University of Alberta

Development of mass spectrometric techniques for protein sequencing and metabolome analysis

by

Nan Guo ©

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

Department of Chemistry

Edmonton, Alberta
Spring 2006

Your file Votre référence
ISBN: 0-494-13818-1
Our file Notre référence
ISBN: 0-494-13818-1

Abstract

The objective of this research is to develop new mass spectrometric methods for protein and metabolite characterization. One major challenge in protein characterization by mass spectrometry (MS) is the study of protein modifications. In this work, a technique involving gel electrophoresis separation of proteins, protein extraction, microwave-assisted acid hydrolysis of the extracted protein, and MS analysis of the polypeptides is developed. This technique is used to characterize a hemoglobin variant from a human blood sample. For human metabolite analysis, one major challenge is to identify the metabolites in a complex sample. A tandem MS technique is developed for metabolite identification based on the match of acquired MS/MS spectra with those in a MS/MS spectral library. The spectral library is created by using a triple quadrupole mass spectrometer and consists of 215 human metabolites. It is demonstrated that this technique can be used to identify metabolites in human urine.

Acknowledgement

First of all and foremost, I would like to thank my supervisor, Dr. Liang Li, for the precious opportunity to pursue my studies in his research group for the past three years. I truly appreciate his encouragement, inspiration, guidance and enthusiasm through my research work. What I have learned from him, without any doubt, will continue to benefit my future life and career.

I would like to thank my committee members, Dr. John S. Klassen and Dr. David S. Wishart, for their time, their comments and their constructive criticism on my thesis.

I would also like to thank everyone in our research group, past and present, Dr. Nan Zhang, Dr. Nan Li, Dr. Rui Chen, Dr. David Craft, Dr. Huai-zhi Liu, Dr. Hongying Zhong, Dr. Cheng-jie Ji, Dr. Ying Zhang, Mr. Chris McDonald, Ms. Jing Zheng, Ms. Xin-lei Yu, Ms. Tien Quach, Ms. Li-dan Tao, Mr. Jacek Stupak, Mr. Mulu Gebre, Mr. Andy Lo, Ms. Melisa Clements, Ms. Andrea De Souza, Mr. Kevin Guo, Mr. Leon Lau, Ms. Helen (Nan) Wang, Mr. Zhi-hui Wen, Mr. Bryce Young.

As well, I am very grateful to Dr. Randy M. Whittal and Don Morgan from Mass Spectrometry Laboratory for their invaluable assistance. Their knowledge and expertise were critical to the completion of my thesis work.

My gratitude also goes to Dan Tzur from Dr. David S. Wishart's lab for the human metabolites supply, Mr. Trefor N. Higgin from Dynacare Kasper Medical Laboratories for the cooperation on hemoglobin analysis, Mr. Micheal Carpenter for his revision on my thesis, and Dr. Sandra Marcus for her assistance of protein extraction.

Finally, I would like to thank my family, my parents and my brothers, for their support, faith, trust, encouragement and their sacrifice. My special thanks to my dearest, my best friend, Hongbin Li. I deeply appreciate what he did in the past several years, love and understanding.

Table of Contents

Chapter 1

Introduction 1

1.1 Introduction to Mass Spectrometry for Proteomics and Metabolomics 1

1.1.1 ESI-TripleQ system 2

1.1.1.1 Electrospray Ionization (ESI) 2

1.1.1.2 Triple Q system 3

1.1.2 MALDI-MS system 5

1.1.2.1 Matrix-assited Laser Desportion/Ionization (MALDI) 5

1.1.2.2 MALDI-TOF MS system 9

1.1.2.3 MALDI-QqTOF MS system 14

1.2 Introduction to Proteomics and Metabolomics 16

1.2.1 Protein Characterization 17

1.2.2 Metabolomics 22

1.3 Overview of the Thesis 25

1.4 Literature Cited 26

Chapter 2

De Novo Protein Sequencing by Microwave-enhanced acid-hydrolysis combined with SDS-PAGE 32

2.1 Introduction 32

2.2 Experimental 35

2.2.1 Materials and Reagents 35

2.2.2 SDS-PAGE 36

2.2.3 Gel Excision & Protein Extraction 36

2.2.4 Protein Digestion 36

2.2.5 Sample Preparation for MALDI-TOF MS Analysis 37

2.2.6 MALDI-TOF analysis 37

2.3 Results & Discussion 38
2.4 Conclusions 56
2.5 Literature Cited 56

Chapter 3
Identification of Hemoglobin G Coshatte Using MALDI-TOF MS Combined With Microwave-Assisted Acid Hydrolysis of Proteins 60

3.1 Introduction 60
3.2 Experimental 62
3.2.1 Materials and Reagents 62
3.2.2 Hb cell extraction 62
3.2.3 Hb purification by RP-HPLC 62
3.2.4 Microwave-assisted acid hydrolysis of Hemoglobin 62
3.2.5 Protein Preparation for Proteolytic Digestion 62
3.2.6 Proteolytic Digestion 64
3.2.7 MALDI-TOF MS analysis 64
3.2.8 MALDI QqTOF analysis of tryptic peptides 65
3.3 Results & Discussion 65
3.4 Conclusions 80
3.5 Literature Cited 81

Chapter 4
Creation and Application of Human Metabolomics MS/MS Database 83

4.1 Introduction 83
4.2 Experimental 86
4.2.1 Reagents and Materials 86
4.2.2 Sample Preparation 87
4.2.3 Instrumentation 88
4.3 Results and Discussion 92

4.3.1 Library Creation ... 92
4.3.2 Investigation of reproducibility parameters ... 99
4.3.3 Limit of Detection Analysis ... 105
4.3.4 Preliminary Human Urinary Analysis ... 110
4.4 Conclusions ... 135
4.5 Literature Cited ... 135

Chapter 5

Conclusions and Future Work ... 139

List of Figures

Figure 1.1. Schematic representation of the ESI process in the positive ion mode. 2

Figure 1.2. Schematic diagram of a Triple Quadrupole mass spectrometer. 4

Figure 1.3. Schematic representation of the MALDI process.. 6

Figure 1.4. Principle of linear (above) and reflectron (bottom) TOF mass spectrometer in the positive ion detection mode. 10

Figure 1.5. Schematic illustration of the time-lag focusing principle in MALDI-TOF MS. (A) Ions with same m/z but different initial kinetic energy expand away from the repeller. (B) Ions with lower initial energy move further from the repeller that is applied to an extraction pulse after a certain time delay. (C) Ions reach the detector at the same time due to the energy compensation.. 13

Figure 1.6. Schematic Diagram of the QqTOF mass spectrometer................................ 15

Figure 1.7. Relationship in a biological system. 16

Figure 1.8. Nomenclature for peptide fragmentation pattern under CID........................ 19

Figure 1.9. MAP protein sequencing for top-down proteomics...................................... 21

Figure 2.1. Effect of ACN on protein extraction. (a) 50%ACN, (b) 60%ACN, (c) 75%ACN. With the increasing CAN percentage, more and more impurity peaks in the low mass range were observed, as shown in the circle. 40

Figure 2.2. Effect of extraction time on protein extraction. (a) 10 min (b) 20 min, (c) 30 min. Apparently, 20 min is the optimal extraction time to obtain high extraction efficiency as well as avoid producing impurity peaks in the low mass range...42

Figure 2.3. Effect of extraction solution volume on protein extraction. (a) 40 μL (b) 50μL

(c) 60μL (d) 80μL. 44

Figure 2.4. Mass spectra of MAAH of 20ug myoglobin protein extracted from gel piece. (a) Full Spectrum, (b) Mass Range of 1800-3000 Da, (c) Mass Range of 3000-4500 Da, (d) Mass Range of 4500-6450Da, (e) Mass Range of 6450-8400Da. Amino acids sequence given at the upper is C-terminal polypeptides while amino acids sequence given lower is N-terminal polypeptides. 46

Figure 2.5. Sensitivity of amino acid sequence reading by MAAH of horse myglobin protein. (a) 0.3 nmol, (b) 0.1 nmol, (c) 0.06 nmol, (d) 0.03 nmol and (e) 0.01 nmol. In this discussed mass range, the less the protein amount loading on the gel and extraction, the less the number of amino acids was identified. 50

Figure 2.6. Mass spectra of MAP of bovine α-S1-casein over 2000 – 7500 Da............. 53

Figure 2.7. Mass spectra of α-S1-casein (a) and de-phosphorylated α-S1-casein (b) over 5000-6000 Da.. 55

Figure 3.1. Steps taken to identify the amino acid variant by MALDI MS analysis....... 66

Figure 3.2. RP-LC sparation chromatograph of Hemoglobin G Coushatta.................. 67

Figure 3.3. MALDI-TOF MS of intact normal and variant β-chain proteins................. 68

Figure 3.4. MALDI-TOF MS of MAP for variant β-chain (A) and normal β-chain (B) over 1200 Da–3300Da. 70

Figure 3.5. MALDI-TOF MS of tryptic digest of Hb protein fraction............................ 76

Figure 3.6(a). CID MS/MS of normal (above) and variant (bottom) peptide to determine the variant position.. 78

Figure 3.6(b). MASCOT search result of normal peptide 1314 Da. The score above 45 indicates identity. 79

Figure 4.1. Ion chromatograph of metabolites separation in positive mode................. 106

Figure 4.2. Ion chromatograph of metabolites separation in negative mode. 108

Figure 4.3. Ion chromatograph of urinary analysis in positive mode. 111

Figure 4.4(a). Product ion spectra of standard (left) and urinary (right) uracil at low level energy. 113

Figure 4.4(b). Product ion spectra of standard (left) and urinary (right) uracil at medium level energy. 113

Figure 4.5(a). Product ion spectra of standard (left) and urinary (right) Xanthine at low level energy. 114

Figure 4.5(b). Product ion spectra of standard (left) and urinary (right) Xanthine at medium level energy. 114

Figure 4.6(a). Product ion spectra of standard (left) and urinary (right) Hypoxanthine at low level energy. 115

Figure 4.6(b). Product ion spectra of standard (left) and urinary (right) Hypoxanthine at medium level energy. 115

Figure 4.7(a). Product ion spectra of standard (left) and urinary (right) Asparginine at low level energy. 116

Figure 4.7(b). Product ion spectra of standard (left) and urinary (right) Asparginine at medium level energy. 116

Figure 4.8(a). Product ion spectra of standard (left) and urinary (right) Serine at low level energy. 117

Figure 4.8(b). Product ion spectra of standard (left) and urinary (right) Serine at medium level energy. 117

Figure 4.9(a). Product ion spectra of standard (left) and urinary (right) Betaine at low level energy. 118

Figure 4.9(b). Product ion spectra of standard (left) and urinary (right) Betaine at medium level energy. 118

Figure 4.10(a). Product ion spectra of standard (left) and urinary (right) Threonine at low level energy. 119

Figure 4.10(b). Product ion spectra of standard (left) and urinary (right) Threonine at medium level energy. 119

Figure 4.11(a). Product ion spectra of standard (left) and urinary (right) Glutamine at low level energy. 120

Figure 4.11(b). Product ion spectra of standard (left) and urinary (right) Glutamine at medium level energy. 120

Figure 4.12(a). Product ion spectra of standard (left) and urinary (right) Creatine at low level energy. 121

Figure 4.12(b). Product ion spectra of standard (left) and urinary (right) Creatine at medium level energy. 121

Figure 4.13(a). Product ion spectra of standard (left) and urinary (right) Creatinine at low level energy. 122

Figure 4.13(b). Product ion spectra of standard (left) and urinary (right) Creatinine at medium level energy. 122

Figure 4.14(a). Product ion spectra of standard (left) and urinary (right) Histidine at low level energy. 123

Figure 4.14(b). Product ion spectra of standard (left) and urinary (right) Histidine at medium level energy. 123

Figure 4.15. Ion chromatograph of urinary analysis in negative mode. 124

Figure 4.16(a). Product ion spectra of standard (left) and urinary (right) Citric Acid at low level energy. 126

Figure 4.16(b). Product ion spectra of standard (left) and urinary (right) Citric Acid at medium level energy. 126

Figure 4.17(a). Product ion spectra of standard (left) and urinary (right) Aconitic Acid at low level energy. 127

Figure 4.17(b). Product ion spectra of standard (left) and urinary (right) Aconitic Acid at medium level energy. 127

Figure 4.18(a). Product ion spectra of standard (left) and urinary (right) Glutaconic Acid at low level energy. 128

Figure 4.18(b). Product ion spectra of standard (left) and urinary (right) Glutaconic Acid at medium level energy. 128

Figure 4.19(a). Product ion spectra of standard (left) and urinary (right) Hippuric Acid at low level energy. 129

Figure 4.19(b). Product ion spectra of standard (left) and urinary (right) Hippuric Acid at medium level energy. 129

Figure 4.20(a). Product ion spectra of 3-Ureidopropionic acid(left) and Asparagine(right) at low level energy. 132

Figure 4.20(b). Product ion spectra of 3-Ureidopropionic acid(left) and Asparagine(right) at medium level energy. 132

Figure 4.21(a). Product ion spectra of Homoserine(left) and Threonine(right) at low level energy. 133

Figure 4.21(b). Product ion spectra of Homoserine(left) and Threonine(right) at low level energy. 133

List of Tables

Table 1.1. Contaminant concentration tolerated in MALDI MS technique [26] 8

Table 1.2. Basic classification of metabolites. 25

Table 3.1. N-terminus polypeptide of normal Hb β-chain. 72

Table 3.2. N-terminus polypeptide of variant Hb β-chain. 73

Table 3.3. C-terminus polypeptide of normal Hb β-chain. 74

Table 3.4. C-terminus polypeptide of variant Hb β-chain. 75

Table 4.1. Summary of instrument parameters 89

Table 4.2. Gradient used in LOD analysis. 90

Table 4.3. Gradient used in urinary analysis. 91

Table 4.4. List of metabolites analyzed for the construction of MS/MS human metabolomics database. 93

Table 4.5. Different LC system used to verify independence of the database and their results in positive mode. 101

Table 4.6. Different LC system used to verify independence of the library and their results in negative mode. 103

Table 4.7. Reproducibility investigation on sample concentration in positive mode. 107

Table 4.8. Reproducibility Investigation on sample concentration in negative mode. ... .109

Table 4.9. Identified urinary metabolites in positive mode. .... .112

Table 4.10. Identified urinary metabolites in negative mode. 125

Table 4.11. Comparison of reslut searching score for the peaks at 10.4 min and 11.5 min. 134

List of Abbreviations

CID	collision-induced dissociation
CHCA	α-cyano-4-hydroxycinnamic acid
CE	capillary electrophoresis
DHB	2, 5-dihydroxybenzoic acid
ESI	electrospray ionization
FIA	flow Injection Analysis
FT-ICR	fourier-transform ion cyclotron resonance
FT-IR	fourier-transform infrared spectroscopy
FWHM	full width at half maximum
GC	gas chromatography
HPLC	high performance liquid chromatography
LOD	limitation of detection
LLE	liquid-liquid extraction
MS	mass spectrometry
m/z	mass to charge
MALDI	matrix-assisted laser desorption/ionization
MRM	multiple reaction monitoring
MAP	mass analysis of polypeptide ladders
MAAH	microwave-assisted acid hydrolysis
MW	molecular weight
NMR	nuclear magnetic resonance
PTM	post-translational modification

QqTOF	quadruple time-of-flight
r.f.	radio frequency
RI	retention time Index
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SDS	sodium dodecyl sulfate
S/N	signal to noise ratio
TFA	trifluoroacetic acid
Triple Q	triple quadrupole
Tris	tris (hydroxymethyl) aminomethane
TOF	time-of-flight
SPE	solid phase extraction
UV	ultraviolet
ppm	part(s) per million
m	milli- (10^{-3})
µ	micro- (10^{-6})
n	nano- (10^{-9})
p	pico- (10^{-12})

Chapter 1

Introduction

1.1 Introduction to Mass Spectrometry for Proteomics and Metabolomics

In the post-genomic era, increasing efforts have been made to describe the relationship between the genome and the phenotype in cells and organisms. It is clear that a comprehensive understanding of the state of the genes, RNA transcripts, proteins and metabolites in a living system is necessary to reveal its phenotype. On the analytical point of view, mass spectrometry (MS) is a major tool. Electrospray ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI) are two soft ionization methods for the ionization of bio-molecules, ranging in size from small molecules, such as metabolites, to large molecules, such as proteins and peptides, without significant in-source sample degradation. Therefore, the advent of these two soft ionization methods combined with improved mass spectrometers makes it possible for protein and peptide analysis [1-5] as well as metabolome study [6-8]. Considering their importance and uses in my thesis work, the fundamentals of these two ionization methods and the principles of the related mass spectrometry methods will be introduced first, followed by an introduction to some of the current general methods used in the proteomics and metabolomics research fields.

1.1.1 ESI TripleQ Mass Spectrometry

1.1.1.1 Electrospray Ionization (ESI)

ESI-MS was introduced as a means of introducing samples into a mass spectrometer in the 1980's by Yamashita and Fenn [9, 10] and Aleksandrov et al. [11]. ESI has become an important method for biochemical applications [12-14].

In brief, ESI is a method to transfer ions from solution to the gas phase in which the ions can be subjected to mass spectrometric analysis. It is a process whereby ionized species in the gas phase are created from an analyte-containing solution via highly charged fine droplets by means of spraying the solution from a narrow-bore needle tip at atmospheric pressure in the presence of a high electric field. This technique affords ion transfer of a wide variety of ions dissolved in a variety of solvents. Figure 1.1 shows a schematic representation of the ESI process.

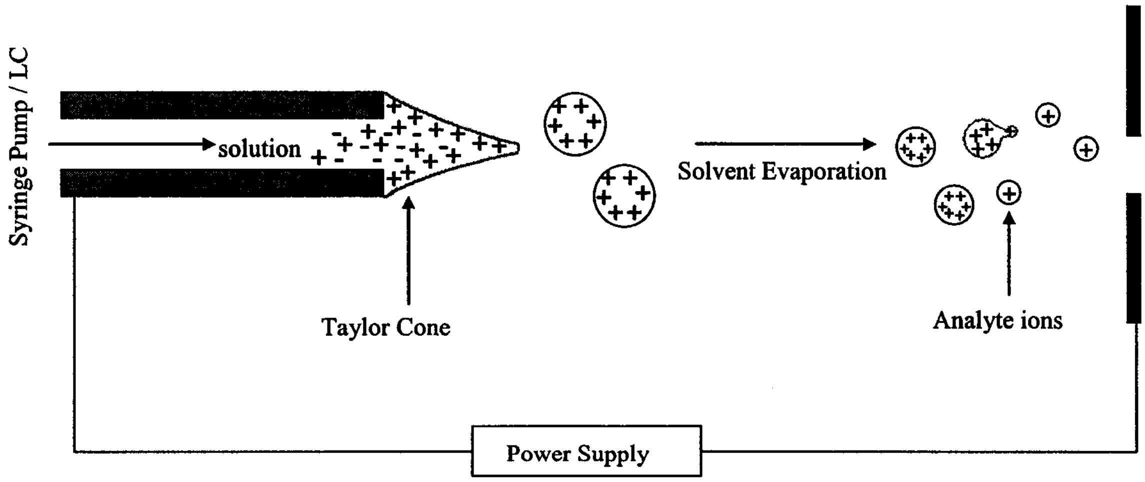


Figure 1.1. Schematic representation of the ESI process in the positive ion mode.

As described by Kebarle and Tang [15], there are four major steps in the production of gas-phase ions by electrospray from electrolyte ions in solution: 1) Production of charged droplets at the ES capillary tip. The penetration of an imposed electric field into the liquid leads to the formation of an electric double layer in the liquid. 2) Shrinkage of the charged droplets by solvent evaporation. 3) Repeated droplet disintegrations leading ultimately to very small, highly charged droplets capable of producing gas-phase ions. 4) Production of gas phase ions. Enrichment of the surface of the liquid by positive electrolyte ions leads to destabilization of the meniscus and formation of cone and jet emitting droplets with an excess of positive ions, then shrinkage of the charged droplets by evaporation and splitting into smaller droplets and finally gas-phase ions are formed. Finally, the gas-phase ions are introduced into the mass spectrometer for analysis.

This is a very soft method of ionization in which very little internal energy is imparted to the analyte upon ionization. Therefore, ESI can provide detailed information regarding molecular weights and structures. One distinguishing feature of ESI is that sample is introduced in solution, which makes ESI compatible with many types of separation techniques. Because ESI produces ions continuously, it can be combined with a scanning mass analyzer, such as quadrupole mass spectrometer.

1.1.1.2 Triple Q system

Triple-stage quadrupole mass filters (i.e. Triple Q) combined with the ESI technique play an important role in the study of drug metabolism [16-18], trace detection [19] and endogenous biologically active substances [20-22]. Because it can easily perform MS/MS analysis with collision-induced dissociation (CID) to provide molecular fragmentation information, it is a good choice for small biological molecule structure analysis. Figure 1.2 shows the schematic principles of Triple Q. When the MS/MS function is chosen, the parent ion is selected in the 1st quadrupole, and the 2nd quadrupole works as a place for fragmentation with CID by flowing nitrogen gas, and finally fragment ions produced by CID are separated based on their m/z in the 3rd quadrupole and travel to the detector. As well, there are several MS/MS operation modes available, such as product scan, precursor scan, neutral loss scan and Multiple Reaction Monitoring (MRM) in a single Triple Q system. Additionally, this instrument is compact and inexpensive, making it the most popular mass spectrometer for pharmaceutical analysis.

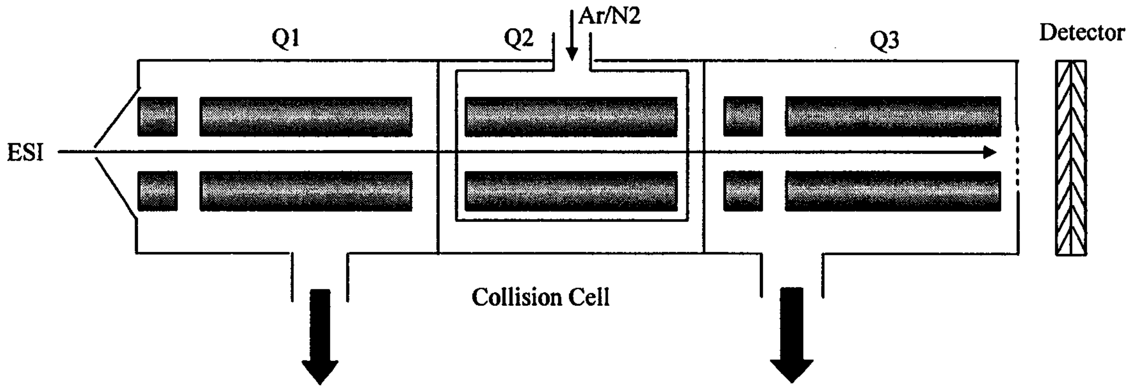


Figure1.2. Schematic diagram of a Triple Quadrupole mass spectrometer.

1.1.2 MALDI-MS system

1.1.2.1 Matrix-Assisted Laser Desorption/Ionization (MALDI)

MALDI technique was developed to transfer relatively large, labile molecules, such as polypeptides and proteins, into the gas phase as intact ions. MALDI combined with MS was introduced in the late 1980's by Hillenkamp and Tanaka [23, 24] when it was demonstrated that adding a small molecular weight organic matrix to an analyte could overcome molecular photo-dissociation of the sample ions induced by direct laser irradiation.

The MALDI process is shown in Figure 1.3. Typically, the analytes are mixed with a matrix solution, which is usually a highly ultraviolet absorbing compound, such as α-cyano-4-hydroxycinnamic acid (CHCA), sinapinic acid (SA), or 2, 5-dihydroxybenzoic acid (DHB). An aliquot of the mixture is taken and deposited onto a MALDI target plate. The mixture is allowed to dry into a crystalline deposit. After inserting the plate to a mass spectrometer, a pulsed laser beam irradiates the co-crystals on the target plate and causes the energy accumulation within the co-crystals, vaporizing both the matrix and the analytes. Ionized molecules, such as proteins and peptides, are transported from the ion source to a mass analyzer. They are separated according to their m/z ratio and finally detected. The incorporation of matrix materials has effects: 1) the matrix strongly absorbs the laser light at a wavelength at which the analytes are only weakly absorbing; then thermal relaxation of excited matrix molecules leads to the evaporation of the matrix and transfers the non-volatile analyte molecules into the gas phase. 2) The matrix acts as a protonating (positive ion detection) or deprotonating (negative ion detection) agent either in solution/solid phase or in the gas phase and is therefore essential in the ion formation process. All these effects combine to provide high ion yields of the intact analyte, giving rise to high sensitivity.

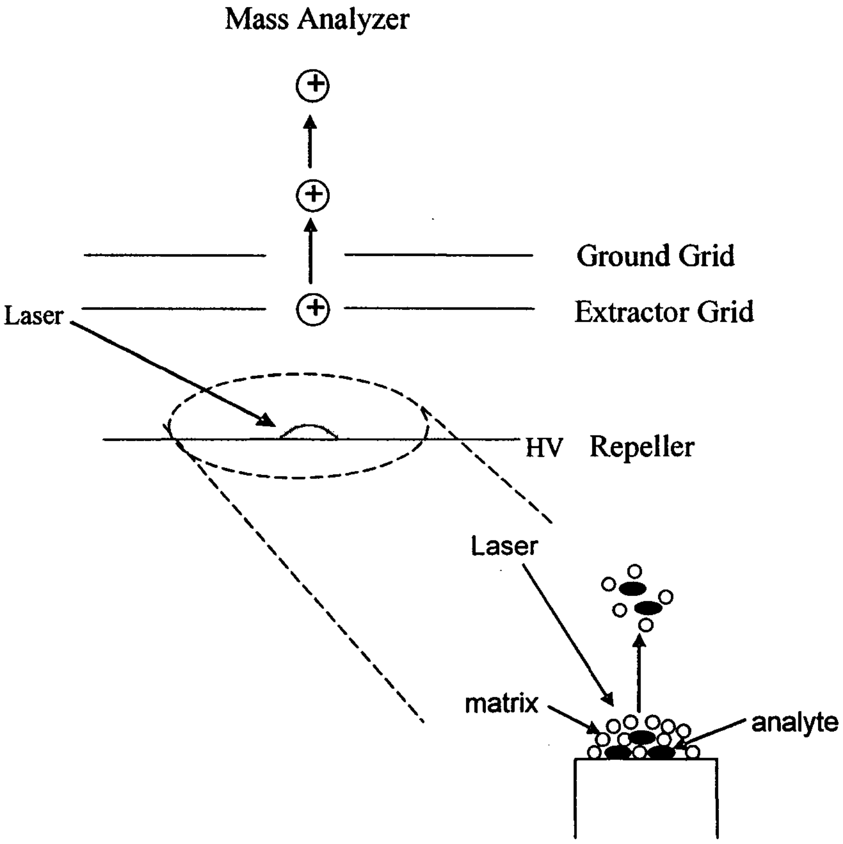


Figure 1.3. Schematic representation of the MALDI process.

Sample preparation is the most crucial step in MALDI mass analysis of peptides and proteins. So far, different on-target sample preparation methods have been developed, among which the most frequently used are the "dried-droplet" [23] and "two-layer" [25] methods. In the dried-droplet method, the solution mixture of matrix and sample is deposited on the target plate directly and air-dried in minutes. In the two-layer method, the first layer is deposited on the spot to form thin, uniformly tiny matrix crystals, and then the mixture of sample and 2nd layer matrix solution is deposited onto the first layer. Generally, the two-layer method can provide uniform sub-micrometer sized crystals and result in high resolution, sensitivity and reproducibility, while the dried-droplet method is quick and easy to perform.

Many variables influence the integrity of a good, homogeneous MALDI sample, including the concentrations of the matrix and analyte, choice of matrix, analyte sample purity (e.g., exposure to strong ionic detergents, formic acid, or strong acid), analyte hydrophobicity or hydrophilicity, contaminants and compatible solubilities of matrix and analyte solutions [26]. Compared with ESI, MALDI is more tolerant of impurities in the analyte samples, but contamination in samples may prevent the generation of homogeneous co-crystals on the target plate and cause peak broadening, sensitivity reduction and mass accuracy reduction. Table 1.1 shows contaminant concentrations tolerated in MALDI MS that do not result in significant degradation of the analyte signals.

Table 1.1. Contaminant concentration tolerated in MALDI MS technique [26].

	Maximum concentration (approx.)
Urea	0.5M
Guanidine-HCl	0.5M
Dithiothreitol (DTT)	0.5M
Glycerol	1%
Alkali metal salts	<0.5M
Tris buffer	0.05M
NH4HCO3	0.05M
Phosphate buffer	0.01M
Detergents (not SDS)	0.1%
SDS	0.01%

One way to remove the impurities is to perform on-probe washing. On-probe washing with water will remove most of the contaminants from sample spots prepared by SA or CHCA matrix, but it is not suitable for the water-soluble matrix DHB. In addition, very hydrophilic peptides or proteins bound to a high concentration of strong ionic detergent will be lost when washed.

When complex sample mixtures are analyzed, the use of a proper method of sample preparation is essential for obtaining good results, due to differential mass discrimination/ion suppression effects. In short, sample preparation should be handled with great care in MALDI MS [27].

1.1.2.2 MALDI-TOF MS system

MALDI mainly produces singly charged ions for macromolecules and uses pulsed laser irradiation. Therefore it is typically connected to a time-of-flight (TOF) mass spectrometer which can be used to detect the ions generated in a pulsed mode. To date, MALDI-TOF MS is a very common technique for protein and peptide analysis in proteomics research [28].

In TOF, the mass-to-charge (m/z) of an ion can be measured by determining its velocity after acceleration in an electrical field. The principle of TOF MS is to accelerate an ion electrostatically to a defined kinetic energy and measure its flight time through a field-free region. A detector determines the flight time for each ion with a particular m/z. At a fixed kinetic energy, larger ions have a longer flight time to arrive the ion detector while small ions arrive to the detector earlier. The relationship between mass-to-charge ratio m/z and flight time t is given below,

$$E_{kin} = U \cdot z = ½\, mv^2$$

$$\text{Therefore, } v = (2U \cdot z/m)^{1/2}$$

$$\text{And } t = L \cdot (m/2Uz)^{1/2}$$

where E_{kin} is kinetic energy, U is electrostatic potential, v is linear velocity of the ion, and L is the flight tube length of TOF. Theoretically, it is possible to accept a large number of ions into the analyzer without any limitations and to detect all ions flying in the flight tube. So TOF has no theoretical upper limit to the m/z ratio. A spectrum over a wide mass range can be obtained with a high sensitivity due to the high ion transmission.

There are two types of operational modes in MALDI-TOF MS: linear mode and reflector mode. Figure 1.4 shows the basic principle of these two modes of TOF mass spectrometer operation.

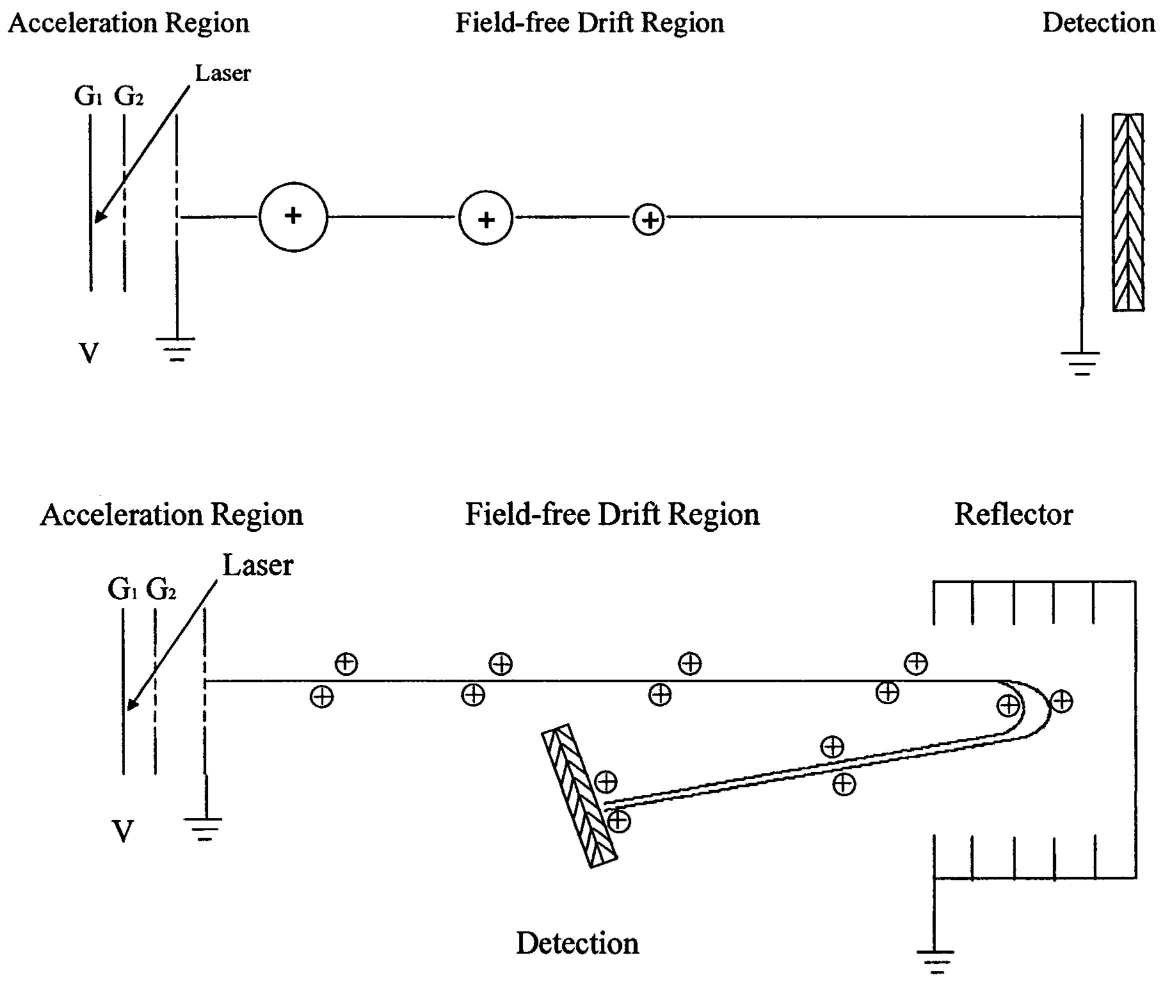


Figure 1.4. Principle of linear (above) and reflectron (bottom) TOF mass spectrometer in the positive ion detection mode.

In the linear TOF-MS mode, ions are accelerated into the field-free drift region and hit the detector positioned at the opposite end. It is the simplest configuration for a TOF MS and usually used for protein analysis. The reflectron mode provides better resolution due to energy focusing in the reflector which is most effective for smaller molecules, such as peptides (see below).

In general, mass resolution is defined as m/Δm. In a TOF mass spectrometer, in which ions are accelerated by a constant energy, the resolution can be expressed as,

$$m / \Delta m = t / 2\Delta t$$

where Δm and Δt are measured as the full width at half maximum (FWHM). In conventional MALDI-TOF MS, continuous ion extraction mode is used. As a result, a range of different flight times will be produced for the ions with the same m/z due to different initial kinetic energy, giving rise to decreased resolution. To improve the performance, a mass reflectron or mirror, located at the end of the flight tube, is used to compensate for the initial energy distribution and focus ions having the same m/z value. The reflectron makes use of an electrostatic field to reflect ions through a small angle towards a second detector, as shown in Figure 1.4. Ions with the same m/z but higher kinetic energy penetrate into the reflector deeper, delaying their arrival time to reflector detector relative to the slower low-energy ions. This effect gives rise to better resolution and mass accuracy. This technique is widely used for peptide detection.

In addition, commercial MALDI-TOF systems are also equipped with time-lag focusing (TLF) to further improve the mass resolution and accuracy. In TLF, the ions are extracted by the application of a high voltage pulse after a time delay, and accelerated into the drift tube at the same time [29-32]. Figure 1.5 shows the schematic illustration of the time-lag focusing concept applied to MALDI-TOF MS. The sample on the MALDI repeller plate is ionized with a very short laser pulse. The ions with the same m/z, expanding away from the repeller, display a broad velocity distribution and initial kinetic energy distribution. After a certain time delay, the repeller and the first extraction grid have the same potential and the ions are separated according to their initial energy, then an extraction pulse is applied to the repeller to extract the ions into the flight tube. The amplitude of the pulse can be adjusted so that the initially less-energetic ions that are closer to the repeller will catch up to the initially more-energetic ions when they both arrive at the detector. As a result, resolving power and mass accuracy of a linear or reflectron TOF is significantly improved.

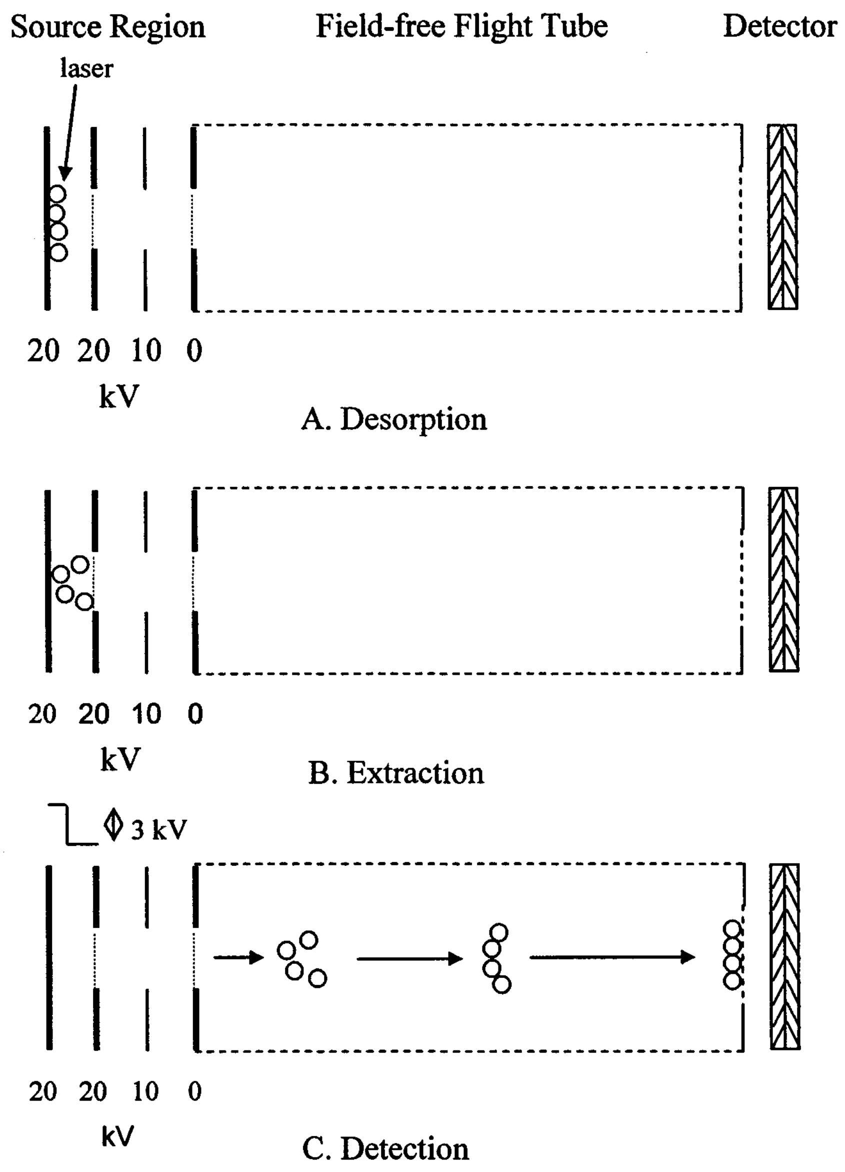


Figure 1.5. Schematic illustration of the time-lag focusing principle in MALDI-TOF MS. (A) Ions with same m/z but different initial kinetic energy expand away from the repeller. (B) Ions with lower initial energy move further from the repeller that is applied to an extraction pulse after a certain time delay. (C) Ions reach the detector at the same time due to the energy compensation.

