Advancement of LC-MS Based Proteomic and Metabolomic Techniques

By

Zhendong Li

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

University of Alberta

Abstract

The simultaneous quantification of all biological molecules, from metabolites to proteins, holds the key to discovering molecular signatures that can diagnose diseases before irreversible damages have occurred in the body.

Recent advances in the field of liquid chromatography coupled mass spectrometry (LC-MS) have fueled the development of two fields pursuing disease diagnosis: metabolomics, the study of metabolites; and proteomics, the study of proteins. Prior to LC-MS, it was difficult to quantify more than several metabolites or proteins in a single analysis; now, thousands of molecules can be monitored in a single LC-MS assay, making it possible to detect minute abundance patterns that are characteristic of diseases. The key for metabolomics and proteomics is sensitivity and efficient processing of large amounts of data.

The goal of my research was to improve LC-MS based methodology for the detection of both metabolites and proteins. For method development in proteomics, my thesis focused on two areas: improvements to a database search engine used to identify peptides, and increased peptide ion intensity by introducing chemical vapours. Firstly, by implementing a machine learning algorithm that better distinguishes correct and incorrect peptide identifications from database searching results, more valid peptide were recovered in proteomic experiments. Next, vapours of various chemicals were introduced to the source region of a mass spectrometer; both the absolute ion intensities and number of peptide detected were increased.

In the field of metabolomics, my work focused on solving the problem of unknown identification and improving sensitivity of chemical-isotope-labeling (CIL) metabolomics. With thousands of unknown metabolites being measured in every metabolomic experiment, confident

high-throughput identification is required. A high-quality human metabolite reference tandem MS spectral library was created for 800 compounds. The data in this library has higher resolution and higher accuracy than any other available library; therefore, LC-MS identification can be more confident. Reversed phase retention time information was also added to this library, which helps to distinguish isomers and further improve identification confidence.

Lastly, nanoliter flow rate LC-MS was optimized for the analysis of CIL metabolomics. CIL metabolomics, is a metabolite labeling strategy that uses stable isotope encoded labeling to improve quantification by LC-MS. The labels used in this technique already improved sensitivity of metabolite detection; however, when samples are limited or dilute, more sensitive instrumentation is needed. By reducing the flow rate and column dimensions, sensitivity of LC-MS was improved, and more metabolites can be detected per sample.

Preface

Chapter 2 of this thesis has been published as "Combining Percolator with X!Tandem for Accurate and Sensitive Peptide Identification." Xu, M.; Li, Z.; Li, L. J. Proteome Res. 2013, 12, 3026-3033. Mingguo Xu designed the experiments, processed the data, and wrote the manuscript. I contributed towards design, programming and partially towards data processing. Professor Liang Li was the supervisor and was involved with concept discussion, experimental design and manuscript editing.

Chapter 3 of this thesis has been published as "Chemical-Vapor-Assisted Electrospray Ionization for Increasing Analyte Signals in Electrospray Ionization Mass Spectrometry." Li, Z.; Li, L. Anal. Chem. 2014, 86, 331-335. I was responsible for the design, data collection, analysis, as well as the manuscript writing. Professor Liang Li was the supervisor and was involved with concept discussion, experimental design and manuscript editing.

Chapter 4 was an international collaboration with Professor Wishart's group and with Dr. Aiko Barsch from Bruker Daltonics (Bremen, Germany). Professor Liang Li, Dr. Aiko Barsch, Mingguo Xu, and I were responsible for the initial library discussions and method development. Edison Dong supplied the metabolite standards. Mingguo Xu, I, Jaspaul Tatlay, Tran Tran Ngoc, Tao Huan, Dr. Yiman Wu, Dr. Ruokun Zhou, Dr. Chiao-Li Tseng, and Wei Han contributed towards the data acquisition and data processing. I processed the urine data and built the library with help from Dr. Aiko Barsch. For Chapter 5, I processed the samples and collected all the data.

Jaspaul Tatlay and I contributed to the experimental design and data interpretation of Chapter 6. Jaspaul prepared samples, collected the data, processed the data and edited the manuscript. I wrote the manuscript and helped with data collection and data processing. Professor

Liang Li was the supervisor and was involved with concept discussion, experimental design and manuscript writing. A manuscript has been submitted for review on August 20th, 2015.

Acknowledgements

I am grateful for the guidance and education I was given by my supervisor, Professor Liang Li. It was through my experience in his world-class laboratory, at the University of Alberta, that I have become the researcher that I am today. No doubt that everything I do in my future will remind me of my experience here. I aim to pass the knowledge I have learnt from Professor Li to future generation of mass spectrometrists.

I would also like to thank my supervisory committee members, Professor Charles A. Lucy and Professor Robert E. Campbell, for offering me their wisdom and keeping me on track for the last five years. Special thanks to Professor Lucy for his dedication and time for improving my teaching abilities through the GTL program. The skills I have learnt from him have been valuable and will help me into the future. I would like to thank Professor Todd Lowry, Professor Jonathan Curtis for participating in my oral examinations and reviewing my thesis. Thanks to Professor Derek Wilson from York University for serving as my external examiner.

I am also grateful for the terrific group members I had the honors of learning together and being friends with for the past five years. In no particular order, thank you to: Dr. Mingguo Xu (my other mentor), Jaspaul Tatlay, Tao Huan, Dr. Helen Wang, Chad Iverson, Dr. Yiman Wu, Dr. Ruokun Zhou, Tran Tran Ngoc, Dr. Chiao-Li Tseng, Wei Han.

The HMDB MS^2 library work would not have been possible without the wonderful collaboration with Dr. Aiko Barsch from Bruker Daltonics in Bremen, Germany. Aiko's expertise and tolerance toward working across large time differences are very much appreciated.

Thanks to Professor David Wishart and Edison Dong for providing the invaluable HMDB standards that was used to create the MS^2 library. Special thanks to Edison for preparing all of the 800 standards.

Thanks to the mass spectrometry lab, Dr. Randy Whittal, Jing Zheng, and Béla Reiz, for their training and friendship. Also thanks to the electronics shop, Al Chilton and Kim Do, for letting me use their workspace and guiding me with all things electronic. Thanks to machine shop, Dirk Kelm, Vince Bizon, Dieter Starke, and Paul Crothers, for the projects that they helped to complete. Also thanks to the support staff in the Department of Chemistry for getting me through my degree.

Lastly, I would like to thank my wife, Mrs. Lily Trieu, for her unwavering support. To my parents, Dr. Tingsheng Li and Mrs. Zhi Ju, no words can express my gratitude towards the sacrifices you two have made so that I may be where I am today. I hope I have made you proud.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

%RSD	Percent relative deviation
AC	Alternating current
APCI	Atmospheric pressure chemical ionization
API	Atmospheric pressure ionization
APPI	Atmospheric pressure photoionization
BnOH	Benzyl alcohol
BSA	Bovine serum albumin
CAD	Collisional activation dissociation
CID	Collision-induced-dissociation
CIL	Chemical-isotope-labelling
CRM	Charge residue model
DC	Direct current
DDA	Data dependent acquisition
DmPA	*p*-dimethylaminophenacyl
DMSO	Dimethyl sulfoxide
EI	Electron impact

ESI	Electrospray ionization
EIC	Extracted ion chromatogram
FIA-MS	Flow injection analysis coupled mass spectrometry
FDR	False-discovery-rate
FT-ICR	Fourier Transform Ion Cyclotron Resonance
FWHM	Full-width-at-half-maximum
GC	Gas chromatography
GC-MS	Gas chromatography coupled mass spectrometry
Glu-Fib	[Glu1]-fibrinopeptide
HCD	Higher-energy collisional dissociation
HILIC	Hydrophilic interaction liquid chromatography
HMDB	The Human Metab olite Database
HPLC	High performance liquid chromatography
ICAT	Isotope-coded affinity tag
IEM	Ion evaporation model
IT-MS	Ion trap
iTRAQ	Isobaric tags for relative and absolute quantification
LC	Liquid chromatography

LC-MS	Liquid chromatography coupled mass spectrometry
Leu-Enk	Leu-enkphalin
MS	Mass spectrometry
MS^2	Tandem mass spectrometry
mLC	Microbore liquid chromatoraphy
nESI	Nano electrospray ionization
nLC	Nano-liquid chromatography
NMR	Nuclear magnetic resonance spectroscopy
PAGE	Polyacrylamide gel electrophoresis
PCA	Principal-component-analysis
PCR	Polymerase chain reaction
PEP	Posterior error probability
pI	Isoelectric Point
PLS-DA	Partial-least-squares discriminant-analysis
PSM	Peptide-spectrum-match
PTM	Post translation modification
QC	Quality control
QTOF	Hybrid quadrupole time-of-flight

RP	Reversed phase
RP-LC	Reversed phase liquid chromatography
RT	Retention time
TIC	Total ion count
TMT	Tandem Mass Tag
TOF	Time-of-flight
UHPLC	Ultrahigh performance liquid chromatography
UV	Ultraviolet

List of Symbols

$[M+H]^+$	Singly charged protonated molecular ion
$[M+Na]^+$	Singly charged sodiated molecular ion
$[M-H]^-$	Deprotonated single charged molecular ion
°C	Degrees Celsius
A	Aqueous mobile phase
B	Organic mobile phase
CF	Retention time correction factor
Da	Daltons
eV	Electron volts
Hz	Hertz
k	Chromatographic retention factor
l	Intensity of the library peaks
L_{eff}	Effective ion flight path length
m/z	Mass to charge ratio
min	Minutes
N	Theoretical chromatographic plate number

p	Probability
pH	Power of hydrogen
ppm	Parts per million
R^2	Square of the correlation coefficient
R_s	Chromatographic resolution
RT_{guard}	Retention time of the standard using the column with guard column
$RT_{no\ guard}$	Retention time of the standard using the column without guard column
S/N	Signal to noise ratio
SCX	Strong cation exchange
t	Time
t_r	Retention time of the peak in minutes or seconds
U	Accelerating voltage, or DC voltage
u	Intensity of the experimental peaks
V	AC voltage
v/v	Volume/volume percentage
$W_{0.5}$	Full width at half maximum of the peak in minutes or seconds
α	Chromatographic selectivity
φ	Percent organic mobile phase

ω	Angular frequency of alternating current

Chapter 1

Introduction

1.1 History of Mass Spectrometry

The age of the atom has shaken the foundation of humanity. This era has brought with it innovations, but also tragedies. Despite trials and tribulations, humanity is thriving today because of the discoveries that were made in this era. Modern analytical mass spectrometry (MS) evolved through a winding path of incremental breakthroughs in the study of the atom.

There were few indications from those first experiments on charge particles that the ability to manipulate these particles would be a powerful tool for the study of molecules. Due to the pioneering work of early mass spectrometrists, such as Alfred Nier,[1] mass spectrometry spread beyond the confines of the physics laboratory and into fields like biology and chemistry. Since that time MS has been used to study chemical systems that were once impossible to observe—in the human body, underwater,[2] or on the surface of planets hundreds of millions of kilometers away.[3] New developments in mass spectrometry promises lifesaving discoveries for understanding and treating diseases.

The following sections will introduce the concepts of MS based "omics" that is the foundation of this thesis.

1.2 Proteomics

All living organism rely on proteins as the engine of life. Proteins are 3D polymers made of amino acids monomers; the sequences of amino acids are encoded according within the genetic

materials of organisms.[4] Proteins take on crucial roles such as providing structural integrity, signaling, and catalyzing biochemical reactions.[5]

As with all engines, there must be the correct number of parts—with the correct specifications—in order for the system to run smoothly. Proteins that do not perform their function cause problems in the system that eventually become diseases. Therefore, monitoring the level of proteins in the body can help diagnose diseases. However, proteins shouldn't be studied individually, as they are part of a complex machinery of other proteins.[6] The identity, quantity, and interactions of all the proteins within a system at one point in time can give us an accurate picture of the condition of that system. For this purpose, the field of proteomics was born.[7]

Proteomics is the comprehensive study of all proteins in an organism. To achieve this goal, there are several principal challenges. Firstly, proteins cannot be amplified through a process like polymerase chain reaction (PCR) that can make exponential copies of the DNA fragment of interest.[8] Therefore, proteomic techniques are limited by the sensitivity of the chosen method of detection. Secondly, the structure, function, and identity of a protein are determined by its amino acid sequence, and this sequence cannot be easily determined compared to genomic sequencing techniques.[9] Lastly, all the proteins in an organism are encoded in their genes; therefore, genomics can already provide a comprehensive list of all the proteins that can be produced in the organism. With these challenges in mind, why do we still pursue the study of the proteome?

