAEQLQR | 0 | -0.521 |
| Q15759 | Mitogen-activated protein kinase 11 | LLQGKALFPGSDY<br>TPSPEVLAKISSEH | 0 | -0.325 |
| O60610 | Diaphanous protein homolog 1 | ILMATMLNGAAVMD<br>AKKEMASLSAAAITV | 0 | -0.555 |

a) Pep: Identified peptides.
b) TMD: transmembrane domain.
c) GRAVY: Grand average of hydrophobicity.

### IV. 2. 4. Conclusion

The above example illustrates that, even with only one-dimensional HPLC separation of the hydrolysate, we can already identify a number of membrane proteins from the MCF7 membrane-enriched fraction. This method is very simple and quick; the acetone-precipitated sample is merely suspended in 25% TFA, followed by microwave irradiation for 10 minutes. The hydrolysate can be directly injected into HPLC for MS and MS/MS analysis.

### IV. 2. 5. Cited Literature

(1) Wallin, E. & Von Heijne, G. *Protein Sci.* **1998,** *7,* 1029-1038 .

(2) Blonder, J. et. al. *J. Proteome Research* **2002,** *1,* 351-360.

(3) Santoni, V., Kieffer, S., Desclaux, D., Masson, F. & Rabilloud, T. *Electrophoresis* **2000,** *21,* 3329-3344.

(4) Santoni, V., Molloy, M. & Rabilloud, T. *Electrophoresis* **2000**, *21*, 1054-1070.

(5) Wilks, M. R., Gasteiger, E., Sanchez, J. C., Bairoch, A. & Hochstrasser, D.F. *Electrophoresis* **1998,** *19,* 1501-1505.

(6) Patterson, S. D. & Aebersold, R. *Electrophoresis* **1995,** *16,* 1791-1814.

(7) Bell, A. W. et. al. , *protein Sci.***1998,** *7,* 758-764.

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

(9) Dai, Y. Q.; Whittal, R. M.; Li, L. *Anal. Chem.*, **1999,** *71*, 1087-1091.

# Part V. Perspectives and Challenges

Mass spectrometry has been widely applied to identify proteins by searching protein and nucleotide sequence databases. This technique has advanced protein expression studies and the identification of proteins in complexes that are essential in understanding the networks of proteins involved in biological processes. Efficient identification of protein primary sequence, detection of posttranslational modifications and location of amino acid mutations are fundamentally important.

A method for peptide *de novo* sequencing using low energy collision induced dissociation was developed to identify unknown proteins isolated from species without DNA or protein databases. The simplified calculation of this method limits the generation of sequence candidates and improves the accuracy of similarity searching. Another advantage of peptide *de novo* sequencing over database searching is the effective identification of modifications. There are more than 200 different kinds of modifications. Because these posttranslational modifications cannot be predicted from DNA databases, difficulties are often associated with searching homologous nonidentical matches to database-derived sequences even if the database search program has been informed some modifications. The more modification possibilities have been considered, the greater the number of database sequences with correct mass that will be obtained and the greater the possibilities to get false positive matches. *De novo* peptide sequencing is based on the calculation of experimental data and thus should be more effective in this case. Methodology development for more accurate or unique *de novo* peptide sequencing is achieved through $^{15}$N *in vivo* isotope labeling and chemical derivation at the terminus of peptides. Even when extra sample preparation is needed, they are indispensable for the situations where only limited peptides are available and are not enough for similarity searching.

For species that have complete DNA databases, challenges stem from the detection of low abundance protein, posttranslational modifications and single-nucleotide polymorphism. Low abundance proteins usually result in incomplete MS/MS spectra with low S/N ratio and in many cases there is only one peptide detected. Additional information is needed for efficient protein identification. 2D-MS method based on uniform $^{15}N$ labeling was developed in this work. The specificity of each peptide and fragment ions was increased by correlating m/z value with the corresponding nitrogen number. It was demonstrated that this 2D-MS method is effective for data interpretation of low quality spectra and single peptide based protein identification resulting from low abundance proteins.