1.1.2.3 MALDI-QqTOF MS system

By combining quadrupole MS with TOF MS, instruments capable of generating high quality MS/MS data that can be used to identify peptide sequences and post-translational modifications (PTMs) in proteins have been developed. In my study, the MS/MS work was done on an API Sciex XL QSTAR QqTOF system (Concord, ON, Canada). A description of this instrument is as follow.

In the QqTOF system, Q refers to a mass-resolving quadrupole for selecting the precursor ion, q refers to an r.f.-only (radio frequency-only) quadrupole collision cell where fragment or product ions are generated, and TOF refers to a time-of-flight mass spectrometer for detecting the fragment ions. This configuration can be regarded as the replacement of the third quadrupole in a triple Q system by a TOF mass spectrometer. Therefore, high sensitivity, mass resolution and mass accuracy in both precursor (MS) and product ion (MS/MS) detection can be obtained [33, 34]. Figure 1.6 shows the basic principle of QqTOF. It consists of three quadrupoles, Q0, Q1 and Q2, followed by a reflecting TOF mass analyzer with orthogonal injection of ions.

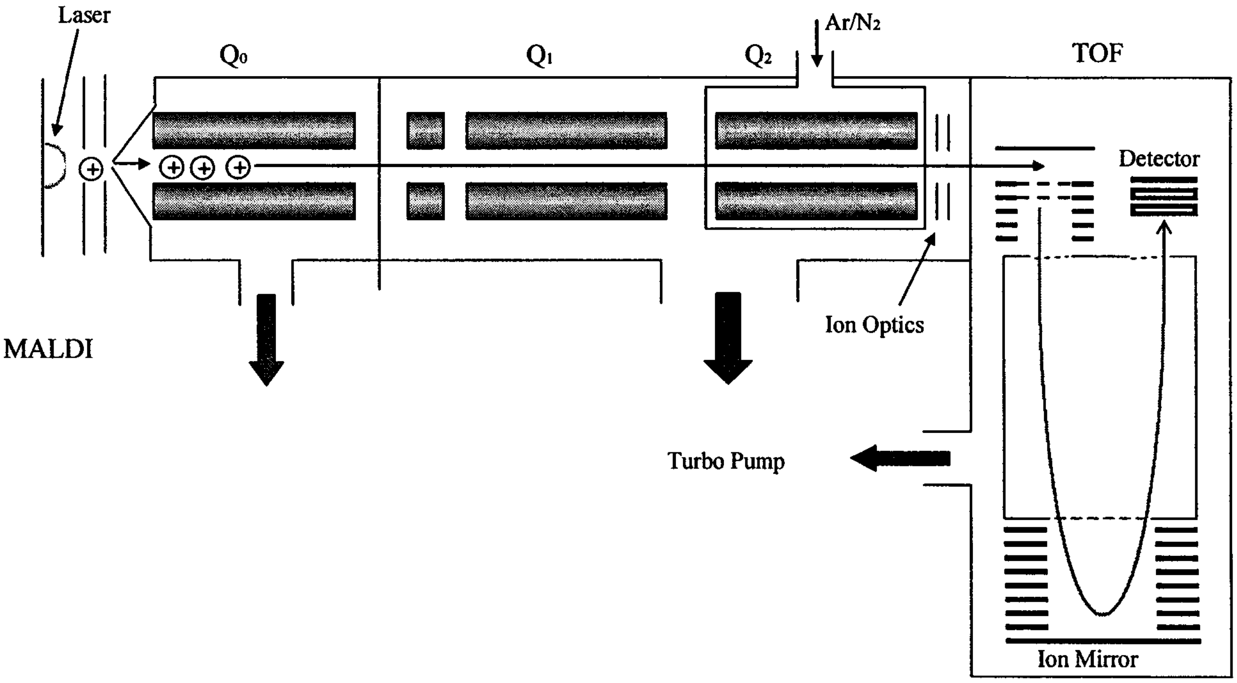


Figure 1.6. Schematic Diagram of the QqTOF mass spectrometer.

In single MS mode, the Q0, Q1 and Q2 are all operated in the r.f.-only mode while TOF is used to collect mass spectra. Q0 is used for collisional cooling and focusing of the ions entering the instrument; Q1 serves as a transmission element, and Q2 provides a potential well for both radial and axial collisional damping of ion motion. Finally, they are re-accelerated in the axial direction to the necessary energies into the TOF and detected at the end.

In the MS/MS mode, Q1 is operated as a mass filter to transmit only the parent ion of interest, then the ion is accelerated to enter the collision cell Q2, where it undergoes collision-induced dissociation (CID) by colliding with neutral gas molecules (usually Ar or N_2). The resulting fragment ions are cooled and focused, then re-accelerated to the required energy and enter into the field-free drift space of the TOF mass analyzer. Finally all ions are detected with a detector such as a multichannel plate (MCP).

1.2 Introduction to Proteomics and Metabolomics

With the accelerated understanding of cellular and organismal biology, proteomics and metabolomics have become two increasingly important research areas. Figure 1.7 shows the relationship in a biological system and their related research areas.

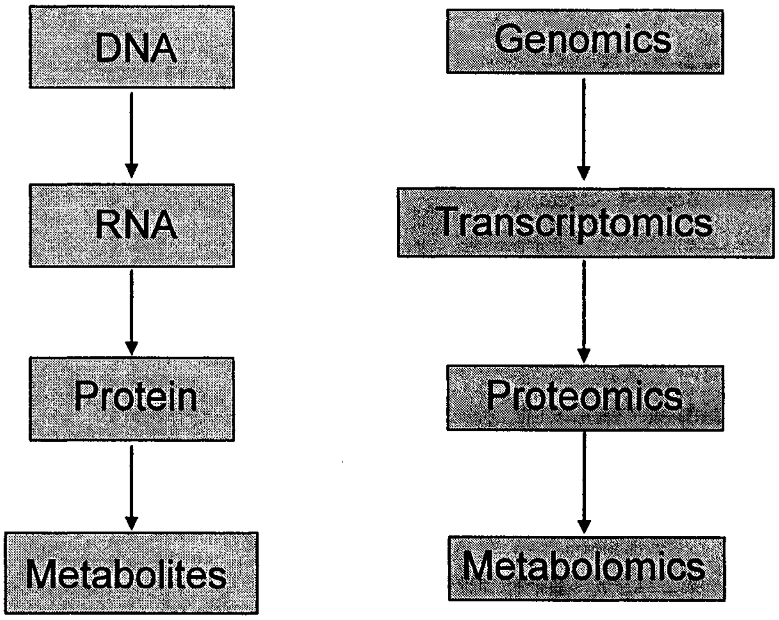


Figure 1.7. Relationship in a biological system.

1.2.1 Protein Characterization

Recent advances in genomics, as a consequence of the availability of detailed genomic information for an increasing number of species, have brought about a revolution in the understanding of cellular and organismal biology. However, this acceleration of understanding makes it evident that higher eukaryotes have many differences in the mechanisms used in their control of cellular function, while lower organisms have similar mechanisms. Furthermore, more and more evidence shows that a majority of genes are subject to such variation in their resulting protein isoforms [35], and different post-translational modifications can play different roles in the control of cellular function. Therefore, an understanding at the level of protein, including which proteins may be present, or how they are modified in specific situations, is essential and complementary to genomics for providing a clearer understanding of biological processes and systems.

Proteomics is defined as the direct qualitative and quantitative analysis of the full complement or subset of the proteins present in an organism, tissue, or cell under a given set of physiological or environmental conditions. It involves the detection, identification, and characterization of protein expression, function, activity, regulation and post-translational modifications (PTMs). All kinds of MS techniques combined with separation/purification techniques, such as LC, electrophoresis, and bioinformatics techniques for data processing, have become preeminent tools for proteomics analysis.

Basically, there are two approaches applied in proteomics research: "bottom-up" and "top-down". Bottom-up strategies, at the peptide level, involve cleaving the proteins into peptide fragments by enzymes or chemicals, then analyzing the peptides by MS and computer algorithms to make the definitive identifications of the original proteins. As a traditional method, the peptide mass fingerprinting involves the mass measurement of a set of peptide digestion products from a protein. This set of peptide masses is used as a "fingerprint" to identify the original protein in a database search. The technique has already been widely applied [36-40]. Reliable protein identification is based on several factors, such as the mass accuracy of peptide mass measurement, the number of peptides, the mass distribution of the query masses in the candidate protein, the number of matched peptides and the size of the sequence database. In many cases, this method cannot provide reliable identification of proteins, especially when the protein sample is complex or a small amount of protein sample is available. Therefore, additional information, such as MS/MS of peptide fragments, is required.

Specific peptides from the protein digest are chosen for fragmentation by tandem mass spectrometry. Collision-induced dissociation (CID) is the most widely used fragmentation method. Under CID, the fragmentation is not random; the pattern of breakage is dependent upon the parameters of the CID, such as collision energy, collision gas, and the amino acid sequence of the peptide. And the backbone cleavage results in the production of a, b, c, x, y, and z ions. Figure 1.8 gives the nomenclature for peptide fragmentation patterns [41]. Thereby, all MS/MS information is collected and searched in a database using a search engine, such as MASCOT [42] or SEQUEST [43], for identification of the protein. If the database search cannot provide positive results, *de novo* sequencing can be used, in which amino acid sequence can be read out manually from MS or MS/MS spectra.

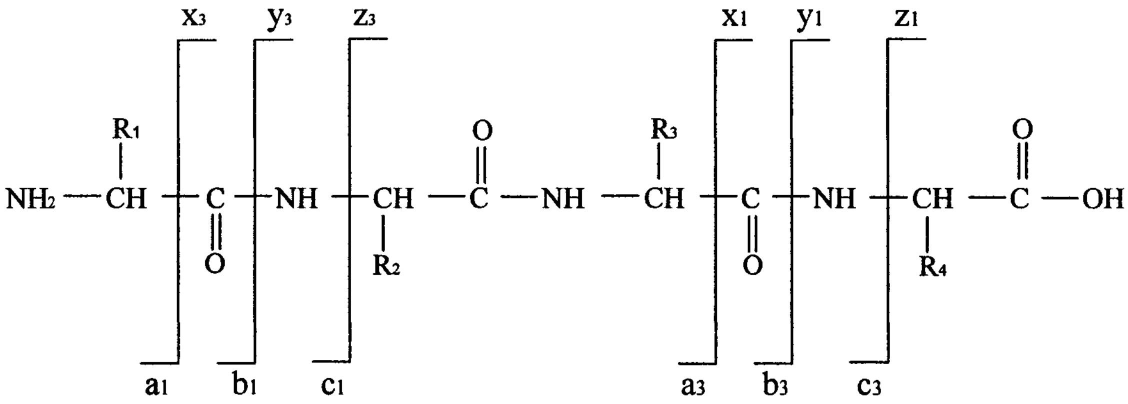


Figure 1.8. Nomenclature for peptide fragmentation pattern under CID.

Emerging as a popular method, shotgun proteomics takes advantage of the information obtained by MS/MS and has been used in many proteomics applications [43]. In this method, complex protein mixtures are extracted from a biological source and digested into peptides that are subsequently separated using multidimensional chromatographic techniques, and then analyzed by MALDI [42] or ESI [14] MS/MS. Using a set of mathematical algorithms, the resulting peptide sequence data generated from MS/MS are searched and mapped in the database to determine the original component of the protein mixture.

In parallel, there are a small but growing number of research groups developing intact protein level strategies for proteomics analysis, referred to as top-down approach

[44-49]. With traditional bottom-up methods, complete sequence coverage of proteins is rarely achieved, the PTMs and mutations observed on different peptides cannot be correlated and the data obtained from mixtures cannot reconstruct the complex set of multiple modifications present in an individual protein. In top-down strategies, the intact protein or large polypeptide is injected into a mass analyzer to obtain the molecular weight and fragmentation information, and when combined with a database search or *de novo* sequencing, information including sequence, mutations and PTMs of the target protein can be achieved. There are many aspects to consider in top-down approaches, such as mass limits of the mass analyzer, mass accuracy and mass resolution of mass spectra for distinct identification of proteins. Recent instrumental development on mass spectrometers, such as FTICR [50, 51], quadrupole ion traps [52-55], and TOF-TOF [45], and separation techniques, such as LC, CE, and gel electrophoresis, has made the top-down method a promising tool for proteome analysis.

Among these top-down proteomics approaches, there is a quick and sensitive approach. Mass Analysis of Polypeptide ladders (MAP), which was developed in our laboratory recently [56], involves the production of two series of polypeptide ladders containing predominantly either the N-terminal or C-terminal amino acid of the protein by the partial hydrolysis of the protein in aqueous acid heated by microwave irradiation for a short time. Mass analysis of these hydrolysates in one simple mass spectrum allows direct reading of the amino acid sequence and gives information on protein PTMs. Figure 1.9 shows the schematic of the MAP protein sequencing technique.

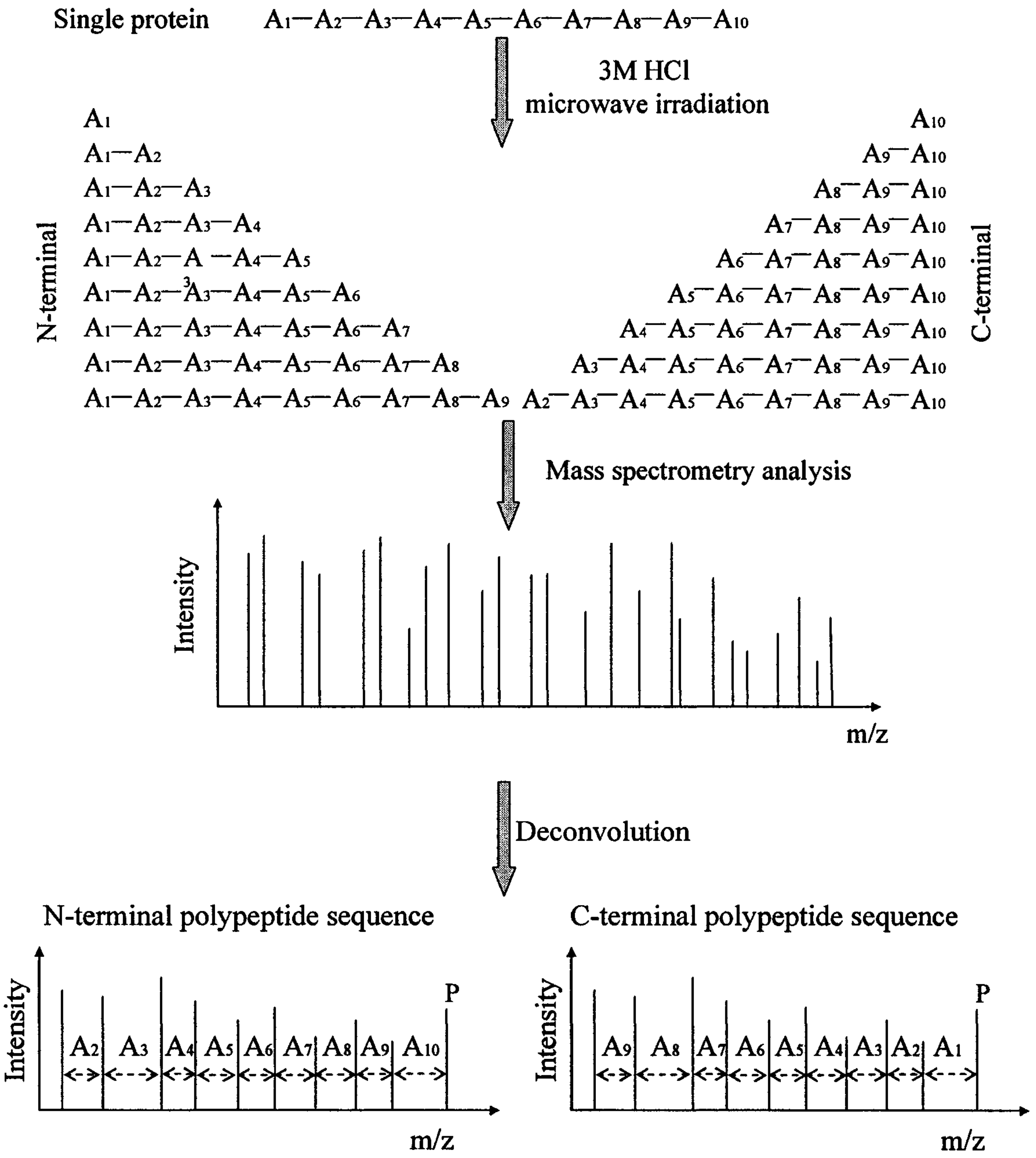


Figure 1.9. MAP protein sequencing for top-down proteomics.

Accelerated acid hydrolysis of a protein by 3 M HCl with microwave irradiation for a very short time results in specific polypeptide ladders, either N-terminal or C-terminal amino acid sequence, probably because of microwave-induced rapid heating and conformational or structural changes of proteins along the peptide bonds. Therefore, it is likely that in an intact protein consisting of many peptide bonds, the first hydrolysis process is a set of parallel reactions breaking every peptide bond at one time to produce many terminal peptides.

Once the C-terminal and N-terminal polypeptides are obtained, from the mass difference of consecutive peaks, each amino acid can be calculated in sequence order in one single mass spectrum, while information on PTMs and amino acid variants can be obtained simultaneously. While this method is simple for protein sequencing, it requires a relatively pure protein sample and high mass accuracy over a wide mass range.

1.2.2 Metabolomics

While there is much current interest in the genome-wide analysis of cells at the level of transcription (to define the 'transcriptome') and translation (to define the 'proteome'), a third level of analysis, 'metabolomics', has come to the fore in recent years [57-59]. The term 'metabolomics' refers to the studies of all the small molecular weight metabolites inside a biological sample of interest, such as urine, saliva, or blood plasma. Whereas genes and proteins set the stage for what happens in the cell, much of the actual activity is at the metabolite level: cell signaling, energy transfer, and cell-to-cell communication are all regulated by metabolites. Furthermore, gene and protein expression are closely linked, but metabolite behavior more closely reflects the actual cellular environment, which is itself dependent on nutrition, drug or pollutant exposures, and other exogenous factors. Metabolites reflect the combined effects of many influences on physiological function and phenotype. And very likely, measurement of the metabolome in different physiological states will in fact be much more discriminating for the purposes of functional genomics. It can be used in many applications, such as determination of metabolic biomarkers, determination of the effect of biochemical or environmental stresses, bacterial characterizations, human health assessments and metabolic engineering.

As with genomics and proteomics, high throughput metabolic profiling or metabolomic analysis typically involves using FT-IR, NMR and MS combined with separation techniques, such as GC or LC.

NMR provides rapid analysis and direct structure information but suffers from low sensitivity, while GC-MS or LC-MS can provide good sensitivity, good selectivity and relatively low sample consumption, all of which make MS techniques good choices for metabolome analysis. Many research groups have already developed strategies in this field such as a GC-TOF MS-generated MS/RI (molecular mass vs. retention time) index database in the area of plant biotechnology [63].

GC/MS or GC-GC MS is used for the analysis of volatile and thermally stable compounds. Derivatization of the sample is usually needed for non-volatile compounds in which sample stability is always a concern.

LC/MS provides a complementary approach, in which sample volatility is not required, higher molecular weight metabolites can be analyzed and the analysis is processed at lower temperatures than with traditional GC/MS. Metabolites are generally detected in one but not both ion modes, so wider metabolome coverage can be obtained by analysis in both negative and positive ion detection modes. However, for large studies, sample preparation and long LC runtimes make analyses both time-consuming and laborious. Minimal separation runtime usually results in lower sensitivity and reduced ability to identify metabolites. Co-elution is an apparent problem and not all of the ions may represent individual compounds; some may result from in-source fragmentation or adduct formation. And isomeric configurations can not be differentiated, which is of great importance in biological studies. Therefore, tandem MS or MS/MS generated by CID techniques combined with high resolution separation methods are needed to obtain further structure information and identify the metabolites with high confidence.

The human metabolome is a complicated system. The number of different metabolites in the human body is unknown. Of particular interest to metabolomics researchers are small, low molecular weight compounds that serve as substrates and products in various metabolic pathways. These "small molecules" include compounds such as lipids, vitamins, fatty acids, carbohydrates, carboxylic acids and amino acids that can provide important clues about an individual's health status. They have widely different structures, functional groups, physicochemical properties and concentrations, and they belong to a wide variety of metabolic pathways. Table 1.2, showing the basic classification of metabolites, demonstrates the challenge for analytical research.

Table 1.2. Basic classification of metabolites.

⟵ Non-polar Polar ⟶

Lipids fatty acids	Carotenoids Steroids	Phenolics alcohols	Alkaloids Organic acids Organic amines	Sugars Nucleotides
Waxes Terpenes volatiles	flavenoids	Catecholamines Polar organics	Nucleosides Amino acids	Metals Salts Ionic

Recently, an increasing number of research groups and facilities including our research group are paying more attention to a complete understanding of human metabolites. The main objective of our work is to develop analytical tools to facilitate the analysis of complicated mixtures of metabolites or metabolome.

1.3 Overview of the Thesis

In this thesis, *de novo* sequencing by combined microwave-assisted acid hydrolysis (MAAH) of proteins with SDS-PAGE separation was investigated using protein standards for amino acid sequencing and characterization of PTMs. Then, MAAH combined with HPLC was used for amino acid variants identification in the hemoglobin β-chain. The results show that in a selected mass range, definitive amino acid identifications can be obtained, making it possible to identify amino acid variants in a protein. Also, all these efforts prove that further improvement of instrumentation and separation techniques would likely increase the protein sequence coverage for characterizing high molecular mass proteins using the top-down proteomics method.

In the second part of my thesis, an MS/MS database for human metabolomics was constructed from 215 acquired metabolites. Three sets of libraries in the positive ion mode and negative ion mode at three different fragmentation energies were constructed to provide comprehensive information on analyte fragmentation. Their reproducibility and concentration limitation of detection were also investigated. Combining all the fragmentation information obtained from the three fragmentation energies for each metabolite provides an essential tool for metabolite identification and structure elucidation which could be very important for understanding biological activities. Finally, this database was applied to the analysis of human urine by combining LC separation and MS/MS.

1.4 Literature Cited

(1) Wysocki, V.H.; Resing, K.A.; Zhang, Q.; Cheng, G. *Methods* **2005**, *35*, 211.

(2) Thadikkaran, L.; Siegenthaler, M.; Crettaz, D.; Queloz, P.A.; Schneider, P.; Tissot, J-D. *Proteomics* **2005**, *12*, 3019-3034.

(3) Wright, M.E.; Han, D.; Aebersold, R. *Mol Cell Proteomics* **2005**, *4*, 545.

(4) Cantin, G.T.; Yates III, J.R. *J Chromatogr A.* **2004**, *1053*, 7.

(5) Thiede, B.; Hohenwarter, W.; Krah, A.; Mattow, J.; Schmid, M.; Schmidt, F.; Jungblut, P.R. *Methods* **2005**, *35*, 237.

(6) Villas-Boas, S.G.; Sandrine, M.; Akesson, M.; Smedsgaard, J.J. *Mass Spectrometry Review* **2005**, *24*, 613.

(7) Brown, S.C.; Gary, K.; Dasseux, J.L. *Mass Spectrometry Review* **2005**, *24, 223.*

(8) Dunn, W.B.; Baile, N.J.; Johnson, H.E. *Analyst* **2005**, *130*, 606.

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

(10) Yamashita, M.; Fenn, J.B. *J. Phys. Chem.* **1984**, *88*, 4671.

(11) Aleksandrov, M.L.; Gall, L.N.; Krasnov, V.N.; Nikolaev, V.I.; Pavlenko, V.A.; Shkurov, V.A. *Dokl. Akad. Nauk SSSR* **1984**, *277*, 379.

(12) Mirza, U.A.; Steven, L.C.; Chait, B.T. *Anal. Chem.* **1993**, *65*, 1.

(13) Winger, B.E.; Light-Wahl, K.J.; Rockwood, A.L.; Smith, R.D. *J. Am. Chem. Soc.* **1992**, *114*, 5897.

(14) Yates III, J.R.; Speicher, S.; Griffin, P.R.; Hunkapiller, T. *Anal. Biochem.*, **1993**, *214*, 397.

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

(16) Mano, N.; Naryi, T.; Nikaido, A.; Goto, J. *Drug Metabol. Pharmacokin.* **2002**, *17*, 142.

(17) Bu, H.Z.; Knuth, L.K.; Magis, L.; Teitelbaum, P. *Rapid Commun. Mass Spectrom.*

2000, *14*, 1943.

(18) Bu, H.Z.; Magis, L.; Knuth, L.K.; Teitelbaum, P. *Rapid Commun. Mass Spectrom.* **2001,** *15*, 741.

(19) Reemtsma, T. *J Chromatogr A.* **2003,** *1000*, 477.

(20) Ikegawa, S.; Yanagihara, T.; Murao, N.; Watanabe, H.; Goto, J.; Niwa, T. *J. Mass spectrum.* **1997,** *32*, 401.

(21) Borts, D.J.; Bowers, L.D. *J. Mass spectrum.* **2000,** *35*, 50.

(22) Falany, C.N.; Fortinberry, H.; Leiter, E.H.; Barnes, S. *J. Lipid Res.* **1997,** *38*, 1139.

(23) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988,** *60*, 2299.

(24) Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. *Rapid Commun. Mass Spectrom.* **1988,** *2*, 151.

(25) Dai, Y.; Whittal, R.M.; Li, L. *Anal. Chem.* **1996,** *68*, 2494.

(26) Coligan, J.E.; Dunn, B.M.; Ploegh, H.L. *Current Protocols in Protein Science* **1995,** *Volume 1*, Unit 16.12.

(27) Cohen, S.L.; Chait, B.T. *Anal. Chem.* **1996,** *68*, 31.

(28) 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. USA* **1996,** *93*, 14440.

(29) Brown, R.S.; Lennon, J.J. *Anal. Chem.* **1995,** *67*, 1998.

(30) Colby, S.M.; King, T.B.; Reilly, J.P. *Rapid Commun. Mass Spectrom.* **1994,** *8*,

865.

(31) Vestal, M.L.; Juhasz, P.; Martin, S.A. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1044.

(32) Whittal, R.M.; Li, L. *American Laboratory* **1997**, 30.

(33) Morris, H.R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R.S.; Hoyes, J.; Bateman, R.H. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 889.

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

(35) Roberts, G.C.; Smith, C.W. *Curr. Opin. Chem. Biol.* **2002**, *6*, 375-383.

(36) Henzel, W.J.; Billeci, T.M.; Stults, J.T.; Wong, S.F.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5011-5015.

(37) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58-64.

(38) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338.

(39) Pappin, D.J.; Hojrup, P.; Bleasby, A.J. *Curr. Biol.* **1993**, *3*, 327.

(40) Yates III, J.R.; Griffin, S.; Hunkapiller, P.R. *Anal. Biochem.* **1993**, *214*, 397.

(41) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.

(42) Zhang, N.; Li, N.; Li, L. *J. Proteome Res.* 2004, *3*, 719.

(43) Yates III, J.R. *J. Mass Spectrom.* **1998**, *33*, 1.

(44) Loo, J.L.; Edmonds, G.G.; Smith, R.D. *Science* **1990**, *248*, 201.

(45) Demirev, P.A.; Feldman, A.B.; Kowalski, P.; Lin, J.S. *Anal. Chem.* **2005**, *77*,

7455.

(46) Senko, M.W.; Speir, J.P.; McLafferty, M.W. *Anal. Chem.* **1994,** *66,* 2801.

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

(48) Kelleher, N.L.; Lin, H.Y.; Valaskovic, G.A.; Aaserud, D.J.; Fridriksson, E.K.; McLafferty, F.W. *J. A. C. S.* **1999,** *121,* 806.

(49) Kelleher, N.L. *Anal. Chem.* **2004,** *76,* 187A.

(50) Bogdanov, B.; Smith, R.D. *Mass Spectrometry Reviews* **2005,** *24,* 168.

(51) Meng, F.; Forbes, A.J.; Miller, L.M.; Kelleher, N.L. *Mass Spectrometry Review* **2005,** *124,* 126.

(52) Cargile, B.J.; McLuckey, S.C.; Stephenson, J.L. *Anal. Chem.* **2001,** *73,* 1277.

(53) Reid, G..E.; Shang, H.; Hogan, J.M.; Lee, G.U.; McLuckey, S.A. *J. A. C. S.* **2002,** *124,* 7353.

(54) Reid, G.E.; McLuckey, S.A. *J. Mass Spectrom.* **2002,** *37,* 663.

(55) Coon, J.J.; Ueberheide, B.; Syka, J.E.; Dryhurst, D.D.; Ausio, J.; Shabanowitz, J.; Hunt, D.F. *Proc. Natl. Acad. Sci. USA* **2005,** *102,* 9463.

(56) Zhong, H.; Zhang, Y.; Wen, Z.; Li, L.; *Nature Biotechnology* **2004,** *22,* 1291.

(57) Fiehn, O. *Comparative and Gunctional Genomics* **2001,** *2,* 155

(58) Dunn, W.B.; Ellis, D.I.; *Trends in Analytical Chemistry* **2005,** *24,* 285

(59) Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. *Trends in Biotechnology* **2004,** *22,* 245

(63) Wagner, C.; Sefkow, M.; Kopka, J. *Phytochemistry*, **2003**, *62*, 887

Chapter 2

De Novo Protein Sequencing by Microwave-Assisted Acid Hydrolysis Combined with SDS-PAGE

2.1 Introduction

Many techniques have been reported for the identification of protein sequences and characterization of post-translational modifications (PTMs) of proteins. The most direct method of reading the amino acid sequence of a protein is by automated Edman degradation, in which amino acids are removed sequentially from the N-terminus by chemical reaction and identified by an analytical HPLC technique [1]. MS-based techniques, such as ladder sequencing [2-9], in-source fragmentation [10, 11], and chemical derivatization [12, 13], have been reported for peptide sequencing with varying degree of success. Compared with the traditional Edman method, the MS approach has the advantage of higher sensitivity and the ability to generate structural information for intact polypeptide chains. Usually, a mass spectrometric peptide mapping technique involves enzymatic cleavage of the protein by specific proteases followed by mass determination of the generated peptides and database searching [14-17]. Acid hydrolysis of a protein provides another important tool for protein structural characterization and amino acids analysis. Acid hydrolysis has the advantage for the analysis of insoluble proteins and proteins resistant to proteolytic degradation, because limited accessibility of the substrate molecules to the enzyme would result in low sequence coverage and difficulty in localization of PTMs. Recently, partial protein hydrolysis approaches using acid vapor in combination with mass spectrometry have been reported [5-7]. Cleavage specificity was obtained by hydrolysis of protein standards and then the optimized procedure was used for C-terminal or N-terminal sequencing of peptides. However, low cleavage specificity resulting in very complex mixtures of hydrolytic product means that lengthy work is needed on spectral interpretation. There is a great need of more rapid and versatile protein sequencing methods.

Microwave is non-ionizing radiation that interacts in the liquid phase with ion pairs and with molecules that have a dipole moment. Microwave energy is directly transferred to these species and the energy profile is different from that involved in traditional conduction/convection heating [18]. In a recent study in our laboratory, microwave irradiation was applied to acid hydrolysis of a protein [19] to accelerate the reaction rate and proved to be successful for single protein analysis. This Mass Analysis of Polypeptides (MAP) sequencing technique allows direct sequencing of protein with high sensitivity and speed. In general, after exposure of a single protein sample to microwave irradiation for a short time (< 5 min), two series of polypeptide ladders containing the N- and C-terminal amino acids are predominantly produced. All peptide bonds of the intact protein are found to be susceptible to cleavage and only small variations in relative intensity of adjacent terminal peptides are found. Terminal backbone cleavages are predominant, while internal cleavages randomly produced are of low abundance and can be ignored. Therefore, the polypeptides detected on

MALDI-TOF MS are exclusively terminal peptides and the amino acid sequence of a protein is read out in a single operation quickly and easily.

However, for unpurified protein, acid hydrolysis of impure proteins will produce unknown impurity peaks as well, resulting in the difficult detection of N-terminal or C-terminal polypeptide peaks from the analyte. Therefore, this MAP technique requires a relatively pure protein for amino acid sequencing and determination of PTMs. To improve this technique, gel-based and solution-based methods have been applied to protein sample purification with each having its own strengths. An in-solution method has already been used to purify the proteins from *E. coli* K12 for MAP [19] and proved to be successful in sequencing small proteins (< 15,000 Da). In this work, a gel-based method for protein purification is described.

The SDS-PAGE (Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis) technique remains as a good choice for the protein separation due to its high resolution. The combination of gel electrophoresis with mass spectrometric analysis has been a powerful tool for protein identification [20, 21]. The gel acts as a molecular sieve in which protein molecules are embedded within the gel pores and separated under a suitable voltage. As a result, the reproducible separation of proteins with a single or a few amino acids difference can be achieved. Furthermore, during the sample preparation for SDS-PAGE and separation, protein conformation is changed and protein molecules become unfolded which leads to better acid hydrolysis of the protein exposing to the microwave radiation.

One of the technical challenges in combining SDS-PAGE with MALDI MS for MAP is to extract the whole protein molecules from the gel. Various protein extraction techniques, such as electroelution [22, 23], electroblotting [24, 25], direct desorption from ultrathin gels [26, 27] and passive elution [28-30], have been reported. While all these methods have been used successfully, the passive elution technique has the following advantages: no special equipment is required, gels of standard thickness can be used, and the elution procedure is relatively simple and fast. As a negative stain, copper stain is a simple, fast procedure, and the proteins are reversibly fixed within the gel, so the gel band of proteins can be destained then eluted for the further study. In this study, a mass spectrometric method combining passive elution of proteins from copper-stained/destained gels and MAP for direct amino acid sequence reading and protein PTM identification is reported.

2.2 Experimental

2.2.1 Materials and Reagents

Tris-HCl, acrylamide/bisacrylamide (electrophoresis purity grade), sodium dodecyl sulfate (SDS), glycerol, 2-mercaptoethanol and Copper stain/destain kit were purchased from Bio-Rad Laboratories (Canada), Ltd. HPLC grade acetonitrile (ACN), acetone, methanol and analytical grade HCl were obtained from Fisher Scientific Canada.

α-Cyano-4-hydroxycinnamic acid (CHCA) was purchased from Sigma Aldrich Canada, and was recrystallized before use. Water used in all experiments was from a Millipore NANOpure water system. Horse myoglobin and bovine α-S1-casein were obtained from Sigma Aldrich Canada.

2.2.2 SDS-PAGE

SDS-PAGE was carried out according to Schagger and von Jagow [31] using Tris-Tricine 10% or 12% gels. Prior to electrophoresis, protein samples were treated at 95°C for 5 min in sample buffer containing 2% mercaptoethanol (v/v), 2% SDS, 12% glycerol, 50 mM Tris and 0.01% bromophenol blue. 10 μL sample or less was loaded in the sample wells. The proteins were stacked at 3 mA for 15 min, followed by separation on the resolving gel at 15 mA for 30 min. Copper staining was employed to display the protein bands (Instruction Manual 161-0470).

2.2.3 Gel Excision & Protein Extraction

After the copper staining, protein bands of interest were excised using a scalpel, then destained for 15 min in copper destaining solution and water consecutively, as described by Qin et al [32]. Protein extraction was carried out by Vortex shaking at 4 °C in a solution of 3 M HCl in ACN/H_2O. The supernatant was then used for MALDI MS analysis and microwave-assisted acid hydrolysis.

2.2.4 Protein Digestion

The supernatant was transferred to a non-siliconized Eppendorf centrifuge tube (1.5 or 0.6 mL), capped and placed in a 900 W output household microwave oven. A bottle of H_2O was put beside the sample to absorb extra microwave energy and typically the sample was hydrolyzed by irradiation for 2 min. Then the sample solution was dried down completely using a Speed Vac concentrator and was ready for MALDI MS analysis.

2.2.5 Sample Preparation for MALDI-TOF MS Analysis

The two-layer method with CHCA was used in MALDI MS analysis [33]. The first layer was prepared as a saturated solution of CHCA in 25% methanol/acetone and the second layer was a saturated solution of CHCA in 50% ACN/ H_2O. For the protein extraction study, the supernatant was mixed with the 2nd layer directly at a 1:1 ratio. After 0.8 μL of the first layer was deposited on the sample probe and air-dried, about 0.4 μL of sample solution was deposited on the top of the first layer and allowed to air-dry without any following on-spot washing. For the MAP mapping analysis, the dried polypeptide mixture sample was dissolved in the second layer to a final concentration of 1 μg/μL and was deposited on the 1st layer to air dry without further on-spot washing.

2.2.6 MALDI-TOF analysis

MS analysis was performed on a Bruker Reflex III MALDI-TOF mass spectrometer equipped with a SCOUT 384 multiprobe inlet (Bremen/Leipzig, German) or an Applied Biosystems Voyager Elite equipped with a 100 multiprobe inlet (Framingham, MA). Both instruments are equipped with a pulsed 337nm nitrogen laser and operated with delayed extraction positive ion mode using linear mode. The spectra were externally calibrated with bovine cytochrome c peak. The data were then reprocessed using the Igor Pro software package (WaveMetrics Inc., Lake Oswego, OR). All mass spectra shown in the figures were normalized to the most intense signal in the mass range displayed.

2.3 Results & Discussion

Influence of Different Elution Conditions on Extraction Efficiency

There are several possible factors or components affecting protein extraction efficiency, such as extraction time, ACN concentration in the extraction solvent, extraction solvent volume and pore size of gels, all of which were investigated. The extraction conditions were optimized to obtain relatively high efficiency and low impurity levels.

1) Acetonitrile concentration

The purpose of introducing ACN in the extraction process is to enhance the efficiency of protein extraction by improving the solubility of protein molecules in the extraction solution. But when the ACN percentage is too high, the gel shrinks and proteins are trapped in the gel network and are hard to dissolve into the extraction solvent. So a suitable ACN concentration in the extraction solution is needed for protein extraction.

Figure 2.1 shows the mass spectra of proteins extracted from gels using different ACN concentrations in the extraction solution. 1 nmol of myoglobin (20 μg) was extracted from gel pieces by 50 μL solution of 3 M HCl in 50% ACN, 3 M HCl in 60% ACN and 3 M HCl in 75% ACN, respectively. 50% and 60% ACN made no difference and 50% ACN was usually used. However, when the ACN percentage in the extraction solution was increased to 75% or even higher, more unknown impurity peaks in the low mass range appeared and a signal-to-noise ratio reduction was observed. Higher ACN concentration facilitates more gel-induced impurities to dissolve into the extraction solution, resulting in the higher noise. In addition, ACN possibly accelerates the internal acid hydrolysis of protein under strong acidic condition during the extraction process, resulting in the production of internal polypeptides. The loss of intact protein molecules will reduce the concentrations of terminal polypeptides generated in the microwave hydrolysis process.