While genomics provides sequence information for all possible proteins, it cannot tell us the quantity and identity of proteins at any given time. Diseases often occur after an organism's DNA is already set, and disturbances to the proteome is not reflected in the genome. Another issue is the prevalence of post translation modification (PTM) of proteins,[10] which also cannot be predicted by genomics. PTM are small functional groups that are added to the amino acid residues

of proteins, and they change the function of the protein as needed by the organism. The most common, and most studied, is the phosphorylation.[11,12] PTM is a highly dynamic process under the control of other proteins, and a large variety of functional groups can be attached; as a result, PTM analysis cannot be studied by genomics. For these reasons, researchers have taken up the challenge of analyzing the proteome of cells and higher organisms to understand the role of proteins in diseases.[13-15]

The tools used to piece together the proteome has evolved in sophistication over time—and progressively increased in cost. The first system wide studies of proteins were completed using polyacrylamide gel electrophoresis (PAGE), where the entire protein content was separated on a slab of polyacrylamide gel based on charge or size of the proteins and visualized by staining with dyes.[16] The pattern on the gel gave information on the quantity and the types of proteins present; this allowed for comparison of the proteome between different samples. However, this technique was labour intensive and had poor reproducibility between analyses.[17] Rapid identification of large number of proteins on the gel was impossible, which added to the challenges of the technique.

In the past two decades, advances in the field of mass spectrometry and increases in computing power allowed researchers to overcome the deficiencies in PAGE.[18] In MS based proteomics, the first major step is to digest all of the proteins with proteases, such as trypsin, into small peptides. Because the starting analyte are peptides instead of proteins, this technique is termed bottom-up proteomics; analyzing the protein without digestion is termed top-down proteomics. The resulting mixture of peptides are injected onto a liquid chromatography instrument (LC) with a reversed phase (RP) column, and separated according to hydrophobicity (Section 1.5). The purpose of this separation is to reduce the sample complexity, and the retention mechanism is not important. The small peptides are more compatible with separation on RP-LC columns than

intact proteins, and can be detected within the optimal mass range of most MS. Peptides eluting from the HPLC are ionized by an atmospheric pressure ionization (API) source and enter the MS. Within the MS, peptides are fragmented by collision-induced-dissociation (CID MS^2) process in a collision cell to determine the sequence of the peptides and proteins (Section 1.7.1). Quantification of proteins can be done by measuring the intensity of the peptide signals,[19-21] counting the number of times a peptide is identified by a fragment spectrum,[22-24] or through the use of isotope encoded labels.[25-28]

Biological samples typically contain thousands of proteins. After protease digestion, the number of peptides increase to create a complex mixture. In a typical liquid chromatography coupled mass spectrometry (LC-MS) experiment, more than 5000 fragment spectra can be recorded. If samples are fractionated, and each fraction analyzed separately, the total number of spectra can be in the 10's to 100's of thousands.[29] Each fragment spectrum needs to be interpreted to determine the peptide sequence, which makes manual data processing impossible, due to the large number of data. Therefore, majority of proteomic studies rely on automated search engines to quickly process and identify the large amount of fragment data from each experiment (Section 1.8.1).

The development of proteomic technology has enabled the in-depth study of proteins in many biological systems, including human diseases. Continued breakthroughs depend on improvements made in all steps of the workflow outlined above. One of the most important aspect is the computer processing of LC-MS data. Chapter 2 outlines a method of improving the accuracy and sensitivity of peptide and protein identification, by implementing a machine-learning algorithm.

1.3 Metabolomics

MS has not only lead the revolution in the investigation of proteins, it has also fueled the developments in the field of metabolism. While proteins make up the biological engine, small molecules under 1000 Da are the fuel and the building blocks. A single metabolite is usually utilized in an entire pathway comprised of multiple metabolic reactions, catalyzed by enzymes; the metabolite is transformed at each step in order to perform some useful work for the organism. For example, there are approximately 50 reactions in central carbon metabolism that transforms glucose into cellular building blocks.[30] These crucial pathways are often highly conserved between organisms, changing very little throughout evolution. Knowledge on model systems can be extrapolated to human metabolism.

Since metabolism and metabolites are the foundation of cellular activity, dysregulation in the metabolic flux can result in life-threatening conditions. Due to progress made in this field, hundreds of treatable metabolic diseases can be tested for in the blood of newborn babies, and corrective actions can be taken. These tests have now become routine, and contain measurements for anywhere between 10-30 metabolites depending on the country.[31]

Metabolic pathways are highly diverse and interconnected, with a large variety of metabolites taking part in this system.[30,32] Targeting a few metabolites for fundamental research on metabolic diseases is difficult; with so many possible metabolites, the chances of finding the few that plays dominant roles in diseases are vanishingly low. Also, much like proteomics, genomic data plays a useful part in modeling the metabolic phenotypes in cells,[33,34] but they are still far too inaccurate to make significant contribution to our understanding of complex diseases in higher organisms like humans. Therefore, the ability to directly quantify and identify large numbers of

metabolites in the entire system—the metabolome—is crucial for major discoveries in metabolism research. This field is now known as metabolomics.

Prior to the proliferation of MS in labs that study metabolism, metabolites were measured on at a time using traditional chemical reactions and titrations. Often, these tests were laborious and time consuming; furthermore, they could only target one metabolite at a time, and cannot give a system wide perspective on the metabolic processes. The introduction of nuclear magnetic resonance spectroscopy (NMR) allowed researchers to quickly quantitate 10-100 metabolites within a sample in one analysis.[35,36] The advantages of NMR are: universal response towards all chemical classes, and usually highly robust and reproducible due to low instrument drift. However, signals from metabolites are often convoluted, which limits the number of metabolites that can be accurately analyzed. Researchers have to rely on spectral pattern recognition tools to quantify metabolites.[37,38] Furthermore, sensitivity of NMR is low when comparing with MS. As a result, while NMR is a powerful tool for metabolomics, it cannot cover the entire metabolome sufficiently.

LC-MS changed the depth at which the metabolome can be studied. It advanced the metabolomics field in similar ways it advanced proteomics, by quantifying thousands of metabolites with one analysis. This allowed studies to collect comprehensive data on disease metabolism across large number of samples, so conclusions can be based on sound statistical basis. For these reasons, LC-MS has mostly replaced NMR in metabolomic laboratories around the world.

The workflow of LC-MS based metabolomics is similar to that of LC-MS based proteomics, but with several differences in sample preparation, data acquisition, and data processing. Since proteins are not the target for metabolomics, their presence complicate the interpretation of collected data—they must be removed prior to analysis. Samples are first deproteinated using

organic solvent.[39] The proteins precipitate and are packed into a pellet with micro-centrifugation. The supernatant, containing metabolites, are collected for study.

Total metabolite concentrations in biological fluids (especially urine) can be highly variable due to water intake, and does not reflect a dysregulation of metabolism. In order to compare metabolite levels between individuals, the metabolite concentrations must be first normalized. The total metabolite concentration can be normalized to an endogenous reference compound (e.g. creatinine[40]) or by total signal (e.g. total ion count (TIC),[41] ultraviolet (UV) absorption area[42]). Following normalization, samples are ready to be analyzed by LC-MS.

Samples are first injected on an LC for metabolite separation, which reduces sample complexity and provide identification information based on retention time (RT). Different column types can be used to alter the RT behavior of LC columns, allowing different classes of metabolites to be targeted (Section 1.5). The LC is coupled to the MS by an API source, and mass to charge ratio (m/z) of ionized metabolites are measured as they elute from the LC. Data extraction software then aligns the metabolite abundance information across multiple samples based on accurate m/z, correcting for small retention time drifts, and reports intensity data for all samples.[43-46] The intensity information for all common metabolites found in the sample set are then input into a multivariate analysis software that will find the largest variation in the data. Most commonly, multivariate analysis, such as principal-component-analysis (PCA) or partial-least-squares discriminant-analysis (PLS-DA), is used to find metabolites that can be used to distinguish between healthy and diseased patients.[47] These significant metabolites can then be identified by using tandem mass spectrometry, or by accurate mass.

Without identification, researchers cannot place the significant metabolites within the context of the metabolic pathway, and answer the question of how these metabolites cause or

indicate diseases. This context is important because LC-MS data are complex multidimensional data, with many sources of noise. These noises can be either instrumental—such as contamination—or biological—such as patient treatments being mistaken for biomarkers of diseases. Without the logical validation based on correct identity of the metabolites, the chances of wasting resources on spurious noise as biomarkers increase. The projects reported in Chapters 4 and 5 utilized the large library of human metabolite standards at the University of Alberta to create a large metabolite spectral and retention time library to improve metabolite identification in future studies.

1.4 Chemical Isotope Labeling

An extension of metabolomics described in the section above is to use derivatizing reagents to improve the quantification and sensitivity of metabolites. The derivatizing reagents can carry enriched carbon or hydrogen isotopes to create a heavy label, while derivatizing reagent without enriched isotopes are light labels. A sample can be labeled with the heavy label, and a reference can be labeled with the light label; after combining the two, sample and reference can be measured in the same MS spectrum—reducing the analytical variation. This type of metabolomics analysis has been referred to as chemical-isotope-labeling (CIL) metabolomics. Our group has been at the forefront of this technique in the field of metabolomics.

CIL metabolomics is conducted by labeling a sample with a derivatizing reagent targeting a broad submetabolome, such as amines and phenols when using dansyl chloride[48] or carboxylic acids using DmPA.[49] In parallel, a reference sample of very similar composition but distinct from the sample (most commonly made by pooling all available samples) is labeled with a $^{13}C_2$ isotopologue of the derivatizing reagent. The derivatized sample and reference are then mixed together and injected into an LC-MS. Each distinct metabolite appears as a doublet of peaks of

exactly 2.00672 m/z units apart (Figure 1.1). Accurate relative and absolute quantification results for thousands of metabolites can be obtained from a single experiment. Related isotopic labeling techniques exist for proteomics; up to 10 peptide samples can be quantified in a single mass spectrum with isobaric labeling (iTRAQ,[50] ICAT,[51] TMT[27]).

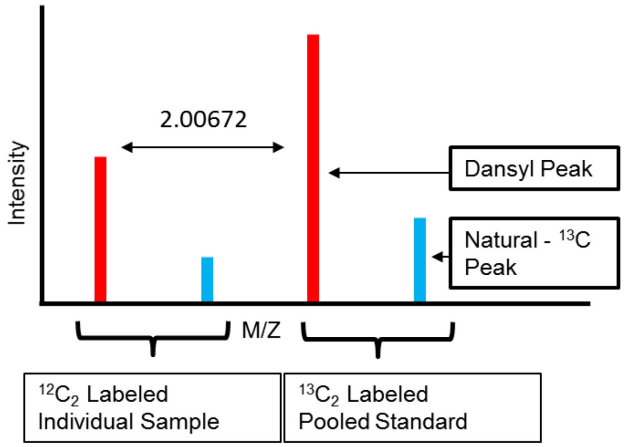


Figure 1.1 An illustration of a CIL labeled metabolite's peak doublet. Relative quantification is calculated as a ratio of individual sample peak height over pooled standard peak height.

The elegance of CIL lies in its ability to improve metabolome quantification through not just one but several key mechanisms. Signal intensity is enhanced through the inclusion of a hydrophobic naphthalene moiety that increases ESI surface activity and a tertiary amine that increases protonation. For certain analytes, the dansyl label was shown to increase signal intensity between 10 to 1000-fold.[48] The hydrophobic moiety also normalizes retention time behaviour on RP-LC columns so that polar compounds and non-polar compounds can be separated on the same C_{18} column, allowing for column standardization.

With differential isotope labeling there is an internal isotopic standard for each detected metabolite; the resulting peak doublet further improves quantification, and can be used for filtering out chemical and instrument noise. Finally, this technique is easily accessible to any mass

spectrometry lab. Standard labeling procedures, sample normalization procedures,[42] reagents, and data processing software (IsoMS[52]) are available and can be integrated into existing workflows. All of these factors improve detection sensitivity and boosts data quality of metabolomic studies.