Though tremendous success has been achieved during the past few years for detection of posttranslational modifications and location of single-nucleotide polymorphism by "bottom up" approach based on proteolytic peptides using tandem MS/MS data, people are still trying to find a more efficient way. It can be imagined that several hundreds of peptides may be produced by enzymatic digestions. One problem is to detect all these peptides in order to cover the whole sequence. Ion suppression and the dynamic range of tandem mass spectrometers make it difficult. Furthermore, even when all these peptides have been detected, the modifications are not predicted by the DNA database so the database searching cannot release a significant hit in many cases. Recently developed "top down" approaches using high accuracy and high resolution FT-ICR mass spectrometer circumvent this problem to some extent and have successfully detected many kinds of posttranslational modifications. In this work, a simpler and efficient intact protein *de novo* method was developed based on sequence ladder generation from C- and N-terminus of proteins by microwave irradiation in strong acid. The co-existing two series of laddes from the C- and the N-terminus complement each other and produce much higher sequence coverage and much clearer and simplified spectra than that of current techniques. With the advantage of MALDI-TOF detection, there are less multiple charge states so the spectra are easier to interpretate. MALDI-TOF also has higher sensitivity than that of ESI, so it is possible for this technique to perform much more sensitive detection. Currently, the main limitation is the resolution and ion suppression in high mass region with MALDI-TOF detection. Side reactions can be avoided by controlling the reactions. Future instrumentation is needed to improve the resolution. HPLC should be used to separate the produced polypeptides and decrease the complexity of the sample thus decreasing the ion suppression. In this work, theoretical and experimental data have demonstrated that this technique is practical though more future works is needed for improvement. One approach is H/D exchange that should be investigated for accurate identification of cross-links such as disulfide bonds. Another interesting experiment is the development of histidine tag affinity chromatography for selectively isolating phosphorylated proteins. It is expected that this should become a new, exciting and simple method for the study of protein phosphorylation. Automation and miniaturation of this technique for high-throughput and sensitivity is also in the future plan.

As a special topic in proteomics, membrane proteins remain a challenge because the established mass spectrometric methods are not compatible with them due to the poor solubilization, low abundance and extensive modifications. The detergent-free technique based on microwave-assisted acid hydrolysis provides a strategy to deal with this situation with high speed and high efficiency. The unique cleavage characterization makes it possible to detect many protein isoforms results from posttranslational modifications. The main disadvantage is the complexity of the produced peptides. Multidimensional separation is needed.

Application of the techniques developed in this work to clinical screening of human disease is the goal. I hope that the hard work during these years will be useful in promoting health and happy lives in the future.