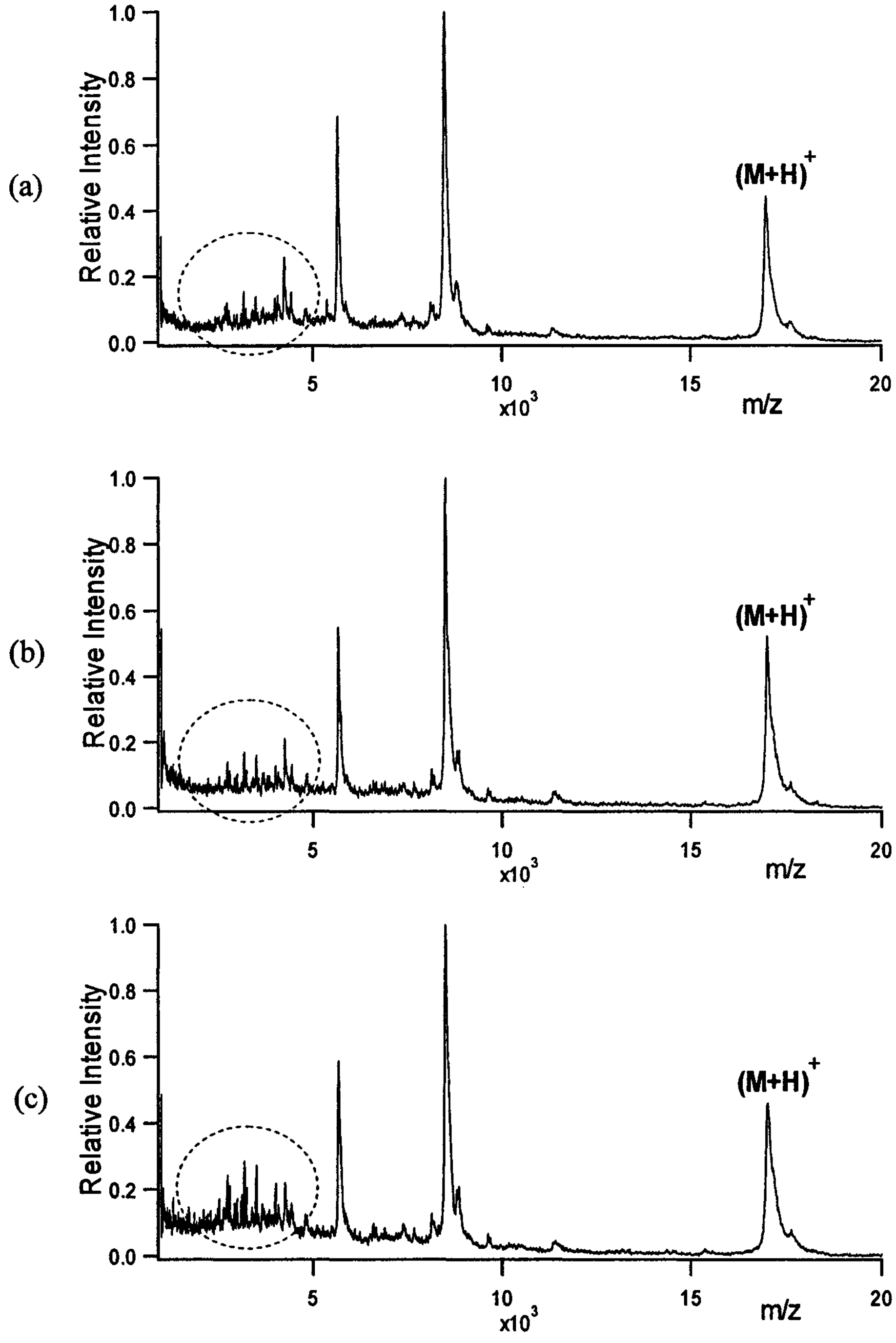


Figure 2.1. Effect of ACN on protein extraction. (a) 50% ACN, (b) 60% ACN, (c) 75% ACN. With the increasing ACN percentage, more and more impurity peaks in the low mass range were observed, as shown in the circle.

2) Extraction time

Extraction time influences extraction efficiency directly. Obviously, an optimized extraction time should be used to extract protein molecules. Figure 2.2 shows the effects of different extraction times on protein extraction efficiency. 1 nmol myoglobin (20 μg) was extracted from gel pieces in 50 μL solution of 3 M HCl in 50% ACN for 10 min, 20 min, and 30 min. Clearly, 10 minutes was not long enough to extract most of the protein, as seen by the weak molecular ion signal in the mass spectrum. Also, a long extraction time was not suitable, because impurities in the gel or gel-related components were extracted together with the proteins. The impurities suppressed the MS signal of the protein, resulting in reduction of the signal-to-noise ratio. As well, with a prolonged extraction time, such as 30 minutes, more random internal hydrolyzed polypeptides are produced because of the strong acid exposure. From this study, we concluded that 20 min gave the best combination of extraction efficiency with low level of impurities introduced into the extracted protein sample solution.

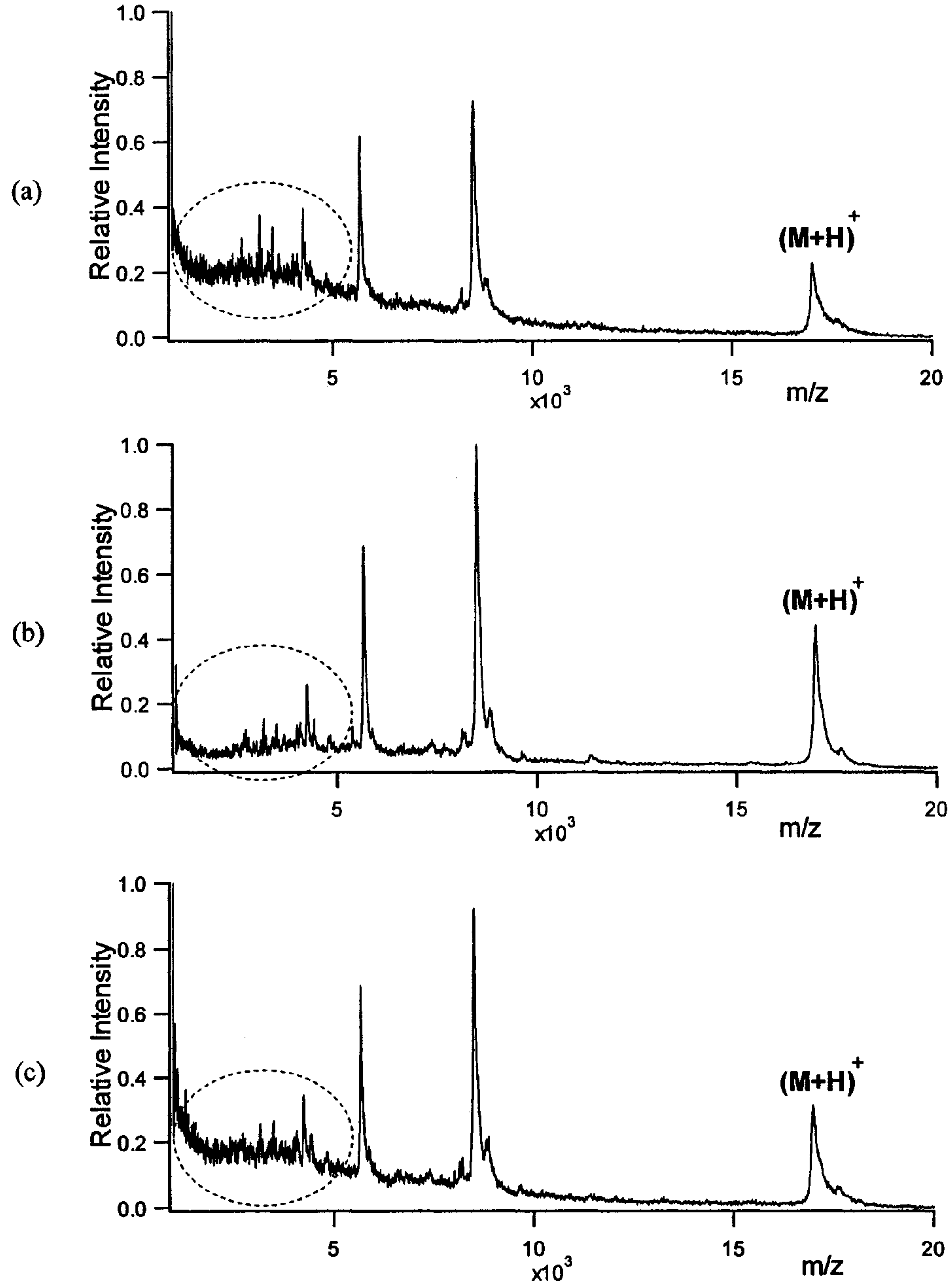


Figure 2.2. Effect of extraction time on protein extraction. (a) 10 min, (b) 20 min, (c) 30 min. Apparently, 20 min is the optimal extraction time to obtain high extraction efficiency as well as avoid producing impurity peaks in the low mass range.

3) Volume of extraction solution

Although the extraction solution volume might not have a very strong influence on protein extraction technically, the optimum volume for the extraction needs to be determined.

Figure 2.3 shows the effect of different extraction solution volumes on protein extraction. To extract 1 nmol of myoglobin from a gel band, a 50 μL solution of 3 M HCl in 50% ACN was found to be the preferred volume. Here different volumes at 40 μL, 60 μL and 80 μL were all studied under the same conditions. If the volume was too large, there was more protein loss, more internal peptides were produced and more impurities were extracted from the gels. Therefore, ion suppression due to impurities and internal cleavage of proteins became more evident.

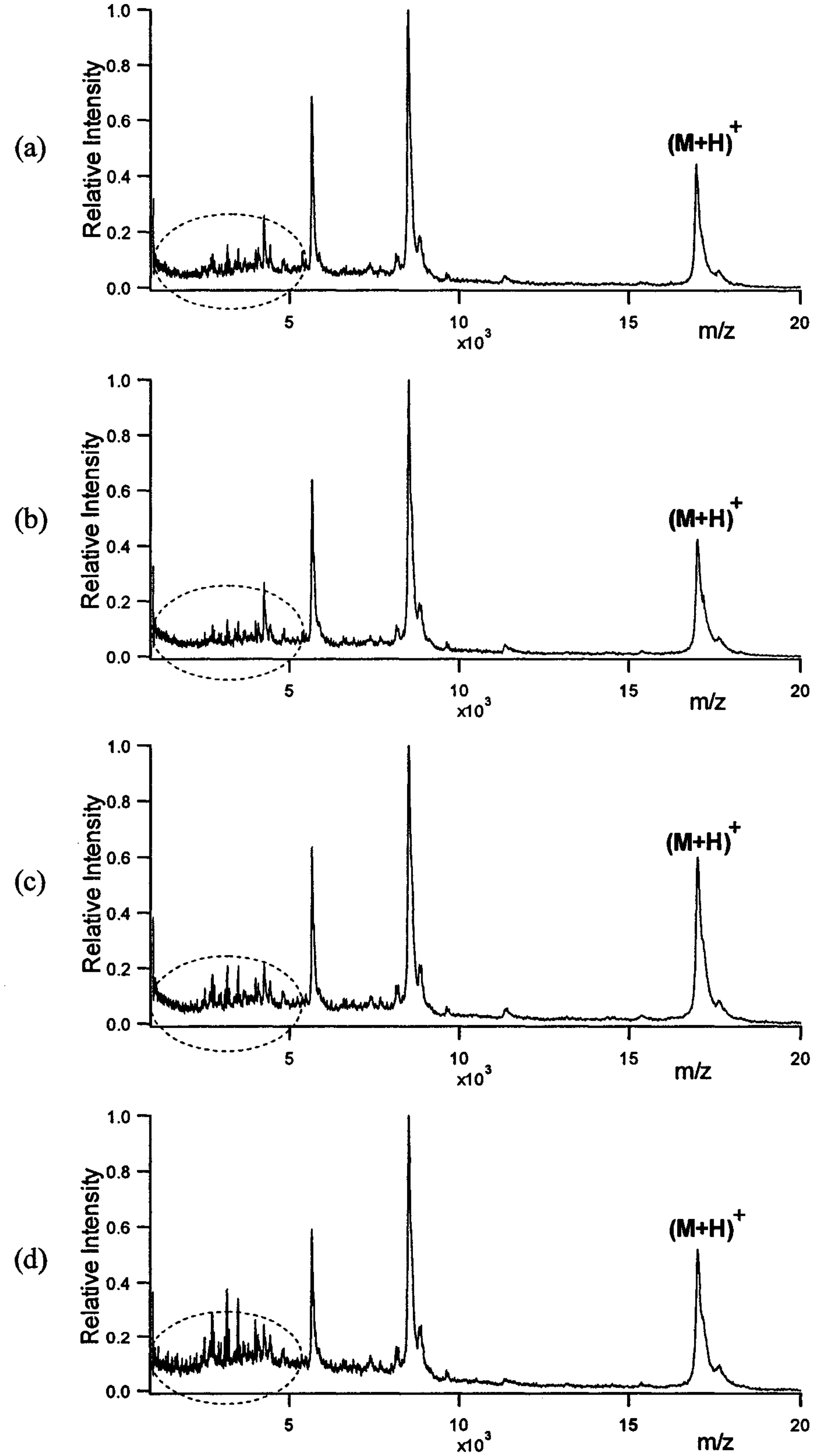


Figure 2.3. Effect of extraction solution volume on protein extraction. (a) 40 μL, (b) 50 μL, (c) 60 μL, (d) 80 μL.

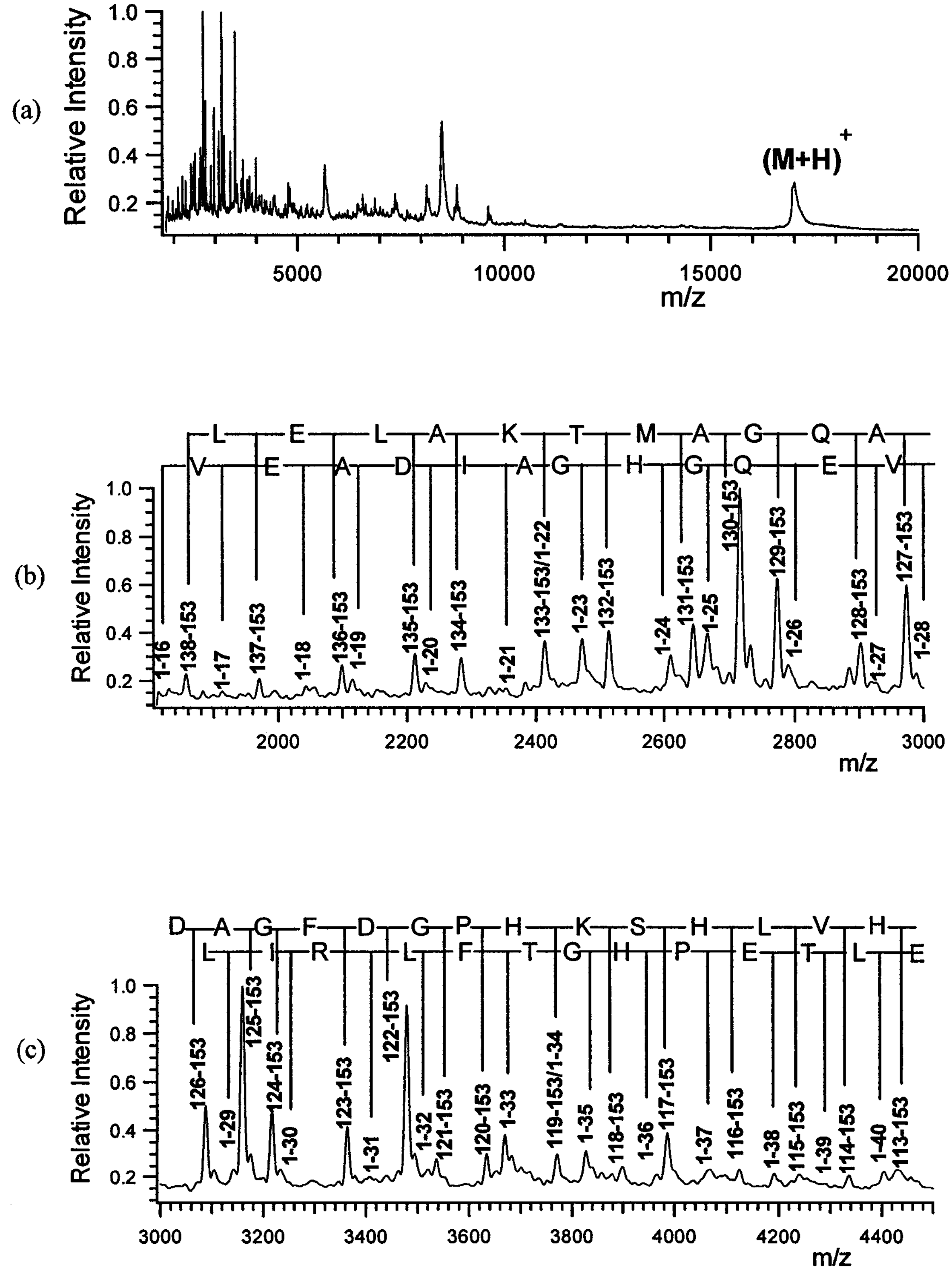
4) Effects of gel pore size and temperature

The pore size of the separation gel used in SDS-PAGE also affected protein extraction efficiency. Low percentage polyacrylamide gels with large pore sizes give rise to high extraction efficiency. This is because it is easy to remove protein molecules from bigger gel networks. To separate large proteins, it is better to use low percentage gels if the required separation resolution can be achieved. For small proteins, gels with large pore size cannot provide good separation as small proteins are not retained by the gel network, so higher percentage gels should be used. Overall, the gel percentage should be chosen to balance extraction efficiency and separation resolution based on the protein sizes.

Temperature also plays an important role in protein extraction. Acid hydrolysis of proteins can occur during exposure at room temperature under strong acid condition for several minutes and produce internal polypeptides. In order to avoid this random hydrolysis, the protein extraction was carried out at 4 °C.

Microwave-assisted acid hydrolysis of horse myoglobin

The initial study of microwave-assisted acid hydrolysis (MAAH) of protein analysis combined with SDS-PAGE separation was carried out on 1 nmol horse myoglobin. The mass spectrum of the hydrolysates from MAAH of the extracted myoglobin is shown in Figure 2.4. Figures 2.4 (b)-(e) show the expanded mass spectra of polypeptide sequence ladders for the myoglobin digest. After 2 minutes of exposure to microwave irradiation, the generation of polypeptide sequence ladders by acid hydrolysis from the protein was complete.


(a)
Relative Intensity
1.0
0.8
0.6
0.4
0.2
(M+H)+
5000
10000
15000
20000
m/z
(b)
L E L A K T M A G Q A
V E A D I A G H G Q E V
1-16
138-153
1-17
137-153
1-18
136-153
1-19
135-153
1-20
134-153
1-21
133-153/1-22
1-23
132-153
1-24
131-153
1-25
130-153
129-153
1-26
128-153
1-27
127-153
1-28
2000
2200
2400
2600
2800
3000
m/z
(c)
D A G F D G P H K S H L V H
L I R L F T G H P E T L E
126-153
1-29
125-153
124-153
1-30
123-153
1-31
122-153
1-32
121-153
120-153
1-33
119-153/1-34
1-35
118-153
1-36
117-153
1-37
116-153
1-38
115-153
1-39
114-153
1-40
113-153
3000
3200
3400
3600
3800
4000
4200
4400
m/z

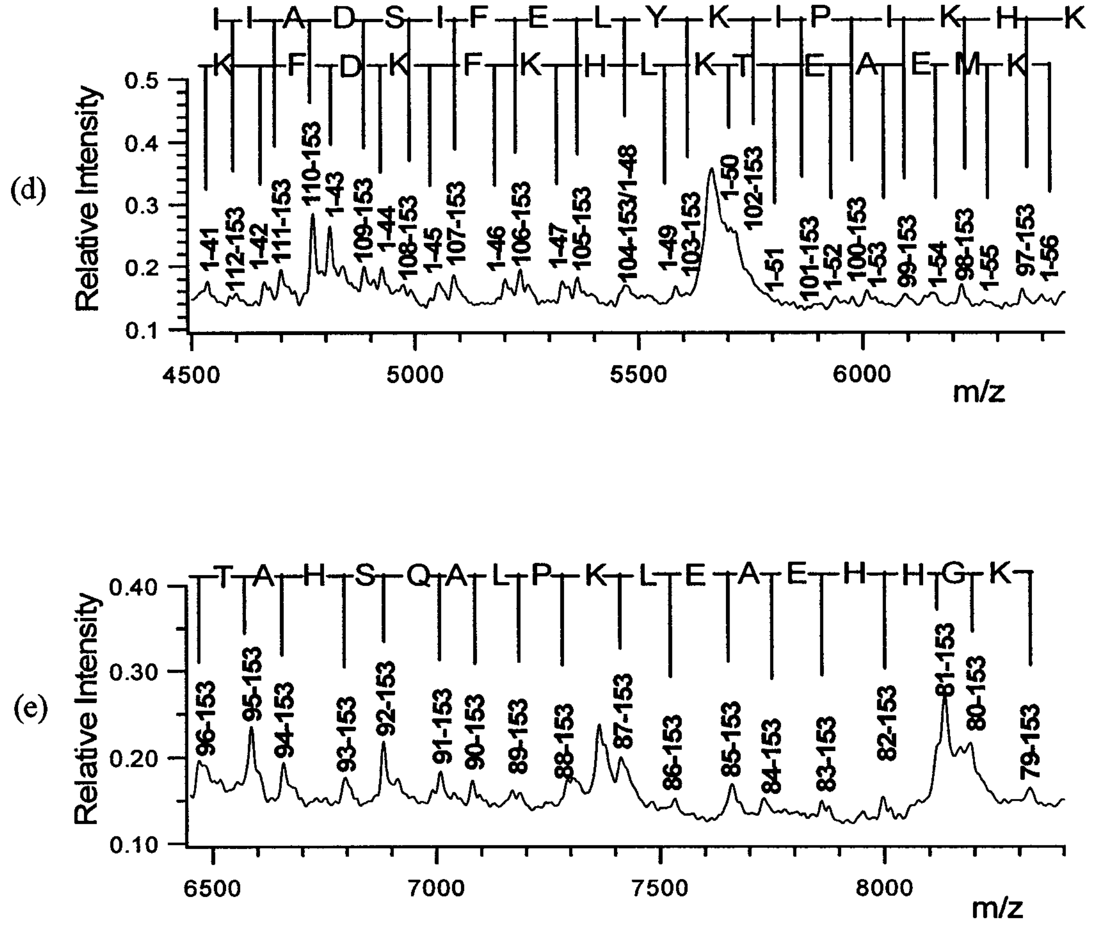


Figure 2.4. Mass spectra of MAAH of 1 nmol myoglobin protein extracted from gel piece. (a) Full Spectrum, (b) Mass Range of 1800-3000 Da, (c) Mass Range of 3000-4500 Da, (d) Mass Range of 4500-6450 Da, (e) Mass Range of 6450-8400 Da. Amino acids sequence given at the upper is C-terminal polypeptides while amino acids sequence given at the lower is N-terminal polypeptides.

In total, 99 out of 153 amino acids are read out quickly and easily in this spectrum and the polypeptide mass mapping obtained has sequence coverage of 65%. Overall, C-terminal polypeptides give stronger signals than N-terminal polypeptides. It has been reported that acid-hydrolyzed myoglobin produces exclusively C-terminal polypeptides [7]. Because microwave irradiation offers uniform energy to break each amide bond at the same time, N-terminal peptides as well as C-terminal peptides were observed. As a result, the amino acid sequence information from both C-terminal peptide peaks and N-terminal peptide peaks can be read in the same mass range compatibly.

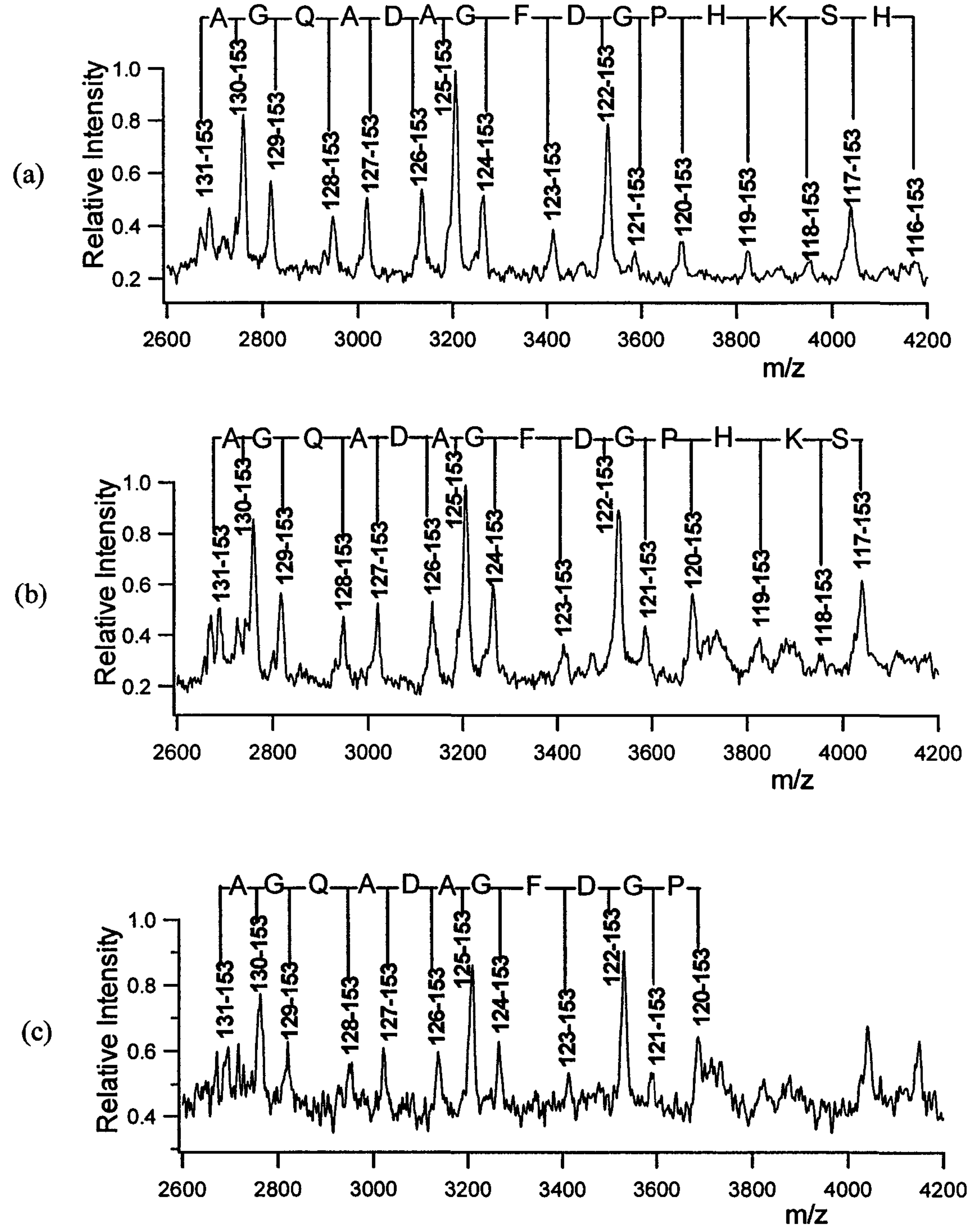
Good resolution and mass accuracy are important for amino acids sequence reading. In this study, linear mode was empolyed to obtain a mass spectrum of polypeptides in a wide mass range. The resulting resolution and sensitivity are acceptable for amino acids identification, especially for polypeptides in the low mass region. In this region, each amino acid can be read out continuously and accurately because small polypeptides are crystallized with the matrix relatively very well, and are easily ionized in MALDI MS. However, in the high mass region above 8000 Da, the identification accuracy is reduced. In this particular case, it is very difficult to obtain a full sequence coverage because hydrolyzed polypeptides fall into an extremely wide mass range over 2000 ~ 10,000 Da and they have different ionization efficiencies. Acid hydrolysis of a protein produces a large number of polypeptides. In the crystallization process of analytes with the matrix in MALDI-TOF experiments, the limited amount of matrix cannot be provided for all the polypeptides, especially large polypeptides which may need more matrix molecules. Moreover, the gel-related components extracted together with protein sample, such as salts, especially SDS, have negative effects on the amino acid identification [30, 34-36]. SDS binds with large polypeptides, so on-spot washing cannot be used for MALDI-MS analysis as sample loss would take place. All these factors will result in the reduction of the signal-to-noise ratio in the high mass region of the MALDI-TOF spectrum.

With the improvement on MALDI-TOF instrumentation or the use of advanced MALDI MS equipment, such as FT-ICR, mass accuracy and detection sensitivity would be increased and it should be possible to overcome these problems.

Sensitivity

The sensitivity of the technique was investigated with horse myoglobin. Figure 2.5 (a)-(e) shows mass spectra (2400 - 4200 Da) of the myoglobin digest obtained by loading 0.3 nmol, 0.1 nmol, 0.06 nmol, 0.03 nmol and 0.01 nmol protein, respectively, on SDS-PAGE. Only C-terminal polypeptides are discussed in this mass range. The number of amino acids identified is 15, 14, 11, 11, and 10, respectively.

Gel bands were handled as described in the experimental section. With the decreasing sample amount loaded on the gel, the signal-to-noise ratio was reduced and the sensitivity to identify amino acid was affected. Particularly, when the sample loading amount was decreased to less than 0.1 nmol, the ion signal was significantly reduced compared with the higher loading one. Ion suppression due to gel-induced components such as SDS on MS signals probably is the main cause of the signal reduction. With decreasing protein sample loading on SDS-PAGE, the protein amount extracted from gel pieces also decreases and the protein-to-SDS ratio is reduced. As a result, the ion signal suppression becomes more severe. However, there was no significant difference for 0.06 nmol, 0.03 nmol and 0.01 nmol, in which the identified amino acid sequence was found to be similar.



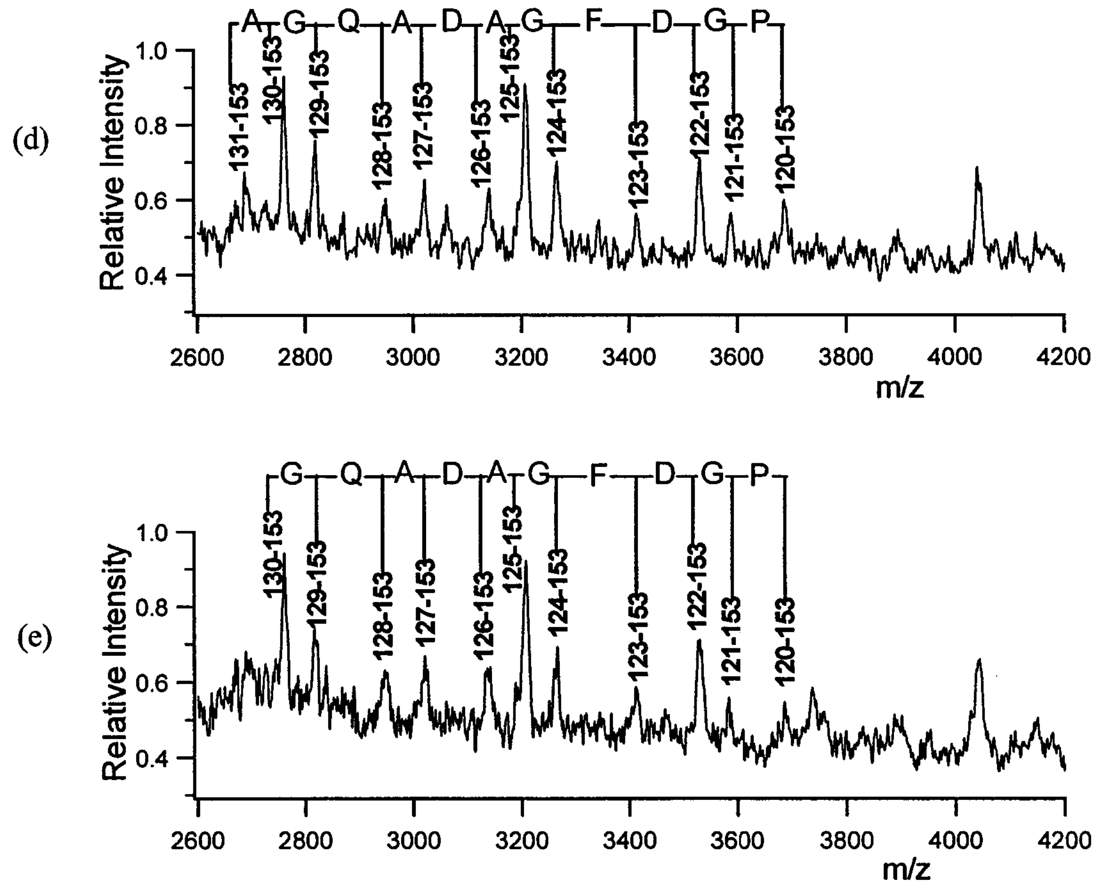


Figure 2.5. Sensitivity of amino acid sequence reading by MAAH of horse myglobin protein. (a) 0.3 nmol, (b) 0.1 nmol, (c) 0.06 nmol, (d) 0.03 nmol and (e) 0.01 nmol. In this discussed mass range, the less the protein amount loading on the gel and extraction, the less the number of amino acids was identified.

Application to α-S1-casein

The study of protein sequencing and protein modification was carried out on bovine α-S1-casein. As purchased, the protein has a purity of 85% that includes 15% α-S2-casein. It is difficult to purify the protein α-S1-casein using traditional in-solution separation approaches. α-S1-casein has six phosphorylation sites, making it a good model for the study of protein PTMs. In my study, SDS-PAGE is employed for a better separation and purification of α-S1-casein, and the MAP sequencing method provides the information needed to characterize PTMs.

The α-S1-casein protein band was excised, destained, and extracted for microwave digestion and MS analysis. Figures 2.6 (a)-(c) show the mass spectra of α-S1-casein microwave digest. 45 N-terminal polypeptides are read out clearly. Polypeptide peaks from α-S2-casein were not found in the mass spectra indicating the separation of the two proteins by SDS-PAGE was complete. Thus the accuracy for reading the amino acid sequence of α-S1-casein was much improved after SDS-PAGE purification.

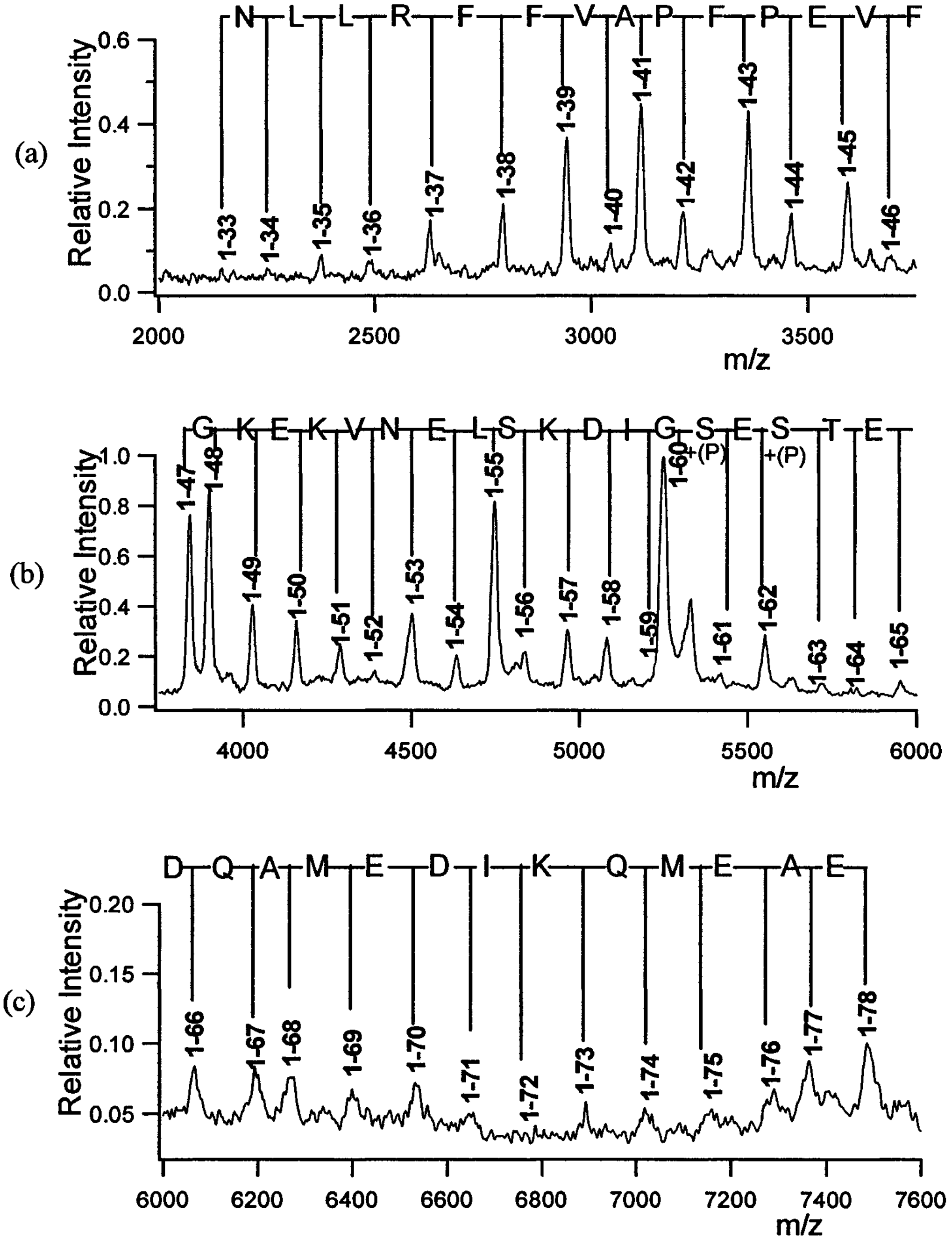


Figure 2.6. Mass spectra of MAP of bovine α-S1-casein over 2000-7500 Da.

Figures 2.7 (a) and (b) show the mass spectra of α-S1-casein and de-phosphorylated α-S1-casein in which 2 out of 6 phosphorylation sites (61S, 63S) are identified. Note that the phosphopeptide has a phosphate group with an 80 Da increment compared with the spectrum of dephosphorylated α-S1-casein sample purified by SDS-PAGE.

With microwave digestion of α-S1-casein, the N-terminal polypeptides have a stronger signal than the C-terminal polypeptides for some unknown reasons. Here, only the N-terminal polypeptides are labeled. Phosphorylation modification is difficult to identify in proteomics due to its weak signal in positive ionization mode and its easy loss of phosphorylation group in the ion-source region. Using the MAP method combined with SDS-PAGE purification, this modification can be identified easily and quickly by comparison with the mass spectra from a de-phosphorylated protein sample. However, as in the case of myoglobin analysis, amino acids are more difficult to identify in the mass region above 7500 Da for α-S1-casein. Gel-related components, like SDS and salt, degrade the mass spectra quality significantly.

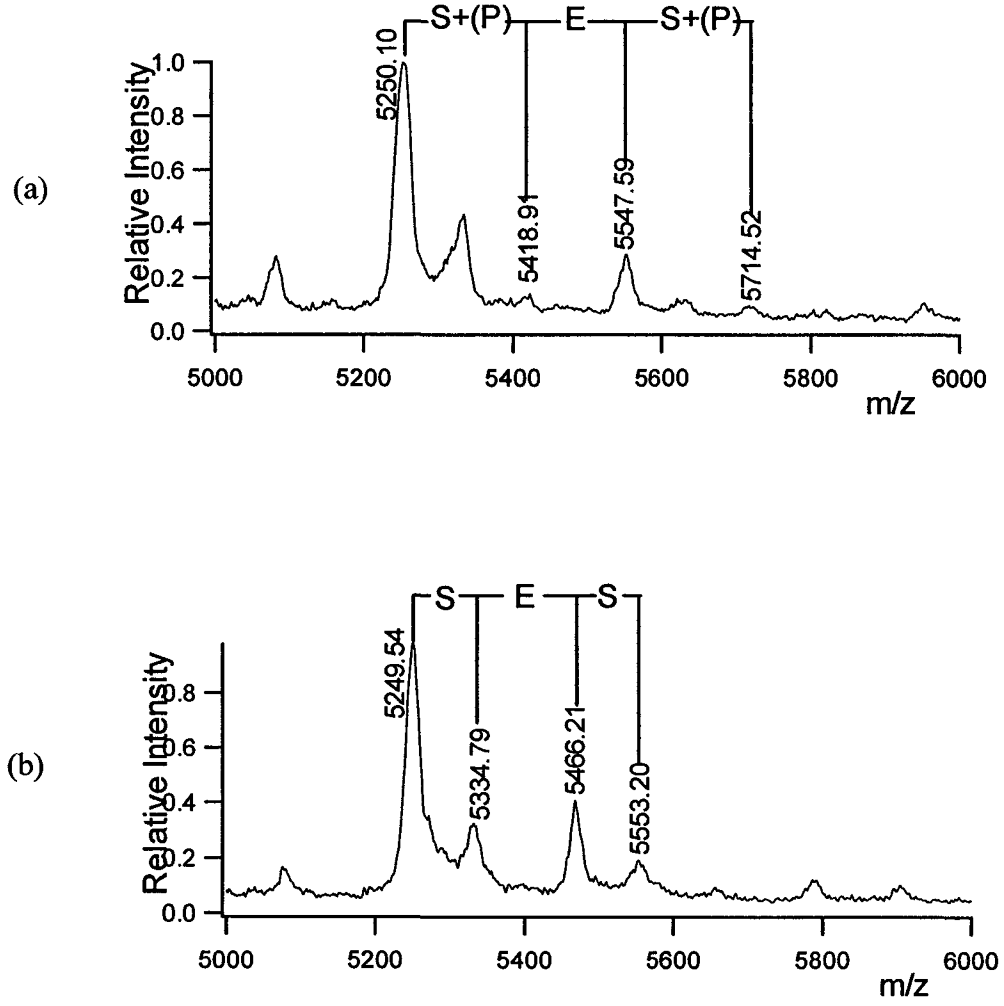


Figure 2.7. Mass spectra of α-S1-casein (a) and de-phosphorylated α-S1-casein (b) over 5000-6000 Da.