1.5 Liquid Chromatography

Prior to the ionization of proteomic or metabolomic samples, an LC system is used to separate the metabolites or peptides in the sample. In this crucial step, sample is pushed by the mobile phase solvent from a liquid pump onto a LC column. The analytes then interact with the stationary phase, which are anchored on micrometer-sized particles packed in the column.

The mobile phase and the stationary phase compete for the analyte; if there is a stronger interaction with the mobile phase than stationary phase, the analyte will be carried along by the solvent and eluted from the column quickly. If interaction with the stationary phase is stronger, then the analyte will be held in the stationary phase more and elute much later in time. Separation occurs because different analytes have different amount of interaction between the two phases, and will migrate through the column at different rates.

Due to the additional analysis time required to perform LC separation, certain applications such as lipidomics[53,54]—where the samples are pretreated and have reduced complexity—can be directly analyzed by MS without LC separation. This is not the case for most complex biological samples, and the primary purpose of LC separation is to reduce the chemical complexity.

Without reducing complexity, high abundance analytes can cause the suppression of lower abundance analytes in API sources.[55] By separating with the LC, high abundance analytes can be eluted away from low abundance ones, reducing suppression. Furthermore, because analytes of

interest are retained on the RP-LC column, they elute far away from biological matrix components such as inorganic salts that can cause suppression.

Analytes elute from the LC column as Gaussian peaks, and the goal of LC is to ensure that these peaks are separated from each other. Although it is impossible to completely separate all metabolites or peptides in a complex mixture, it is still desirable to minimize the amount of overlap. The degree of LC separation is often described as resolution (R_s), described by Equation 1.1.[56]

$$R_s = \frac{1}{4}\left[\frac{k}{1+k}\right]\left[\frac{\alpha-1}{\alpha}\right]N^{0.5} \tag{1.1}$$

k is the retention factor; larger k indicates the analyte is retained more on the column. α is the selectivity; greater retention time difference results in larger α. N is the plate number; the narrower the peaks the larger the N. Equation 1.1 shows that resolution can be improved in several ways: increase in retention (k), increase in selectivity (α), and increasing N. These parameters can be optimized by changing the mobile phase solvent composition and column selection and stationary phase type. The detailed methods of optimizing these parameters have been thoroughly covered by Snyder et al.[56]

In the field of metabolomics and proteomics, there are two main types of LC column stationary phases, each with different retention mechanisms. The first type, RP-LC, is the primary stationary phase type used for proteomics, and the majority of metabolomics. It retains hydrophobic compounds and does not retain highly hydrophilic compounds. The second is hydrophilic interaction liquid chromatography (HILIC), which retains hydrophilic compounds that are prevalent in metabolomics analysis. Due to the wide variation of analyte polarity, the two stationary phase types can offer complementary data, and the choice of one over the other depends on the

hydrophobicity of the analytes of interest. The following two sections describes the two phases in more detail.

1.5.1 Reversed Phase

The base for the majority of modern day LC column packing material are porous fused silica particles of 1.7 to 5 μm in diameter; they act as a solid support for the stationary phase. Porous organic polymer solid supports have also been tested and they are more resistant to hydroxide attacks than normal silica particles.

The stationary phase is first chemically bonded onto the support.[57] Careful control during the chemical bonding step is crucial in determining the retention behavior, because it can affect surface coverage of the particles by stationary phase.[58] After successful bonding of stationary phase with the solid support, the finished packing material is then packed into stainless steel columns at high pressures to ensure the even distribution of particles within the column.

For RP-LC stationary phases, alkyl chains of varying lengths are commonly used (C_4, C_8, C_{18}). The alkyl chains form a thick layer in which the hydrophobic analytes can partition or adsorb, depending on size.[56] C_{18} shows good retention of peptides and hydrophobic metabolites; as a result, it has become ubiquitous in the field of metabolomics and proteomics.

In the reversed phase separations, the strong mobile phase solvent that competes for the hydrophobic analytes is commonly acetonitrile. By adjusting the percentage of acetonitrile in water (φ), the strength of the mobile phase can be increased to elute more hydrophobic analytes. Due to the wide range of polarities in proteomic and metabolomic samples, very hydrophobic analytes takes too long to elute from the column when lower φ is used. If higher percentages are used, the

hydrophobic analytes will elute in the optimal time, but the less hydrophobic analytes will be unretained and co-elute at the beginning of the column.

To solve this issue, φ is often ramped up during the analysis, and this is referred to as gradient elution. The low φ at the beginning of the analysis ensures that less hydrophobic analytes are retained and elute far away from the unretained sample matrix. As φ increases during the analysis, more hydrophobic analytes elute from the column. By using gradient elution, analysis time can be shortened and can be applied to the analysis of a wide variety of analytes. For these reasons, the majority of LC-MS omics research take advantage of gradient elution.

Compounds with different chemical properties elute at different times; therefore, elution times can be used as a parameter for the identification of a compound. The challenge for using retention time for this purpose is the lack of reproducibility between labs and even between LC instruments. Small changes in tubing volume of the LC, or changes in gradient generation at the pump, can all have large impacts on the retention time. For these reasons, the use of retention time as an identification tool is still limited in the field of metabolomics. Chapter 5 describes the development of a library of retention times for human metabolites, and the creation of retention time normalization protocol to help make this library more portable to other labs.

1.5.2 HILIC

Metabolites have wide-ranging hydrophobicity; therefore, it is not possible to apply only reversed phase columns to separate them all. Hydrophilic compounds (e.g. sugars, nucleic acids, and amino acids) are biologically important and are important in metabolomics analysis. These Hydrophilic compounds cannot be analyzed by RP-LC, and crucial metabolic information would be missing. To retain and separate these classes of metabolites, polar stationary phase can be used

in HILIC mode. Due to its complementary separation mechanism, HILIC is often used in addition to RP-LC separation to provide comprehensive analysis of the metabolome.

In HILIC mode, the polar stationary phases commonly used is bare silica. Other polar bonded phases like amino, diols, zwitter ionic, and amides, have also been used with slightly different selectivity.[59,60] The mobile phase composition is the reverse of RP-LC. In HILIC, the strong solvent is water and the weak solvent is often acetonitrile. The acetonitrile percentage is lowered as the gradient proceeds and eluting the strongly retained hydrophilic compounds. Evidence has shown that the most likely mechanism of this retention is the formation of a thin water layer over the polar stationary phase, and analytes retain on the stationary phase by either partitioning into or adsorbing onto the water layer. Polar hydrophilic compounds retain strongly on this water layer, while non-polar hydrophobic compounds are not retained.

While HILIC separations are already widely used in metabolomics studies,[60,61] the large variety of stationary phase chemistries that have slightly different selectivity makes the selection of a "standard" HILIC phase will be unlikely in the future. Without a standard column, inter-laboratory sharing of data will be difficult.

1.5.3 Nano-LC

The most wide spread LC technique is high performance liquid chromatography (HPLC). Analytical columns used in HPLC are typically between 2 mm to 4.5 mm in inner diameter, and are referred to as microbore columns. Flow rates can be anywhere between 200 µL/min to 1 mL/min. When there are sufficient amount of sample that can be injected onto the HPLC, this dimension of columns is perfectly suitable for metabolomic and proteomic analysis. However, there are situations in which samples are limited, HPLC is not sensitive enough. In order to increase

the sensitivity, the dimensions of the HPLC columns can be reduced to μm sizes, and flow rate down to nL. At these dimensions, HPLC becomes nano-LC (nLC)

The typical internal diameter of nLC columns is 75 μm, made possible by packing stationary phase in fused silica capillaries. The length and packing materials are kept the same as a standard analytical column. Flow rate is slowed down in proportion to the decrease in column dimensions, to 200 to 500 nL/min. The lower flow rate and smaller inner diameter reduces analyte dilution by the mobile phase, which increases the detection sensitivity. The linear velocity is kept the same as an analytical column, so column efficiency is similar to conventional HPLC columns. Due to the slower flow rates, the transfer of analyte ions into the gas phase becomes more efficient (Section 1.6.2), which also improves sensitivity.

Unlike HPLCs, nLC are usually performed with a sample trapping step prior to the analytical separation. In this step, samples in the microliter range are pushed through a short trap column and into the waste at high flow rates—usually in the μL/min range. Analytes retain on the stationary phase in the trap column, while matrix is diverted to waste. The analytes are slowly eluted onto the analytical column as the analysis begins. The purpose of this step is twofold: first, if microliter sized samples are injected directly onto the analytical column, it would take more than 10 minutes to load the sample onto the column at nanoliter flow rates—which increase analysis time; secondly, the trapping can remove most matrix material prior to ionization, which reduces ion suppression and source contamination. Therefore, sample trapping is a crucial step in nLC analysis.

nLC is used almost exclusively in the field of proteomics,[62,63] because of its higher sensitivity. The downside of small column and tubing dimensions is the fragility of nLC systems for day-to-day use. For this reason, nLC has not seen routine use outside of research institutions.

However, the sensitivity of nLC can of great use in fields other than proteomics. Chapter 6 leverages the advantages of nLC to develop a sensitive method for CIL metabolomics.

1.6 Mass Spectrometry

1.6.1 ESI Source

As analytes elute from the LC they are still dissolved in the mobile phase; before they can be analyzed by the mass spectrometer, they must be ionized and transferred into the gas phase. Due to the LC being in atmospheric pressure, the ionization source for the mass spectrometer are usually operated at atmospheric pressure. The majority of API sources utilizes electrospray ionization (ESI).[64] ESI's ability to produce intact ions of analytes (such as peptides and metabolites) and be compatibility with RP-LC and HILIC mobile phases, have been the reasons for its dominance as the main ionization source for the proteomic and metabolomic research.

The goal of ESI is to generate fine droplets that can transfer analyte ions into the gas phase. The ESI source is simple in design, consisting of a thin inner metal capillary housed within a larger metal tube (Figure 1.2). LC eluent flows through the inner metal capillary, and N_2 gas passes through the larger metal tube to aid in solvent evaporation. A high voltage potential (3-5 kV) is applied between the inner needle and the inlet of the mass spectrometer. In the metal capillary, the LC eluent contains protonated positive ions from the higher pH generated by formic acid mobile phase additive, and also from the electrochemical reaction at the electrode.[65,66] The applied electric field causes the ions to build up at the tip of the metal capillary, and distorts the liquid in the metal capillary to form the characteristic "Taylor cone" (Figure 1.2). A jet of ion rich liquid is emitted from the tip of the Taylor cone, towards the MS skimmer. Due to the columbic repulsion between ions, the thin jet soon destabilize and break apart into smaller micrometer sized droplets.

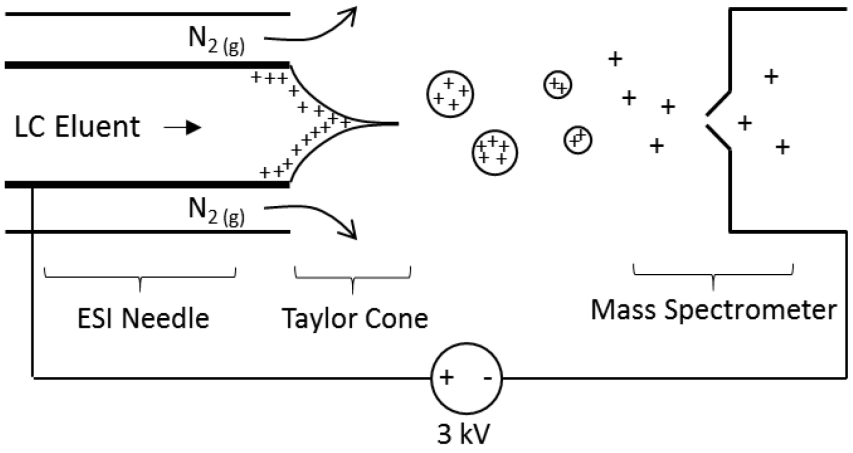


Figure 1.2 Electrospray ionization process in the positive mode. Adapted from Kerbarle et al.[55]

The exact mechanism of how ions are transferred from micrometer sized droplets into the gas phase and ultimately end up in the MS, has been the subject of many studies and reviews.[55,65,67] There are two prevailing theories on the matter, and both have been shown to occur depending on the type of analytes. The first theory is the ion evaporation model (IEM). As the solvent in the µm sized droplets evaporate, the droplets decrease in size and the ions contained in the droplets come in closer proximity to each other. Once the sizes have shrunk down to the Rayleigh Limit, the repulsive forces causes droplet to break apart and ions to be ejected from the droplets.