## Appendix I. Flow Chart of Extraction, Separation and Purification of Bioactive Components from Shark Cartilage.

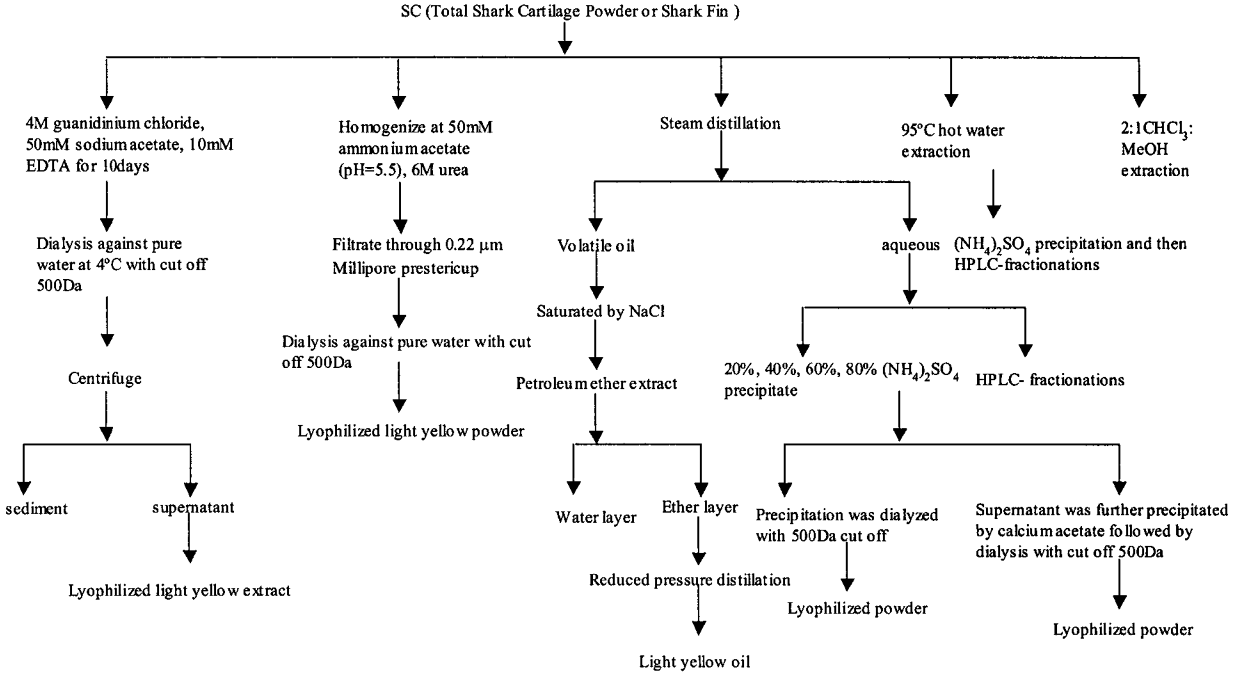

SC (Total Shark Cartilage Powder or Shark Fin )
4M guanidinium chloride, 50mM sodium acetate, 10mM EDTA for 10days
Dialysis against pure water at 4°C with cut off 500Da
Centrifuge
sediment
supernatant
Lyophilized light yellow extract
Homogenize at 50mM ammonium acetate (pH=5.5), 6M urea
Filtrate through 0.22 μm Millipore prestericup
Dialysis against pure water with cut off 500Da
Lyophilized light yellow powder
Steam distillation
Volatile oil
Saturated by NaCl
Petroleum ether extract
Water layer
Ether layer
Reduced pressure distillation
Light yellow oil
aqueous
20%, 40%, 60%, 80% $(NH_4)_2SO_4$ precipitate
Precipitation was dialyzed with 500Da cut off
Lyophilized powder
HPLC- fractionations
Supernatant was further precipitated by calcium acetate followed by dialysis with cut off 500Da
Lyophilized powder
95°C hot water extraction
$(NH_4)_2SO_4$ precipitation and then HPLC-fractionations
2:1$CHCl_3$: MeOH extraction

Figure-Appendix-1. Flow chart of the extraction, separation and purification of bioactive components from shark cartilage.

## Appendix II. GC/MS Analysis of Volatile Components Isolated from Shark Cartilage.

Instrument conditions:
7070E VG Analytical Organic Mass Spectrometry
Varian Vista 6000 Gas Chromatograph
ZB-5 7HG-G002-11(5%phenyl polysiloxane) 30m 0.25mm i.d.
Injector: 250°C
Source: 270°C
EI,70eV
Initial Temp: 80°C
Final Temp : 250°C
Rate: 3°C /min