2.4 Conclusions

In this chapter, 1-D SDS-PAGE separation technique, a traditional approach to purifying protein samples, was combined with the MAP technique for amino acid sequence analysis and characterization of post-translational modifications. A passive gel extraction method was studied to bridge the two techniques. Several factors affecting protein extraction were investigated and the optimal extraction conditions were obtained. Horse myoglobin, as a protein model, was studied to test the efficiency of the two techniques in combination. In addition, a sensitivity study of this method was also presented. Finally, application of the technique to the study of post-translational modifications of proteins was demonstrated for the identification of phosphorylation sites on α-S1-casein. Overall, in the low mass region below 7000 Da, this combined technique proved to be efficient, but for amino acid sequence identification and identification of PTMs in a wide mass range, this method was found to have some limitations. The main problem comes from gel-related components, SDS, salts, etc., and gel staining materials, which deteriorate the mass spectra quality. In the future, a purification procedure may be developed to clean up the polypeptide samples which may improve the detection sensitivity and mass resolution in MALDI MS. In addition, high resolution instruments, such as FT-ICR, may also improve the performance of the described technique.

2.5 Literature Cited

(1) Edman, P.; Begg, G. *Eur.J.Biochem.* **1967**, *1,* 80.

(2) Patterson, D.H.; Tarr, G.E.; Regnier, F.E.; Martin, S.A. *Anal. Chem.* **1995**, *67*, 3971

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

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

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

(6) Zubarev, R.A.; Chivanov, V.D.; Hakansson, P.; Sundqvist, B.U. *Rapid Commun. Mass Spectrom.* **1994,** *8*, 906

(7) Gobom, J.; Mirgordskaya, E.; Nordhoff, E.; Hojrup, P.; Roepstorff, P. *Anal. Chem.* **1999**, *71*, 919

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

(9) Lin, S.H.; Tornatore, P.; Weinberger, S.R. *Eur. J. Mass Spectrom.* **2001**, *7*, 131

(10) Reiber, D.C.; Brown, R.S.; Weinberger, S.; Kenny, J.; Bailey, J. *Anal. Chem.* **1998**, *70*, 1214

(11) Lennon, J.J.; Walsh, K.A. *Protein Sci.* **1997**, *6*, 2446

(12) Keough, T.; Youngquist, R.S.; Lacey, M.P. *Proc. Natl. Acad. Sci.* USA **1999,** *96*, 7131

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

(14) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass. Spectrom.* **1993,** *22*, 338

(15) Pappin, D.J.; Hojrup, P.; Bleasby, A. *J. Curr. Biol.* **1993**, *3*, 327-332.

(16) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58-64.

(17) Henzel, W. J.; Billeci, T.M.; Stults, J.T.; Wong, S.F.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5011-5015.

(18) Pramanik, B.N.; Mirza, U.A.; Ing, Y.H.; Liu, Y.H.; Bartner, P.L.; Weber, P.C.; Bose, A.K. *Protein Sci.* **2002**, 11, 2676

(19) Zhong, Y.; Zhang, Y.; Wen, Z.; Li, L. *Nature Biotech.* **2004**, *22*, 1291

(20) Patterson, S.D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791

(21) Jungblut, P.; Thiede, B. *Mass spectrum. Rev.* **1997**, *16*, 145

(22) Schuhmacher, M.; Glocker, M.O.; Wunderlin, M.; Przybylski, M. *Electrophoresis,* **1996**, *17,* 848

(23) Clarke, N.J.; Li, F.; Tomlinson, A.J.; Naylor, S. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 88

(24) Vestling, M.M.; Fenselay, C. *Anal. Chem.* **1994**, *66*, 471

(25) Liang, X.; Bai, J.; Liu, Y.H.; Lubman, D.M. *Anal. Chem.* **1996**, *68*, 1012

(26) Ogorzalek Loo, O.; Stevenson, T.L.; Mitchell, C.; Loo, J.A.; Andrews, P.C. *Anal. Chem.* **1996**, *68*, 1910

(27) Ogorzalek Loo, O.; Mitchell, C.; Stevenson, T.I.; Martin, S.A.; Hines, W.M.; Juhasz, P.; Patterson, D.H.; Peltier, J.M.; Loo, J.A.; Andrews, P.C. *Electrophoresis.* **1997**, *18*, 382

(28) Ehring, H.; Stromberg, S.; Tjernberg, A.; Noren, B. *Rapid commun. Mass Spectrom.* **1997**, *11*, 1867

(29) Cohen, S.L.; Chait, B.T. *Anal. Biochem.* **1997**, *247*, 257

(30) Jeannot, M.A.; Zheng, J.; Li, L. *J. Am. Soc. Mass. Spectrum.* **1999,** *10*, 512

(31) Schagger, H.; von Jagow, G. *Anal. Biochem.* **1987,** *166*, 368

(32) Qin, J.; Fenyo, D.; Zhao, Y.; Hall, W.W.; Chao, D.M.; Wilson, C.J.; Young, R. A.; Chait, B.T. *Anal Chem.* **1997**, *69*, 3995

(33) Dai, Y-Q.; Whittal, R.M.; Li, L. *Anal. Chem.* **1996**, *68*, 2494

(34) Rosinke, B.; Strupat, K.; Hillenkamp, F.; Rosenbusch, J.; Dencher, N.; Kruger, U.; Galla, H.J. *J. Mass Spectrom.* **1995**, 30, 1462

(35) Puchades, M.; Westman, A.; Blennow, K.; Davidsson, P. *Rapid Commun. Mass Spectrom.* **1999,** *13*, 344

(36) Fenesan, I.; Popescu, R.; Supuran, C.T.; Nicoara, S.; Culea, M.; Palibroda, N.; Moldovan, Z.; Cozar, O. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 721

Chapter 3

Identification of Hemoglobin G Coshatte Using MALDI-TOF MS Combined With Microwave-Assisted Acid Hydrolysis of Proteins

3.1 Introduction

The first analysis of a tryptic peptide mixture of globin chains from human hemoglobin (Hb) using field desorption mass spectrometry was reported by Matsuo [1]. Since then, MS has been rapidly applied to the characterization of a great variety of Hb's. Recently, electrospray ionization mass spectrometry (ESI-MS) [2, 3, 5, 8-11] and matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) [4, 6, 7] have become two main analysis tools for the detection and identification of hemoglobin variants.

Most researchers begin with the use of ESI-MS to obtain accurate molecular masses of the intact normal and variant globin chains [2-5]. Peptide mass mapping after enzymatic digestion of Hb is then carried out to determine the variant peptide [2, 3, 5]. Finally, MS/MS analysis of the selected peptides is used to identify the variant amino acid sites. In MS/MS, the obtained CID spectra can provide fragmentation information of variant peptides, and *de novo* sequencing technique can also be applied to identify the amino acid mutation sites based on the obtained information such as the assigned *b* or *y* ions [5]. However, data interpretation in variant analysis by MS and MS/MS requires extensive expertise and working experience. In most cases, other biological techniques, such as gel electrophoresis and PCR, are also involved to provide complementary information. Instead of manually carrying out *de novo* sequencing, MS/MS spectra can be analyzed with the help of modified algorithm of data analysis software [3, 8-11] which includes possible amino acid variants predicted from the study of genomics and transcriptomics.

Due to its high mass accuracy, MALDI-TOF MS was used to determine the molecular mass of intact hemoglobins [13] as well as for peptide mass mapping [4, 6]. However, MALDI-TOF itself cannot be used for CID analysis of peptide ions. Thus, other tandem mass spectrometers are needed.

To obtain definitive identification of amino acid variants of hemoglobin chains, multi-step procedures are involved, such as heme group removal, derivatization or reduction of the cysteines, overnight digestion with trypsin or other enzymes, LC separation of peptides prior to MS, and time-consuming data processing and interpretation of complex multi-charged MS spectra obtained by ESI. Any of those steps may directly affect the identification of amino acid variants. Therefore, each step needs to be carried out carefully.

Recently, it has been demonstrated that exposure of a purified protein solution to a short time microwave irradiation increases the efficiency of acid hydrolysis. Such a treatment generates two series of polypeptide ladders containing either the N- or C-terminal polypeptides predominantly. MS analysis of hydrolysate produces a simple mass spectrum consisting of polypeptide peaks, which allows direct reading of the amino acid sequence of a protein quickly [12]. This microwave-assisted acid hydrolysis (MAAH) of proteins followed by MALDI MS analysis was termed as mass analysis of polypeptide ladders (MAP) sequencing technique.

In this work, accurate determination of the amino acid variant site of Hb G Coushatta by using the MAP sequencing technique is reported. After exposure of the purified Hb β-chain fraction from LC separation to microwave irradiation for 1.5 min, polypeptide sequence ladders in the mass region of 1200-3300 Da were analyzed by reflector mode MALDI-TOF and each amino acid could be read unambiguously on the basis of the mass difference of consecutive polypeptide peaks in a single mass spectrum. By comparison with the resulting polypeptide ladder of normal Hb β-chain obtained under the same experimental conditions, the site of the amino acid variant could be determined accurately. To confirm the results, the conventional peptide mapping technique and *de novo* sequencing of MS/MS of selected trypsin digest peptides were also employed.

3.2 Experimental

3.2.1 Materials and Reagents

Deionized water from a Millipore NANOpure water system was used in all cases. HPLC grade acetonitrile, methanol, TFA and analytical grade HCl were purchased from Fisher Scientific Canada. 2, 5-dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHCA) and hemoglobin standard were obtained from Sigma Aldrich. DHB and CHCA were re-crystallized before use.

3.2.2 Hb cell extraction

The blood sample was washed three times with 0.85% NaCl to remove serum proteins. The remaining red cells were lysed in four volumes of cold H_2O and vortexed. Cell membranes were removed by centrifugation for 10 min at 14000 g.

3.2.3 Hb purification by RP-HPLC

Hemoglobin separation was performed by a reversed phase HPLC procedure. Hemoglobin samples were applied to a C_8 column developed with a linear gradient between solvent A (0.1%TFA/0.4%ACN/H_2O) and solvent B (0.1%TFA/10%MeOH/ACN) at a flow rate of 40μL/min. The gradient started with 0%B and ended with 90%B in 100 min. Peaks were monitored at 214 nm in the UV detector.

3.2.4 Microwave-assisted acid hydrolysis of hemoglobin

Purified hemoglobin β-chain fraction was dried down completely and 10 μL H_2O was added to dissolve the protein sample, then 10 μL 6 M HCl was mixed with the protein solution. It is worthwhile to mention that when the hemoglobin solution was mixed with the HCl the heme group was removed simultaneously. The sample solution was irradiated in a 900W household microwave oven for 1.5 min. Finally, the hydrolyzed sample was dried down in a vacuum centrifuge before MALDI analysis.

3.2.5 Protein preparation for proteolytic digestion

20 μL extracted hemoglobin solution was added to 20 volumes of ice-cold 0.6% HCl/acetone. The precipitated globin was washed three times with cold acetone and dissolved into 100 μL water. 50 μL of globin solution was adjusted to pH 8.5 with 1 M

$NaHCO_3$, then 5 µL of 90 mM DTT was added to the protein sample and incubated for 1 h at 37 °C. 5 µL of 8.4 mg/mL iodoacetamide was added to the reduced sample and kept in the dark for 1 h.

3.2.6 Proteolytic digestion

1 µL of trypsin (1 µg/µL) was added to the protein solution and the mixture was incubated for 8 h at 37 °C. The reaction was quenched with 1% TFA.

3.2.7 MALDI-TOF MS analysis

All MS analyses were performed on a Bruker Reflex III MALDI-TOF mass spectrometer (Bremen/Leipzig, German) with a SCOUT 384 multiprobe inlet and a pulsed 337 nm nitrogen laser. It was operated with delayed extraction positive ion mode. The linear mode of operation was used for protein detection while the reflectron mode was used for peptide detection. The data were then reprocessed using the Igor Pro software package (WaveMetrics Inc., Lake Oswego, OR).

For the mass measurement analysis of intact Hb proteins, the two-layer method with CHCA was used to provide good resolution. 1 µL of saturated solution of CHCA in 25% methanol/acetone was deposited on the target and was allowed to dry. 0.5 µL of protein solution was mixed with the 2nd layer (saturated solution of CHCA in 50% ACN/H_2O) at a 1:1 ratio and 0.4 µL of this mixture was deposited on the top of the first layer and allowed to dry, followed by on-spot washing three times. External calibration was performed with bovine myoglobin (MW=16951.61 Da).

For the polypeptide ladder analyses, the dried-droplet method with DHB was used. The dried sample was dissolved in 1 µL water and mixed with matrix solution of 100 mg/mL DHB in 50%ACN/H_2O at a 1:1 ratio and 1 µL was loaded on the spot.

For tryptic protein digest analysis, the sample was prepared as above except the tryptic digest solution was mixed directly with DHB matrix solution at a ratio of 1:1.

3.2.8 MALDI QqTOF analysis of tryptic peptides

MS/MS analyses of tryptic peptides were performed on an Applied Biosystems/MSD-Sciex QSTAR Pulsar QqTOF instrument equipped with an orthogonal MALDI source employing a 337 nm nitrogen laser (Concord, ON, Canada). The instrument was operated in positive mode and collision-induced dissociation (CID) was achieved with Ar as collision gas. The spectra were then re-processed using the Igor Pro software package. Samples were prepared using the dried-droplet method with DHB, as described above.

3.3 Results & Discussion

The amino acid variant of hemoglobin G Coushatta, a mutation of 22E→22A, was identified by MS analysis with high accuracy. Two methods were used: MAP and traditional tryptic digestion. The workflow of the MS strategies is given in Figure 3.1. The MAAH method was shown on the left, while the traditional *de novo* sequencing combined with tryptic protein digestion was shown on the right.

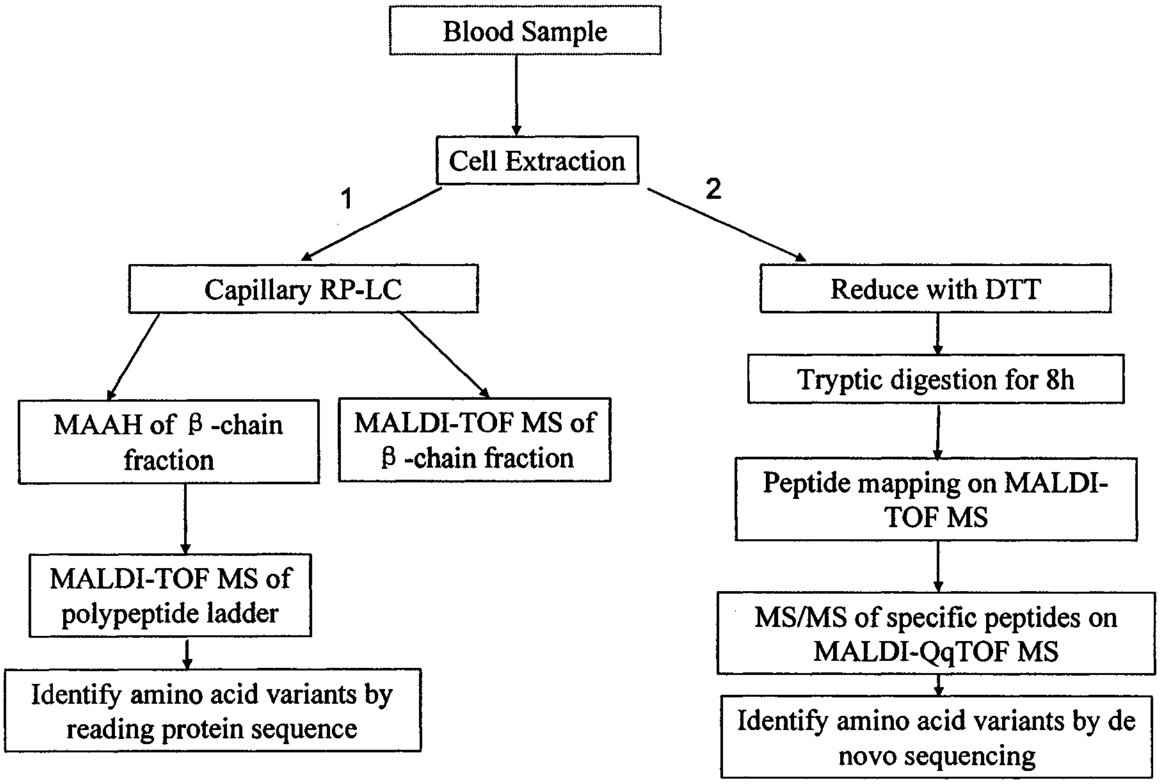

Blood Sample
Cell Extraction
1
2
Capillary RP-LC
Reduce with DTT
MAAH of β -chain fraction
MALDI-TOF MS of β -chain fraction
Tryptic digestion for 8h
Peptide mapping on MALDI-TOF MS
MALDI-TOF MS of polypeptide ladder
MS/MS of specific peptides on MALDI-QqTOF MS
Identify amino acid variants by reading protein sequence
Identify amino acid variants by de novo sequencing

Figure 3.1. Steps taken to identify the amino acid variant by MALDI MS analysis.

In the MAAH method, Hemoglobin was purified first by RP-LC after red blood cell extraction. Figure 3.2 shows the separation chromatogram of hemoglobin extraction. Normal and variant *β*-chain eluted together while the α-chain protein eluted out separately. Although single pure variant β-chain fraction was not obtained, the mixture of the two β-chain proteins could be employed for direct analysis to obtain their molecular mass information and for subsequent digestion because their similar sequence has little influence on the digestion.

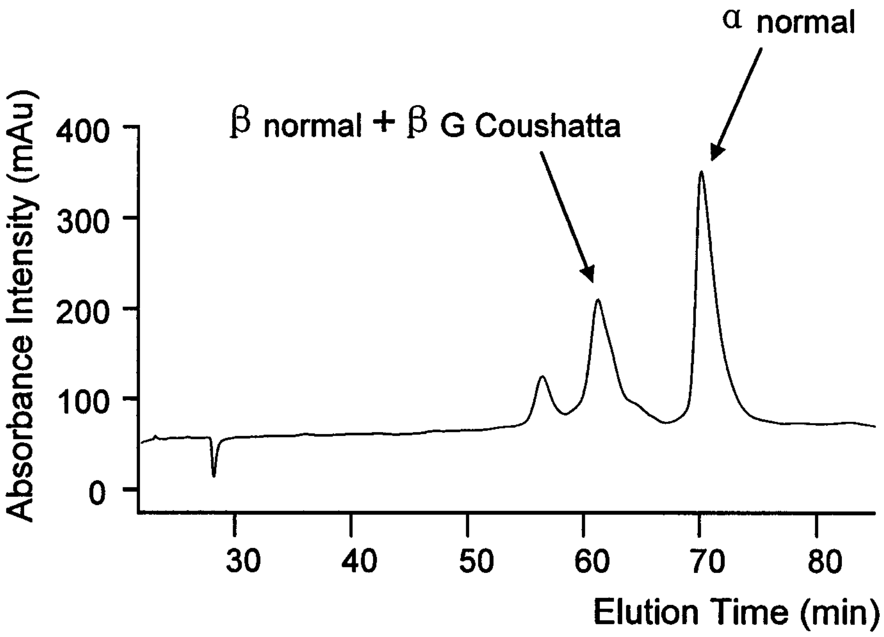


Figure 3.2. RP-LC Separation Chromatograph of Hemoglobin G Coushatta.

The accurate molecular mass information of the β-chains is shown in Figure 3.3. Unlike the ESI technique which generates multi-charged species, singly-charged molecular ion peak is usually obtained in MALDI MS spectra and can be readily interpreted to generate the molecular weight information. The normal β-chain co-eluted with the variant provides an internal standard for evaluating the accuracy of the molecular mass obtained for the variant globin chain. For the normal β-chain protein, the molecular mass is 15867.5 Da (± 0.3 Da) and, for the variant β-chain protein, the obtained molecular mass is 15809.2 Da (± 0.1 Da). It is very clear that there is a mass shift of 58 Da between the variant β-chain protein and the normal one. This indicates the presence of an amino acid variant, but the variant site is unknown and needs further investigation.

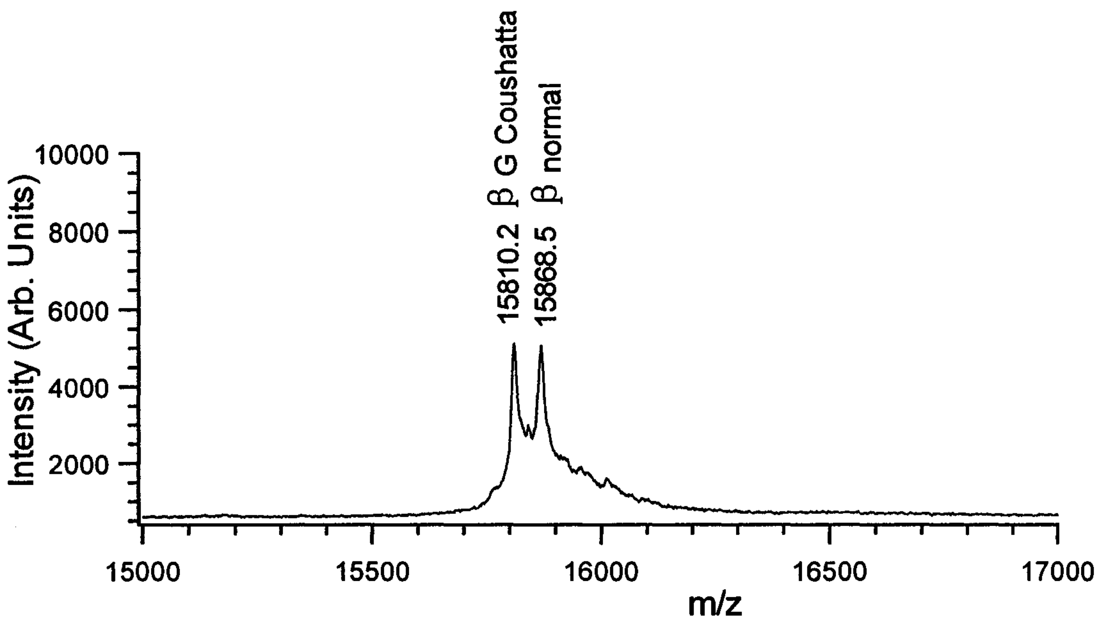


Figure 3.3. MALDI-TOF MS of intact normal and variant β-chain proteins.

In the second experiment, the protein fraction containing the normal and variant β-chains was subjected to MAAH. As shown in Figure 3.4, the amino acid sequence can be read completely over a mass range of 1200-3300 Da for both the normal β-chain and the variant G-Coushatta β-chain. Table 1-4 shows the N- and C-terminal polypeptide information for the variant and normal β-chain, including polypeptide sequences, m/z, and each amino acid calculated by the mass difference between consecutive peptides. The error values are also provided to evaluate the mass accuracy. The errors are all below 30 ppm, ensuring accurate determination of the amino acid sequence. By comparing the amino acids one by one in the mass spectra of polypeptide ladders of the normal and variant β-chains, it is easy to find that each C-terminal polypeptide is identical in both spectra, while for the N-terminal polypeptides, there is a mass shift happening at the 22nd site. And starting with the polypeptide peak 1-22, each N-terminal polypeptide peak has a mass shift of 58 Da. There is only one possibility corresponding to a mass reduction of 58 Da to amino acid residue β22 Glu (129 Da): Ala (71 Da) was substituted for Glu.

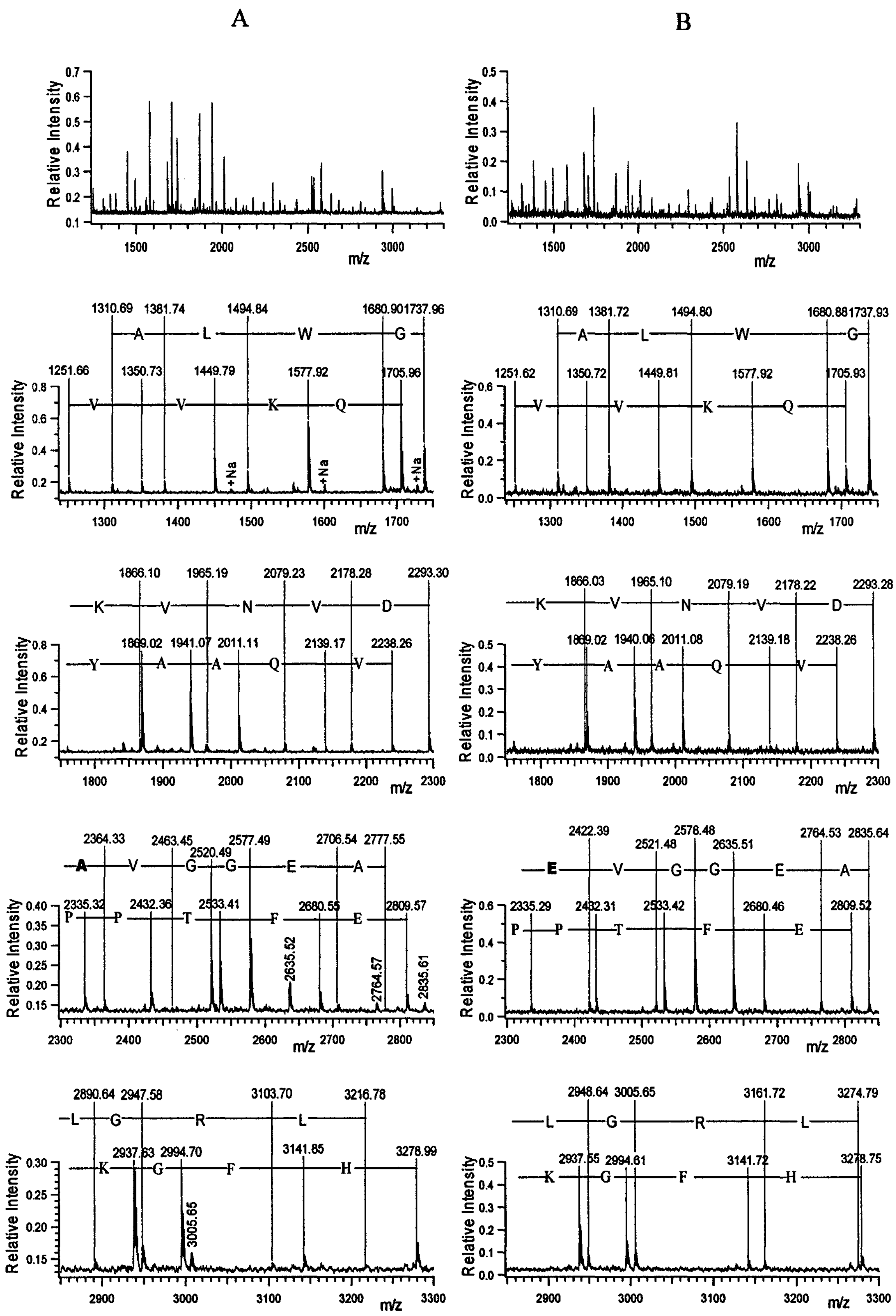

A
B
Relative Intensity
m/z
1310.69 1381.74 1494.84 1680.90 1737.96
A L W G
1251.66 1350.73 1449.79 1577.92 1705.96
V V K Q
+Na
1310.69 1381.72 1494.80 1680.88 1737.93
1251.62 1350.72 1449.81 1577.92 1705.93
1866.10 1965.19 2079.23 2178.28 2293.30
K V N V D
1869.02 1941.07 2011.11 2139.17 2238.26
Y A A Q V
1866.03 1965.10 2079.19 2178.22 2293.28
1869.02 1940.06 2011.08 2139.18 2238.26
2364.33 2463.45 2520.49 2577.49 2706.54 2777.55
A V G G E A
2335.32 2432.36 2533.41 2680.55 2809.57
P P T F E
2635.52 2764.57 2835.61
2422.39 2521.48 2578.48 2635.51 2764.53 2835.64
E V G G E A
2335.29 2432.31 2533.42 2680.46 2809.52
2890.64 2947.58 3103.70 3216.78
L G R L
2937.63 2994.70 3141.85 3278.99
K G F H
3005.65
2948.64 3005.65 3161.72 3274.79
2937.55 2994.61 3141.72 3278.75

Figure 3.4. MALDI-TOF MS of MAP for variant β-chain (A) and normal β-chain (B) over 1200 Da - 3300 Da. To read out amino acid sequence and further identify amino acid variant site, expanded MS are given below each one.

Amino acid sequence of hemoglobin β-chain: VHLTP EEKSA VTALW GKVNV DE(A)VGG EALGR LLVVY PWTQR......VCVLA HHFGK EFTPP VQAAY QKVVA GVANA LAHKYA. Identified amino acids are underlined.

In the expanded MS spectra, the amino acid sequence given at the upper corresponds to N-terminus while the amino acid sequence given at the lower corresponds to C-terminus. In the mass spectra of variant β-chain, labeled peaks 2635.52 Da, 2764.57 Da, 2835.61 Da, and 3005.65 Da are from normal β-chain because this fraction actually is a mixture of normal and variant β-chain protein.

Table 3.1. N-terminus polypeptide of normal Hb β-chain.

Position	Peptide sequence	m/z (Da)	Δm	AA identified	Error (ppm)
1-12	VHLTPEEKSAVT	1310.69			
1-13	VHLTPEEKSAVTA	1381.72	71.06	A	14
1-14	VHLTPEEKSAVTAL	1494.80	113.08	L/I	0
1-15	VHLTPEEKSAVTALW	1680.88	186.08	W	0
1-16	VHLTPEEKSAVTALWG	1737.93	57.05	G	17
1-17	VHLTPEEKSAVTALWGK	1866.03	128.10	K	5
1-18	VHLTPEEKSAVTALWGKV	1965.10	99.07	V	0
1-19	VHLTPEEKSAVTALWGKVN	2079.19	114.09	N	24
1-20	VHLTPEEKSAVTALWGKVNV	2178.22	99.03	V	18
1-21	VHLTPEEKSAVTALWGKVNVD	2293.28	115.06	D	13
1-22	VHLTPEEKSAVTALWGKVNVDE	2422.39	129.11	E	28
1-23	VHLTPEEKSAVTALWGKVNVDEV	2521.48	99.09	V	8
1-24	VHLTPEEKSAVTALWGKVNVDEVG	2578.48	57.00	G	8
1-25	VHLTPEEKSAVTALWGKVNVDEVGG	2635.51	57.03	G	4
1-26	VHLTPEEKSAVTALWGKVNVDEVGGE	2764.53	129.02	E	7
1-27	VHLTPEEKSAVTALWGKVNVDEVGGEA	2835.64	71.11	A	24
1-28	VHLTPEEKSAVTALWGKVNVDEVGGEAL	2948.64	113.00	L/I	27
1-29	VHLTPEEKSAVTALWGKVNVDEVGGEALG	3005.65	57.01	G	3
1-30	VHLTPEEKSAVTALWGKVNVDEVGGEALGR	3161.72	156.07	R	9
1-31	VHLTPEEKSAVTALWGKVNVDEVGGEALGRL	3274.79	113.07	L/I	3

Table 3.2. N-terminus polypeptide of variant Hb β-chain.

Position	Peptide sequence	m/z (Da)	Δm	AA identified	Error (ppm)
1-12	VHLTPEEKSAVT	1310.69			
1-13	VHLTPEEKSAVTA	1381.74	71.05	A	7
1-14	VHLTPEEKSAVTAL	1494.84	113.1	L/I	13
1-15	VHLTPEEKSAVTALW	1680.90	186.06	W	12
1-16	VHLTPEEKSAVTALWG	1737.96	57.06	G	23
1-17	VHLTPEEKSAVTALWGK	1866.10	128.14	K	26
1-18	VHLTPEEKSAVTALWGKV	1965.19	99.09	V	10
1-19	VHLTPEEKSAVTALWGKVN	2079.23	114.04	N	0
1-20	VHLTPEEKSAVTALWGKVNV	2178.28	99.05	V	9
1-21	VHLTPEEKSAVTALWGKVNVD	2293.30	115.02	D	4
1-22	VHLTPEEKSAVTALWGKVNVDA	2364.33	71.03	A	4
1-23	VHLTPEEKSAVTALWGKVNVDAV	2463.45	99.12	V	20
1-24	VHLTPEEKSAVTALWGKVNVDAVG	2520.49	57.04	G	8
1-25	VHLTPEEKSAVTALWGKVNVDAVGG	2577.49	57.00	G	7
1-26	VHLTPEEKSAVTALWGKVNVDAVGGE	2706.54	129.05	E	4
1-27	VHLTPEEKSAVTALWGKVNVDAVGGEA	2777.55	71.01	A	11
1-28	VHLTPEEKSAVTALWGKVNVDAVGGEAL	2890.64	113.09	L/I	3
1-29	VHLTPEEKSAVTALWGKVNVDAVGGEALG	2947.58	56.94	G	27
1-30	VHLTPEEKSAVTALWGKVNVDAVGGEALGR	3103.70	156.12	R	6
1-31	VHLTPEEKSAVTALWGKVNVDAVGGEALGRL	3216.78	113.08	L/I	0

Table 3.3. C-terminus polypeptide of normal Hb β-chain.

Position	Peptide sequence	m/z (Da)	Δm	AA identified	Error (ppm)
135-146	AGVANALAHKYH	1251.62			
134-146	VAGVANALAHKYH	1350.72	99.10	V	22
133-146	VVAGVANALAHKYH	1449.81	99.09	V	14
132-146	KVVAGVANALAHKYH	1577.92	128.11	K	13
131-146	QKVVAGVANALAHKYH	1705.93	128.01	Q	29
130-146	YQKVVAGVANALAHKYH	1869.02	163.09	Y	16
129-146	AYQKVVAGVANALAHKYH	1940.06	71.04	A	0
128-146	AAYQKVVAGVANALAHKYH	2011.08	71.04	A	0
127-146	QAAYQKVVAGVANALAHKYH	2139.18	128.10	Q	28
126-146	VQAAYQKVVAGVANALAHKYH	2238.26	99.07	V	0
125-146	PVQAAYQKVVAGVANALAHKYH	2335.29	97.03	P	8
124-146	PPVQAAYQKVVAGVANALAHKYH	2432.31	97.02	P	12
123-146	TPPVQAAYQKVVAGVANALAHKYH	2533.42	101.11	T	23
122-146	FTPPVQAAYQKVVAGVANALAHKYH	2680.46	147.04	F	11
121-146	EFTPPVQAAYQKVVAGVANALAHKYH	2809.52	129.06	E	7
120-146	KEFTPPVQAAYQKVVAGVANALAHKYH	2937.55	128.03	K	20
119-146	GKEFTPPVQAAYQKVVAGVANALAHKYH	2994.61	57.07	G	16
118-146	FGKEFTPPVQAAYQKVVAGVANALAHKYH	3141.72	147.11	F	13
117-146	HFGKEFTPPVQAAYQKVVAGVANALAHKYH	3278.75	137.03	H	9

Table 3.4. C-terminus polypeptide of variant Hb β-chain.

Position	Peptide sequence	m/z (Da)	Δm	AA identified	Error (ppm)
135-146	AGVANALAHKYH	1251.66			
134-146	VAGVANALAHKYH	1350.73	99.07	V	7
133-146	VVAGVANALAHKYH	1449.79	99.06	V	7
132-146	KVVAGVANALAHKYH	1577.92	128.13	K	25
131-146	QKVVAGVANALAHKYH	1705.96	128.04	Q	11
130-146	YQKVVAGVANALAHKYH	1869.02	163.06	Y	0
129-146	AYQKVVAGVANALAHKYH	1940.07	71.05	A	5
128-146	AAYQKVVAGVANALAHKYH	2011.11	71.04	A	0
127-146	QAAYQKVVAGVANALAHKYH	2139.17	128.06	Q	0
126-146	VQAAYQKVVAGVANALAHKYH	2238.26	99.09	V	9
125-146	PVQAAYQKVVAGVANALAHKYH	2335.32	97.06	P	4
124-146	PPVQAAYQKVVAGVANALAHKYH	2432.36	97.04	P	4
123-146	TPPVQAAYQKVVAGVANALAHKYH	2533.41	101.05	T	0
122-146	FTPPVQAAYQKVVAGVANALAHKYH	2680.55	147.14	F	26
121-146	EFTPPVQAAYQKVVAGVANALAHKYH	2809.57	129.02	E	7
120-146	KEFTPPVQAAYQKVVAGVANALAHKYH	2937.63	128.04	K	20
119-146	GKEFTPPVQAAYQKVVAGVANALAHKYH	2994.70	57.07	G	16
118-146	FGKEFTPPVQAAYQKVVAGVANALAHKYH	3141.85	147.15	F	25
117-146	HFGKEFTPPVQAAYQKVVAGVANALAHKYH	3278.99	137.14	H	24

This result was further confirmed by traditional tryptic protein digestion experiments. Figure 3.5 shows the mass spectrum of the tryptic peptides generated by 8 h digestion. Peptide peaks at 1256.38 Da and 1314.36 Da correspond to the peptide sequences 18-30 [VNVDE(A)VGGEALGR] of the normal and variant β-chains, respectively, which have a mass difference of 58 Da. As they have similar intensity, the concentration ratio of the variant and normal globins is likely to be 1:1 in the human red blood cell. However the definite variant site is still unknown and there are many possibilities, such as Hb Tripoli (β26 Glu→Ala) [4] and Hb Connecticut (β21 Asp→Gly), so further information is needed.

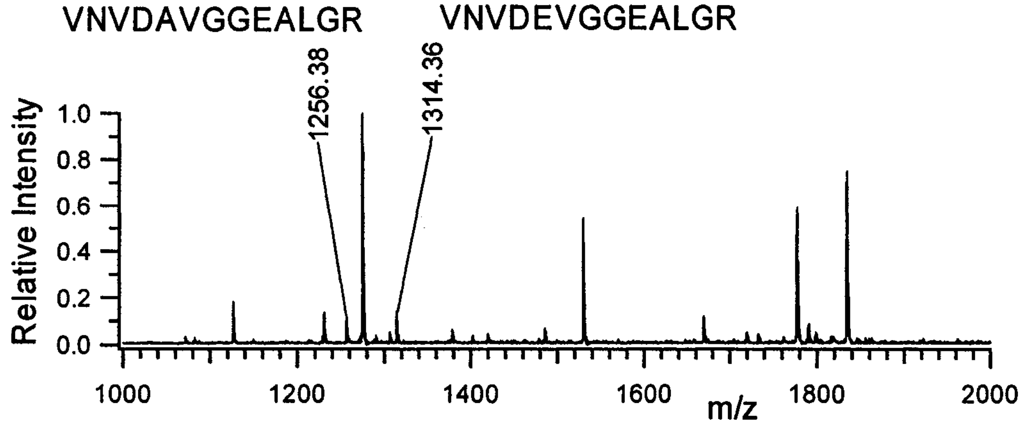


Figure 3.5. MALDI-TOF MS of tryptic digest of Hb protein fraction.

Next, the two peptide ions were analyzed by MALDI QqTOF to obtain the MS/MS spectra show in Figure 3.6(a). The MS/MS result of the normal β-chain peptide 18-30 can be searched against the protein database by MASCOT. As shown in Figure 3.6(b), the high matching score was obtained. The variant one cannot be searched against an unmodified database. Therefore, the b and y ions are both labeled in the MS/MS spectrum for *de novo* sequencing as given in Figure 3.6(a). The yl to y7 ions in both CID spectra are identical, which indicates there is no mutation from yl to y7 (-GGEALGR). As well, the b2 to b4 ions are the same, so there is no mutation from b1 to b4 (VNVD-). The peptide sequence deduced is VNVD_ _GGEALGR, indicating that the mutation happens at b5/y9 or b6/y8. In fact, there is no b5 ion (m/z= 557 Da) of normal peptide in the mass spectrum of variant peptide while a peak of 557 Da minus 58 Da (499 Da) was observed. This indicates the amino acid variant position happened at b5. Subsequently, there is a mass shift of 58 Da observed for the b6 and y9 ions. This result confirmed that the 22Glu of the Hb β-chain protein was substituted to become 22Ala.