The second theory is the charge residue model (CRM) which was found to be more applicable to large polymers or biomolecules. In this theory, as the solvent evaporate from the droplets, charges adhere to the analytes. After all of the solvents evaporate, the charged analyte is left in the gas phase. This theory better explain the ionization process for larger biomolecules because it is energetically unfavorable for them to undergo IEM.[67]

1.6.2 Nano-ESI Source

In order to cope with the much lower flow rate of the nLC, the normal ESI source is shrunk down to smaller dimensions. The typical ESI metal capillary are usually in the hundreds of μm range; this large diameter will add too much post-column volume at nLC flow rates of around 300 μL/min—causing severe peak broadening. In nano-ESI (nESI), the metal capillary is switched with a pulled fused silica tip that can reach a tip diameter of less than 10 μm. This not only reduces the volume in the nESI source, it also improves the ionization of analytes. Smaller tips and lower flow rates can generate smaller droplets that evaporate quickly and yield more ions than ESI.[68,69]

Studies have shown that additives in the analyte solvent affects the nESI ionization process, mostly changing the charge of the analyte.[68,70,71] Chapter 3 improved the nESI process, by increasing ionization efficiency through the use of gas phase chemical in nano-spray ESI.

1.6.3 High-Resolution Mass Analyzers

The ionized analytes enters the low-pressure region inside the mass spectrometer, where m/z is measured and recorded. The m/z values are measured as Gaussian peaks; the resolution of the peaks determine the accuracy of the m/z value and the ability to distinguish similar masses. The accuracy and resolution of mass analyzers play a large part in the correct identification of peptides and metabolites.

High-resolution mass analyzers,[72] as defined in this thesis, are instruments capable of less than 5 ppm of mass error and greater than 30,000 full-width-at-half-maximum (FWHM) resolving power. Time-of-flight[73] (TOF), Fourier Transform Ion Cyclotron Resonance[74] (FT-ICR), and Orbitrap[75] instruments fit the definition of high resolution. However, all work in this thesis were conducted on a hybrid quadrupole time-of-flight (QTOF) instrument.[76]

QTOFs are, as their name suggests, a combination of a quadrupole and TOF in one instrument (Figure 1.3). The first quadrupole (Figure 1.3b) resides in the front of the instrument and acts as a gatekeeper, controlling the ions that enter; it does not function as a mass analyzer and is not used to record any data.

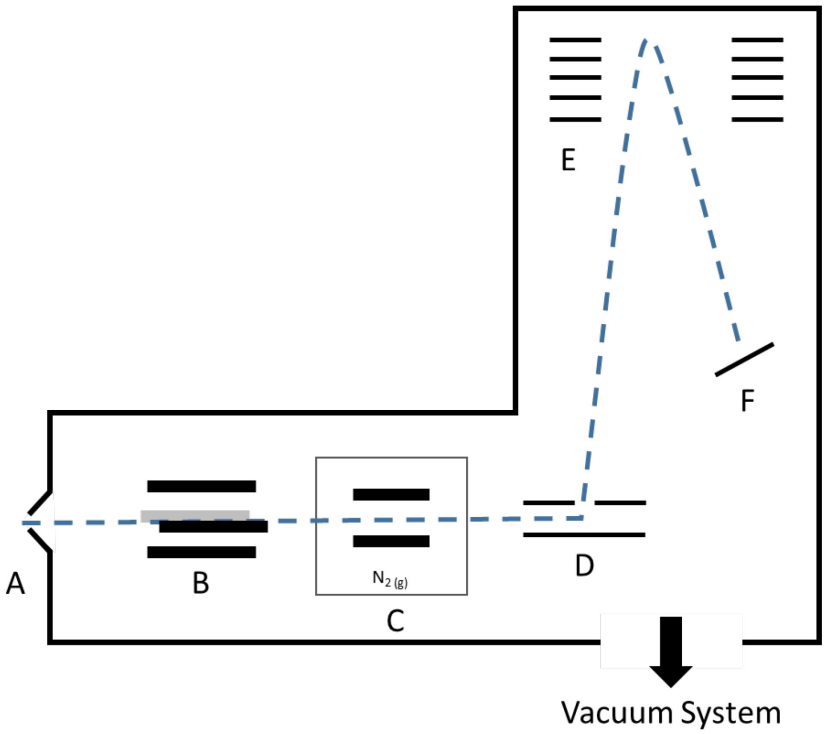


Figure 1.3 Simplified schematic of a QTOF mass spectrometer. Ion focusing funnels, and ion transfer hexapoles and lenses are not shown for simplicity. (A) Ion Inlet (B) Analytical quadrupole (C) Collision cell filled with nitrogen (D) Ion pusher/puller assembly (E) Reflector (F) Detector. Dotted line represents the ion path.

The quadrupole is a section of four parallel metal rods that are machined to be precisely uniform in dimensions, and held in parallel by ceramic spacers to create the quadrupole electric field; ions travel along the length of the quadrupole, in its center. A positive superimposed direct current (DC) and alternating current (AC) are applied to a pair of rods that are facing each other on the y-axis (Figure 1.4a). The combined potential takes on the formula of $U+V\cos(\omega t)$, where U is the DC voltage, V is the AC voltage, and ω is the angular frequency of the AC component.

Assuming we are operating in positive mode with positive ions; the constant positive DC voltage pushes all the ions to the center as they travel along the z-axis. The positive AC current destabilizes the smaller ions by pulling ions toward the rods, while heavier ions travel through along the z-axis undisturbed. These two rods act as the high-pass filter, removing all light ions.

The other two rods have the same superimposed AC and DC potentials, but of the opposite polarity. The negative DC voltage destabilizes heavier ions, while the negative AC voltages stabilizes lighter ions; this forms a low-pass filter that removes heavier ions, allowing the lighter ions to pass. By controlling the DC and AC potentials, the combined effect of the two polarities is that only a small selection of m/z are allowed to pass the quadrupole,[76,77] as shown in Figure 1.4b. This is how a single analyte can be isolated and fragmented to give a MS^2 spectrum. If no DC potential is applied (U=0), ions of all m/z are able to pass through the quadrupole.

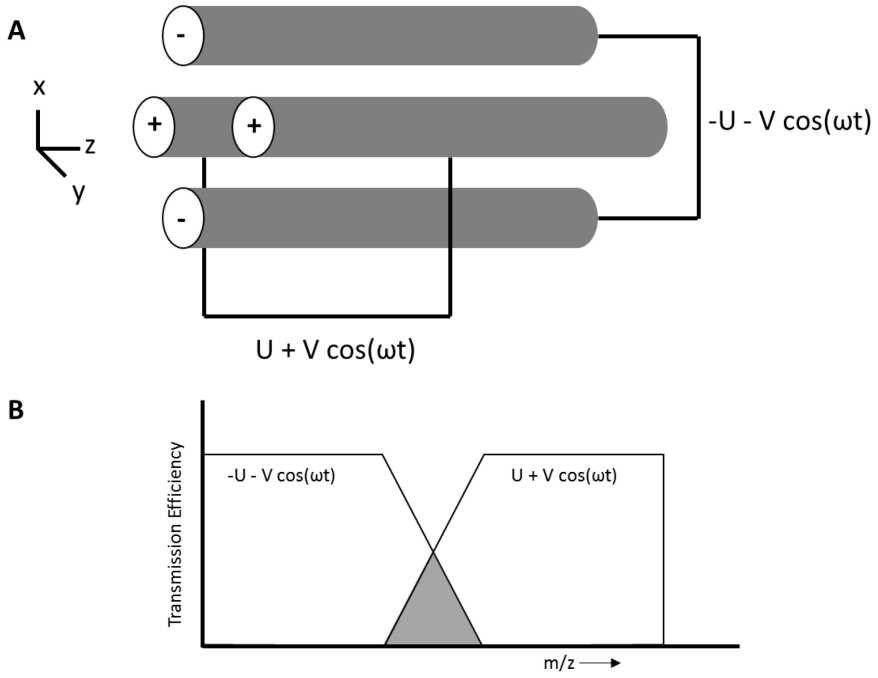


Figure 1.4 Simplified schematic of a quadrupole and its mechanism. (A) Schematic of a quadrupole. A superimposed AC and DC potential is applied to rods opposing each other. The rods adjacent to each other are of the opposite polarity. (B) A simplified

stability diagram illustrating the effect of the two sets of rods. The rods with the net positive polarity (U + V cos(ωt)) transmit high masses only. The negative rods (-U - V cos(ωt)) transmit low masses only. The grey area shows the masses that are stable and are pass through the quadrupole – due to the combined effect of the quadrupole.

As the ions exit the quadrupole they enter the collision cell (Figure 1.3c). The collision cell is another quadrupole held at higher pressure, between two accelerating plates. The cell is usually filled with either $N_{2\,(g)}$ or $Ar_{(g)}$. Ions are accelerated through the length of the collision cell, towards the exit and into the TOF analyzer. At low accelerating voltages the ions pass through unaffected, but at higher acceleration the ions collide with the gas molecules with enough force to fragment into smaller ion pieces (Section 1.7).

Once the ions enter the TOF region of the MS, they are pushed orthogonally by an accelerating voltage into the flight tube (Figure 1.3d). The amount of time (t) required by the ion to reach the detector is determine by the mass to charge ratio (m/z), the accelerating voltage (U), and length of the entire flight path (L_{eff}), as described by Equation 1.2.[76]

$$t = \frac{L_{eff}}{\sqrt{2eU}}\sqrt{m/z} \tag{1.2}$$

Most modern day QTOF MS includes a section of closely spaced metal plates, referred to as the reflectron, at the end of the flight tube to reflect ions to the detector (Figure 1.3e); this increases the resolution by increasing the flight path and reducing the spread of ions. The reflectron first decelerate a packet of ions of the same m/z, collapsing them into a tighter formation. The reflectron then accelerate the tight packet to the detector. Because the ions are packed tighter, a peak of reduced width is detected. With the reflectron, modern QTOFs have reached greater than 80,000 FWHM resolving power.

Once the ion reaches the detector, the flight time is recorded and converted into the m/z of the measured ion. Because flight time is dependent on the length of the flight tube, small changes in tube length will affect the accuracy of the measured m/z. Temperature changes in the lab changes the length by minute amounts, but they affect the QTOF's accuracy. In order to achieve less than 5 ppm of mass accuracy, every analysis needs to be internally calibrated. This can be accomplished by the periodic infusion of calibrant solution into the source region, as done by the Water's "Lockmass" system. Alternatively, a plug of calibrant solution can be injected at the beginning of every LC-MS run with a switching valve and additional pumps.

The success of QTOF in the scientific community is due to a combination of mass accuracy, fast scan speed, high resolution, and low initial and maintenance costs. QTOFs have been successfully used for both metabolomics and proteomics.

1.7 Tandem Mass Spectrometry

The accurate mass of the intact metabolite or peptide is not enough to ensure identification, even at below 1 ppm accuracy.[78] Tandem mass spectrometry (MS^2) is often performed to identify molecules through interpreting the unique fragment spectra after CID fragmentation. This is a simple and convenient identification method, because most modern MS have MS^2 function built into the instrument. The name and resulting spectra appears slightly different for each class of MS; CID for QTOF, collisional activation dissociation (CAD) for ion traps, and higher-energy collisional dissociation (HCD)[79,80] for Orbitraps. CID was used in this thesis because all data were acquired on QTOF instruments with CID.

In most cases, MS^2 is generated in the data dependent acquisition mode (DDA); the MS first scans the intact ions eluting off the LC (the precursor scan), then the analytical quadrupole

isolates a chosen ion for fragmentation. The resulting fragment spectrum contains only the fragments of the chosen ion, and not from any other sources. Only the target ion passes through the analytical quadrupole with a stable trajectory and enters the collision cell; it is accelerated by a user-defined energy (10-50 eV) and collides with the gas molecules. Kinetic energy is converted into internal energy, and the bonds break. The resulting fragment ions are then pushed into the TOF region to generated fragment spectra.