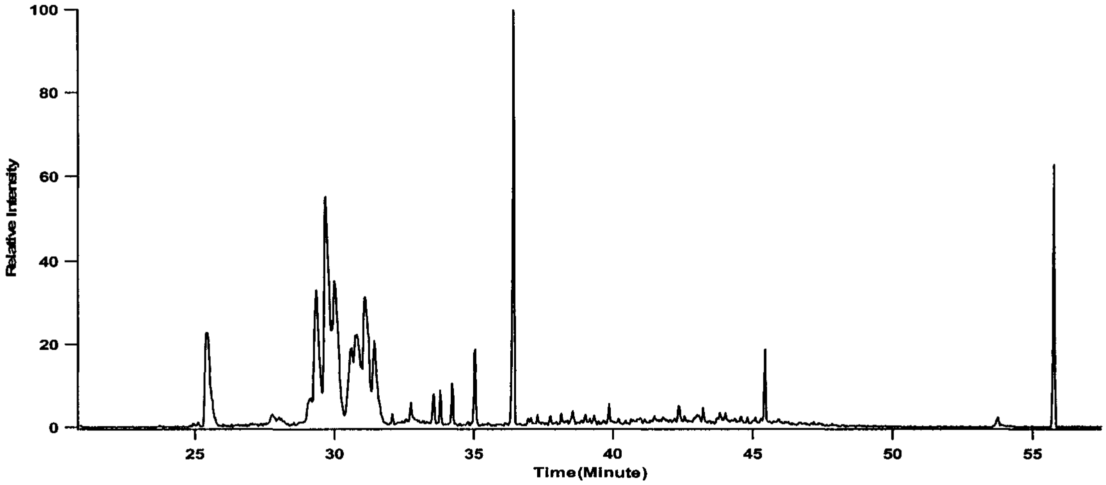
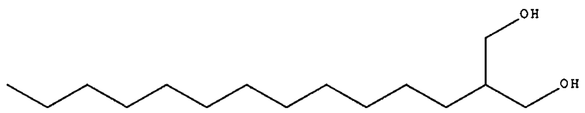
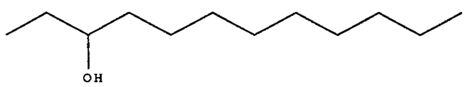
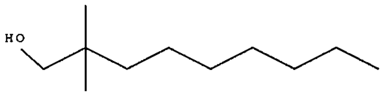


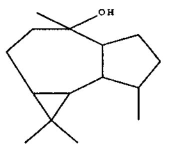
Figure-Appendix-2. Gas chromatography of volatile components from shark cartilage.

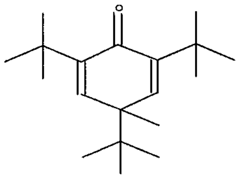
Identified components by NIST database searching:

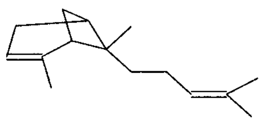


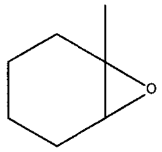


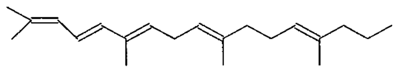


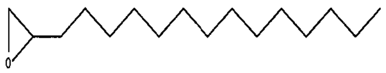


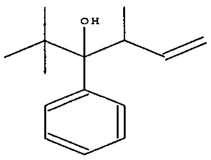


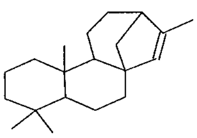


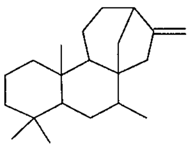


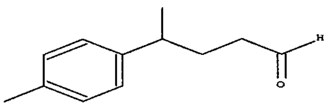


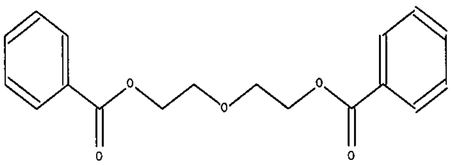


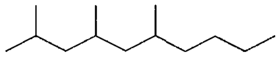


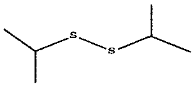


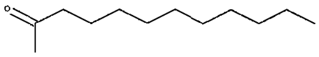


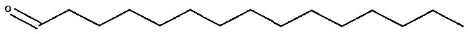


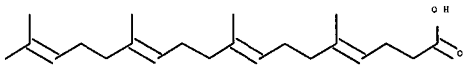


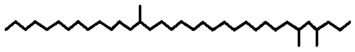


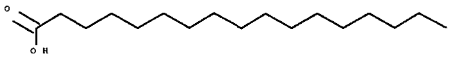

S
S

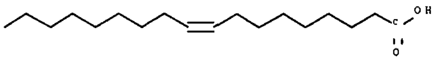

O

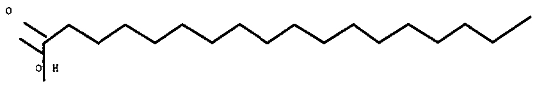

O

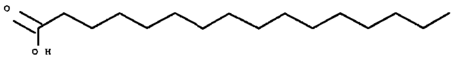

O H
O

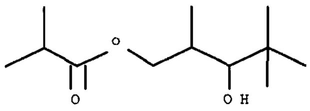


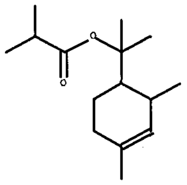

O
O H

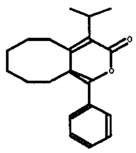

O H
C
O

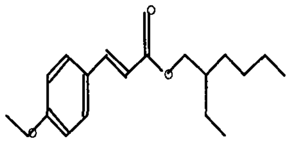

O
O H


O
O H


O
O
O H


O
O


O
O


O
O
O

The following is one typical mass spectrum (* represent the peaks that match with NIST database)

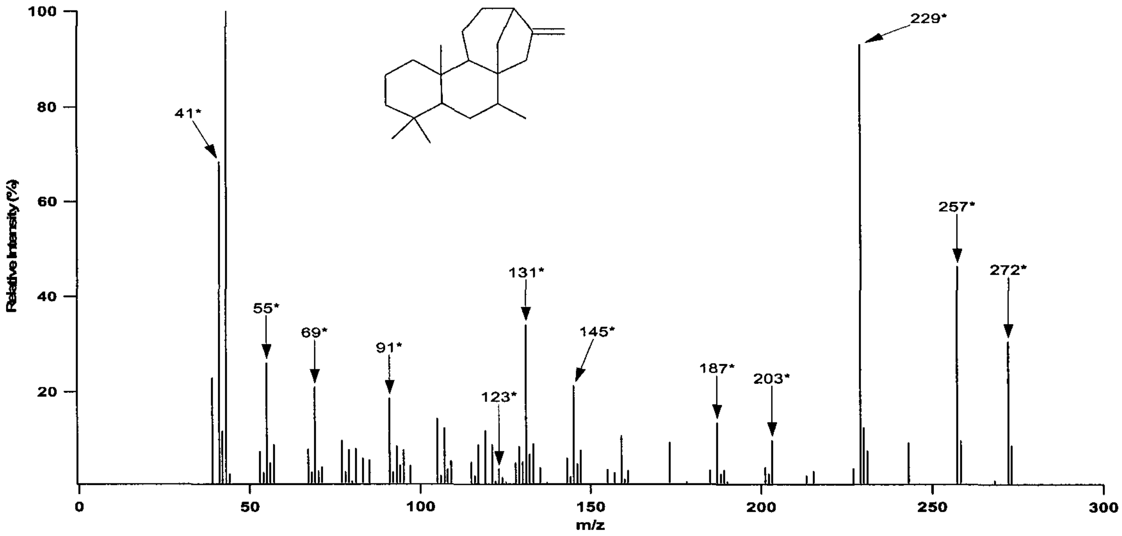
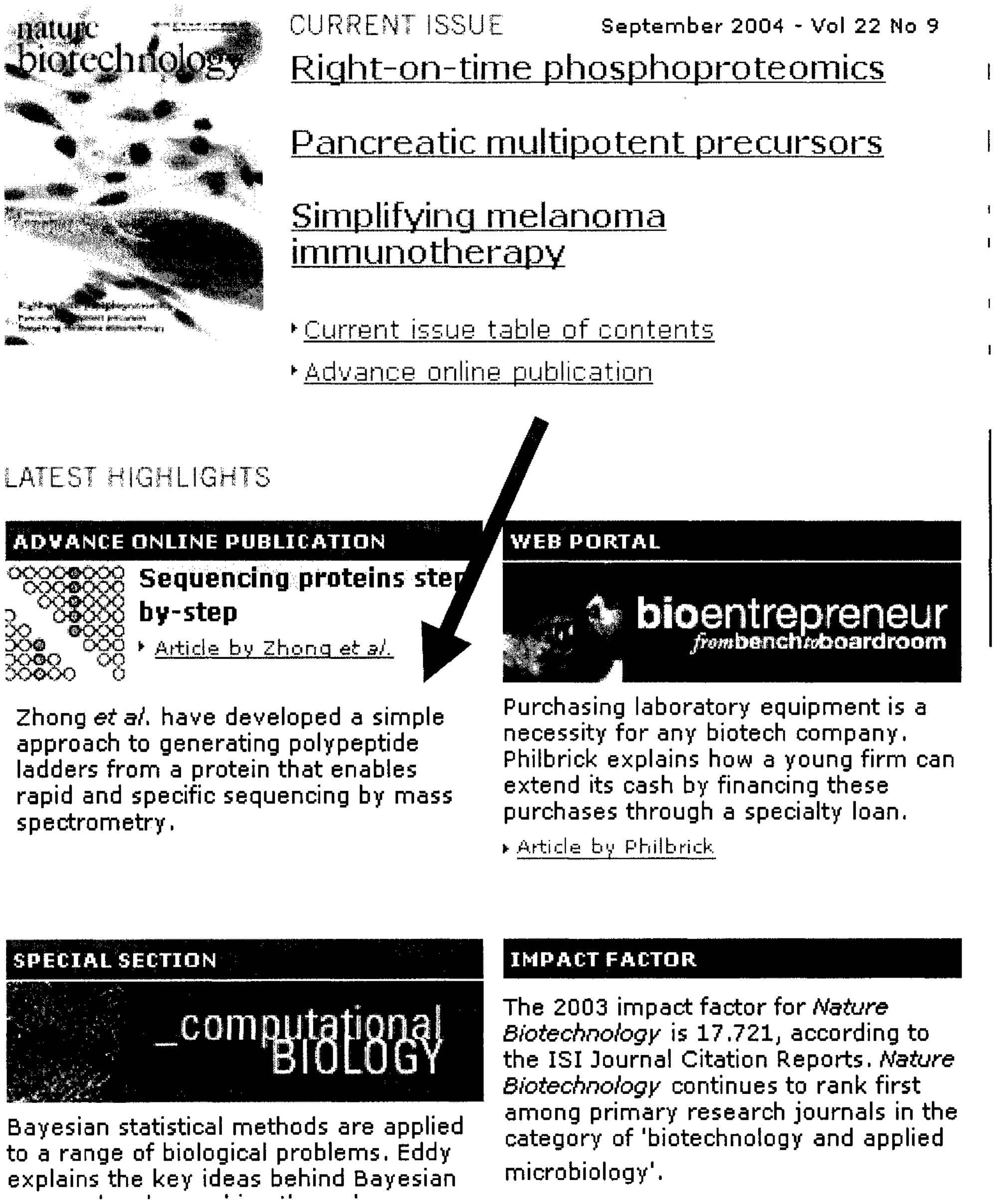