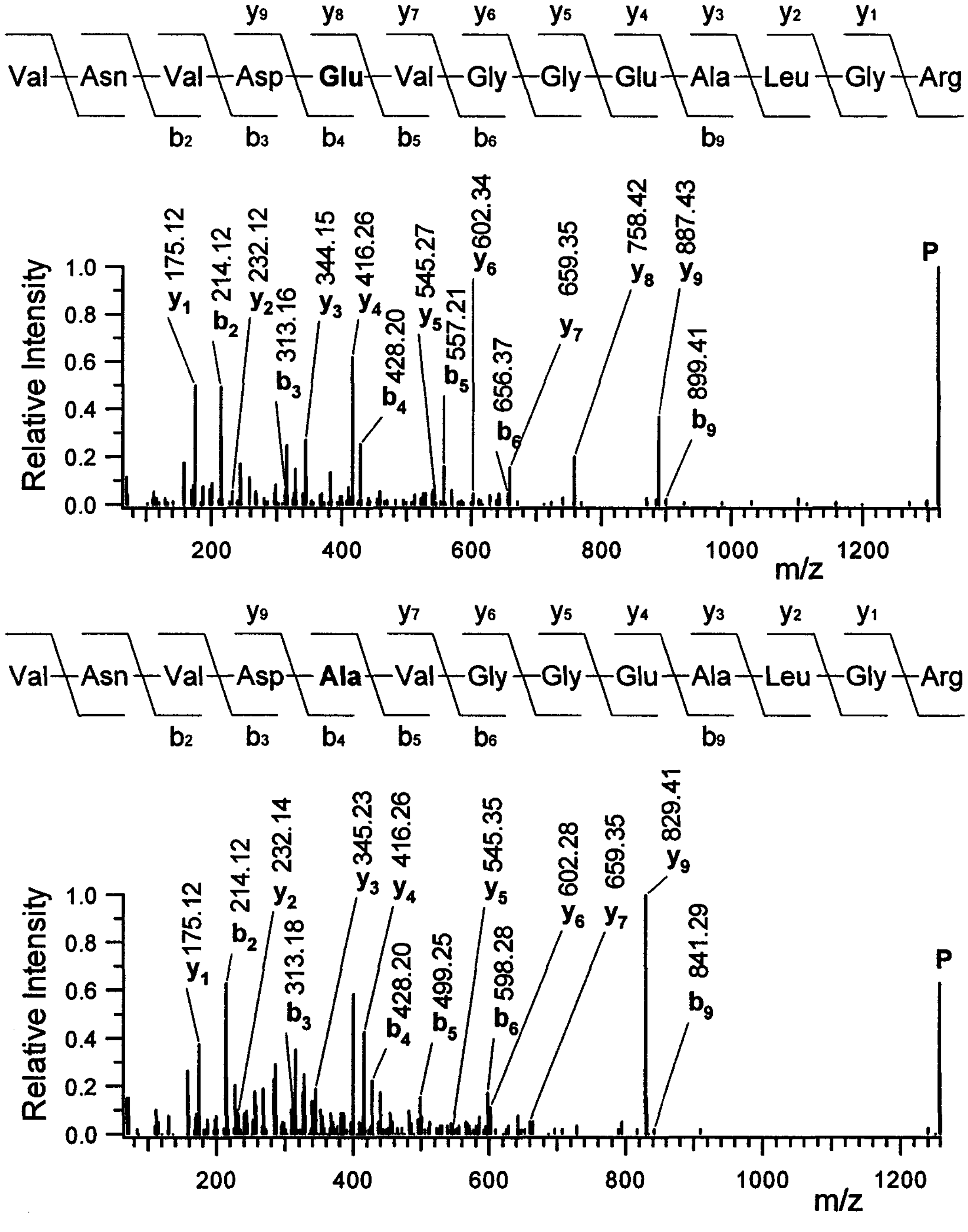


Figure 3.6(a). CID MS/MS of normal (above) and variant (bottom) peptide to determine the variant position.

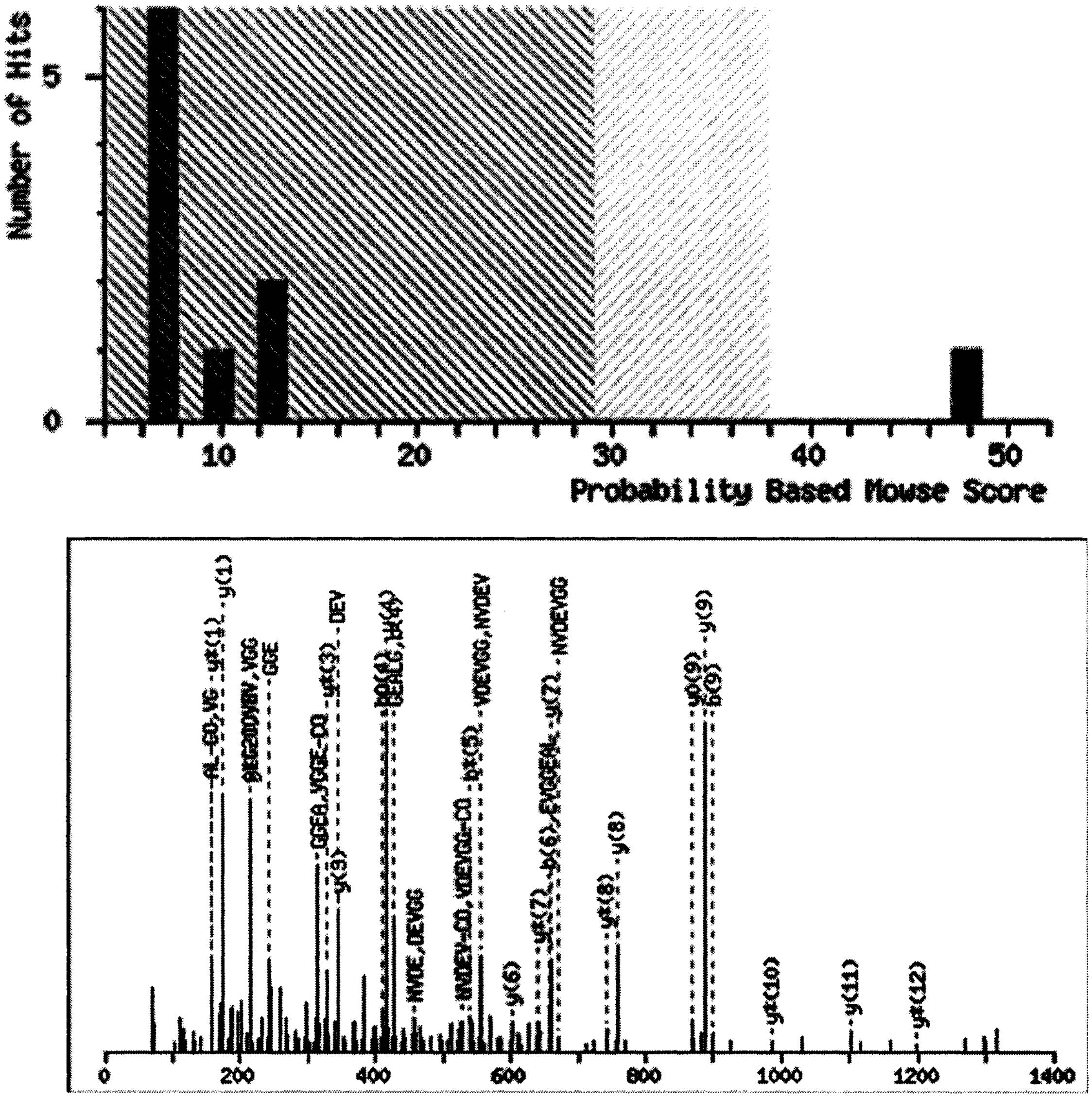


Figure 3.6(b). MASCOT search result of normal peptide VNVDEVGGEALGR.

The score above 38 indicates positive identity.

3.4 Conclusions

In this work, variant hemoglobin G-Coushatta was characterized by two mass spectrometric approaches. In the MAP sequencing method, high mass accuracy and good spectral resolution by MALDI-TOF MS made possible to detect the mass difference of 58 Da between the normal and the variant hemoglobin. This method provides a simple approach for direct diagnostic analysis. Compared with traditional MS analysis methods, there is no need to remove the heme group from the globin chain. The acid digestion process takes only a couple of minutes, and time-consuming peptide separation is not necessary for the mass spectrometric analysis. All the data are acquired in a single spectrum, so data processing can be done quickly and easily. Overall, the process is very simple and fast. The obtained mass accuracy and resolution are good enough for exact amino acid identification, which make it suitable for amino acid mutation study such as variant hemoglobin. Due to the simplicity of the experiment, the amount of sample consumed is very small. In practice, 1 μL blood sample (after cell extraction) is sufficient. Compared with traditional methods of DNA sequencing, this method provides an inexpensive and time-saving proteomic analysis for the amino acid mutation identification. However, it should be pointed out that the experimental mass range for high resolution and mass accuracy analysis is limited in MALDI-TOF. With the employment of a more advanced mass spectrometry, such as FTICR MS [14], the coverage of the amino acid sequence could be improved and variants appearing over a wider mass region could be identified.

Finally, conventional tryptic digestion of the hemoglobin combined with *de novo* sequencing, although it is more complicated and takes longer experimental time than the

MAP technique, was also carried out and the results obtained confirmed the conclusion obtained by the MAP sequencing method.

3.5 Literature Cited

(1) Matsuo, T.; Matsuda, H.; Katakuse, I.; Wada, Y.; Fujita, T. *Biomed. Mass Spectrom.* **1981**, *8*, 25-30.

(2) Wild, B. J.; Green, B.; Cooper, E.K.; Lalloz, M.R.; Erten, S.; Stephens, A.D.; Layton, D.M. *Blood Cells Mol Dis.* **2001**, *27*, 691.

(3) Caruso, D.; Luca, D.R.; Giavarini, F.; Giovanni, G.; Simona, B.; Paola, L.; Carlo, F. *Hemoglobin* **2002**, *26*, 197-199.

(4) Lacan, P.; Michel, B.; Zanella-Cleon, I.; Martine, A. *Hemoglobin* **2004**, *28*, 205.

(5) Shimizu, A.; Toyofumi, N.; Kishikawa, M.; Miyazaki, A. *Jouunal of Chromatography B* **2002**, *776*, 15.

(6) Lacan, P.; Mathieu, M.; Becchi, M.; Zanella-Cleon, I. *Hemoglobin* **2005**, *29*, 69.

(7) McComb, M.E.; Chow, A.; Ens, W.; Standing, K.G.; Perreault, H.; Smith, M. *Anal. Chem.* **1998**, *70*, 5142.

(8) Reynolds, T.M.; Tim, C.H.; Green, B.N.; Smith, A.; Hartland, A. J. *Clin. Chem.* **2002**, *48*, 2261.

(9) Gatlin, C.L.; Cross, S.T.; Detter, J.C.; Yates III, J.R. *Anal. Chem.* **2000**, *72*, 757.

(10) Caruso, D.; Crestani, M.; Mitro, N.; Da Riva, L.; Mozzi, R.; Sarpau, S.; Merlotti, C.; Franzini, C. *Clin Lab Haematol.* **2005**, *27*, 111.

(11) Nakanishi, T.; Miyazaki, A.; Shimizu, A.; Yamaguchi, A.; Nishimura, S. *Clin Chim Acta.* **2002**, *323*, 89.

(12) Zhong, H.; Zhang, Y.; Wen, Z.; Li, L. *Nature Biotechnology* **2004**, *22*, 1291.

(13) Houston, C.T.; Reilly, J.P. *Rapid Comm. Mass Spectrom.* **1997**, *11*, 1435.

(14) Bogdan, B.; Smith, R.D. *Mass Spectrometry Reviews* **2005**, *24*, 168.

Chapter 4

Creation and Application of Human Metabolome MS/MS Database

4.1 Introduction

In this post-genomic era, increasing efforts have been put into understanding the relationship between the genome and the phenotype in cells and organisms. Functional studies have focused on analyses at the level of the gene (transcriptomics), proteins (proteomics) and metabolic networks (metabolomics), with a view to a "systems biology" approach to defining the phenotype. As a downstream result of gene expression, metabolomics is complementary to transcriptomics and proteomics, being referred to as the third cornerstone. The metabolome is further down the line from gene to function and connects many different pathways that operate within a living cell, so it reflects more closely the activities of the cell.

Metabolomics, as described by Fiehn [1], is the comprehensive analysis of the whole metabolome including all classes of compounds under a given set of conditions. Its ultimate goal is to identify and quantify every metabolite in a biological system. So far, it is not feasible to use one single technology to provide full detection of all species present in a metabolome, but relevant metabolite profiling using NMR spectroscopy, FT-IR or mass spectrometry (MS) is underway [11, 25-29]. MS-based metabolomics analysis has been used most widely due to its high sensitivity to identify metabolites in a complex sample. In general, mass spectrometric studies on metabolomics involve separation techniques, such as GC [2-7], LC [8, 9] or CE [10], combined with MS. Due to its high separation resolution and reproducibility, GC is combined with MS as a popular technology in plant metabolomics studies currently [2-7]. The combination of a retention time index (RI) obtained from a GC separation and MS information acquired from a mass spectrometer provides an index database of MS/RI for reference metabolites. Two criteria were used for metabolite identification: metabolite-specific fragmentation pattern and relative retention time. Commercially available software and tools, such as AMDIS, can be used for GC data processing and the results can be searched against commercially available GC/MS databases, such as NIST/EPA/NIH.

However, GC-MS has mass limitation (i.e., high molecule weight compounds cannot be analyzed) and is difficult to detect non-volatile, thermo-labile metabolites directly which are important classes of metabolites. The detection of these metabolites usually involves derivatization before separation. Thus the identification of unknown compounds is difficult because they are chemically modified. In addition, data processing is time-consuming and labor-intensive.

As a complementary approach, LC-MS methods are also being used for metabolome analysis [8, 9, 16, 30]. Without the need of carrying out chemical derivatization, sample preparation is simplified, and a wide range of species, including non-volatile and thermolabile metabolites, can be detected directly. LC-MS profiling relies on comparisons with reference compounds. In most cases, it can be used to identify the class of compounds to which the metabolite belongs, but not its exact identity. Further identification techniques should be employed subsequently. Using high accuracy mass spectrometers, metabolite identification can be performed by calculation of definitive molecular formulae. However, for studies of very complex metabolite mixtures, peak overlapping and analyte co-elution in LC are apparent. Thus many analytes may not be detected. In addition, not all the ions detected represent individual compounds because some may result from in-source fragmentation or adduct formation during the MS detection. As a result, data processing can be quite extensive. And the inability to differentiate isomeric configurations, which is of vast importance in biological studies, is another major bottleneck in metabolomics analysis of complex biological systems. Therefore, metabolomic research should include approaches to elucidate chemical structure information as well. MS/MS can be very useful to generate structure information derived from mass spectral fragmentation patterns and chemical databases. Even for isomers that have the same chemical formula but different structures, different fragmentation pathways resulting in different product ions with different intensities can differentiate these isomers.

Early metabolomic studies using LC combined with tandem MS were limited to targeted analysis of metabolites [12-15]. Because product ions obtained from MS/MS can provide structure information, some work on the construction of MS/MS libraries from different mass spectrometers [17-21] and their independency have been reported, and these libraries have been applied to the identification of drug and drug metabolites or pesticides [22-24]. The application of LC combined with MS/MS database to non-targeted metabolite profiling has not been used widely.

In this work, construction of a triple-energy ESI-MS/MS spectral library of human metabolome was initiated. Three sets of MS/MS spectra using three different ion fragmentation energies in negative or positive ion detection mode were generated from each known human metabolite that can be found commercially. This triple-energy MS/MS library provides more complete information on analyte fragmentation, compared to that using one-energy to generate MS/MS spectra. A spectral reproducibility study was undertaken for ten selected metabolites. The concentration limit of detection was determined and it was found that metabolite concentration can be as low as 1 μM to produce good quality MS/MS spectra. Finally, this library was applied to the analysis of human urinary metabolites and 15 metabolites were identified successfully, confirming the matrix independence of this database for real world applications.

4.2 Experimental

4.2.1 Reagents and Materials

Deionized water from a Millipore NANOpure water system was used in all cases. HPLC grade acetonitrile (ACN), methanol (MeOH) and aqueous ammonia (NH_4OH) were purchased from Fisher Scientific Canada. Trifluoroacetic acid (TFA), acetic acid (HAc), formic acid (FA) and ammonium acetate (NH_4Ac) were ordered from Sigma Aldrich. All metabolites were commercially purchased from various sources and made available to us by Dr. David Wishart, Department of Biology Sciences, University of Alberta.

4.2.2 Sample Preparation

Typically, for MS/MS spectral collection by flow injection analysis (FIA), compounds were made up at 1 mM in 80/20 MeOH/H_2O for negative mode or 0.1% HAc in 50/50 MeOH/H_2O for positive mode. For the reproducibility study on the solvent system, in negative mode, Monoethyl Glutarate, Pantothenic Acid, 4-Hydroxy-3-Methoxybenzoic Acid, Aconitic Acid and Glutaconic Acid were selected and made up at 1 mM in various solvent systems of 20%MeOH/water, 50%MeOH/water, 20%ACN/water, 50%ACN/water, 80%ACN/water, 10 mM NH_4OH in 50/50 MeOH/water or 1% NH_4OH in 50/50 MeOH/water. In positive mode, Aspartic Acid, Histidine, Homoarginine, Ornithine and β-Aminoisobutyric Acid were selected and made up at 1 mM in different solvent systems of 0.1%HAc in 20%MeOH/water, 80%MeOH/water, 20%ACN/water, 50%ACN/water, 80%ACN/water, 0.1% TFA in 50/50 MeOH/water or 0.1% FA in 50/50 MeOH/water. For concentration limit of detection (LOD) analysis, concentrations of a mixture solution of Gentisic acid, Hippuric acid, Atrolacetic acid and Salicylic acid (negative mode), and a mixture of Methionine, Tyrosine, Phenylalanine and Tryptophan (positive mode) were both made at 1 mM, and were diluted to final concentrations of 0.1 mM, 0.01 mM and 1 μM for LC separation.

In urine analysis, human urine samples from a healthy donor were centrifuged at 30,000 rpm for 30 min at room temperature, and proteins were removed by passing through a 3000 Da cut-off membrane (Micron), lyophilized overnight to complete dryness and dissolved to two times their original concentration in water.

4.2.3 Instrumentation

Flow Injection Analysis (FIA) and LC separation were performed on an Agilent 1100 HPLC system equipped with a 100-well auto-sampler. Electrospray mass spectrometry was performed on a Waters Micromass Quattro TripleQ (Waters Micromass, UK) equipped with an electrospray ionization source.

For FIA with MS/MS spectral collection, samples were introduced into the mass spectrometer without splitting by automated injection of 50 μL at a flow rate of 50 μL/min. Each sample was analyzed twice at each of three levels of collision-induced dissociation (CID) energy applied: 10 eV, 25 eV and 40 eV. Instrument parameters for the database construction and urinary analysis are given in Table 4.1. All raw data were acquired in centroid mode with MassLynx 3.5 software (Waters) and re-processed using MS Manager 7.0 (ACD, Toronto, ON, Canada). Data smoothing and baseline correction were performed first, then a threshold to obtain noise reduction was introduced, in which typically S/N = 10, 5, 3 were used for low, medium and high CID energy spectra, respectively.

Table 4.1. Summary of Instrument Parameters.

Capillary Voltage	4kV (+) / 2.8kV (-)
Cone Voltage	25V
Desolvation temperature	300 °C
Source Temperature	120 °C
Pressure	1.0E-3
CID energies	10eV / 25eV / 40eV

For the reproducibility study, selected metabolites at 1 mM were run under various conditions: flow rates of 200 μL / min, 100 μL/min and 30 μL/min; different organic modifiers of the mobile phase ACN; different organic modifier percentages of 20% or 50% of MeOH, or 20%, 50% or 80% ACN in the mobile phase in negative mode, and 20% or 80% of MeOH, or 20%, 50% or 80% ACN in the mobile phase in positive mode; different additives in the mobile phase of 1% NH_4OH or 10mM NH_4Ac in negative mode, and 0.1% TFA or 0.1% FA in positive mode.

For the LC-MS analysis, separation was performed on a Zorbax C_{18} 4.6 x 150 mm column with 5 μm particle size, the flow rate was 1 mL/min, the effluent from the

HPLC system was directed with splitting into the source at 50 μL/min. The gradient used in the LOD analysis is shown in Table 4.2.

Table 4.2. Gradient used in LOD analysis.

	Time (min)	%B
Positive Mode	0	0
	2	0
	22	50
	25	50
	30	0
	55	0

Mobile phase A was 100% water and mobile phase B was 100%MeOH.

	Time (min)	%B
Negative Mode	0	0
	20	40
	24	60
	25	0
	30	0

Mobile phase A was 20mMHAc/15%MeOH/water (PH=3.7) and mobile phase B was 100%MeOH.

The human urine sample was directly injected into the LC-MS system for analysis without specific analyte extraction. Urine separation combined with negative ion mode MS detection was performed on an Agilent Zorbax C_{18} column 4.6 x 150 mm at a flow rate of 1 mL/min, and a splitter was used to direct 50 μL/min into the mass spectrometer. Urine separation with MS at the positive ion mode was performed on a Primesep 2.1 x 150 mm column (SIELC) at a flow rate of 0.2 mL/min. The gradient used in urinary analysis is shown in Table 4.3.

Table 4.3. Gradient used in urinary analysis.

	Time (min)	%B
Positive Mode	0	0
	15	0
	30	15
	45	45
	50	100
	70	100
	Mobile phase A was 0.03% TFA in 10%ACN/water and mobile phase B was 0.5%TFA in 40%ACN/water.	
	Time (min)	%B
Negative Mode	0	0
	20	20
	30	40
	35	60
	40	90
	Mobile phase A was 20mMHAc/15%MeOH/water (PH=3.7) and mobile phase B was 100%MeOH.	

4.3 Results and Discussion

4.3.1 Library Creation

The ultimate goal of this study is to construct an MS/MS spectra library for a wide range of human metabolites which can be used to identify unknown metabolites in real world samples by MS/MS spectral matching. There are several criteria to consider in constructing the library. A standard set of conditions should be followed for acquiring MS/MS spectra of a wide range of compounds. The spectra generated should be reproducible at different sample conditions such as flow rate and solvent system used for sample introduction to the mass spectrometer. They should also be independent of metabolite concentrations. Finally, the spectral quality should be sufficient to differentiate between closely related metabolites.

In this work, the product ion spectra of 215 human metabolites were collected, including amino acids, carboxylic acids, amines, sugars, steroids, nucleotides, and vitamins. Based on their chemical properties, they are divided into two groups and analyzed in either negative ion mode or positive ion mode. The names, molecular masses, chemical formula, and ionization mode of the analyzed metabolites are listed in Table 4.4.

Table 4.4. List of metabolites analyzed for the construction of MS/MS human metabolomics database.

Metabolite Name	Molecular Weight	Formula	Ionization Mode
Betaine	117.1	C5H11NO2	Positive
Sarcosine	89.1	C3H7NO2	Positive
Taurine	125.1	C2H7NO3S	Positive
Inosine	268.2	C10H12N4O5	Positive
Pyridoxine	169.2	C8H11NO3	Positive
Ascorbic Acid	176.1	C6H8O6	Positive
Adenosine	267.2	C10H13N5O4	Positive
Myo-Inositol	180.2	C6H12O6	Positive
Adenine	135.1	C5H5N5	Positive
Creatinine	113.1	C4H7N3O	Positive
Tyramine	137.2	C8H11NO	Positive
Thymine	126.1	C5H6N2O2	Positive
Uracil	112.1	C4H4N2O2	Positive
Xanthosine	284.2	C10H12N4O6	Positive
Uric Acid	168.1	C5H4N4O3	Positive
Uridine	244.2	C9H12N2O6	Positive
Galactitol	182.2	C6H14O6	Positive
Xanthine	152.1	C5H4N4O2	Positive
Allantoin	158.1	C4H6N4O3	Positive
3-Methylindole	131.2	C9H9N	Positive
Niacinamide	122.1	C6H6N2O	Positive
Adonitol	152.1	C5H12O5	Positive
Epinephrine	183.2	C9H13NO3	Positive
Riboflavin	376.4	C17H20N4O6	Positive
Lactose	372.3	C12H22O11	Positive
Sorbitol	182.2	C6H14O6	Positive
Sucrose	342.3	C12H22O11	Positive
Glucono 1,5-lactone	178.1	C6H10O6	Positive
D-Melibiose	342.3	C12H22O11	Positive
Cytosine	111.1	C4H5N3O	Positive
Guanosine	283.2	C10H13N5O5	Positive
Thiamine	337.3	C12H17N4OS•Cl	Positive
Estradiol	272.4	C18H24O2	Positive
Thymidine	243.2	C10H14N2O5	Positive

Hypoxanthine	136.1	C5H4N4O	Positive
Nimbosterol	414.7	C29H50O	Positive
Butanone	72.1	C4H8O	Positive
N-Acetyl-b-Glucosamine	221.2	C8H15NO6	Positive
Deoxyguanosine	267.2	C10H13N5O4	Positive
Xanthurenic Acid	205.2	C10H7NO4	Positive
Indole	117.2	C8H7N	Positive
2-Methylacetoacetate	116.1	C5H8O3	Positive
2'-Deoxyuridine	228.2	C9H12N2O5	Positive
3-Methoxytyramine	167.2	C9H13NO2	Positive
17-Hydroxyprogesterone	330.5	C21H30O3	Positive
2'-Deoxycytidine	227.2	C9H13N3O4	Positive
6-Dimethylaminopyrine	163.2	C7H9N5	Positive
1,3-Diaminopropane	74.1	C3H10N2	Positive
1,6-Anhydro-beta-glucopyranose	162.1	C6H10O5	Positive
Histamine	111.2	C5H9N3	Positive
Glyceraldehyde	90.1	C3H6O3	Positive
Pyrocatechol	110.1	C6H6O2	Positive
Pyridine	79.1	C5H5N	Positive
5a-Androstane-3,17-dione	288.4	C19H28O2	Positive
4-Methylcatechol	124.1	C7H8O2	Positive
Testosterone	288.4	C19H28O2	Positive
Serotonin	176.2	C10H12N2O	Positive
6-Hydroxydopamine	169.2	C8H11NO3	Positive
16-Dehydroprogesterone	312.5	C21H28O2	Positive
Corticosterone	346.5	C21H30O4	Positive
Dopa	197.2	C9H11NO4	Positive
Dopamine	153.2	C8H11NO2	Positive
Tryptamine	160.2	C10H12N2	Positive
Dehydroepiandrosterone	288.4	C19H28O2	Positive
Cortisol	362.5	C21H30O5	Positive
7-Dehydrocholesterol	384.6	C27H44O	Positive
Ergosterol	396.7	C28H44O	Positive
5-Methyluridine	258.2	C10H14N2O6	Positive
Vitamin A	286.5	C20H30O	Positive
Putrescine	88.2	C4H12N2	Positive
7-Methylguanosine	297.3	C11H16N5O5	Positive
Octaldehyde	128.2	C8H16O	Positive
Pterin	163.1	C6H5N5O	Positive
Pyridoxamine	168.2	C8H12N2O2	Positive

Purine	120.1	C5H4N4	Positive
Methylguanidine	73.1	C2H7N3	Positive
Neopterin	253.2	C9H11N5O4	Positive
Nicotinic Acid	123	C6H5NO2	Positive
Arabitol	152.2	C5H12O5	Positive
Normetanephrine	183.2	C9H13NO3	Positive
5-Hydroxymethyluracil	142.1	C5H6N2O3	Positive
Trimethylamine Oxide	75.1	C3H9NO	Positive
Mannitol	182.2	C6H14O6	Positive
Quinolinic Acid	167.1	C7H5NO4	Positive
Adrenaline	183.2	C9H13NO3	Positive
Dihydrouracil	114.1	C4H6N2O2	Positive
Biopterin	237.2	C9H11N5O3	Positive
5-Hydroxymethyl-4-methyluracil	156.1	C6H8N2O3	Positive
Cytidine	243.2	C9H13N3O5	Positive
Dihydrothymine	128.1	C5H8N2O2	Positive
Pyridoxal	167.2	C8H9NO3	Positive
Carnitine	161.2	C7H15NO3	Positive
Acetylcarnitine	203.2	C9H17NO4	Positive
Deoxycholic Acid glycine conjugate	449.6	C26H43NO5	Positive
Pantothenic Acid	219.2	C9H17NO5	Positive
Porphobilinogen	226.2	C10H14N2O4	Positive
Glutamine	146.1	C5H10N2O3	Positive
Methionine	149.2	C5H11NO2	Positive
Ornithine	132.2	C5H12N2O2	Positive
Creatine	131.1	C4H9N3O2	Positive
b-Aminoisobutyric Aicd	103.1	C4H9NO2	Positive
Leucine	131.2	C6H13NO2	Positive
Valine	117.1	C5H11NO2	Positive
Glutamic Acid	147.1	C5H9NO4	Positive
Homoarginine	188.2	C7H16N4O2	Positive
Pipecolic Acid	129.2	C6H11NO2	Positive
Serine	105.1	C3H7NO3	Positive
Tyrosine	181.2	C9H11NO3	Positive
Phenylalanine	165.2	C9H11NO2	Positive
Cystine	240.3	C6H12N2O4S2	Positive
Proline	115.1	C5H9NO2	Positive
Homoserine	119.1	C4H9NO3	Positive
Asparagine	132.1	C4H8N2O3	Positive

Aspartic Acid	133.1	C4H7NO4	Positive
Histidine	155.2	C6H9N3O2	Positive
Alanine	89.1	C3H7NO2	Positive
Glycine	75.1	C2H5NO2	Positive
Threonine	119.1	C4H9NO3	Positive
Tryptophan	204.2	C11H12N2O2	Positive
N-Acetyl-alanine	131.1	C5H9NO3	Positive
gamma-Aminobutyric Acid	103.1	C4H9NO2	Positive
Isoleucine	131.2	C6H13NO2	Positive
Aminoadipic Acid	161.2	C6H11NO4	Positive
N-Carbamoyl-Aspartic Acid	176.1	C5H8N2O5	Positive
Alanyl-Norvaline	188.2	C8H16N2O3	Positive
Norleucine	131.2	C6H13NO2	Positive
Prolyl-Isoleucine	228.3	C11H10N2O3	Positive
beta-3,4-Dihydroxy-phenylalanine	197.2	C9H11NO4	Positive
Norvaline	117.1	C5H11NO2	Positive
Alanyl-Valine	188.2	C8H16N2O3	Positive
Citrulline	175.2	C6H13N3O3	Positive
Kynurenine	208.2	C10H12N2O3	Positive
3-Hydroxy-Kynurenine	224.2	C10H12N2O4	Positive
Hydroxy-Proline	131	C5H9NO3	Positive
Arginine	174.2	C6H14N4O2	Positive
Lysine	146.2	C6H14N2O2	Positive
Kynurenic Acid	189.2	C10H7NO3	Positive
Aspartyl-Phenylalanine	280.3	C13H16N2O5	Positive
Homocarnosine	240.3	C10H16N4O3	Positive
Acetylglycine	117.1	C4H7NO3	Positive
Glycyl-leucine	188.2	C8H16N2O3	Positive
3-Methylhistidine	169.2	C7H11N3O2	Positive
Indolelactic Acid	205.2	C11H11NO3	Positive
3-Iodo-Tyrosine	307.1	C9H10INO3	Positive
Alloisoleucine	131.2	C6H13NO2	Positive
Hydroxyindoleacetic Acid	191.2	C10H9NO3	Positive
2-Amino-n-Butyric Acid	103.1	C4H9NO2	Positive
Homocitrulline	189.2	C7H15N3O3	Positive
Cysteine	121.2	C3H7NO2S	Positive
p-Aminobenzoic Acid	137.1	C7H7NO2	Positive
B-Alanine	89.1	C3H7NO2	Positive
Carnosine	226.2	C9H14N4O3	Positive
Cystathionine	222.3	C7H14N2O4S	Positive

2-Aminooctanoic Acid	159	C8H17NO2	Positive
m-Aminobenzoic Acid	137	C7H7NO2	Positive
Acetylcysteine	163.2	C5H9NO3S	Positive
2-Furoylglycine	169.1	C7H7NO4	Positive
Pyroglutamic Acid	129.1	C5H7NO3	Positive
Guanidoacetic Acid	117.1	C3H7N3O2	Positive
N-acetylneuraminic Acid	309.3	C11H19NO9	Positive
Allocystathionine	222.3	C7H14N2O4S	Positive
3-Ureidopropionic Acid	132.1	C4H8N2O3	Positive
Salicyluric Acid	195.2	C9H9NO4	Positive
N6-Acetyl-lysine	188.2	C8H16N2O3	Positive
a-Aminobutyric Acid	103.1	C4H9NO2	Positive
N-Acetylaspartic Acid	175.1	C6H9NO5	Positive
4,5-Dihydroorotic Acid	158.1	C5H6N2O4	Positive
Homocysteine	135.2	C4H9NO2S	Positive
N-Acetyltyrosine	223.2	C11H13NO4	Positive
Glutathione	307.3	C10H17N3O6S	Positive
Indoleacetic Acid	175.2	C10H9NO2	Positive
Urocanic Acid	138.1	C6H6N2O2	Positive

Glutaconic Acid	130.1	C5H6O4	Negative
Fumaric Acid	116.1	C4H4O4	Negative
Mandelic Acid	152.1	C8H8O3	Negative
4-Hydroxybenzoic Acid	138.1	C7H6O3	Negative
Butyric Acid	88.1	C4H8O2	Negative
Hydroxyphenylpyruvic Acid	180.2	C9H8O4	Negative
Pimelic Acid	160.2	C7H12O4	Negative
Glyceric Acid	106.1	C3H6O4	Negative
Glyoxylic Acid	74	C2H2O3	Negative
Suberic Acid	174.2	C8H14O4	Negative
Undecanoic Acid	186.3	C11H22O2	Negative
Maleic Acid	116.1	C4H4O4	Negative
Cholic Acid	408.6	C24H40O5	Negative
Caproic Acid	116.2	C6H12O2	Negative
Tetradecanedioic Acid	258.4	C14H26O4	Negative
Deoxycholic Acid	392.6	C24H40O4	Negative
4-Hydroxy-3-Methoxybenzoic Acid	168.1	C8H8O4	Negative
Palmitic Acid	256.4	C16H32O2	Negative

Aconitic Acid	174.1	C6H6O6	Negative
trans-2-Octenoic Acid	142.2	C8H14O2	Negative
Monomethyl glutarate	146.1	C6H10O4	Negative
Glycolic Acid	76.1	C2H4O3	Negative
Sebacic Acid	202.2	C10H18O4	Negative
Phenylacetic Acid	136.2	C8H8O2	Negative
Pyromucic Acid	112.1	C5H4O3	Negative
Gentisic Acid	154.1	C7H6O4	Negative
Glutaric Acid	132.1	C5H8O4	Negative
Nonanedioic Acid	188.2	C9H16O4	Negative
Levulinic Acid	116.1	C5H8O3	Negative
Succinic Acid	118.1	C4H6O4	Negative
Citric Acid	192.1	C6H8O7	Negative
Orotic Acid	156.1	C5H4N2O4	Negative
Oxoglutaric Acid	146.1	C5H6O5	Negative
Glycocholic Acid	465.6	C26H43NO6	Negative
Malic Acid	134.1	C4H6O5	Negative
Myristic Acid	228.4	C14H28O2	Negative
Adipic Acid	146.1	C6H10O4	Negative
3-Hydroxybutyric Acid	104.1	C4H8O3	Negative
Hippuric Acid	179.2	C9H9NO3	Negative
Atrolactic Acid	166.2	C9H10O3	Negative
Benzoic Acid	122.1	C7H6O2	Negative
Salicylic Acid	138.1	C7H6O3	Negative
Pantothenic Acid	219.2	C9H17NO5	Negative

Because ion fragmentation is strongly dependent on the collision-induced dissociation energy, for each metabolite, three levels of CID energy, i.e., 10 eV, 25 eV and 40 eV, were used for MS/MS spectral generation. This triple-energy database has several advantages over that created using one fixed energy or ramped energy. Most importantly, this database can increase the confidence in analyte identification. A positive identification is made when high spectral matching scores are produced for two of the three levels. While structural information can be obtained at high CID energies, molecular weight information is usually obtained at low energies. In general, the fragmentation patterns obtained at different CID energies are quite different. Low abundance fragment ions are more easily found, when three energies are used for fragmentation. In most cases, it is found that MS/MS spectra obtained at low and medium CID energies can provide enough fragmentation information for good spectral matching.

4.3.2 Investigation of reproducibility parameters

Investigation of the dependence of spectral reproducibility on flow rate and solvent systems determines if different practical LC conditions have an effect on the utility of the created MS/MS library. In this work, ten metabolites were run in positive ion or negative ion MS/MS mode. For these experiments only a single parameter was changed without varying the other parameters. Spectra obtained under various conditions were compared with the library spectra.

In practice, assessing the comparison results in terms of matches of masses and intensities is a challenge. Thus three parameters, Fit, RFit and Purity were chosen as a measure of spectral agreement. Fit score indicates how well the masses and intensities of the library spectrum agree with those found in the acquired spectrum. RFit (Reverse Fit) score indicates how well the masses and intensities of the acquired spectrum agree with those found in the library spectrum. Purity score indicates how well the masses and intensities of the library spectrum and the acquired spectrum agree. They were calculated using an algorithm in which relative intensity of product ions and weighting factors were included. The detailed algorithm is given in the Appendix I. The maximum value for all three is 100%.

A list of different LC conditions including different flow rates and solvent systems used to verify the independence of the library are shown in Table 4.5. The obtained Fit/RFit/Purity values from these experiments are given in Table 4.6. In general, an influence of the eluent composition on the ion intensities is observable; however, the values of Fit/RFit/Purity are nearly independent of the buffer system. Therefore, problems are not expected in the cases of gradient elution, different mobile phase systems or different flow rates. It is worthwhile to mention that a higher content of water has a negative effect on the ionization process and results in ion intensity reduction; but it has little influence on the MS/MS spectral patterns. Different solvents can be used and, as it is well known, MeOH is preferred over ACN. Although different additives cause changes to the eluent pH and the cluster ion formation of [M+Ac$^-$] or [M+OH$^-$], they have little influence on the product ion spectra obtained.

Table 4.5. Different LC system used to verify independence of the database and their results in positive mode.

Positive Mode

	% H_2O	% ACN	% MeOH	Additive	Flow Rate (µL/min)
1	20	0	80	0.1% HAc	50
2	80	0	20	0.1% HAc	50
3	20	80	0	0.1% HAc	50
4	50	50	0	0.1% HAc	50
5	80	20	0	0.1% HAc	50
6	50	0	50	0.1% TFA	50
7	50	0	50	0.1% FA	50
8	50	0	50	0.1% HAc	30
9	50	0	50	0.1% HAc	100
10	50	0	50	0.1% HAc	200

	Homoargininine (40eV)			Aspartic Acid (10eV)			Ornithinine (10eV)			b-Aminoisobutyric Acid (25eV)			Histidine (25eV)		
	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
1	99.90	99.91	99.90	100.00	100.00	100.00	99.95	99.94	99.95	99.91	99.92	99.91	99.95	99.95	99.95
2	99.68	99.68	99.70	99.97	99.97	99.97	99.93	99.94	99.93	99.95	99.95	99.95	99.94	99.95	99.95
3	99.98	99.98	99.98	100.00	100.00	100.00	99.99	99.99	99.99	99.98	99.99	99.98	99.90	99.91	99.91
4	99.94	99.94	99.94	99.96	99.96	99.96	100.00	100.00	100.00	99.99	99.99	99.99	99.92	99.92	99.92
5	99.96	99.96	99.96	99.99	99.99	99.99	100.00	100.00	100.00	100.00	100.00	100.00	99.88	99.90	99.89
6	99.96	99.96	99.96	99.99	99.99	99.99	100.00	100.00	100.00	100.00	100.00	100.00	99.91	99.92	99.91
7	99.91	99.91	99.91	99.99	99.99	99.99	100.00	100.00	100.00	100.00	100.00	100.00	99.90	99.91	99.91
8	99.99	99.99	99.99	99.98	99.98	99.98	100.00	100.00	100.00	100.00	100.00	100.00	99.87	99.88	99.87
9	99.89	99.90	99.90	100.00	100.00	100.00	100.00	100.00	100.00	99.98	99.98	99.98	99.81	99.84	99.82
10	99.84	99.86	99.85	100.00	100.00	100.00	99.99	99.99	99.99	99.97	99.98	99.97	99.83	99.85	99.84

Table 4.6. Different LC system used to verify independence of the library and their results in negative mode.