An alternative method would be fragment all ions coming into the mass spectrometer, and then deconvolute the resulting mixed fragment spectra with software[81]. This was difficult with older, lower resolution MS; however it has been gaining wider use and development with higher resolution data. All MS^2 work done in this thesis was done with DDA acquisition, because it is still the most accurate method of obtaining a fragment spectrum.

CID MS^2 is a crucial step in the metabolomics or proteomics workflow to identify unknowns. The following two sections describes how MS^2 is used for identification in the two omics fields.

1.7.1 MS^2 in Proteomics

In proteomics, the analytes are almost exclusively protonated peptides analyzed in positive mode. These medium sized polymers are made up of a relatively small number of monomer amino acid units; because of this, the fragmentation of peptides are predictable and reproducible. Through experimental and theoretical data, researchers have compiled a list of fragmentation rules[82] for peptides that enabled automated large-scale sequencing of peptides in LC-MS proteomic experiments.

The fragmentation of peptides mostly occurred in the backbone—between the carbon and nitrogen in the amide bond. Often, bond breakage generates useful b and y-ions (Figure 1.5). By comparing the mass difference between successive b or y-ion, the amino acid residue sequence can be determined. Additional fragmentation along the sidechains of the amino acids can be observed, but these fragments do no offer additional sequence information.

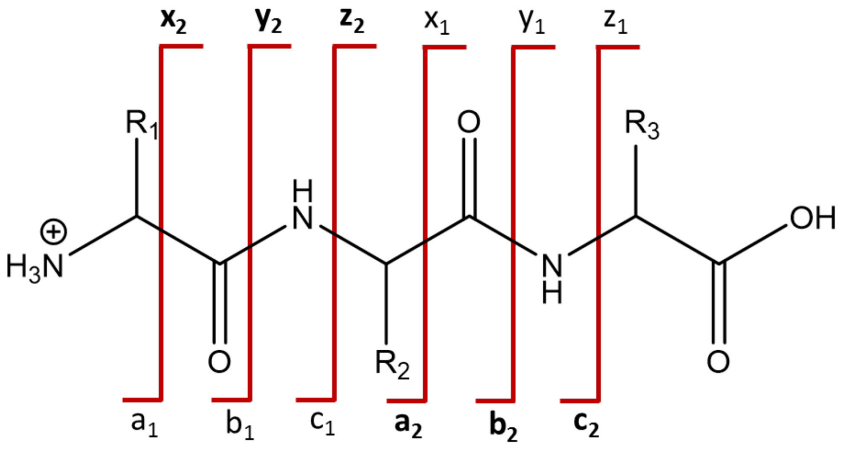


Figure 1.5 Schematic of a peptide fragmentation in MS2. Bond cleavage occurs in the peptide backbone; x,y,z-ions are fragments that contain the C-terminal of the peptide, while a,b,c-ions are contain the N-terminal. In this three-residue peptide example, y1 is the smallest y-fragment containing the C-terminal amino acid, and y2 adds on one additional residue.

The location of the bond breakage along the peptide backbone is described by the mobile proton model.[83] In this model, the extra proton on the peptide ion will migrate to the nitrogen on the amide bond to initiate bond cleavage.[82] The production of either b or y-ion depends on proton affinity of the resulting b and y fragment. This effect occurs most favorably in doubly charged peptides with C-terminal basic residues such as lysine or arginine; one proton is sequestered on the basic residue, while the other proton is free to direct backbone cleavages. Trypsin digestion cleaves at lysine or arginine to leave peptides with a C terminal basic residue, and is the preferred protease for proteomic work.

Most of the time, not all predicted fragments along the backbone amide bonds are observed. This leaves some ambiguity towards the accurate assignment of peptide sequence when trying to interpret the spectra manually. This means that, for most peptides, incomplete b/y-ion ladder prevents automated *de novo* sequencing—obtaining a complete peptide sequence only with the MS^2 data. Therefore, proteomics relies on prior knowledge of protein sequences derived from genome data. The experimental data is then matched against the predicted fragmentation pattern to determine the peptide sequence, as described in Section 1.8.1.

1.7.2 MS^2 in Metabolomics

Small molecules that make up the metabolome in organisms do not have fragment patterns that are as easy to predict as peptides. This is because small molecules have many different chemical structures and functionalities. With a vast diversity of small molecules, it is difficult to develop a universal fragmentation rule that can be used to predict fragmentation patterns, like in proteomics.

Although difficult, some work has been done with fragmentation prediction. Most *in silico* predictions are based on bond splitting algorithms that cleave all possible bonds to generate all possible unique fragment structures.[84-87] These algorithms do not take into account gas phase rearrangements and other reactions of the fragments, and they predict far more fragments than is empirically observed.[88] As the fragmentation rules of CID MS^2 of small molecules become well understood—through the collection of high quality reference data—these algorithms might be improved. However, as of now, these tools are not sufficient to give unambiguous results.

Instead, reference CID MS^2 spectral libraries of small molecule standards are the norm for identification in metabolomic experiments. To create these libraries, synthesized standards are

injected into the MS to generate CID MS^2 spectra. These spectra are then processed and entered into a database, so future experimental MS^2 spectra can be searched against the library (Section 1.8.4). Although this is the only way to obtain accurate fragmentation spectra of small molecules, synthesizing standards is very labour intensive. Despite the difficulty, building high quality libraries for metabolites and other small molecules is valuable for the entire metabolomics community.

1.8 Unknown Identification

Due to the large number of MS^2 spectra generated in one LC-MS experiment for both proteomics and metabolomics, the introduction of computer assisted unknown identification is the only way of dealing this flood of information. Due to the different fragmentation observed between metabolites and peptides (Section 1.7), the approaches for MS^2 spectral identification are very different between metabolomics and proteomics. The latter relies exclusively on genome derived database search engines, while the former relies only on reference library spectral matching. The following subsections explain the nuisances of LC-MS unknown identification in the two fields.

1.8.1 Proteomics

Peptide and protein identification from MS^2 data are almost exclusively done with database search engines. These search engines take predicted protein sequences from genomic data of the organism of interest, and performs *in silico* digestion based on the protease used in the experiment. The *in silico* digestion generates a list of peptides by cleaving according to experimentally determined cleavage rules of the protease. The search engine then matches the precursor mass of an experimental peptide to the list of all *in silico* peptides. Once a list of possible peptides are found, predicted b/y-ions (a/x,c/z-ions as well depending on conditions) are generated and compared to

the experimental MS^2 spectrum. Each MS^2 spectrum is matched to at least one peptide sequence, to generate one peptide-spectrum-match (PSM). Up to this step, the two search engine used in this thesis, Mascot (Matrix Science, London, UK) and X!Tandem[89,90] are very similar. Where the two differ is in the way that each search engine assigns the significance of the above PSM matching.

For Mascot,[91] its scoring algorithm is probability based. The score assigned to each PSM is based on the probability (p) of that match being a random one. The probability is then converted into a readable score with the following equation:$-10 \times \log_{10}(p)$. The probability is derived from factors such as how many MS^2 peaks were matched and their intensities. Confident matches have low probabilities of being random, and their resulting score would be large. However, the exact algorithm showing what parameters were included in the score calculation was never published.

X!Tandem's matching score is referred to as a hyperscore. For each candidate peptide, X!Tandem sums the intensities of the matched fragment ions and multiplied by the factorial of the number of matched ions. A distribution of scores for all matches are made, including all low scoring random matches. From this distribution, the number of random match having a score of the top-scoring match is extrapolated (Figure 1.5). The extrapolated value is referred to as the E-value; the more negative the E-value, the more significant the PSM.

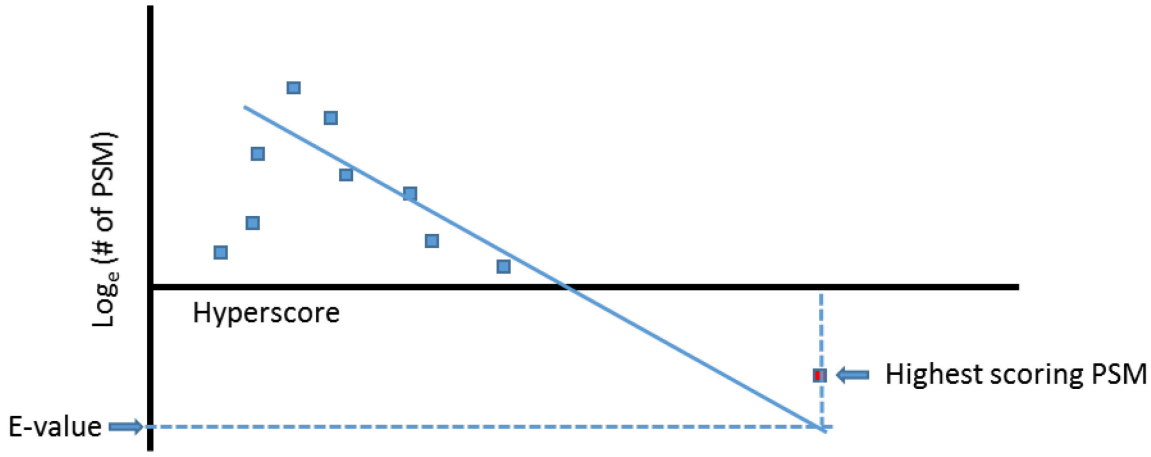


Figure 1.6 Schematic of the X!Tandem scoring algorithm to statistically generate the final E-value score. For one MS^2 spectrum, there are multiple PSMs, most are low scoring random match and a few high scoring "correct" matches. X!Tandem extends the distribution of low scoring random matches (solid line) to predict what is the likely number of random PSM to have the same score as the "correct" PSM. This is the E-value, and the lower the number the less likely the "correct" PSM is random.

While there are small variation in the algorithm for scoring PSMs, both search engine performs similarly in terms of number of matches.[92] However, X!Tandem is free to use and accessible because it is open-sourced—the exact algorithms are known to the user. Mascot is expensive for an academic group, and it is close-sourced. In addition, recent developments allowed X!Tandem to be used with a cluster of computers, greatly increasing the processing power when dealing with large high resolution data files.[93] For these two reasons, X!Tandem remains an important search engine in the proteomic community.

1.8.2 False Discovery Rate in Proteomics

The scoring algorithms described above can give a good indication of matching probability on the single PSM level. It would also be informative to obtain information on the confidence of the entire search result, in terms of the number of false positives (the number of wrong identifications that are scored as correct). Since the correctness of the identifications is not known

with certainty, the false-discovery-rate (FDR) needs to be estimated. A popular method of estimation is the target-decoy strategy.[94]

The target-decoy strategy estimates the FDR by searching a decoy database—a database that is purposely made to contain incorrect protein sequences. Any matches to this decoy database can be assumed to be a false positive, and will provide an estimate of the FDR. The number of matches to the decoy database divided by the total number of matches to the real ("forward") database gives the FDR.

The composition of the decoy database affects the result of the estimated false positives. Studies have shown that the best strategy for generating the decoy database is to reverse the target database, as opposed to randomizing the sequence.[95,96]

The value of performing decoy database searching is that it gives an estimation of the error rate. In addition to this use, decoy searching can be used for increasing the sensitivity of search engine results by using machine learning, as described in the next section.

1.8.3 Percolator

The properties between true and false positives should be quite different. A true match have better scoring statistics and less fragment mass errors than a false positive. Percolator is an algorithm that takes advantage of this property in order to filter out the false positives and to recover the false negatives from a search engine result set.

To accomplish this, Percolator trains a support vector algorithm to recognize the features of a true match by looking at all the higher scoring forward protein sequence database matches. These features are predetermined by the user, and can include the matching score, the number of fragments matched, and so on. The support vector algorithm then looks at the high scoring matches

to the reversed sequence database—theoretically identical to the false positives—and is trained to recognize the false positive matches in the forward database results. The algorithm creates a classifier to separate all of the true positives from the false positives in the forward search result. With the newly filtered list of results, Percolator repeats the above learning and classifying, until the false positive rate has been reduced to an acceptable level that is set by the user.

Chapter 2 optimizes this algorithm for coupling with the popular open-source search engine X!Tandem. Percolator has been shown to improve sensitivity and selectivity in peptide spectrum matching with other search engines, like Mascot[97] and SEQUEST.[98] By implementing Percolator with X!Tandem, there would be more choice for the proteomic community.