Figure-Appendix-3. GC/MS spectrum of one component from shark cartilage

## Appendix III. A Highlighted Research Paper in Nature Biotechnology.



CURRENT ISSUE September 2004 - Vol 22 No 9

Right-on-time phosphoproteomics

Pancreatic multipotent precursors

Simplifying melanoma immunotherapy

- Current issue table of contents
- Advance online publication

LATEST HIGHLIGHTS

**ADVANCE ONLINE PUBLICATION**

**Sequencing proteins step-by-step**

- Article by Zhong *et al.*

Zhong *et al.* have developed a simple approach to generating polypeptide ladders from a protein that enables rapid and specific sequencing by mass spectrometry.

**WEB PORTAL**

bioentrepreneur *from*bench*to*boardroom

Purchasing laboratory equipment is a necessity for any biotech company. Philbrick explains how a young firm can extend its cash by financing these purchases through a specialty loan.

- Article by Philbrick

**SPECIAL SECTION**

_computational BIOLOGY

Bayesian statistical methods are applied to a range of biological problems. Eddy explains the key ideas behind Bayesian

**IMPACT FACTOR**

The 2003 impact factor for *Nature Biotechnology* is 17.721, according to the ISI Journal Citation Reports. *Nature Biotechnology* continues to rank first among primary research journals in the category of 'biotechnology and applied microbiology'.