Negative Mode

	% H2O	% ACN	% MeOH	Additive	Flow Rate (μL/min)
1	50	0	50	none	50
2	80	0	20	none	50
3	20	80	0	none	50
4	50	50	0	none	50
5	80	20	0	none	50
6	20	0	80	10mM NH_4Ac	50
7	20	0	80	1% NH_4OH	50
8	20	0	80	none	30
9	20	0	80	none	100
10	20	0	80	none	200

	4-Hydroxy- 3-Methoxybenzoic Acid (40eV)			Monoethyl Glutarate (10eV)			Glutaconic Acid (10eV)			Pantothenic Acid (25eV)			Aconitic Acid (25eV)		
	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
1	100	100	100	99.89	99.88	99.88	99.97	99.97	99.97	99.96	99.97	99.97	99.99	99.97	99.98
2	100	100	100	99.97	99.97	99.97	99.97	99.98	99.97	99.99	99.99	99.99	99.98	99.98	99.98
3	100	100	100	100	100	100	99.99	99.99	99.99	99.98	99.98	99.98	100	100	100
4	100	100	100	99.99	99.99	99.99	100	99.99	100	99.97	99.97	99.97	99.98	99.97	99.98
5	100	100	100	99.95	99.95	99.95	99.99	99.99	99.99	100	100	100	99.97	99.97	99.97
6	100	100	100	100	99.99	100	100	100	100	99.96	99.96	99.96	99.98	100	99.99
7	100	100	100	100	100	100	100	100	99.99	99.98	99.98	99.98	99.98	100	99.99
8	100	100	100	99.97	99.97	99.97	99.99	99.99	99.99	99.98	99.98	99.98	99.92	99.60	99.75
9	100	100	100	99.93	99.92	99.92	99.97	99.98	99.97	99.96	99.96	99.96	99.99	99.99	99.99
10	100	100	100	99.98	99.98	99.98	99.99	99.99	99.99	99.95	99.95	99.95	100	100	100

4.3.3 Limit of Detection Analysis

In real world samples, analyte concentrations are unknown and can vary greatly. The applicability of the MS/MS spectral library to a wide range of analyte concentrations was studied. In our work, the spectral library was built up using metabolite standards at a concentration of 1 mM, whereas the concentrations of many metabolites in the human body are much lower, e.g., at 1 μM. Different concentrations of metabolites at 0.1mM, 0.01 mM and 1 μM were prepared and analyzed. Figures 4.1 and 4.2 show the ion chromatograms obtained from LC-MS in positive mode and negative mode, respectively. Their resulting Fit, RFit and Purity values are listed in Tables 4.7 and 4.8. It is found that sample concentration plays a very important role in the metabolite identification. Lower concentrations possibly result in the disappearance of low abundance fragment ions. However, at the metabolite concentration of above 1 μM, good Fit/RFit/Purity scores were obtained, thus the product ion spectra generated could provide positive identification due to the appearance of characteristic product ions and their consistent distribution of relative intensities in comparison to the library spectra.

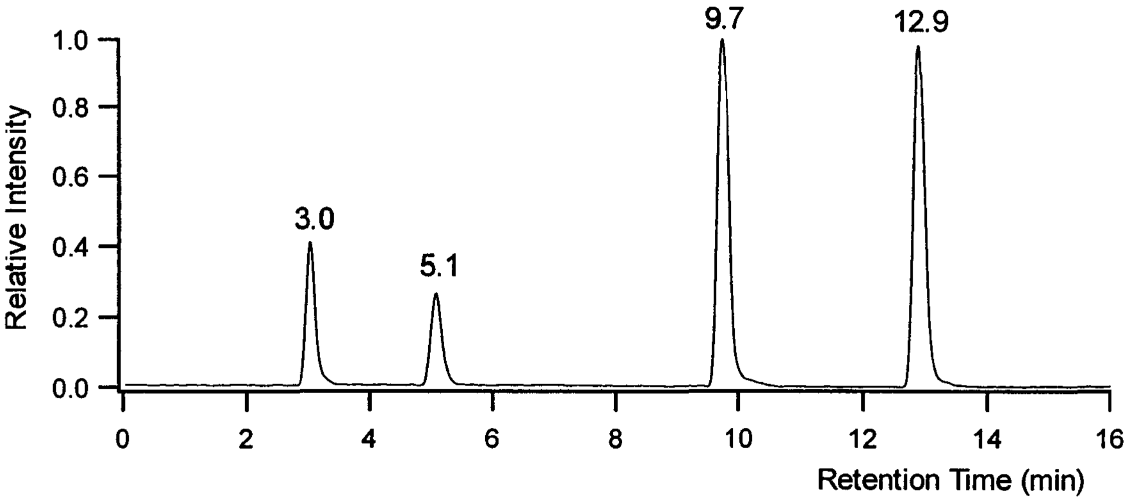


Figure 4.1. Ion chromatograph of metabolites separation in positive mode.

Table 4.7. Reproducibility investigation on sample concentration in positive mode.

Metabolite Name	CID energy	t_R (min)	MW (Da)	100μM			10μM			1μM		
				Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
Methionine	10 eV	3.0	149	99.89	99.86	99.87	99.75	99.43	99.59	99.68	98.31	98.93
Tyrosine	25eV	5.1	181	99.96	99.24	99.60	99.72	99.77	98.12	95.23	93.26	94.18
Phenylalanine	25eV	9.7	165	99.99	99.99	99.99	99.91	95.88	97.77	99.79	86.27	92.32
Typtophan	25eV	12.9	204	99.90	98.80	98.90	99.91	98.21	98.97	98.67	95.42	96.80

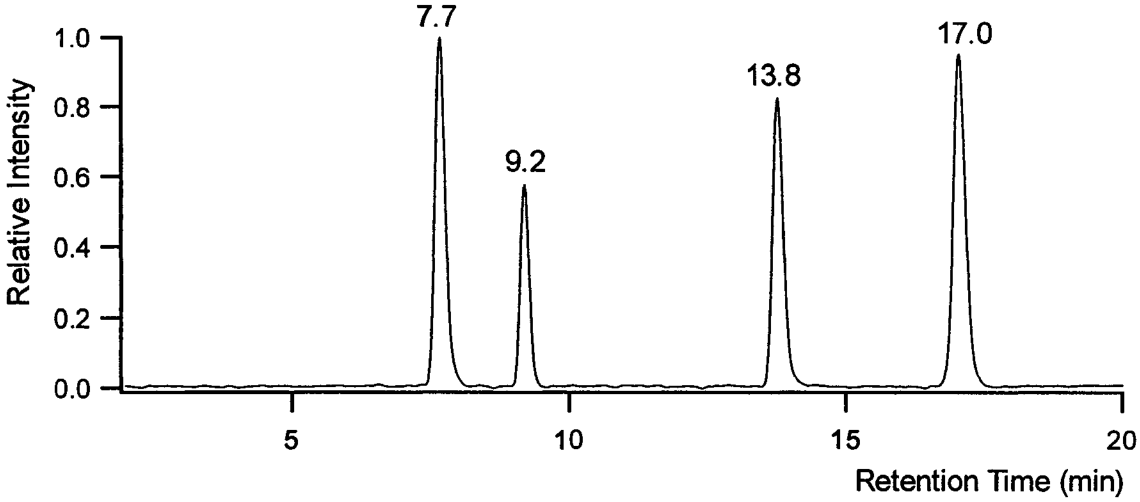


Figure 4.2. Ion chromatograph of metabolites separation in negative mode.

Table 4.8. Reproducibility investigation on sample concentration in negative mode.

Metabolite Name	CID energy	t_R (min)	MW (Da)	100μM			10μM			1μM		
				Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
Gentisic Acid	25 eV	7.7	154.1	99.99	99.99	99.99	99.98	99.80	99.88	99.89	99.85	99.87
Hippuric Acid	25eV	9.2	179.2	100.00	99.99	99.99	99.98	99.99	99.99	99.82	99.71	99.27
Atrolacetic Acid	25eV	13.8	166.2	100.00	100.00	100.00	100.00	100.00	100.00	99.96	99.96	99.83
Salicylic Acid	25eV	17.0	138.1	99.97	99.95	99.96	99.89	99.83	99.86	99.93	99.75	99.83

4.3.4 Human Urine Analysis

As an example of real-world applications of the created MS/MS spectral library, a preliminary analysis of human urinary metabolites was carried out. The MS/MS spectra generated in LC-MS analysis of a urine sample were compared to the library spectra to identify the metabolites in human urine.

In most analyses in metabolomics, specific sample preparation steps, such as solid-phase extraction (SPE), or liquid-liquid extraction (LLE) are involved, which have intrinsic biases for and against chemically different classes of metabolites. In this study, urine samples were directly analyzed without any further extraction procedure. The urinary metabolite profiling was run in both positive and negative ion mode. In this way, broader metabolome coverage can be obtained. Four organic acids were identified in the negative mode and eleven metabolites were identified in the positive mode. In this case, a two-step analysis scheme was used. First, full scan MS was conducted to obtain the molecular masses of metabolites and their retention time. Knowing the masses and retention times for metabolites, LC MS/MS was set up to select and fragment a particular mass of ions at a given retention time. Low and medium CID energies were used in two individual runs. The MS/MS spectra were searched against the library spectra.

Chromatographic separations of the urine sample combined with the positive or negative ion MS detection are shown in Figure 4.3 and Figure 4.15, respectively. The

MS/MS spectral searching results are shown in Tables 4.9 and 4.10. The acquired MS/MS spectra of the identified metabolites from urine and their corresponding library spectra are shown in Figures 4.4 - 4.14 and Figures 4.16 - 4.19, respectively. Finally, LC-MS of the standards of these identified metabolites or urine mixed with these metabolite standards was conducted under the same conditions to check their retention times (data not shown), which led to the confirmation of the identification results.

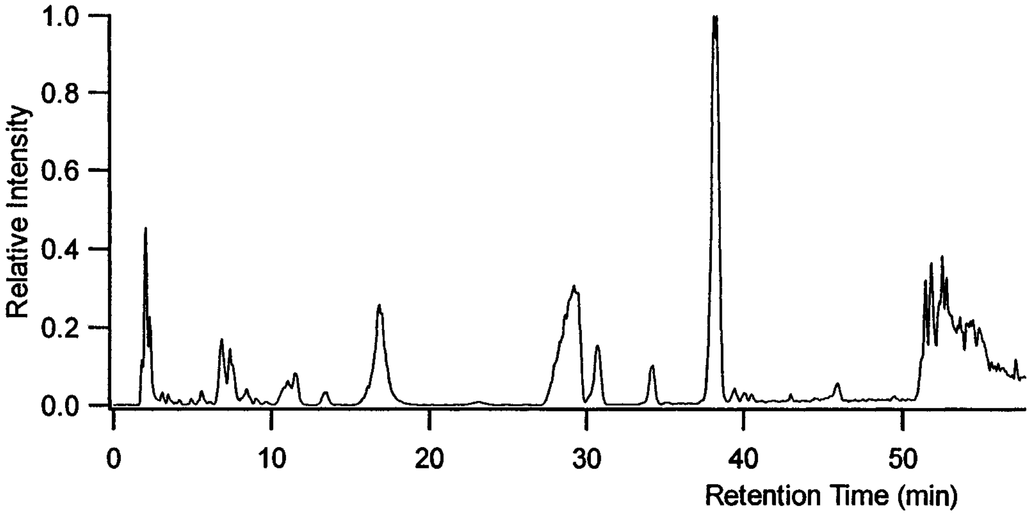


Figure 4.3. Ion chromatograph of urinary analysis in positive mode.

Table 4.9. Identified urinary metabolites in positive mode.

t_R (min)	m/z (Da)	Identified Metabolites	Low CID energy (10eV)			Medium CID energy (25eV)		
			Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
2.5	112.1	Uracil	99.93	99.73	99.82	99.73	99.64	99.68
3.5	152.1	Xanthine	99.50	99.45	99.47	99.75	96.72	98.08
9.1	136.1	Hypoxanthine	99.70	99.74	99.72	99.80	92.45	95.73
10.4	132.1	Asparginine	99.06	92.19	95.09	99.30	96.80	97.98
10.6	105.1	Serine	99.86	99.91	99.88	99.75	96.71	98.22
11	117.1	Betaine	100.00	100.00	100.00	100.00	99.95	99.98
11.5	119.1	Threonine	99.90	99.90	99.90	99.55	92.40	95.78
11.6	146.1	Glutamine	99.96	99.06	99.50	99.88	99.63	99.75
30.4	131.1	Creatine	99.94	99.95	99.94	99.86	99.89	99.88
38.3	113.0	Creatinine	99.74	99.75	99.75	99.99	99.98	99.99
51.2	155.2	Histidine	99.67	99.67	99.67	99.28	92.49	95.30

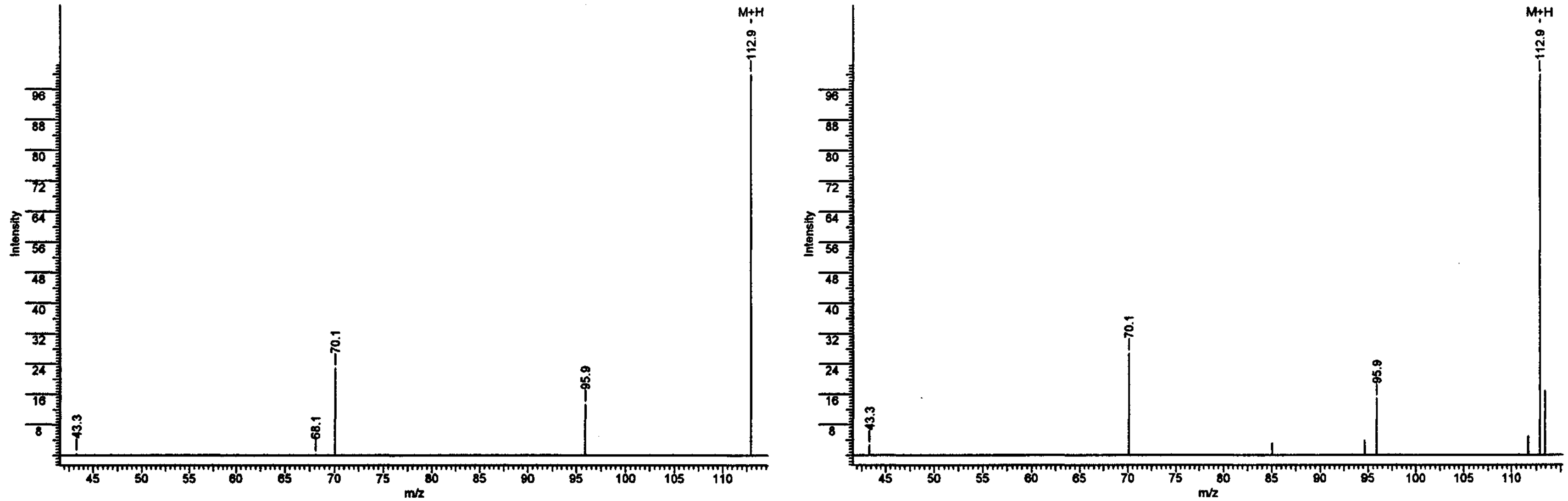


Figure 4.4 (a). Product ion spectra of standard (left) and urinary uracil (right) at low level energy.

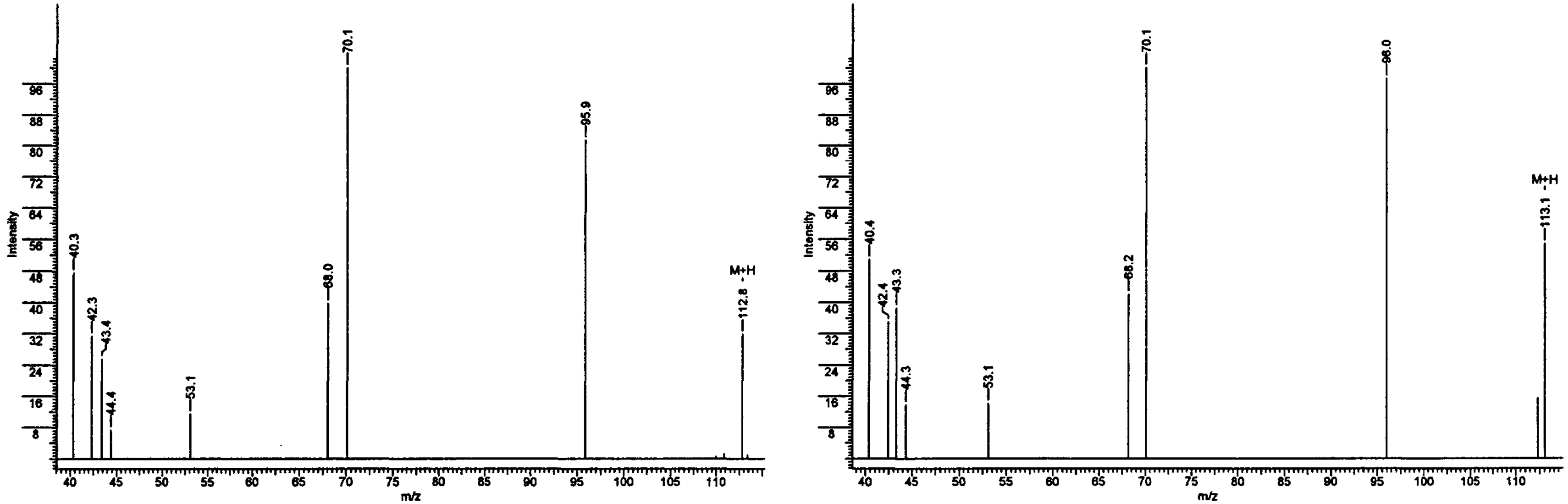


Figure 4.4 (b). Product ion spectra of standard (left) and urinary uracil (right) at medium level energy.

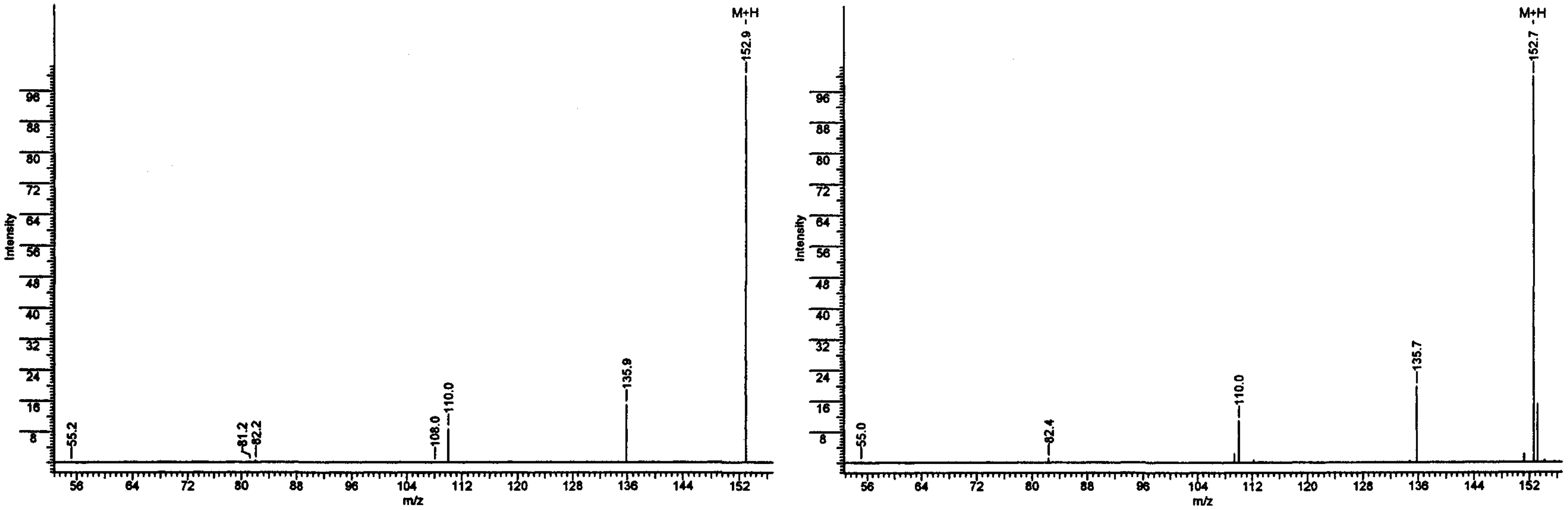


Figure 4.5 (a). Product ion spectra of standard (left) and urinary (right) Xanthine at low level energy.

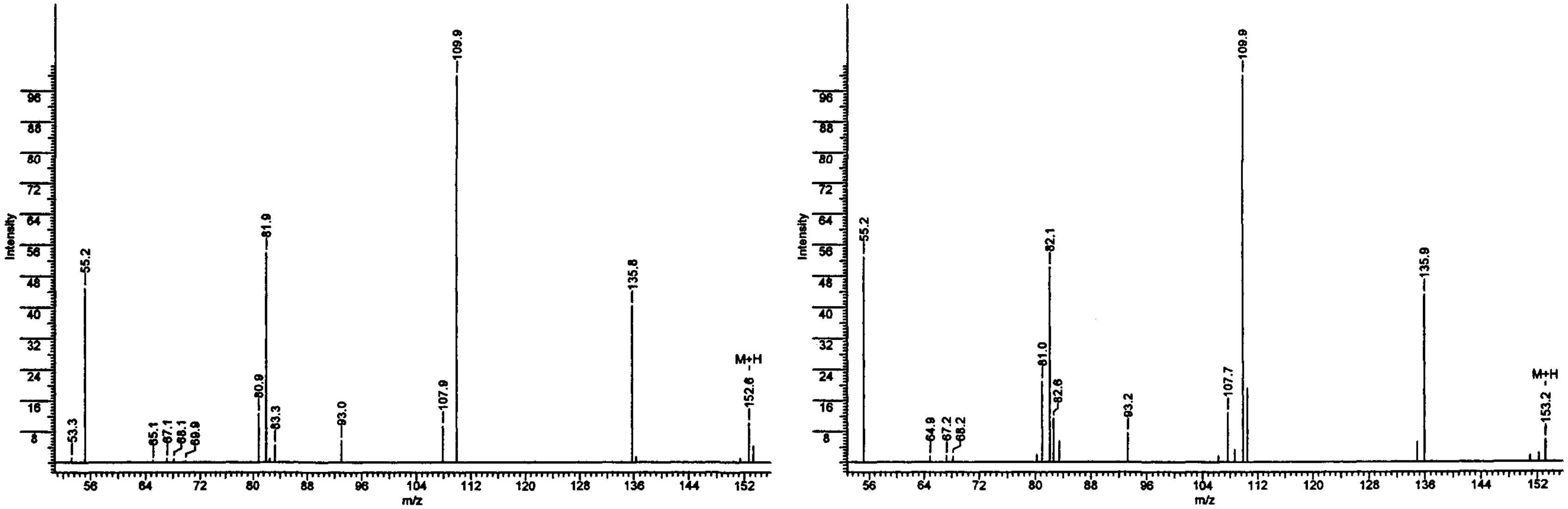


Figure 4.5 (b). Product ion spectra of standard (left) and urinary (right) Xanthine at medium level energy.

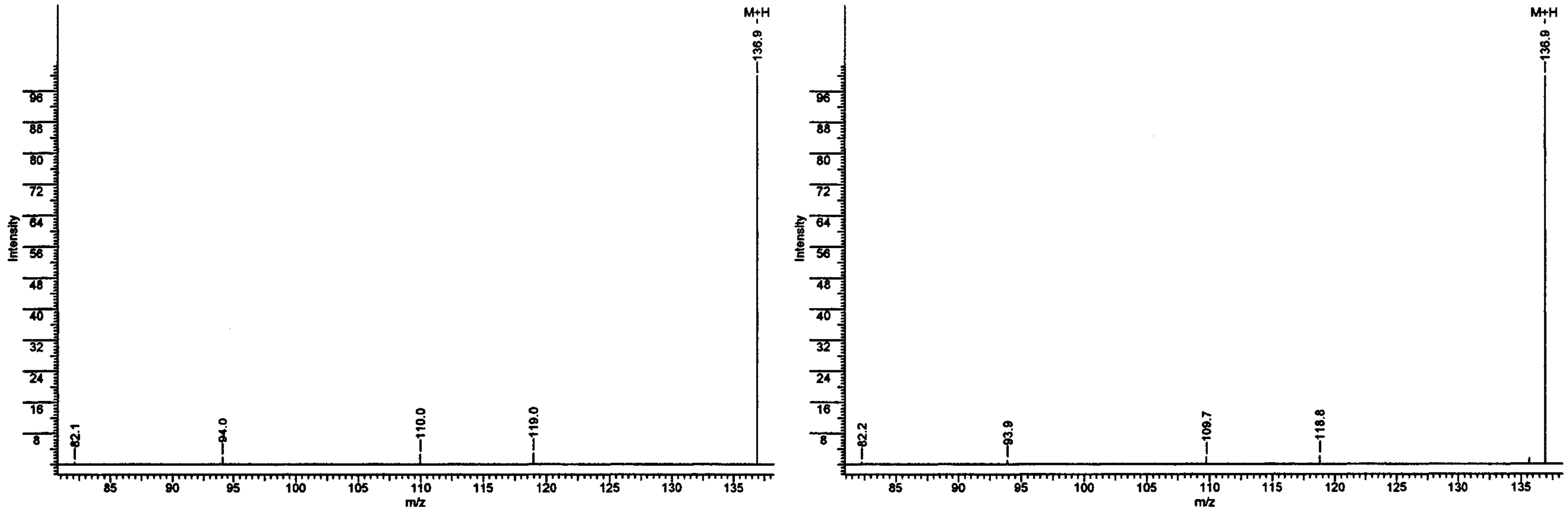


Figure 4.6(a). Product ion spectra of standard (left) and urinary (right) Hypoxanthine at low level energy.

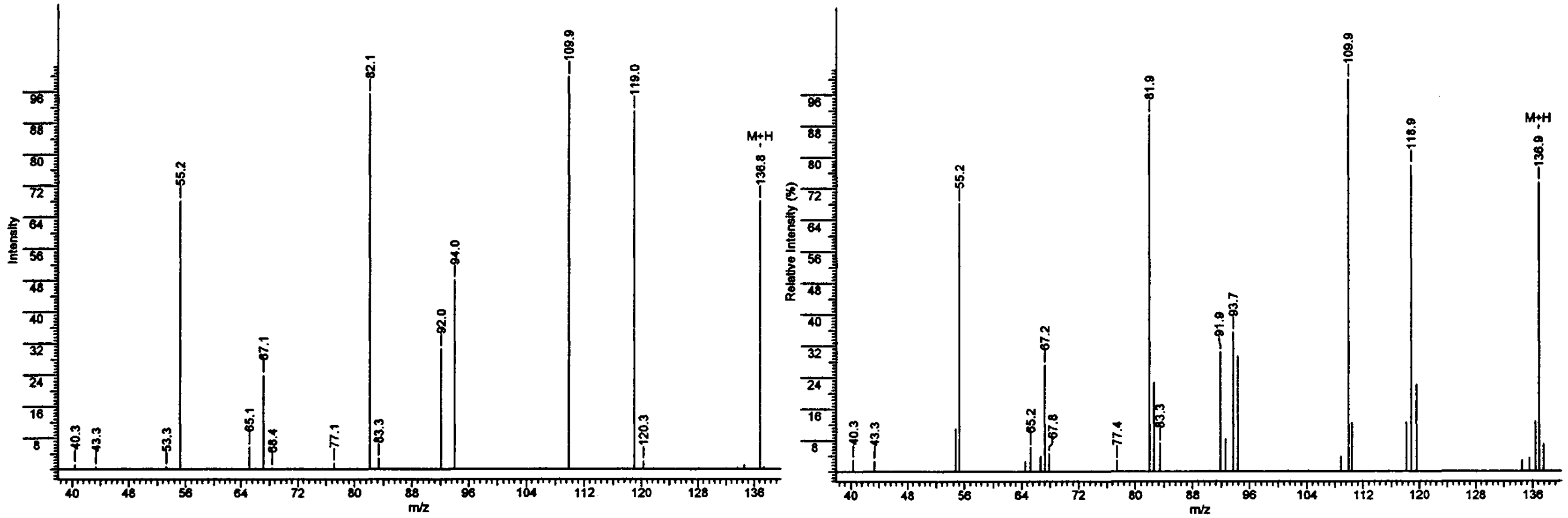


Figure 4.6(b). Product ion spectra of standard (left) and urinary (right) Hypoxanthine at medium level energy.

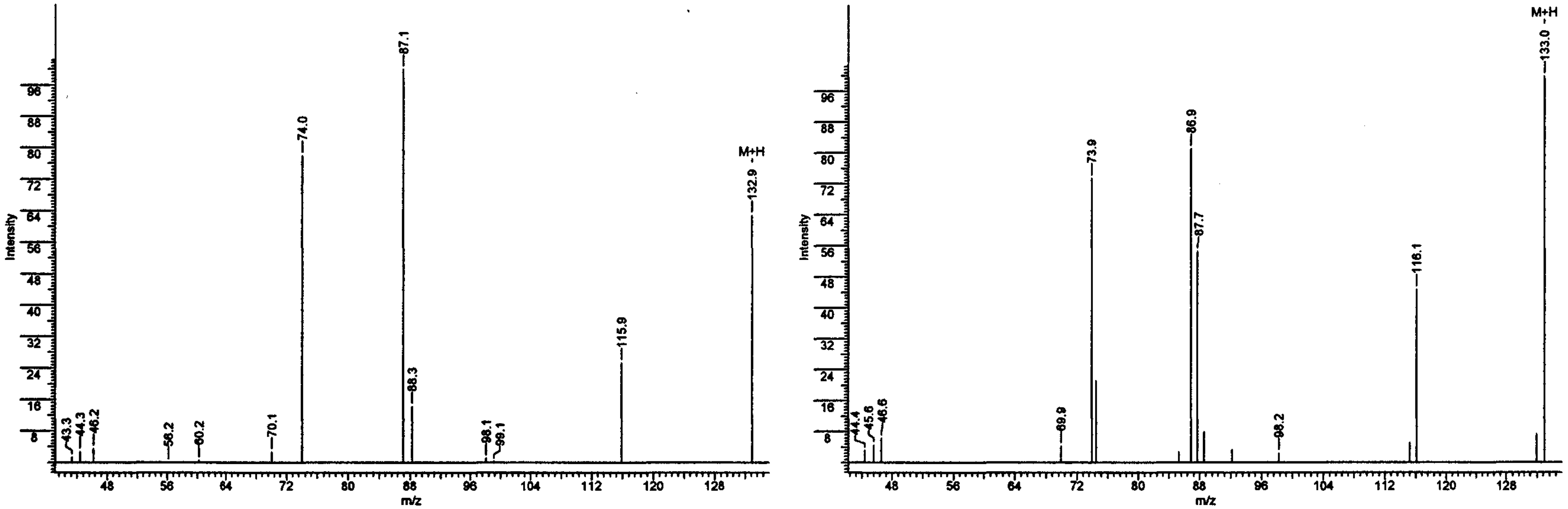


Figure 4.7(a). Product ion spectra of standard (left) and urinary (right) Asparginine at low level energy.

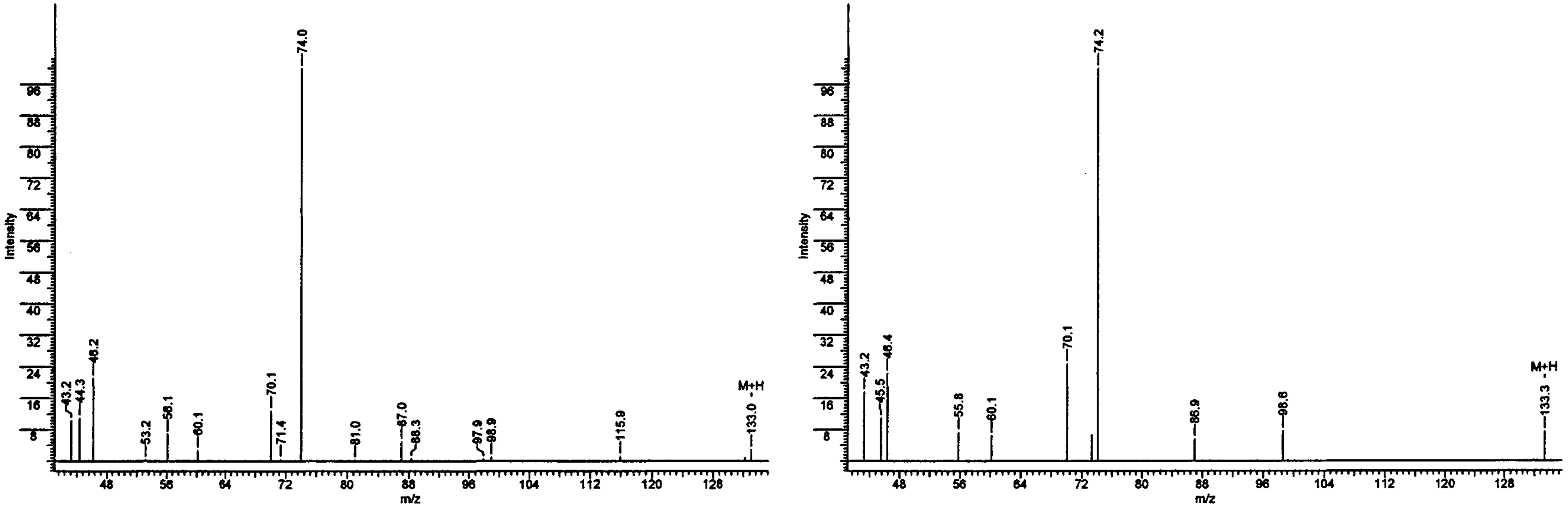


Figure 4.7(b). Product ion spectra of standard (left) and urinary (right) Asparginine at medium level energy.

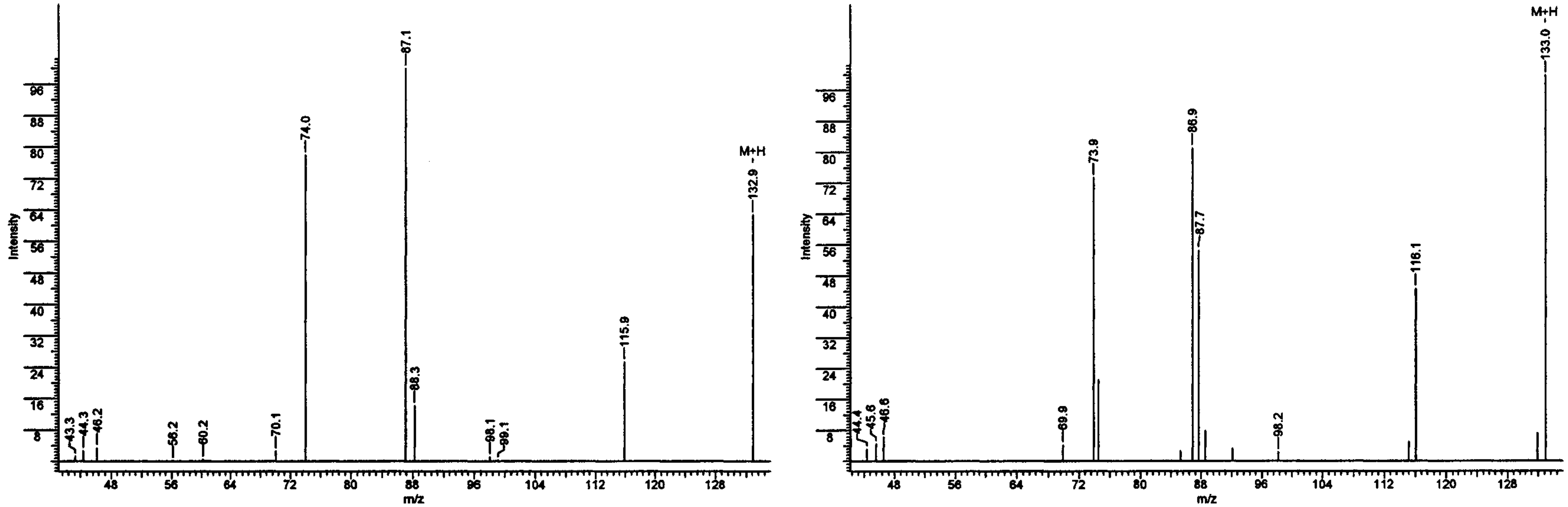


Figure 4.7(a). Product ion spectra of standard (left) and urinary (right) Asparginine at low level energy.

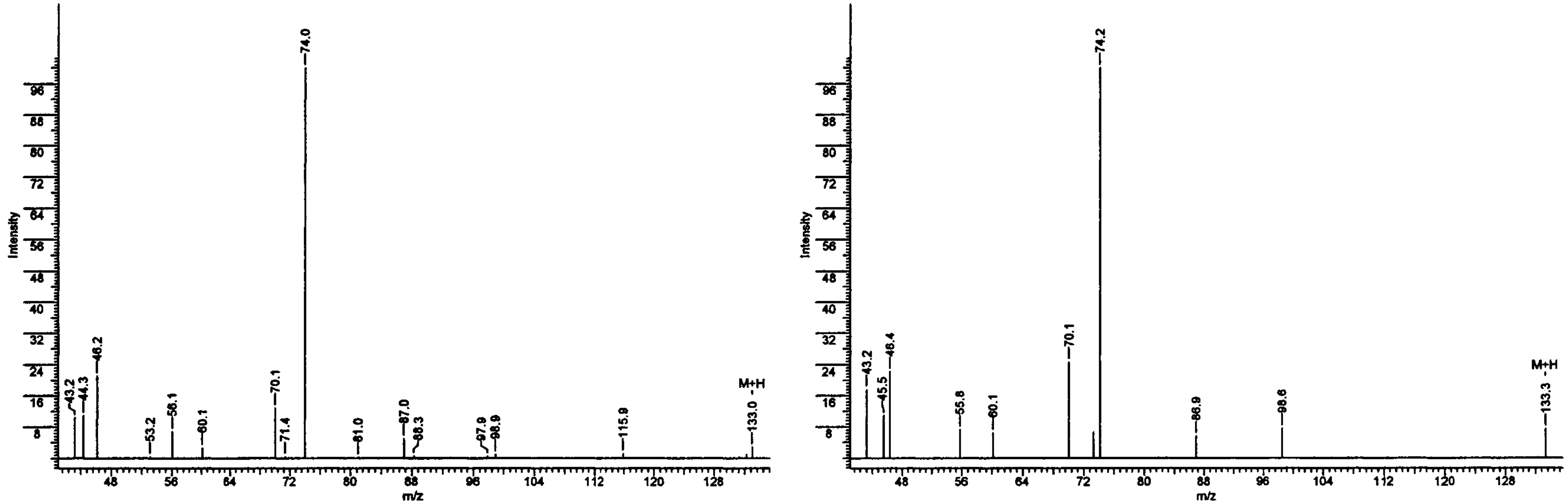


Figure 4.7(b). Product ion spectra of standard (left) and urinary (right) Asparginine at medium level energy.