1.8.4 Metabolomics

Unlike proteomic database searching, spectral identification for metabolomics has seen little change since the first spectral matching algorithms appeared for electron impact (EI) spectra from gas chromatography coupled mass spectrometry (GC-MS) experiments. This is because small molecules do not have predictable fragmentation patterns and associated genome data to predict the structure of metabolites. Currently, metabolomics still relies on spectral matching with reference libraries as the most reliable method of correct identification.

Most algorithms calculate the spectral similarity between the reference spectrum and the experimental spectrum by calculating a dot product of the peaks between the two spectra. The greater the overlap between the two spectra, the higher the dot product, and the higher the final score. Formula 1.3 shows the calculation of the “Purity” score – measuring how exactly the two spectra matches. This type of algorithm is employed in the data processing and library program used in this thesis, DataAnalysis (Bruker Daltonics, Bremen, Germany).

$$Purity = 1000\frac{(\sum u \times l)^2}{\sum u^2 \times \sum l^2} \quad (1.3)$$

u – the intensity of the unknown peak

l – the intensity of the library peak

If there is an exact match between the unknown and the library spectrum, the numerator and the denominator would be equal; a perfect match would have a purity score of 1000.

However, metabolomic samples are complicated and there is often chemical interference in the MS^2 of small molecules, because two or more metabolites can be isolated by the quadrupole. It is not always possible to have high purity scores even though a match is correct. Extra peaks might appear from co-isolated molecules. To counter this effect, two other scores are commonly used that are more robust against interference—the fit and the reversed fit score.

The fit score indicates whether all the peaks in the library spectrum are found in the unknown spectrum. Extra peaks from inference will not reduce the fit score, unlike the purity score. Formula 1.4 shows how the fit score is calculated:

$$Fit = 1000\frac{(\sum u \times l)^2}{\sum_{l>0} u^2 \times \sum l^2} \quad (1.4)$$

u – the intensity of the unknown peak

l – the intensity of the library peak

The only difference, compared to the purity score, is in the summation of the denominator. The fit score will only sum the intensities of unknown peaks if they have a corresponding peak in the library spectrum. If an unknown spectrum contains the entire library spectrum and some extra peaks, the numerator will be larger than the denominator for the Fit score calculation – giving a high score. While in the purity score calculation, the denominator will be larger than the numerator – which gives a low score.

The reversed fit score is calculated in a similar way as Formula 1.4. The difference is that reversed score looks for whether peaks in the unknown spectrum is contained in the library spectrum; any extra peaks in the library spectrum will not affect the score. This is valuable in case the library spectrum is low quality and contains interferences; or if the experimental spectrum is low-intensity and has missing peaks.

Due to the large data quality variation possible in one single LC-MS metabolomic experiment, all three scores are useful for deciding matching confidence. The usual practice is to calculate and report the purity, fit, and reversed fit scores. Figure 1.6 illustrates the differences between the three scoring schemes.

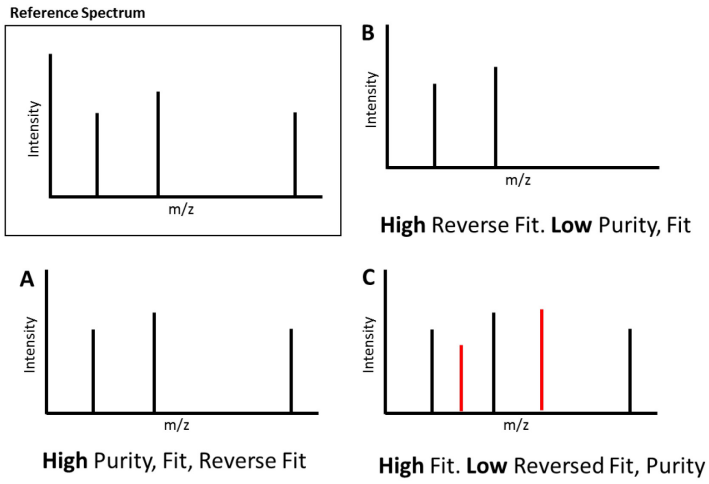


Figure 1.7 Illustration of the purity, fit, and reversed fit scoring schemes. All experimental spectra are of the same compound as the reference spectrum. (A) An experimental spectrum of the pure analyte. Perfect match, so all scores are high. (B) A low quality spectrum, where one of the peak is missing due to low ion abundance. Reversed fit is high, while the other scores are low. (C) Contains extra co-isolated fragments (in red). Only the fit score is high, others are low.

1.9 Scope of thesis

The main objective of this thesis was to develop tools and methods for LC-MS based metabolomics and proteomics, and it spans a wide range of topics from peptide database searching to nLC metabolomics.

In Chapter 2, Percolator algorithm was interfaced with the X!Tandem protein search engine to increase the sensitivity and confidence of protein discovery, while reducing the number of false positives—without having to redo time consuming experiments. Chapter 3 also increased the sensitivity of proteomic experiments by modifying the nESI source of the MS. By adding vapours of common chemical, the number of discovered peptide was increased.

Chapters 4 and 5 moves on to metabolomics, where we developed a large metabolite library that was used to identify unknowns from metabolomic experiments. This library is the highest resolution metabolite library commercially available. It is also the largest retention time library of metabolites available.

Chapter 6 continues with method development for metabolomics. In this chapter, a common proteomic technique (nLC) was combined with our group's CIL method. Metabolite sensitivity was higher than using traditional LC, and less samples were required.

1.10 Literature Cited

(1) Griffiths, J. *Anal. Chem.* **2008**, *80*, 5678-5683.

(2) Wenner, P. G.; Bell, R. J.; van Amerom, F. H. W.; Toler, S. K.; Edkins, J. E.; Hall, M. L.; Koehn, K.; Short, R. T.; Byrne, R. H. *Trends Anal. Chem.* **2004**, *23*, 288-295.

(3) Anderson, D. M.; Biemann, K.; Orgel, L. E.; Oro, J.; Owen, T.; Shulman, G. P.; Toulmin, P.; Urey, H. C. *Icarus* **1972**, *16*, 111-138.

(4) Crick, F. *Nature* **1970**, *227*, 561-563.

(5) Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; Yeates, T. O. *Nature* **2000**, *405*, 823-826.

(6) Baker, M. *Nature* **2012**, *484*, 271-275.

(7) Patterson, S. D.; Aebersold, R. H. *Nat. Genet.* **2003**, *33*, 311-323.

(8) Garibyan, L.; Avashia, N. *J. Invest. Dermatol.* **2013**, *133*, e6.

(9) Grada, A.; Weinbrecht, K. *J. Invest. Dermatol.* **2013**, *133*, e11.

(10) Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2012**, *4*, 565-583.

(11) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301-305.

(12) Lemeer, S.; Heck, A. J. *Curr. Opin. Chem. Biol.* **2009**, *13*, 414-420.

(13) Geiger, T.; Madden, S. F.; Gallagher, W. M.; Cox, J.; Mann, M. *Cancer Res.* **2012**, *72*, 2428-2439.

(14) Musunuri, S.; Wetterhall, M.; Ingelsson, M.; Lannfelt, L.; Artemenko, K.; Bergquist, J.; Kultima, K.; Shevchenko, G. *J. Proteome Res.* **2014**, *13*, 2056-2068.

(15) Wiśniewski, J. R.; Friedrich, A.; Keller, T.; Mann, M.; Koepsell, H. *J. Proteome Res.* **2015**, *14*, 353-365.

(16) Issaq, H.; Veenstra, T. *BioTechniques* **2008**, *44*, 697-698.

(17) Righetti, P. G.; Castagna, A.; Antonucci, F.; Piubelli, C.; Cecconi, D.; Campostrini, N.;

Antonioli, P.; Astner, H.; Hamdan, M. In *J. Chromatogr. A*, 2004, pp 3-17.

(18) Wu, W. W.; Wang, G.; Baek, S. J.; Shen, R.-F. *J. Proteome Res.* **2006**, *5*, 651-658.

(19) Asara, J. M.; Christofk, H. R.; Freimark, L. M.; Cantley, L. C. *Proteomics* **2008**, *8*, 994-999.

(20) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. *Anal. Bioanal. Chem.* **2007**, *389*, 1017-1031.

(21) Mayr, B. M.; Kohlbacher, O.; Reinert, K.; Sturm, M.; Gröpl, C.; Lange, E.; Klein, C.; Huber, C. G. *J. Proteome Res.* **2006**, *5*, 414-421.

(22) Choi, H.; Fermin, D.; Nesvizhskii, A. I. *Mol. Cell. Proteomics* **2008**, *7*, 2373-2385.

(23) Lundgren, D. H.; Hwang, S.-I.; Wu, L.; Han, D. K. *Expert Rev. Proteomics* **2010**, *7*, 39-53.

(24) Qendro, V.; Lundgren, D. H.; Rezaul, K.; Mahony, F.; Ferrell, N.; Bi, A.; Latifi, A.; Chowdhury, D.; Gygi, S.; Haas, W.; Wilson, L.; Murphy, M.; Han, D. K. *J. Proteome Res.* **2014**, *13*, 5031-5040.

(25) Rauniyar, N.; Yates, J. R. *J. Proteome Res.* **2014**, *13*, 5293-5309.

(26) Shiio, Y.; Aebersold, R. *Nat. Protoc.* **2006**, *1*, 139-145.

(27) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. *Anal. Chem.* **2003**, *75*, 1895-1904.

(28) Ross, P. L. *Mol. Cell. Proteomics* **2004**, *3*, 1154-1169.

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

(30) Chubukov, V.; Gerosa, L.; Kochanowski, K.; Sauer, U. *Nat. Rev. Microbiol.* **2014**, *12*, 327-340.

(31) Mak, C. M.; Lee, H.-C. H.; Chan, A. Y.-W.; Lam, C.-W. *Crit. Rev. Clin. Lab. Sci.* **2013**, *50*, 142-162.

(32) Heinemann, M.; Sauer, U. In *Curr. Opin. Microbiol.*, 2010, pp 337-343.

(33) Karr, J. R.; Sanghvi, J. C.; Macklin, D. N.; Gutschow, M. V.; Jacobs, J. M.; Bolival, B.; Assad-Garcia, N.; Glass, J. I.; Covert, M. W. *Cell* **2012**, *150*, 389-401.

(34) Orth, J. D.; Conrad, T. M.; Na, J.; Lerman, J. A.; Nam, H.; Feist, A. M.; Palsson, B. Ø. In *Mol. Syst. Biol.*, 2011.

(35) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protoc.* **2007**, *2*, 2692-2703.

(36) Wishart, D. S. *Trends Anal. Chem.* **2008**, *27*, 228-237.

(37) Antti, H.; Ebbels, T. M. D.; Keun, H. C.; Bollard, M. E.; Beckonert, O.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 139-149.

(38) Gartland, K. P.; Sanins, S. M.; Nicholson, J. K.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C. *NMR Biomed.* **1990**, *3*, 166-172.

(39) Polson, C.; Sarkar, P.; Incledon, B.; Raguvaran, V.; Grant, R. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2003**, *785*, 263-275.

(40) Venta, R. *Clin. Chem.* **2001**, *47*, 575-583.

(41) Hutschenreuther, A.; Kiontke, A.; Birkenmeier, G.; Birkemeyer, C. *Analytical Methods* **2012**, *4*, 1953.

(42) Wu, Y.; Li, L. *Anal. Chem.* **2014**, *86*, 9428-9433.

(43) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *86*, 6931-6939.

(44) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J. *Bioinformatics (Oxford, England)* **2009**, *25*, 758-764.

(45) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9*, 504.

(46) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826-828.

(47) Trygg, J.; Holmes, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469-479.

(48) Guo, K.; Li, L. *Anal. Chem.* **2009**, *81*, 3919-3932.

(49) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82*, 8789-8793.

(50) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Mol. Cell. Proteomics* **2004**, *3*, 1154-1169.

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

(52) Zhou, R.; Tseng, C.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86*, 4675-4679.

(53) Almeida, R.; Pauling, J. K.; Sokol, E.; Hannibal-Bach, H. K.; Ejsing, C. S. *J. Am. Soc. Mass Spectrom.* **2014**, *26*, 133-148.