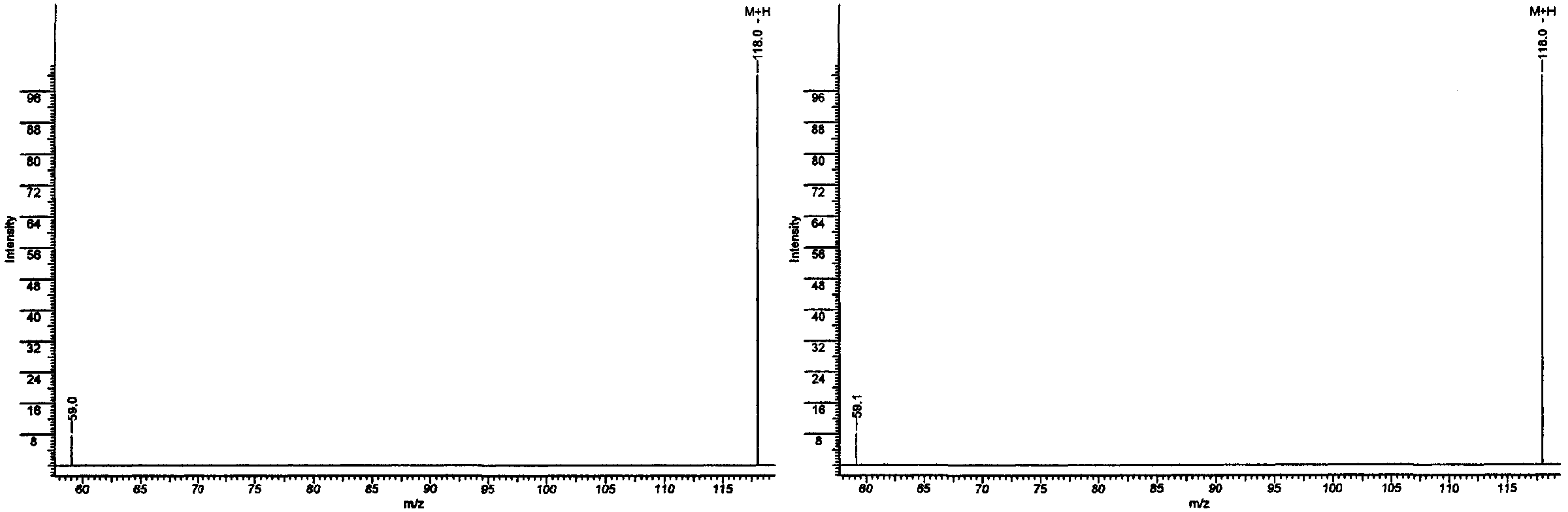


Figure 4.9(a). Product ion spectra of standard (left) and urinary (right) Betaine at low level energy.

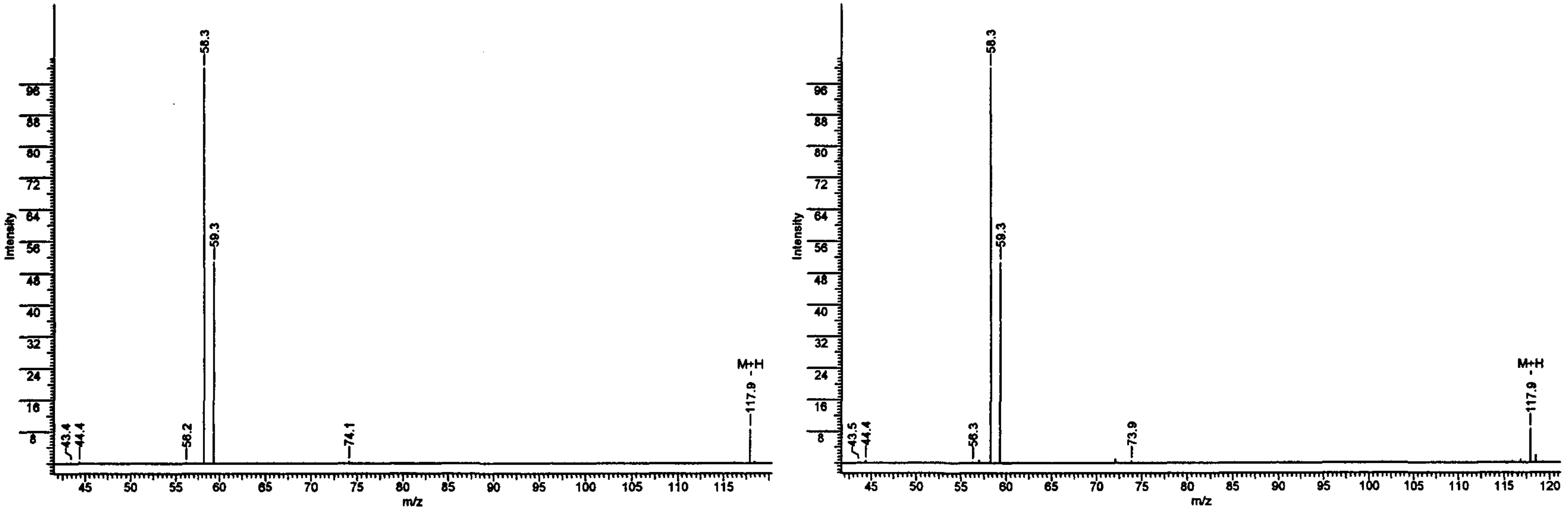


Figure 4.9(b). Product ion spectra of standard (left) and urinary (right) Betaine at medium level energy.

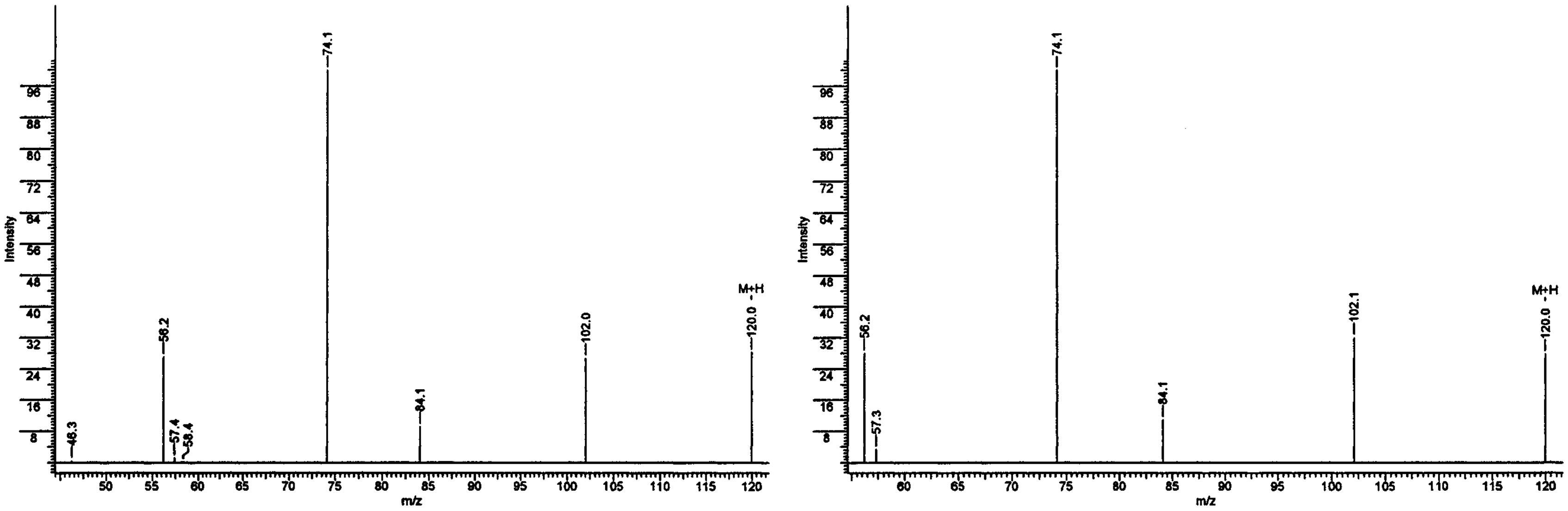


Figure 4.10(a). Product ion spectra of standard (left) and urinary (right) Threonine at low level energy.

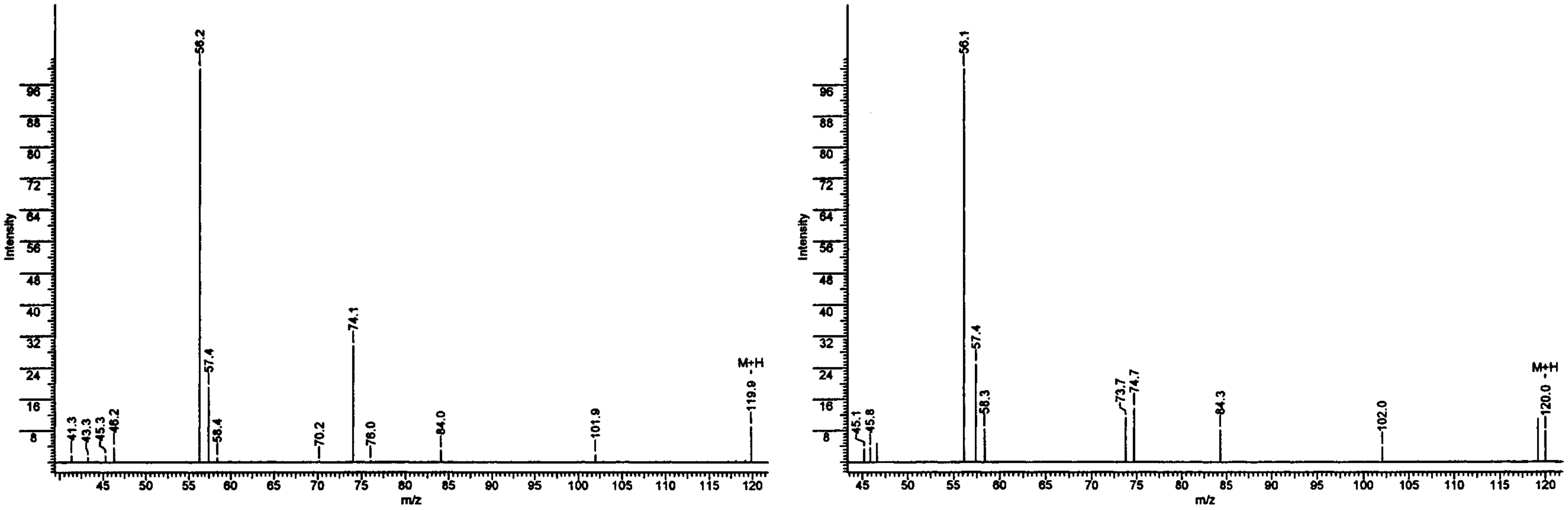


Figure 4.10(b). product ion spectra of standard (left) and urinary (right) Threonine at medium level energy.

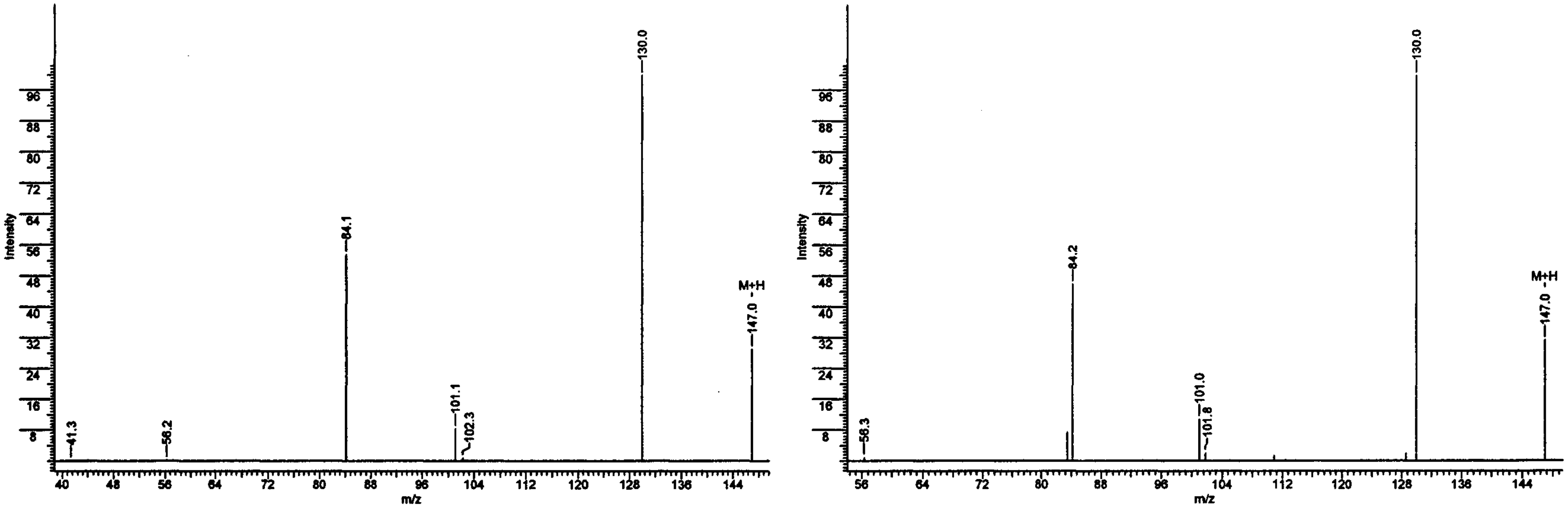


Figure 4.11(a). Product ion spectra of standard (left) and urinary (right) Glutamine at low level energy.

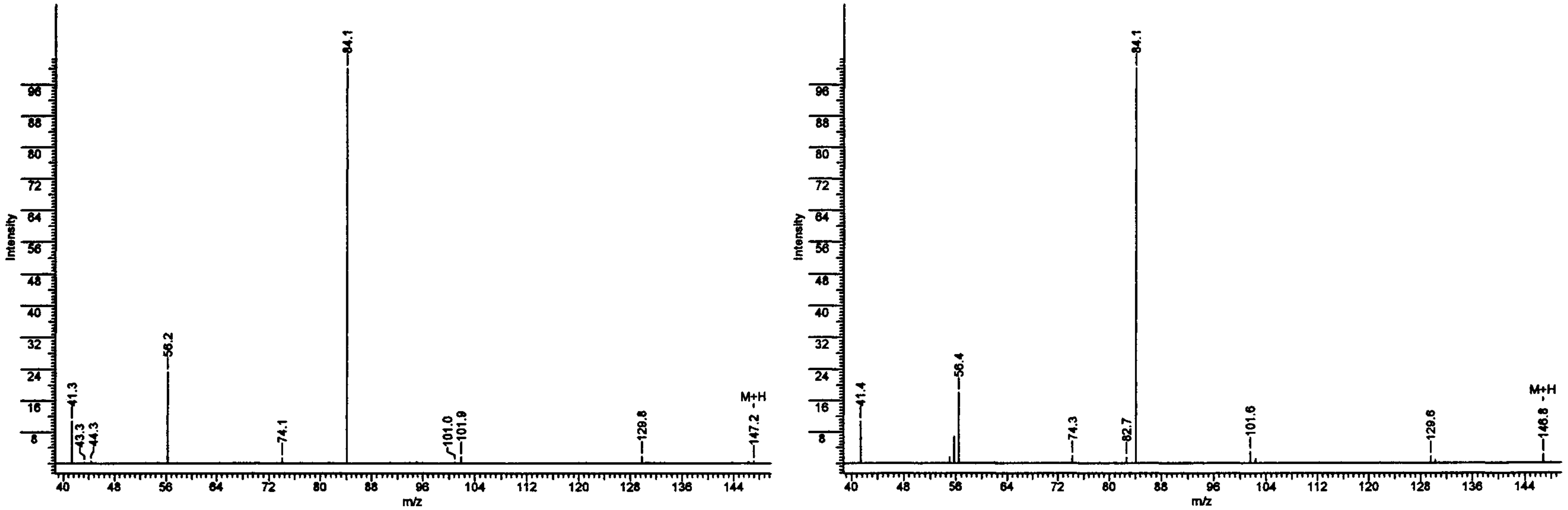


Figure 4.11(b). Product ion spectra of standard (left) and urinary (right) Glutamine at medium level energy.

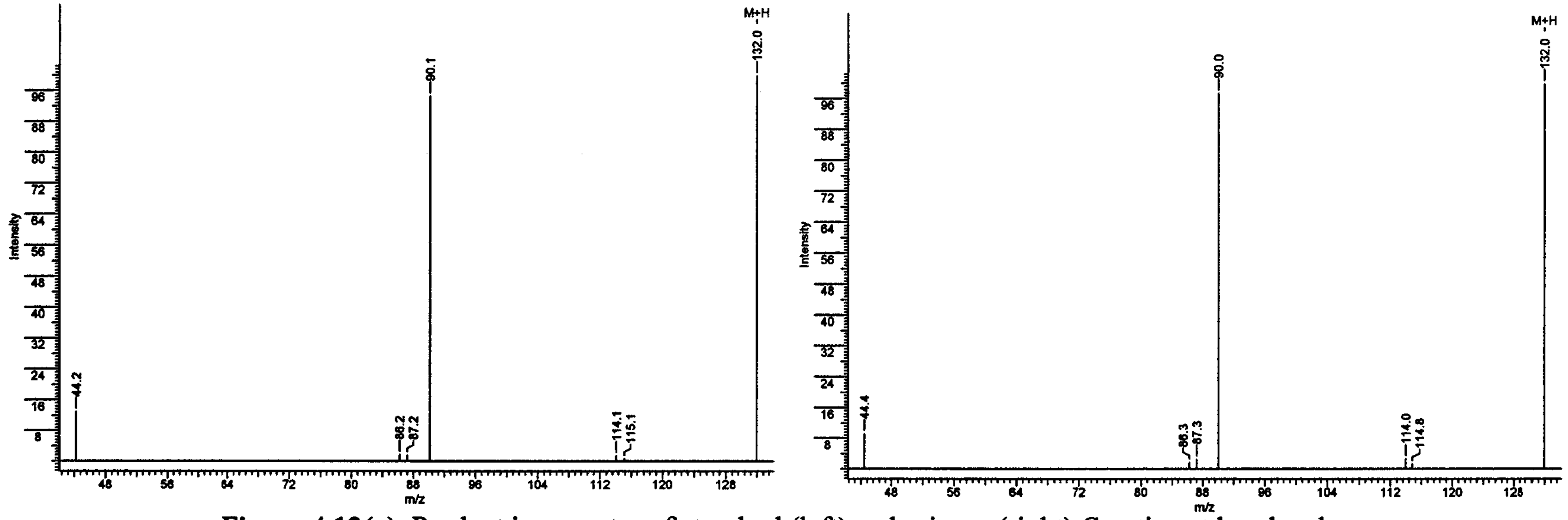


Figure 4.12(a). Product ion spectra of standard (left) and urinary (right) Creatine at low level energy.

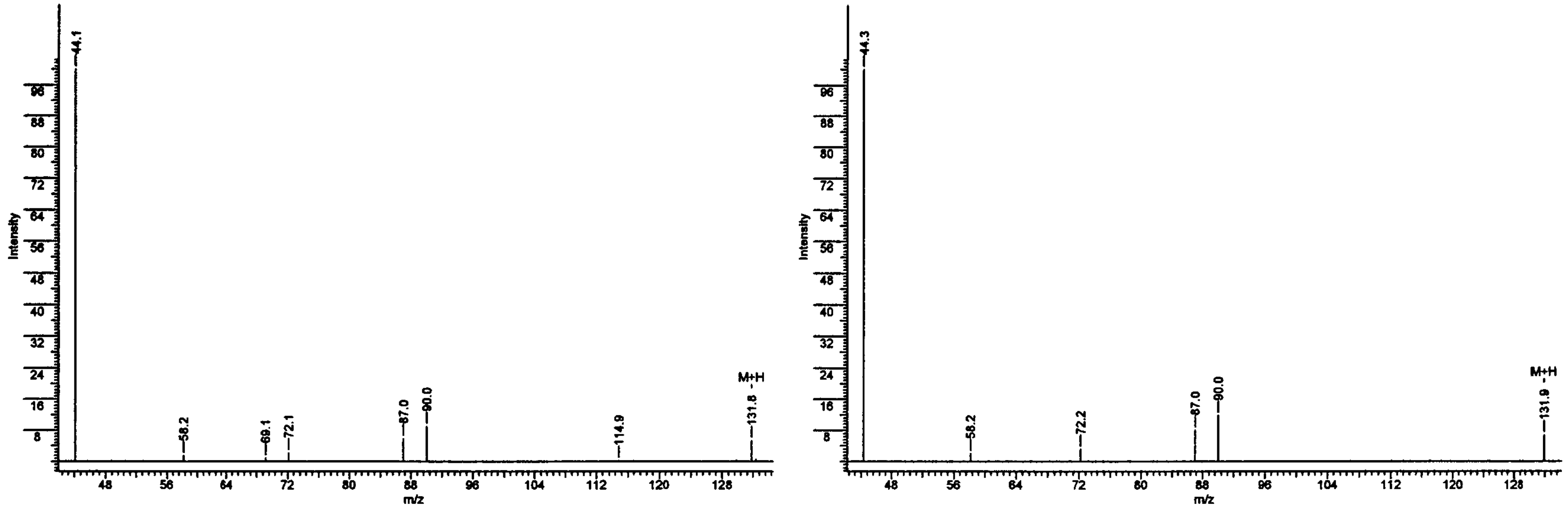


Figure 4.12(b). Product ion spectra of standard (left) and urinary (right) Creatine at medium level energy.

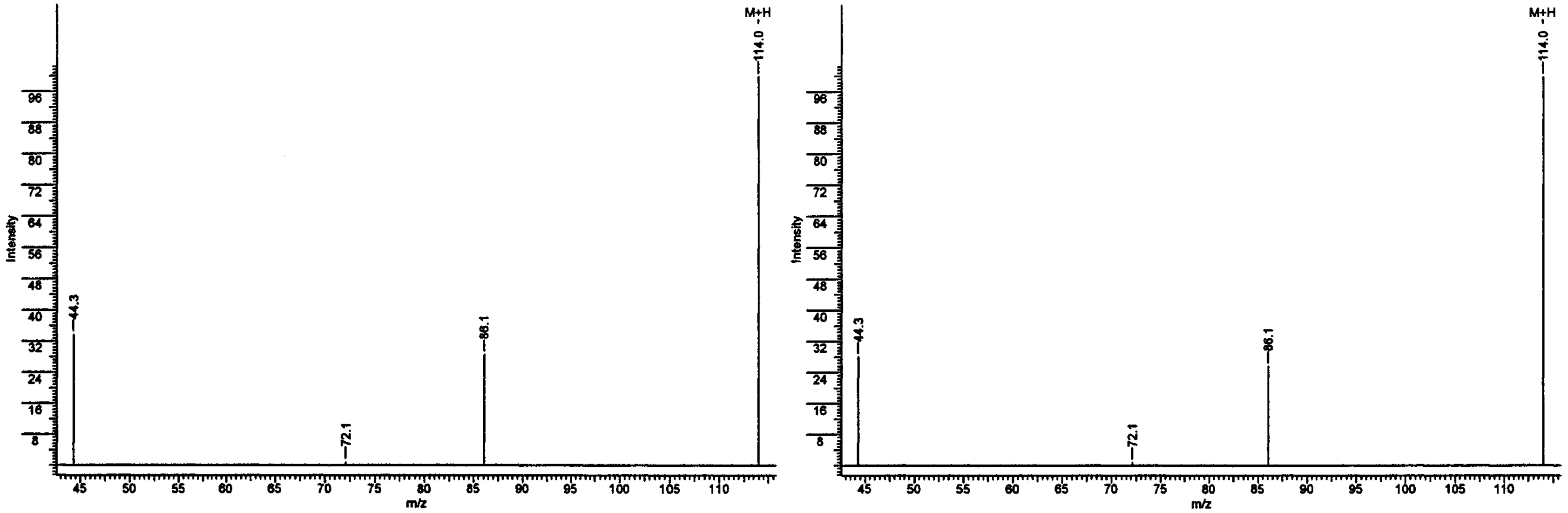


Figure 4.13(a). Product ion spectra of standard (left) and urinary (right) Creatinine at low level energy.

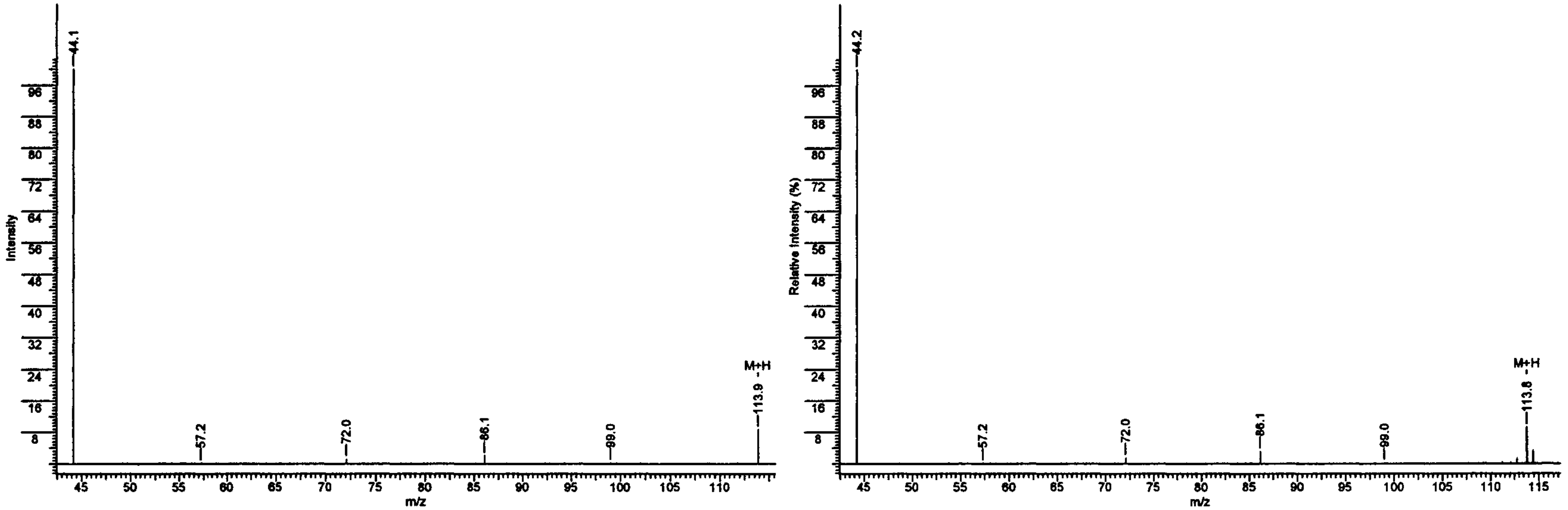


Figure 4.13(b). Product ion spectra of standard (left) and urinary (right) Creatinine at medium level energy.

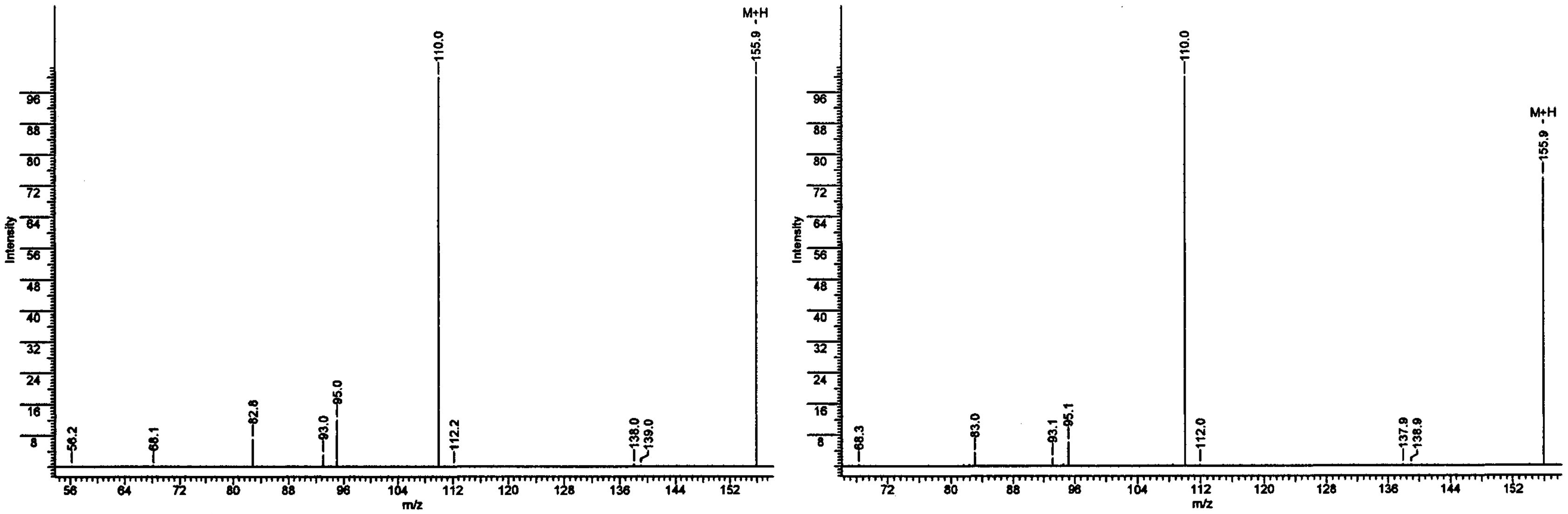


Figure 4.14(a). Product ion spectra of standard (left) and urinary (right) Histidine at low level energy.

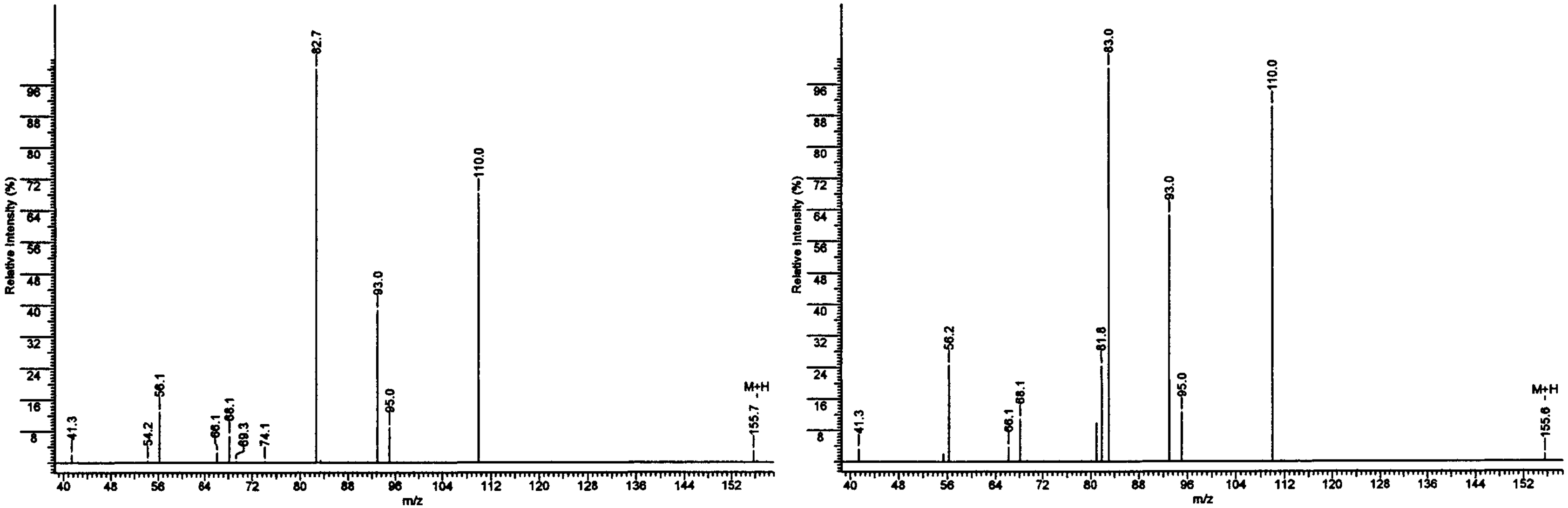


Figure 4.14(b). Product ion spectra of standard (left) and urinary (right) Histidine at medium level energy.

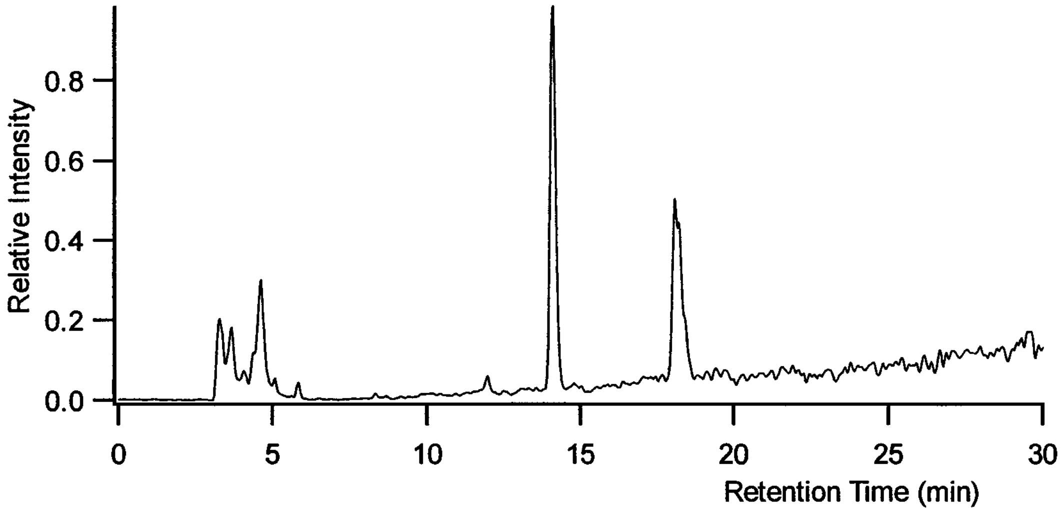


Figure 4.15. Ion chromatograph of urinary analysis in negative mode.

Table 4.10. Identified urinary metabolites in negative mode.

t_R (min)	m/z (Da)	Identified Metabolites	Low CID energy (10eV)			Medium CID energy (25eV)		
			Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
4.6	192.1	Citric Acid	99.92	99.93	99.92	99.89	99.55	99.71
5.2	174.1	Aconitic Acid	99.79	97.19	98.37	99.31	83.76	89.96
5.3	130.1	Glutaconic Acid	99.53	99.44	99.49	98.44	89.22	93.14
14.1	179.2	Hippuric Acid	99.98	99.98	99.98	99.99	99.99	99.99

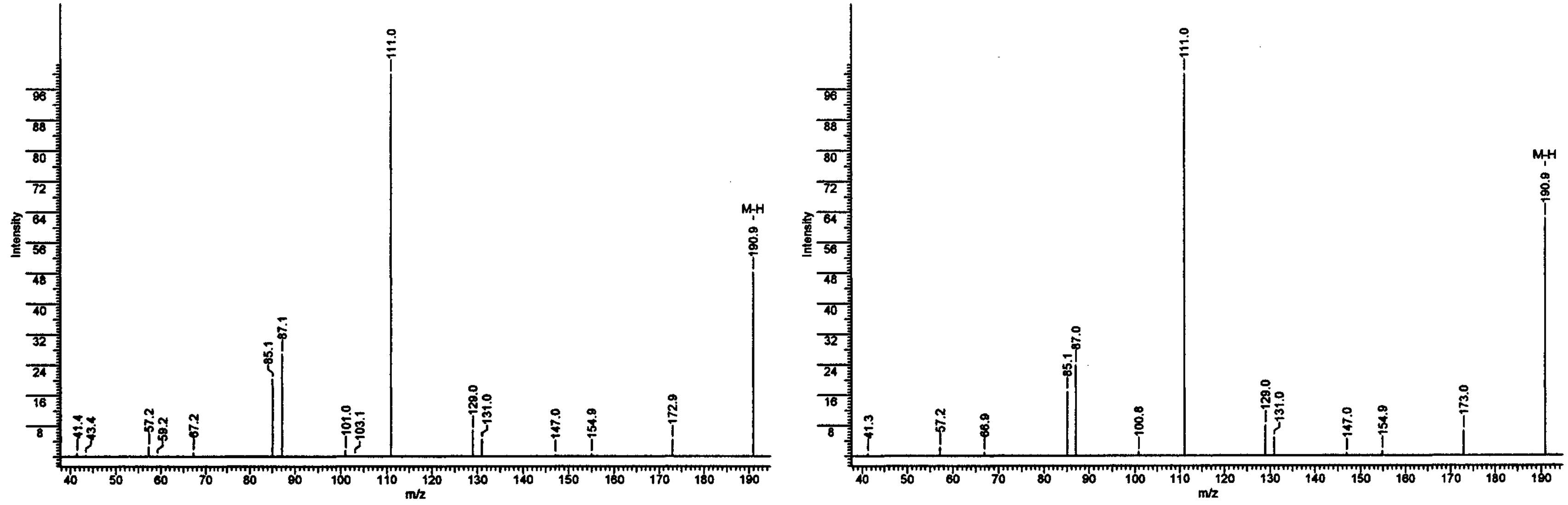


Figure 4.16(a). Product ion spectra of standard (left) and urinary (right) Citric Acid at low level energy.

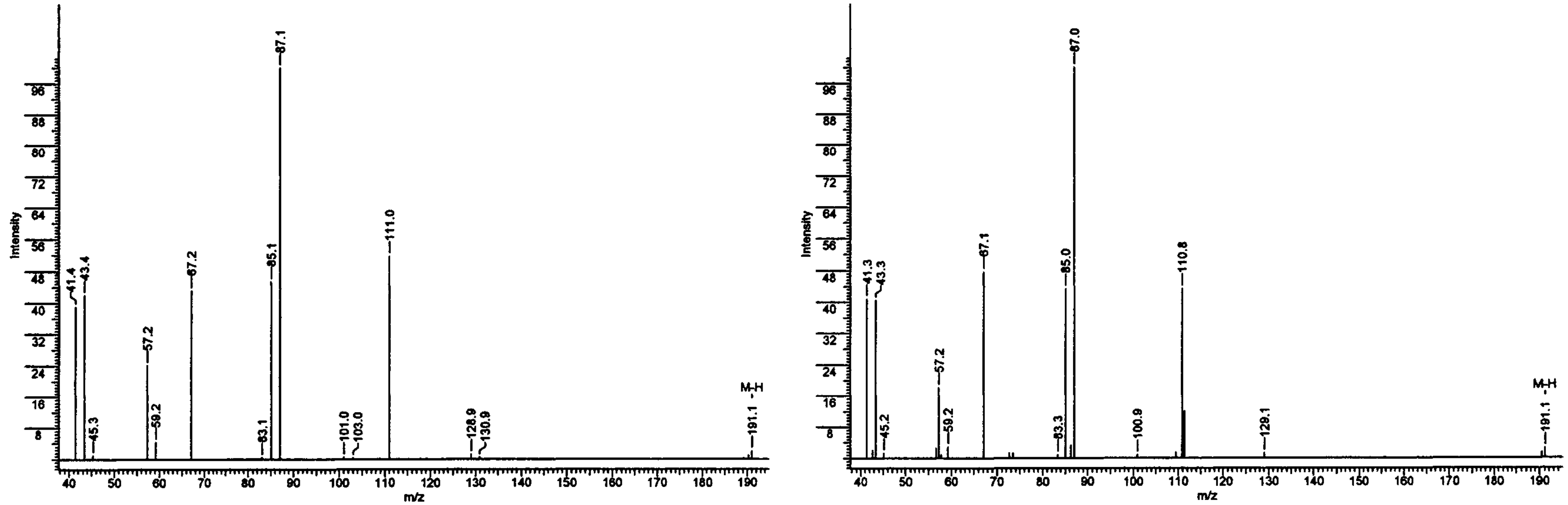


Figure 4.16(b). Product ion spectra of standard (left) and urinary (right) Citric Acid at medium level energy.

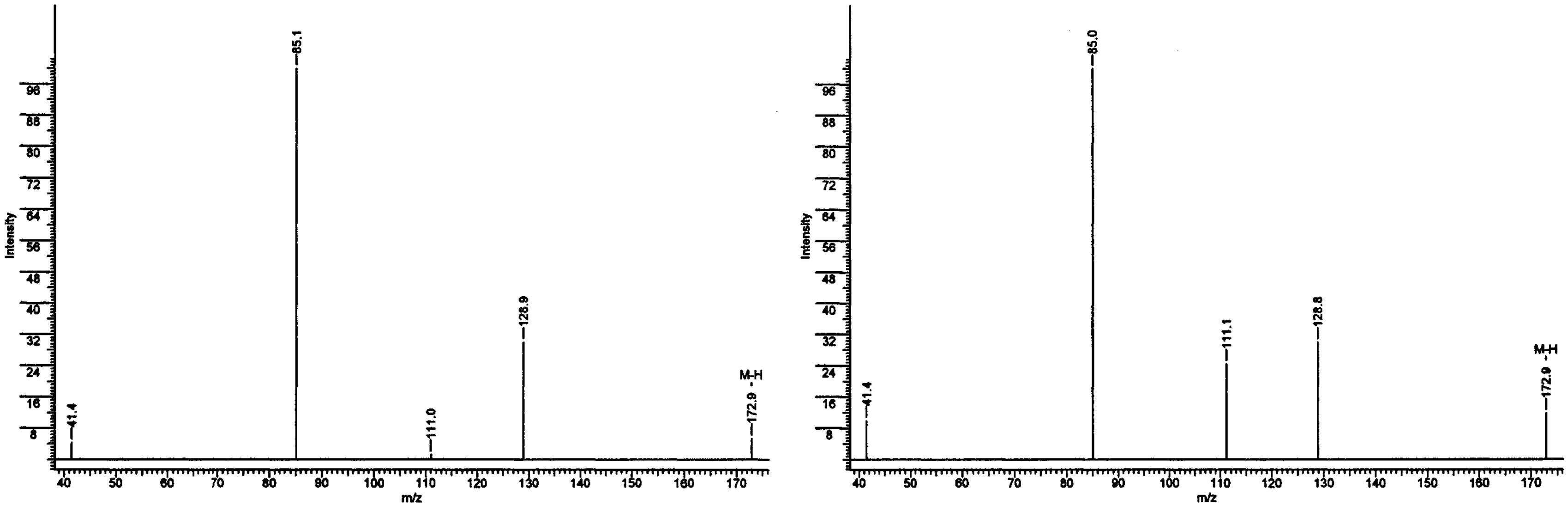


Figure 4.17(a). Product ion spectra of standard (left) and urinary (right) Aconitic Acid at low level energy.

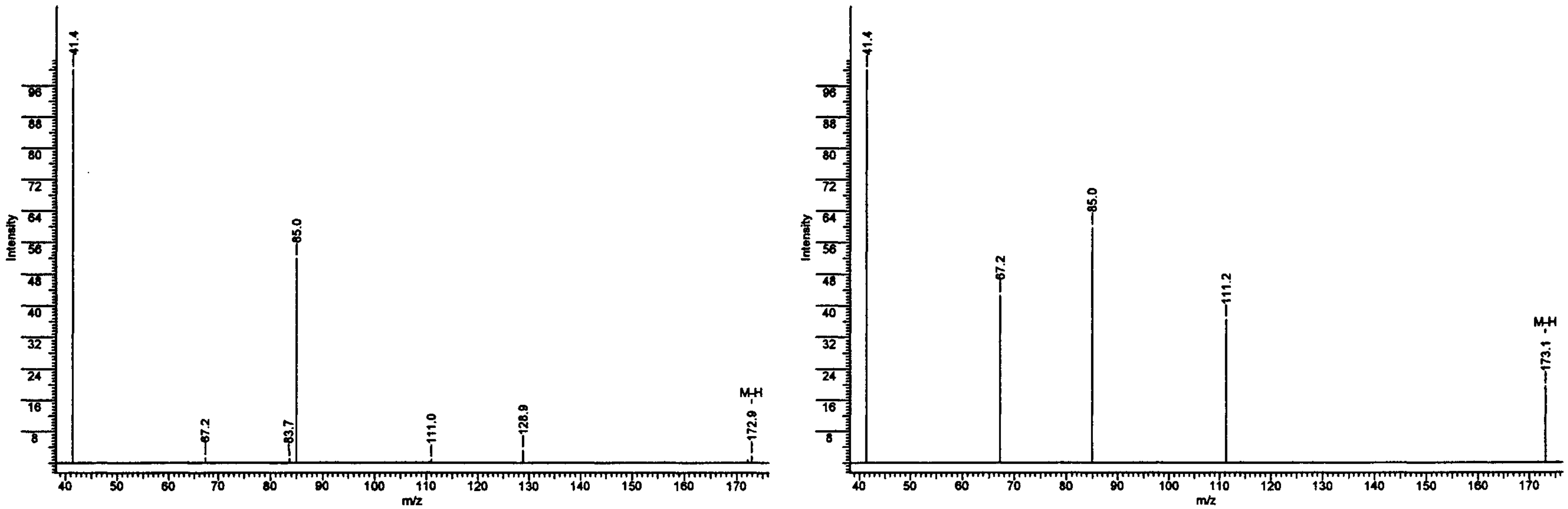


Figure 4.17(b). Product ion spectra of standard (left) and urinary (right) Aconitic Acid at medium level energy.

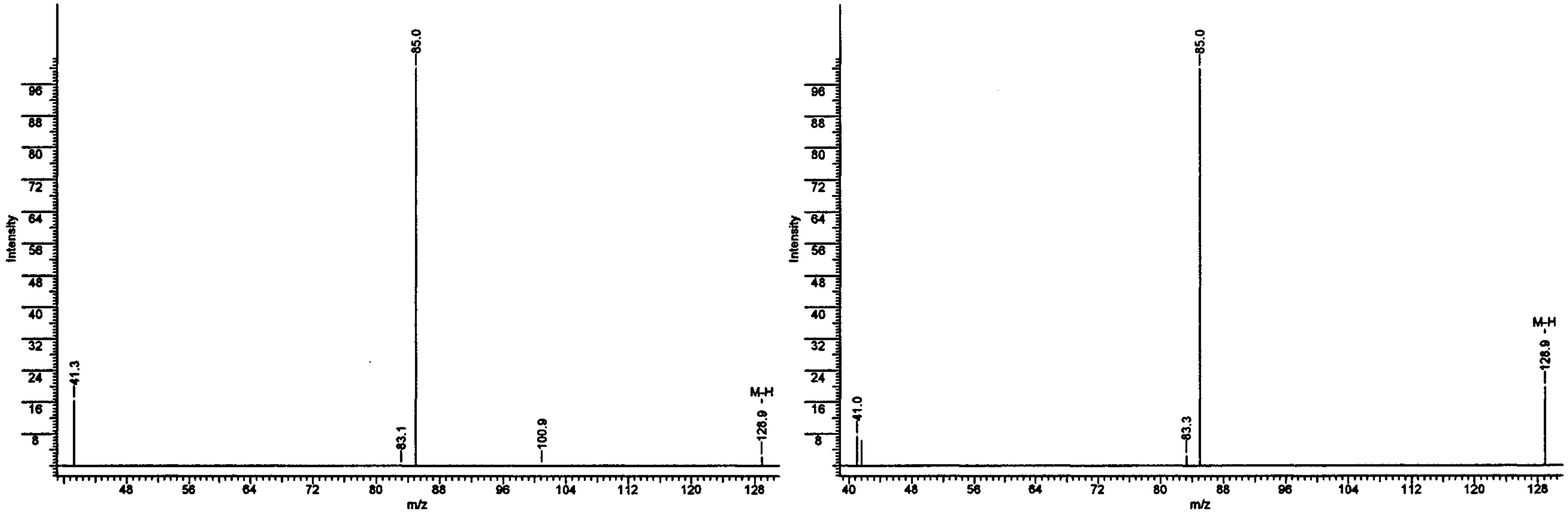


Figure 4.18(a). Product ion spectra of standard (left) and urinary (right) Glutaconic Acid at low level energy.

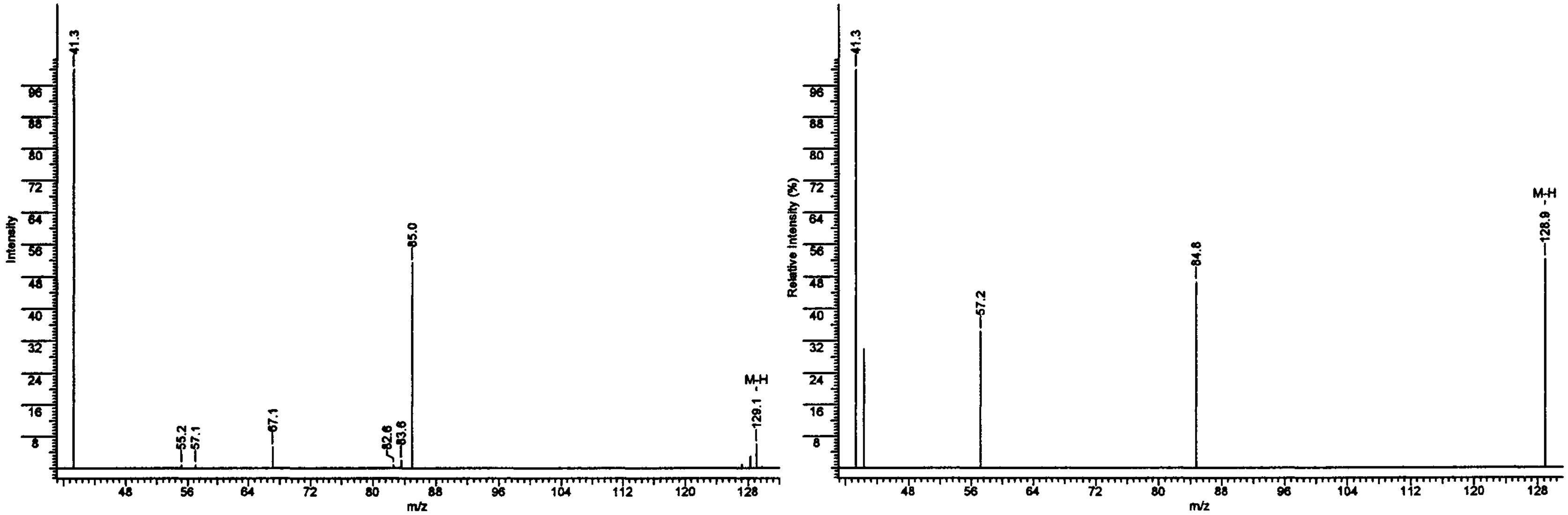


Figure 4.18(b). Product ion spectra of standard (left) and urinary (right) Glutaconic Acid at medium level energy.

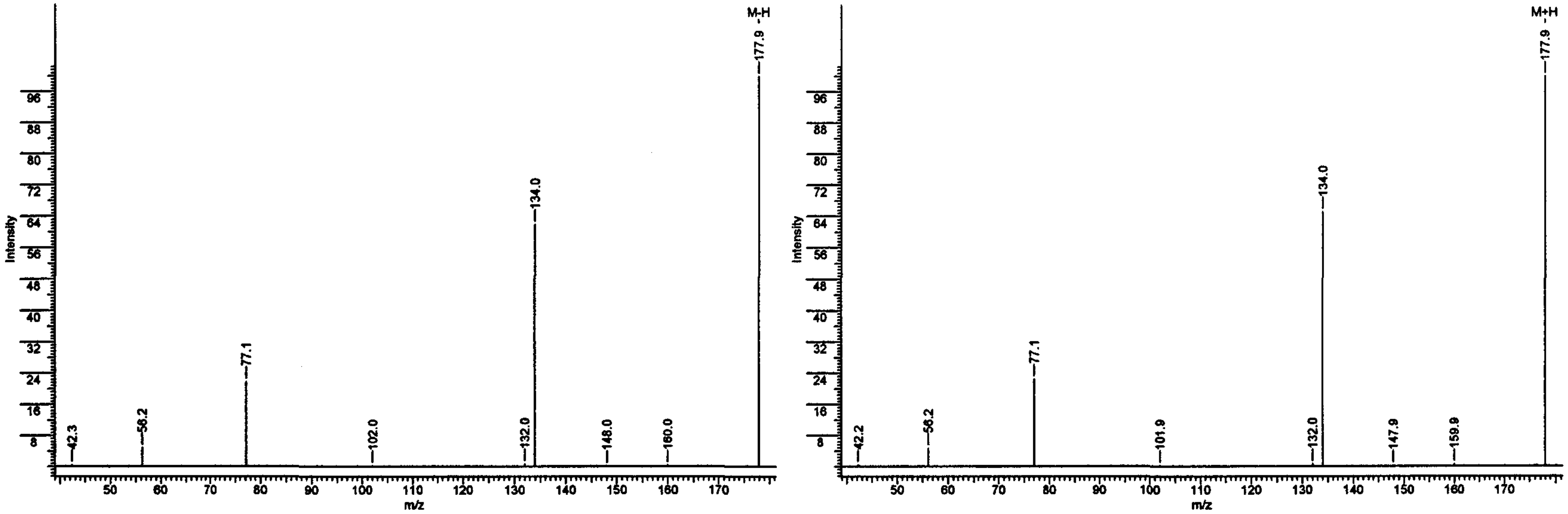


Figure 4.19(a). Product ion spectra of standard (left) and urinary (right) Hippuric Acid at low level energy.

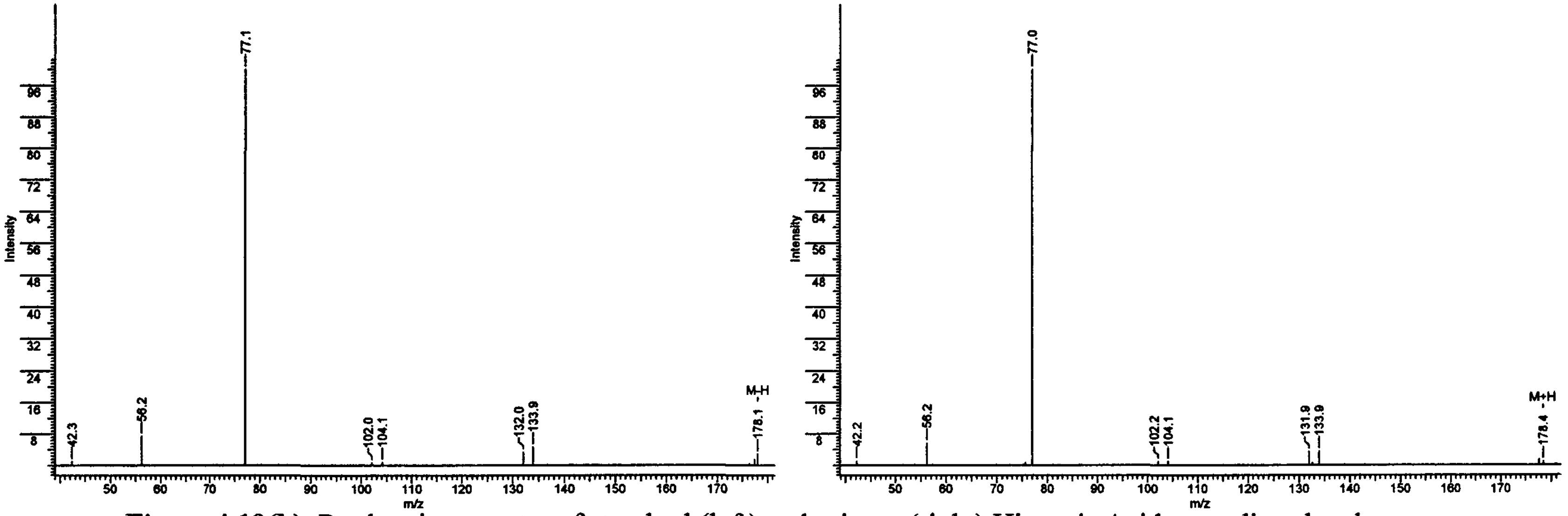


Figure 4.19(b). Product ion spectra of standard (left) and urinary (right) Hippuric Acid at medium level energy.

It should be noted that, a higher concentration of metabolites in a sample ensures the acquisition of good quality MS/MS spectra, resulting in their accurate identification, while a low concentration metabolite usually results in a MS/MS spectrum with the disappearance of the low abundance product ions. However, the high abundance product ions are critical for accurate identification. In some cases, only a few product ions of metabolites are generated, due to the simple structure of parent molecules or difficult in fragmenting the parent ion extensively. But these product ions can still be used for reliable identification.

The confidence for metabolite identification based on the searching results also depends on the mathematical algorithm used for scoring of matches. In the given algorithm, not only m/z of product ions but also their abundance is considered. Based on their relative abundance, different product ions have different contributions to the identification process, which is expressed as a weighing factor. All these make the algorithm more reliable. For example, in the urinary profiling in the positive ion mode, the unknown peak at 10.4 min was identified as asparagine. In the library, 3-ureidopropionic acid has exactly the same chemical composition and molecular weight as asparagine and these two have similar structure. But they are different in fragmentation patterns, as shown in their CID MS/MS spectra (Figure 4.20). If the unknown compound is searched with 3-ureidopropionic acid, the result shown in Table 4.11 would be obtained. Evidently, the low score proves that the unknown peak is not 3-ureidopropionic acid. Another example is the unknown peak at 11.5 min, which is identified as threonine. The isomer of threonine is homoserine in the database and they have very similar structures. From their CID MS/MS spectra shown in Figure 4.21, the differences in their fragmentation can be seen to be very small. Using this algorithm to calculate the Fit/RFit/Purity, shown in Table 4.11, although the scores are high in both cases, the 1st hit is threonine not homoserine. Therefore, the identification process can be achieved reasonably using this mathematical algorithm.

With the enlargement of the MS/MS library, more metabolites will be identified from the urine sample. Nonetheless, this current application has demonstrated that the spectral library can be used for real world application.

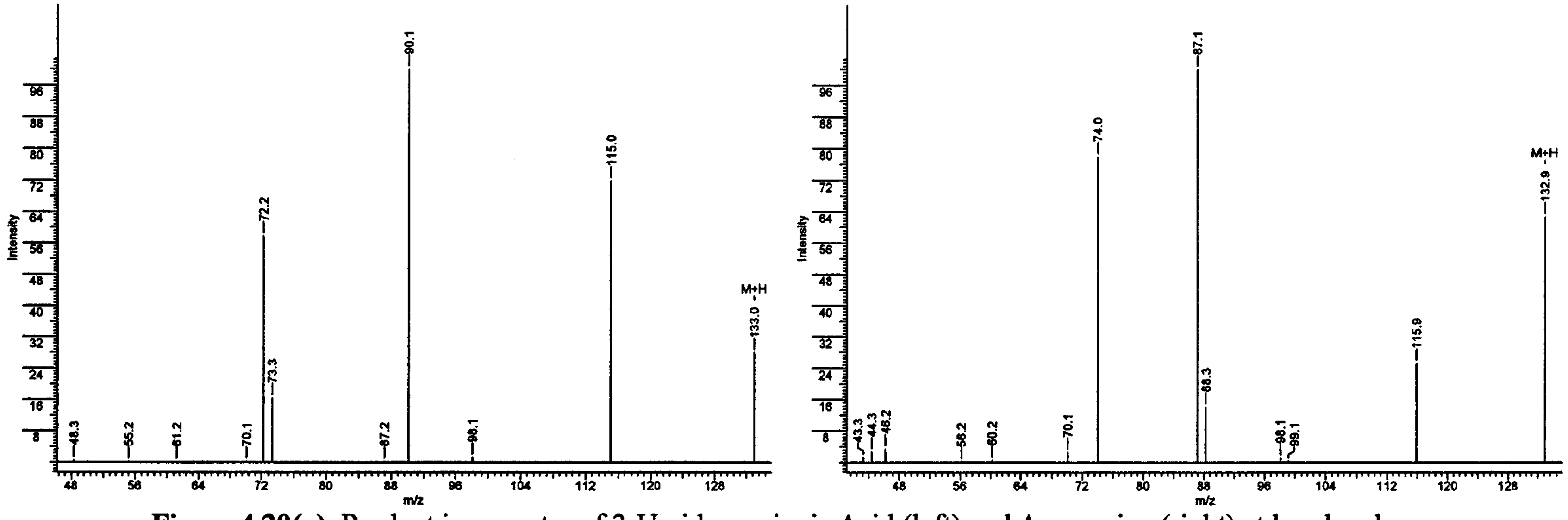


Figure 4.20(a). Product ion spectra of 3-Ureidopropionic Acid (left) and Asparagine (right) at low level energy.

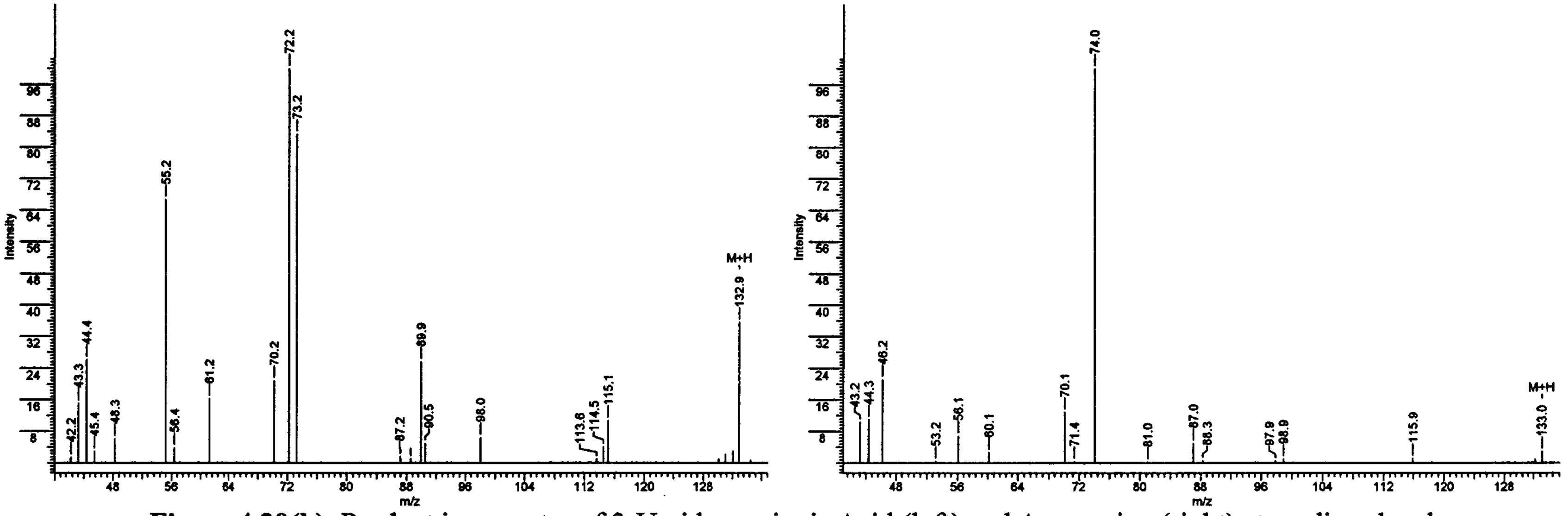


Figure 4.20(b). Product ion spectra of 3-Ureidopropionic Acid (left) and Asparagine (right) at medium level energy.

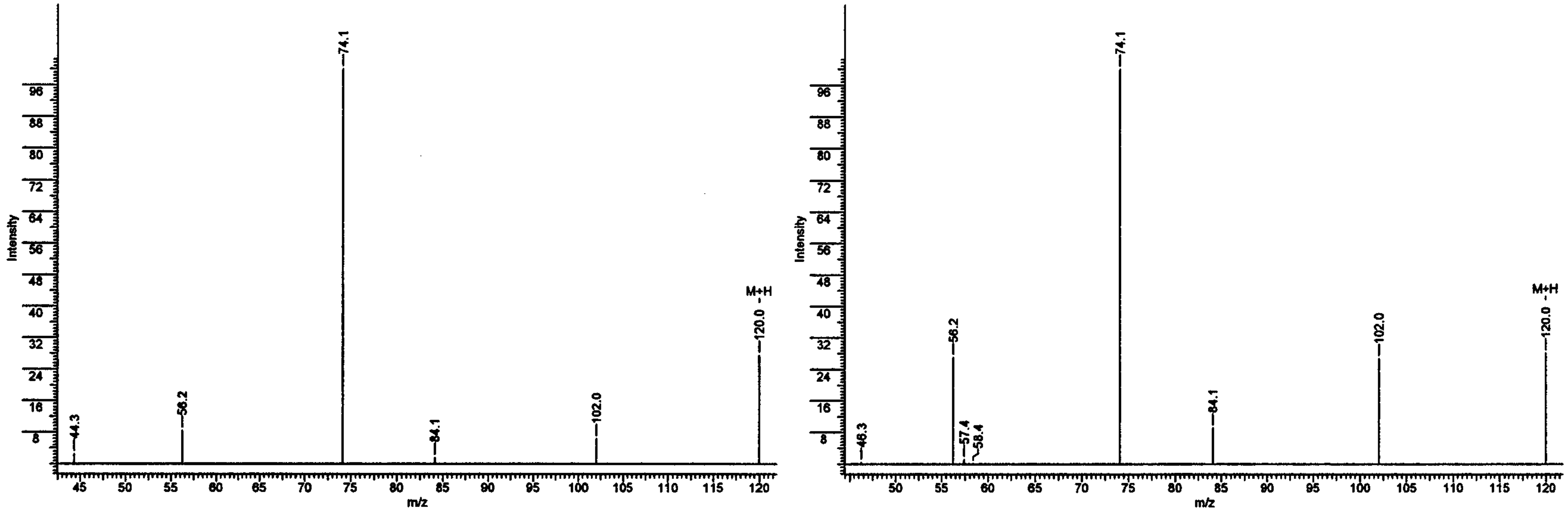


Figure 4.21(a). Product ion spectra of Homoserine (left) and Threonine (right) at low level energy.

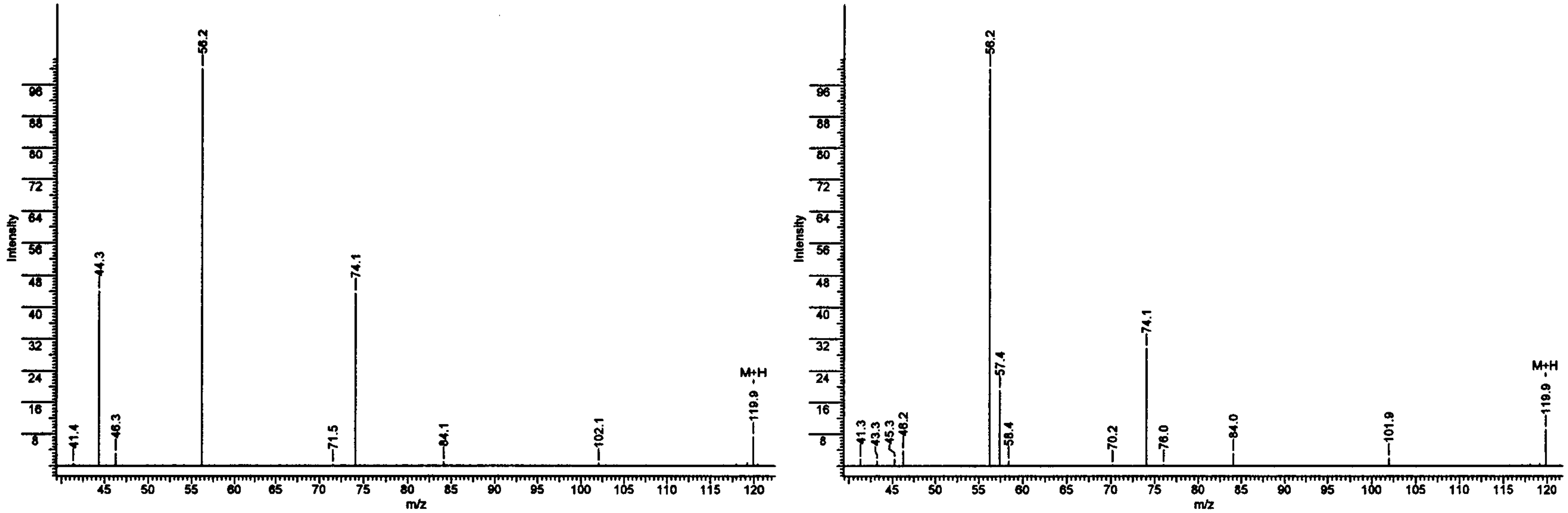


Figure 4.21(b). Product ion spectra of Homoserine (left) and Threonine (right) at medium level energy.

Table 4.11. Comparison of result searching score for the peaks at 10.4 min and 11.5 min.

t_R (min)	m/z (Da)	Assumed Metabolite	Low CID energy (10eV)			Medium CID energy (25eV)		
			Fit (%)	RFit (%)	Purity (%)	Fit (%)	RFit (%)	Purity (%)
10.4	132.1	Asparagine	99.06	92.19	95.09	99.30	96.80	97.98
		3-Ureidopropionic Acid	27.2	28.81	28.1	54.85	49.7	53.08
11.5	119.1	Threonine	99.90	99.90	99.90	99.55	92.40	95.78
		Homoserine	97.62	91.51	93.98	88.90	85.80	87.40

4.4 Conclusions

We have initiated the construction of a MS/MS spectral library of human metabolites. The current library consists of 215 human metabolites. For each metabolite, the product ion spectra were obtained at three CID energy levels using a Triple Q mass spectrometer. Various experimental parameters were investigated for their influence on spectral reproducibility. It was found that, in a wide range of LC conditions and metabolite concentrations, MS/MS spectra were very reproducible. This library was subsequently applied to human urinary analysis and fifteen metabolites were identified by comparing acquired MS/MS spectra with the library MS/MS spectra using a matching algorithm developed in house. Future work will include the expansion of the present library by including more metabolites that are either commercially available or new identified by other techniques. In addition, other types of mass spectrometers, such as Q-TOF and FTICR-MS [31], with improved mass resolution and mass measurement accuracy will be investigated for spectral library creation. We anticipate high quality spectral library will be constructed with these newer mass spectrometers and such a library will improve the identification confidence based on MS/MS spectral matching. Finally, new matching algorithms with improved statistics scoring methods will need to be developed to facilitate data processing and metabolite identification.

4.5 Literature Cited

(1) Fiehn, O. *Comp Funct Genom* **2001**, *2*, 155-168.

(2) Wagner, C.; Sefkow, M.; Kopka, J. *Phytochemistry* **2003**, *62*, 887.

(3) Morgenthal, K.; Wienkoop, S.; Scholz, M.; Selbig, J.; Wechwerth, W. *Metabolomics* **2005**, *1*, 109.

(4) Tikunov, Y.; Lommen, A.; Ric de Vos, C.H.; Verhoeven, H.A.; Bino, R.J.; Hall, R.D.; Bovy, A.G. *Plant Physiology* **2005**, *139*, 1125.

(5) Jonsson, P.; Johansson, A.I.; Gullberg, J.; Trygg, J.; Jiye, A.; Grung, B.; Marklund, S.; Sjostrom, M.; Anttim, H.; Moritz, T. *Anal. Chem.* **2005**, *77*, 5635.

(6) Shellie, R.; Marriott, P.; Morrison, P. *Anal. Chem.* **2001**, *73*, 1336.

(7) Roessner-Tunali, U.; Hegemann, B.; Lytovchenko, A.; Carrari, F.; Bruedigam, C.; Granot, D.; Fernie, A.R. *Plant Physiology* **2003**, *133*, 84.

(8) Dunn, W.B.; Overy, S.; Quick, W.P. *Metabolomics* **2005**, *1*, 137.

(9) Vorst, O.; de Vos, C.H.; Lommen, A.; Staps, R.V.; Visser, R.G.; Bino, R.J.; Hall, R.D. *Metabolomics* **2005**, *1*, 169.

(10) Soga, T.; Ohashi, Y.; Ueno, Y.; Naraoka, H.; Tomita, M.; Nishioka, T. *J. Proteome Research* **2003**, *2*, 488.

(11) Dunn, W.B.; Ellis, D.I. *Trends in Analytical Chemistry* **2005**, *24*, 285.

(12) Rashed, M.S. *J. Chromatography B* **2001**, *758*, 27.

(13) Pitt, J.J.; Eggington, M.; Kahler, S.G. *Clin. Chem.* **2002**, *48*, 1970.

(14) Jemal, M.; Zheng, Q.; Zhao, W.; Zhu, W.; Wu, W.W. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2732.

(15) Ophelia, Q.P.; Yin, S.; Lam, S.L.; Li, C.M.; Chow, M.S. *Rapid Commun. Mass Spectrom.* **2004,** *18,* 2921.

(16) Paik, M.J.; Lee, H.J.; Kim, K.R. *J. Chromatography B* **2005,** *821,* 94.

(17) Hough, J.M.; Haney, C.A.; Voyksner, R.D.; Bereman, R.D. *Anal. Chem.* **2000,** *72,* 2265.

(18) Bristow, A.W.; Webb, K.S.; Lubben, A.T.; Halket, J. *Rapid Commun. Mass Spectrom.* **2004,** *18,* 1447.

(19) Baumann, C.; Cintora, M.A.; Eichler, M.; Lifante, E.; Cooke, M.; Przyborowska, A.; Halket, J.M. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 349.

(20) Gergov, M.; Weinmann, W.; Meriluoto, J.; Uusitalo, J.; Ojanpera, I. *Rapid Commun. Mass Spectrom.* **2004,** *18,* 1039.

(21) Josephs, J.L.; Sanders, M. *Rapid Commun. Mass Spectrom.* **2004,** *18,* 743.

(22) Schreiber, A.; Efer, J.; Engewald, W. *J. Chromatography A* **2000,** *869,* 411.

(23) Weinmann, W.; Lehmann, N.; Muller, C.; Wiedemann, A.; Svoboda, M. *Forensic Science International* **2000,** *2000,* 339.

(24) Weinmann, W.; Gergov, M.; Goerner, M. *Analysis* **2000,** *28,* 934.

(25) Maloney, V. *BioTeach Journal* **2004,** *2,* 92.

(26) Fernie, A.R. *Functional Plant Biology* **2003,** *30,* 111.

(27) Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. *Trends in Biotechnology* **2004,** *22,* 245.

(28) Villas-Boas, S.G.; Mas, S.; Akesson, M.; Smedsgaard, J.; Nielsen, J. *Mass Spectrometry Review* **2005**, 24, 613.

(29) Dunn, W.B.; Bailey, N.J.; Johnson, H.E. *Analyst* **2005**, *130*, 606.

(30) Roepenack-Lahaye, E.; Degenkolb, T.; Zerjeski, M.; Franz, M.; Roth, U.; Wessjohann, L.; Schmidt, J.; Scheel, D.; Clemens, S. *Plant Physiology* **2004**, *134*, 548.

(31) Brown, S.C.; Kruppa, G.; Dasseux, J.L. *Mass Spectrometry Review* **2005**, *24*, 223.

Chapter 5

Conclusions and Future Work

Mass spectrometry is an important tool for proteome and metabolome analysis, which plays a critical role in understanding the relationship between the genome and the phenotype in biological research. In proteome analysis, molecular mass information on proteins and peptides derived from the digestion of proteins as well as the fragment ion masses of peptides obtained from tandem mass spectrometry can be used for the characterization of proteins, such as amino acid sequence analysis and detection of PTMs and amino acid variants. In metabolome analysis, mass information on metabolites generated by MS and fragment ions of metabolites from MS/MS analysis can be used for the identification of metabolites.

In Chapter 2, microwave-assisted acid hydrolysis (MAAH) of proteins combined with 1-D SDS PAGE separation was reported. The SDS PAGE technique has been used widely as a powerful tool for protein separation. Acid hydrolysis of a protein under microwave exposure, followed by MS analysis for protein sequencing, requires a pure protein sample. Therefore, in this chapter a method of combining 1-D SDS-PAGE and MAAH was investigated. As a bridge between PAGE and acid hydrolysis, a passive elution method was used to extract protein molecules from gel pieces. Various influencing parameters were studied to obtain optimal extraction conditions with high efficiency and low impurities; and the amino acid sequences and modifications of horse myoglobin and bovine α-casein were studied. The results indicate that this method can be used for protein identification in the low mass region with acceptable mass accuracy and resolution. However, this method is difficult to be used in a wide mass region, because full sequence coverage is difficult to achieve. High mass accuracy and resolution over a wide mass range is limited due to the detection limitations of the instrument. Furthermore, gel-induced impurities, such as SDS, salts, and gel-staining chemicals, also lead to peak broadening and reduction of the signal-to-noise ratio.

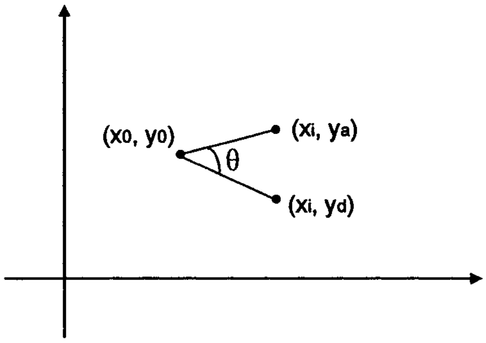
In Chapter 3, MAAH was applied to the identification of the mutation position in a hemoglobin variant. As a real world example, variant hemoglobin G Coushatta was extracted from human red blood cells and analyzed using the MAAH approach. The mutation of 22E→22A with a mass difference of 58 Da was identified. Due to the employment of reflectron mode in MALDI-TOF MS analysis, high spectral resolution and mass accuracy could be obtained to define the variant site. Compared with traditional tryptic digestion of protein followed by peptide mapping and MS/MS analysis of tryptic peptides, the process of MAAH is quick and easy for identifying amino acid mutations.

In Chapter 4, a mass spectrometric tool was developed for metabolomics. An MS/MS spectral library of 215 human metabolites was constructed at three levels of CID energy. Various experimental parameters were investigated to determine the independency of this library. The limits of detection were studied for MS/MS spectral generation and it was found that at a metabolite concentration of as low as 1 μM quality fragment ion spectra could still be obtained. A mathematical algorithm based on the comparison of fragment ion masses and intensities was proposed for scoring the match between an acquired spectrum and those in the library. This scoring method was used to assess the spectral matching reproducibility. The results showed good spectral matching could be obtained at different concentrations of metabolites. The spectral matching method, combined with LC MS/MS, was applied to the identification of human urinary metabolites and 15 metabolites were identified successfully. This result indicates that the method described is a useful tool for reliable identification of metabolites in complicated samples.

In the future, with improvements in mass spectrometric analysis methods and instruments, as a top-down method, microwave-assisted acid hydrolysis of proteins can be employed in wider applications in proteome analysis, particularly in the area of characterizing PTMs.

The metabolite MS/MS spectral library is still at a small scale. Future work will involve the expansion of this library to include all known human metabolites. In addition, the searching algorithm needs to be further optimized for automated data processing and metabolite identification. This method of spectral library creation can be extended to other types of mass spectrometers which may provide better mass resolution and mass measurement accuracy, hence improving the confidence for metabolite identification based on MS/MS spectral matching.

Appendix I: Algorithm used for the Fit/RFit/Purity



x_i is the compared m/z, y_d is the intensity at the compared m/z in the library spectrum, and y_a is the intensity at the compared m/z in the acquired spectrum. x_0 is the average of x_{min} and x_{max}, y_0 is the intensity at m/z = x_0, usually it is zero.

As Cosθ always falls into [1, 0] when θ is within [0°, 90°], we introduced Cosθ to compare the similarity of two spots in two spaces. Herein, two spots in two spaces correspond to the two intensities in acquired spectrum and library spectrum at the same m/z.

The definition of Cosθ (i) and w (i) are given as below,

$$\mathrm{Cos}\theta\ (i) = \frac{(x_i-x_0)^2 + (y_a-y_0)\,(y_d-y_0)}{\sqrt{(x_i-x_0)^2 + (y_a-y_0)^2} \times \sqrt{(x_i-x_0)^2 + (y_d-y_0)^2}}$$

$$w\ (i) = \frac{y\ (i)}{\sum_{i=1}^{N} y\ (i)}$$

Fit is calculated as,

$$\mathrm{Fit} = \sum_{i=1}^{N} [w\ (i)\ \mathrm{Cos}\theta\ (i)]$$

i is every m/z appearing in database spectrum.

RFit is calculated as Fit except that x_i is the m/z appearing in acquired spectrum.

$$\mathrm{RFit} = \sum_{i=1}^{N} [w\ (i)\ \mathrm{Cos}\theta\ (i)]$$

Purity is calculated as Fit and RFit except that x_i is the m/z appearing in either the acquired spectrum or the library spectrum and w_i is the density with respect to the sum of acquired and database density.

$$\mathrm{Purity} = \sum_{i=1}^{N} [w\ (i)\ \mathrm{Cos}\theta\ (i)]$$