(54) Heiskanen, L. a.; Suoniemi, M.; Ta, H. X.; Tarasov, K.; Ekroos, K. *Anal. Chem.* **2013**, *85*, 8757-8763.

(55) Kebarle, P.; Verkerk, U. H. *Mass Spectrom. Rev.* **2009**, *28*, 898-917.

(56) Snyder, L. R.; Kirkland, J. J.; Dolan, J. W. *Introduction to modern liquid chromatography*, 3rd ed.; John Wiley & Sons: Hoboken, 2010.

(57) Buszewski, B.; Jaroniec, M.; Gilpin, R. K. *J. Chromatogr. A* **1994**, *673*, 11-19.

(58) Gritti, F.; Kazakevich, Y. V.; Guiochon, G. *J. Chromatogr. A* **2007**, *1169*, 111-124.

(59) Buszewski, B.; Noga, S. *Anal. Bioanal. Chem.* **2011**, *402*, 231-247.

(60) Spagou, K.; Tsoukali, H.; Raikos, N.; Gika, H.; Wilson, I. D.; Theodoridis, G. *J. Sep. Sci.* **2010**, *33*, 716-727.

(61) Kamleh, A.; Barrett, M. P.; Wildridge, D.; Burchmore, R. J. S.; Scheltema, R. A.; Watson, D. G. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1912-1918.

(62) Chen, A.; Lynch, K. B.; Wang, X.; Lu, J. J.; Gu, C.; Liu, S. *Anal. Chim. Acta* **2014**, *844*, 90-98.

(63) Cutillas, P. R. *Curr. Nanosci.* **2005**, *1*, 65-71.

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

(65) Ikonomou, M. G.; Blades, A. T.; Kebarle, P. *Anal. Chem.* **1991**, *63*, 1989-1998.

(66) Van Berkel, G. J.; Zhou, F. *Anal. Chem.* **1995**, *67*, 3958-3964.

(67) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. *Anal. Chem.* **2013**, *85*, 2-9.

(68) Gangl, E. T.; Annan, M.; Spooner, N.; Vouros, P. *Anal. Chem.* **2001**, *73*, 5635-5644.

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

(70) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Médard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat. Methods* **2013**, *10*, 989-991.

(71) Sterling, H. J.; Prell, J. S.; Cassou, C. A.; Williams, E. R. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1178-1186.

(72) Marshall, A. G.; Hendrickson, C. L. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* **2008**, *1*, 579-599.

(73) Mamyrin, B. a. *Int. J. Mass Spectrom.* **2001**, *206*, 251-266.

(74) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1-35.

(75) Zubarev, R. a.; Makarov, A. *Anal. Chem.* **2013**, *85*, 5288-5296.

(76) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849-865.

(77) Miller, P. E.; Denton, M. B. *J. Chem. Educ.* **1986**, *63*, 617.

(78) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.

(79) Guthals, A.; Clauser, K. R.; Frank, A. M.; Bandeira, N. *J. Proteome Res.* **2013**, *12*, 2846-2857.

(80) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nat. Methods* **2007**, *4*, 709-712.

(81) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, O111.016717.

(82) Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2005**, *24*, 508-548.

(83) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365-8374.

(84) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111-3118.

(85) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. *Anal. Chem.* **2008**, *80*, 5574-5582.

(86) Jarussophon, S.; Acoca, S.; Gao, J.-M.; Deprez, C.; Kiyota, T.; Draghici, C.; Purisima, E.; Konishi, Y. *The Analyst* **2009**, *134*, 690-700.

(87) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148.

(88) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2013**, 1-8.

(89) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(90) Fenyö, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(91) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(92) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. *Mol. Cell. Proteomics* **2008**, *7*, 962-970.

(93) Bjornson, R. D.; Carriero, N. J.; Colangelo, C.; Shifman, M.; Cheung, K.-H.; Miller, P. L.; Williams, K. *J. Proteome Res.* **2008**, *7*, 293-299.

(94) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(95) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2007**, *6*, 392-398.

(96) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43-50.

(97) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(98) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

Chapter 2

Combining Percolator with X!Tandem for Accurate and Sensitive Peptide Identification*

2.1 Introduction

During the past decade MS^2 has progressed to be a popular and powerful tool to generate proteome profiles for studying complex biological systems.[1-3] In conjunction with LC separations, thousands of tandem mass spectra are routinely acquired and then correlated to peptide sequences and eventually protein identifications. Instead of manual interpretation of each spectrum, search engines, such as Mascot[4] and X!Tandem,[5] were developed to match the spectra with peptide sequences by comparing the experimental spectrum with the theoretical fragmentation patterns of individual peptide sequences derived from the protein sequences in a proteome database. To measure the reliability of each individual PSM, some statistical analysis of the match is usually carried out. For example, a probability-based Mascot ion score and identity threshold are implemented in the algorithm of Mascot (www.matrixscience.com). When the significance threshold of 0.05 ($p = 0.05$) is applied, it ensures that the probability of an identified PSM being random in the identification list is no more than 5%. Meanwhile, X!Tandem adopted the concept of reporting expectation values (E-values) of PSMs. In its algorithm, X!Tandem first calculates the number of matched fragment ions between the experimental spectrum and several candidate

*A version of this chapter has been published as "**Combining Percolator with X!Tandem for Accurate and Sensitive Peptide Identification.**" Xu, M.; Li, Z.; Li, L. *J. Proteome Res.* **2013**, *12*, 3026-3033. Also published as Chapter 6 in Mingguo's thesis: Xu, M. (2012). ***Development of a spectral searching strategy for peptide and protein identification*** (Order No. NR98542). Available from ProQuest Dissertations & Theses Global. (1458438348). I contributed towards design, programming and partially towards data processing.

theoretical peptide fragmentation patterns, generates hyperscores (the sum of matched fragment ion intensities multiplied by the N factorial for the number of matched ions), plots a distribution of hyperscores for the spectral search and extrapolates an E-value to provide a statistical evaluation for each identification. E-value is defined as the number of random matches that would be expected to have the same or better scores. This X!Tandem scoring scheme is an empirical measure of whether the match is an outlier.[6]

After the analysis of individual MS^2 matches, the concept of global FDR was proposed[7] as the standard to regulate the reporting of search results. The most common approach to estimate the global FDR of a search result is the target-decoy approach,[8] which is based on the use of randomized decoy proteome databases. The target-decoy approach enables the error control at the peptide level for different results from different search engines. However, the global FDR estimation cannot provide any statistical evaluation on the reliability of each individual PSM.

In addition to the target-decoy strategy, more sophisticated algorithms were developed to provide both global FDR estimation as well as individual assessment of PSMs. They re-evaluate the qualities of PSMs from the original search result and assign new probability to each PSM by examining the properties of correct and incorrect PSMs. For instance, PeptideProphet[9] uses an expectation-maximization algorithm to fit the bimodal distribution formed by discriminant scores of correct and incorrect PSMs in the histogram and thus computes the probability of each PSM and global FDR of the entire result. Alternatively, Percolator[10] implements a different machine learning approach. After searching all the spectra in both target and decoy databases, Percolator extracts a vector of features that are related to the quality of the match (e.g., mass error and PSM score) from both target and decoy PSMs. Assuming that the features of correct matches (represented by high scoring target matches) differ from the features of incorrect matches (represented by decoy

matches), an iterative classification process is applied to find the best separation between correct and incorrect matches. After several iterations, the system converges and generates a robust classifier that can be used to calculate the probability of each PSM being random (posterior error probability, PEP) and the minimal FDR at which a PSM is accepted (q-value).[11,12]

Among all the statistical strategies, Percolator is one of the most sensitive and accurate tools to evaluate PSMs. Due to the adaptive nature of Percolator, Percolator has been successfully extended from the application of SEQUEST and Inspect[13] results to the use of Mascot results,[14] and more recently to X!Tandem.[15] When implementing Percolator with Mascot, the selection of the features used by Percolator was shown to be vitally important to the performance of Percolator. When the authors included extra features that represent information such as intensity and fragment ion mass error, the sensitivity was boosted by 17%.

In this work, Percolator program has been successfully interfaced with X!Tandem using a PHP program. Since it is critical to select the best discriminating features for Percolator as to achieve the best performance, a set of experimentally validated PSMs were used to optimize and validate our choice of features. This approach is different from the other statistical means to extend Percolator to a search engine. For example, Yang et al[15] used a cascade learning to filter the training dataset iteratively to classify PSMs into true and false groups from which feature selection was done and evaluated. The major question here is how one can be sure about the true and false PSM assignments, which can affect the feature separation. In a previous study,[16] Xu and Li described a method of using ^{15}N-labeling for validating the spectrum-to-sequence assignments and constructing a more reliable MS^2 spectral library. Due to the advantages of this experimental validation method, not only are a large set of spectrum-to-sequence assignments justified, the annotation of the spectra are also validated, giving us an opportunity to examine the spectral

characteristics of true peptide identifications. By comparing the features of these experimentally validated PSMs (34,993 MS^2 spectra) with those of false identifications, a comprehensive set of features can be chosen for Percolator in an objective and rational manner. Next, the accuracy of X!Tandem Percolator was demonstrated by comparing the estimated false-discovery rate of the validated dataset with the factual false-discovery rate. By comparing the results from our X!Tandem Percolator and the original X!Tandem, superior sensitivity and specificity of the X!Tandem Percolator result was demonstrated. Lastly, X!Tandem Percolator was applied to results with various search conditions, such as large MS^2 datasets from different species, human (46,494 MS^2 spectra) and *E. coli* (88,306 MS^2 spectra) to examine robustness.

2.2 Experimental Section

2.2.1 Sample Preparation.

Three different MS^2 datasets were used in this study. They were collected using a QTOF Premier mass spectrometer (Waters, Manchester, U.K.) equipped with a nanoACQUITY Ultra Performance LC system (Waters, Milford, MA). The experimental workflows are outlined below while the detailed procedures used to generate these datasets are the same as those reported in a previous publication.[16]

2.2.2 *E. coli* Dataset.

The *E. coli* K12 cells (*E. coli*, ATCC 47076) were cultured, collected and disrupted. They were subsequently subjected to reduction, alkylation, acetone precipitation, trypsin digestion and strong cation exchange (SCX) chromatographic fractionation. All the peptide fractions were then desalted and analyzed by reversed phase (RP) LC-QTOF to generate the MS^2 spectra. In total, 88,306 spectra were collected and searched with both Mascot (version 2.2.1) and X!Tandem (The

Global Proteome Machine Organization, 2007.07.01, version Cyclone) using the same search parameters including: enzyme, trypsin; fixed modifications, carbamidomethylation (C); variable modifications, acetylation (N-term), ammonia-loss (N-term C), pyro-Glu (N-term Q), pyro-Glu (N-term E), and oxidation (M); precursor mass error, 30 ppm, fragment mass error, 0.2 Da, maximum missed cleavages, 2. The lists of peptides and proteins identified from the *E. coli* dataset are provided in Supplemental Table S1.

2.2.3 Human Dataset.

Similar to the *E. coli* dataset, SU-DH-L1 cells[17] (A human lymphoma cell line) were cultured, harvested, disrupted by cell lysis buffer and subjected to acetone precipitation, reduction, alkylation, protein reversed phase chromatographic fractionation and trypsin digestion. Then RP-LC-QTOF-MS^2 analysis was performed on all the desalted peptide fractions to collect the MS^2 spectra. Overall this human dataset contained 46,494 MS^2 spectra and was searched by Mascot and X!Tandem using the same parameters aforementioned in the *E. coli* dataset section. The lists of peptides and proteins identified from the human dataset are provided in Supplemental Table S1.

2.2.4 Validated *E. coli* Dataset.

This dataset consists of 34,993 experimentally validated spectral identifications. Each of them was examined using a ^{15}N-metabolic labeling validation process. This approach has been described in detail in the reference.[16] Briefly, unlabeled and ^{15}N-labeled *E. coli* spectra were collected independently and further compared by overlaying the spectra of unlabeled and labeled matches of the same peptide sequence for validation. In the validation process, two cut-off filters were developed. One was based on the number of common fragment ions found in the overlaid spectra; a minimum of 5 common ions were found to be needed to judge the fragmentation pattern

matches. The second filter was based on the similarity of the fragmentation patterns of the unlabeled and labeled peptide pairs. A similarity score was calculated by using the fragment ion intensity dot-product, and the cut-off score was found to be 0.96 out of 1.00, with 1.00 to be a perfect score. This isotopic labeling approach provided experimental evidence to validate the PSMs generated by sequence-database searching and ensured the reliability of peptide identifications.

2.2.5 Databases.

Target-decoy search strategy proposed by Elias and Gygi in 2007[8] was applied by searching the MS^2 spectra against two separate databases (target and decoy databases) to calculate the q-value. The target databases used for *E. coli* and human datasets were *E. coli* K12 proteome sequences (size ≈ 2 MB, 4,339 sequences) and international protein index human database (IPI human database, version 3.68, size ≈ 48 MB, 87,061 sequences), respectively. The construction of a decoy database in this study was done by reversing all the protein sequences found in target database.

2.2.6 Percolator Processing.

The raw search results from Mascot (*.dat file) and X!Tandem (*.xml file) were imported to Percolator program (version 2.01). The original result files were then parsed. Scoring features were extracted accordingly and sent to Percolator for further training steps.

In Mascot Percolator, the default features were used, including precursor mass, charge, Mascot score, score difference between the best and second best match, precursor mass error, fraction of variable modification sites that was modified, the number of missed cleavages, fragment mass error, total intensity of the spectrum, total intensity of peaks that were used to identify a

peptide, relative total intensity of peaks that were used to identify a peptide, fraction of ions that were matched in an ion series, and others.

X!Tandem has a different scoring scheme than Mascot. Instead of reporting the probability of a PSM being random, X!Tandem first plots a distribution of calculated hyperscores from a specific search and then extrapolates E-values to provide an statistical evaluation for each identification. It was therefore found that the features extracted from X!Tandem were not exactly the same as the ones from Mascot or SEQUEST. All the features used in the X!Tandem Percolator were listed and explained in Table 2.1. As shown in Table 2.1, all these features can be categorized into three different groups, namely spectral quality, scoring and PSM statistics. In the category of spectral quality, all three features represent the intrinsic quality of an MS^2 spectrum regardless of its peptide assignment. In the category of scoring, all of the eight features come from the original X!Tandem scoring algorithm and are used to measure how reliable the sequence to spectrum assignment is. Lastly, all the features in the PSM statistics category are the collective statistical information that are not directly used by X!Tandem but still might indicate the difference between true and false PSMs. They all can be switched on or off based on different requirements of applications.

2.2.7 Comparison.

In order to evaluate the performance of X!Tandem Percolator, various comparisons among Mascot, Mascot Percolator, original X!Tandem and X!Tandem Percolator were carried out. For the experimental validated dataset, empirical q-value was proposed to measure the error rate of search results. Because of experimental validation of sequence assignments, correct and incorrect PSMs can be isolated by comparing the X!Tandem or Percolator result with the validated result. Therefore, empirical q-value were accurately calculated by dividing the number of total PSMs with the number

of incorrect PSMs. For real shot-gun proteomic data, estimated q-values were used instead. By definition, q-value is the minimal global FDR at which a PSM is accepted. To demonstrate the improved performance of X!Tandem Percolator, performance curves were plotted by examining the number of estimated correct PSMs at different levels of q-value for the results obtained from different data processing tools (e.g., Mascot, Mascot Percolator and SEQUEST Percolator).

Protein identifications were inferred from peptide identifications. A protein was considered to be identified when at least one associate unique peptide sequence was identified with q-value of lower than the threshold. The Occam's razor approach[18] was applied to deal with degenerate peptides by finding a minimum subset of proteins that covered all the identified peptides.

2.2.8 X!Tandem Percolator.

X!Tandem Percolator was developed based on a mix of PHP and Perl scripts in an Apache server (the interface program and installation instruction can be downloaded from Supplemental Folder 1 in on-line Supporting Information) and it is integrated directly into X!Tandem. A simple click at the interface was all that needed to begin calculating features, re-ranking peptide identifications, assigning statistical values and exporting results in Percolator. X!Tandem Percolator can calculate features at 57 PSM per second on a quadcore 3.20GHz Phenom II 955 AMD processor.

Table 2.1 List of features extracted from X!Tandem search results.

Index	Feature name	Feature type	Feature description
1	Mass	Spectral quality	The observed mass $[M+H]^+$
2 - 5	Charge	Spectral quality	Four Boolean features indicating the charge state
6	MaxI	Spectral quality	The maximum fragment ion intensity
7	PSMSumI	Spectral quality	The $\log_{10}$ value of the sum of all of the fragment ion intensities
8	Log(E)	Scoring	The $\log_{10}$ value of the expectation value for the peptide identification
9 - 10	IonScore	Scoring	The summed intensities of different types of fragment ions (y, b ions)
11 - 12	IonNo	Scoring	The number of peaks that matched between the theoretical and the test mass spectrum
13	HyperScore	Scoring	X!Tandem's score for the peptide Identification
14	NextScore	Scoring	The HyperScore of the second best peptide match of the spectrum
15	DeltaScore	Scoring	The difference of HyperScore between the best and the second best peptide matches
16	DeltaM	PSM statistics	The difference in calculated and observed mass (Th)
17	RelDeltaM	PSM statistics	The relative difference in calculated and observed mass (ppm)
18 - 19	IonFrac	PSM statistics	The fraction of fragment ions being matched in an ion series (y, b ion series)
20	MissClea	PSM statistics	The number of missed internal enzymatic (tryptic) sites
21	FragError	PSM statistics	The average mass error of all the fragment ions
22	AnnoPeaks	PSM statistics	The fraction of high intensity peaks being annotated as fragment ions
23	ModNo	PSM statistics	The number of variable modifications
24	ModFrac	PSM statistics	The fraction of modifiable residues being found modified (variable)
25	EnzN	PSM statistics	Boolean value: is the peptide preceded by an enzymatic (tryptic) site?
26	EnzC	PSM statistics	Boolean value: does the peptide have a C-terminal enzymatic (tryptic) site?
27	PepLeng	PSM statistics	The length of the peptide identification

2.3 Results and Discussion

2.3.1 Feature Evaluation

Although Percolator is a semi-supervised learning method that does not need to construct a manually curated training set, in order to train a support vector machine, a variety of specific features that are capable of discriminating between true and false PSMs is still required. Understandably, the choice of features is vitally important. In this work, a list of features were composed that were deemed to be capable of differentiating true and false PSMs. As shown in Table 2.1, there are 27 features that are able to measure the intrinsic quality of the spectra and the quality of PSMs. The rationale of doing this is that spectral level information might indicate what types of MS^2 spectra (e.g., precursor charge states and fragment ion intensities) are more likely to lead to correct identifications. Next, the features in the scoring category, such as Log(E) values, measure how reliable PSMs are individually from the perspective of the search engine. Finally, the features in the category of PSM statistics suggest how likely one PSM is to lead to a true identification from a global perspective. Different from the original and Mascot Percolator, it was decided not to include features that exploit protein-level information (e.g., protein E-values) because we found that protein-level information led to a significant increase in the number of spurious PSMs in the search results. This might be attributed to the fact that protein-level feedback changes the score distribution of spurious PSMs, resulting in PSM misclassification.[19,20]

To determine if these features will actually discriminate between true and false positives, a set of experimentally validated PSMs from a previous study on *E. coli* cell lysates were used as standards. The advantage of using this dataset is that every one of the 34,993 PSMs was both correctly annotated and experimentally validated. In addition, unlike the limited number of peptides from several standard proteins, the greater number of peptide identifications from this *E. coli* cell

lysates made the analysis more statistically robust. In terms of the negative training set, the matches from the decoy search would be good representatives of false PSMs. Figure 2.1 shows the discriminatory power of the individual features examined. Using a selected feature as the only scoring metric, each curve in Figure 2.1 was generated by plotting the number of true PSMs as the function of the number of decoy PSMs under different thresholds. The comparison shown in Figure 2.1 was important for all the chosen features, considering that they are the key to differentiate the true and false PSMs.

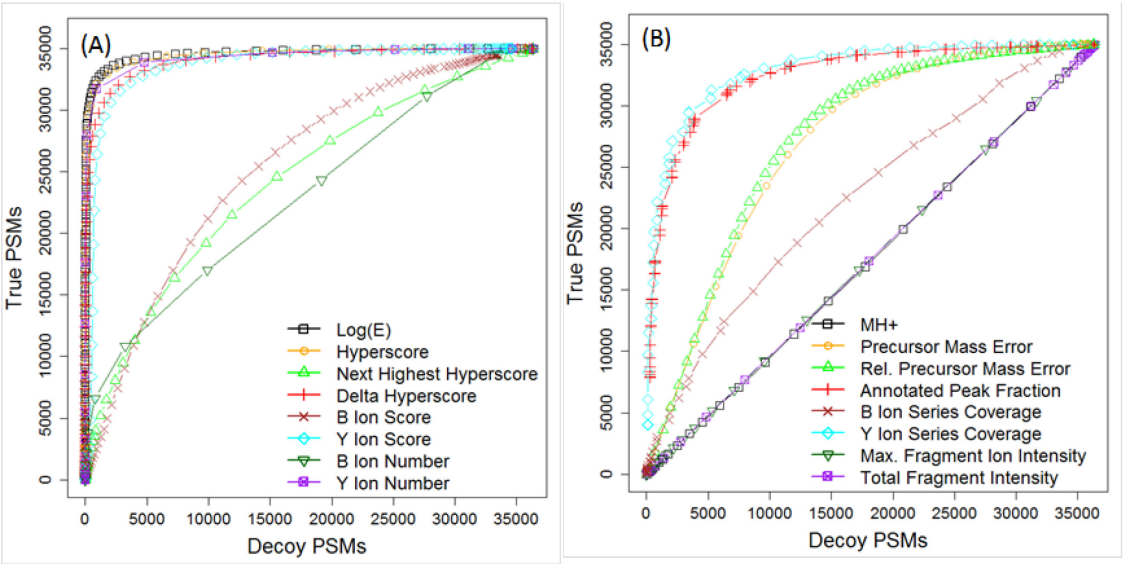


Figure 2.1 Discriminatory power of various features selected.

For all the features in the scoring category, features 8 – 15 in Table 2.1, in principle, should be able to show good discriminatory power. For instance, feature 8, log(E), is the $\log_{10}$ value of the expectation value for the peptide identification calculated by X!Tandem. As the main X!Tandem scoring metric to determine the reliability of PSMs, it was not surprising to see that true PSMs have a distinctively different distribution of log(E) than decoy PSMs. Figure 2.1A shows that log(E)

value is a good candidate to separate the true and decoy PSMs. Similarly, the same conclusion can be drawn for features including hyperscore, delta hyperscore, y ion score and y ion number. On the other hand, the discriminatory power of the rest of the features shown in Figure 2.1A is still visible, but not as great as the other features. However, this does not mean that they should not be included in the X!Tandem or Percolator algorithm. In fact, they all can be explained. For the "next highest hyper score" feature, considering that the next best match is supposed to be a random match, it was reasonable to see little difference between true and decoy PSMs for this feature. As for the B ion number and B ion score, the little difference between true and decoy PSMs can be ascribed to the intrinsic fragmentation preference of tryptic peptides to the Y ions.

Apart from the features that were directly used by the X!Tandem algorithm, a set of features that indicate PSM statistics and a set of spectral features were also used (see Figure 2.1B). Take the feature "fragment error" as an example. In its definition, it represents the average absolute mass error of all the matched fragment ions. Generally, in a sequence database search, in an attempt to fully annotate an MS^2 spectrum, a wide mass tolerance window for fragment ions is used. Unfortunately, it inevitably increases the possibility of random matches from falsely matching fragment ions. However, in principle, the true fragment ions should have a smaller average absolute mass error than the false one, especially when MS^2 spectra were acquired with good accuracy (e.g., 30 ppm in QTOF data). Based on this concept, the fragment error should be able to provide a unique perspective to pin down false PSMs, which are primarily identified by detection of false fragment ions in MS^2 spectra. As illustrated in Figure 2.1B, true PSMs have a distinctively different distribution of fragment ion mass errors than decoy PSMs do. In addition, a new feature called AnnoPeaks was created, which was defined as the fraction of high intensity peaks (at least 70% intensity of the most intense peak) that were matched as fragment ions. Since peptides fragment in

a reasonably predicable manner, most high intensity peaks in an MS^2 spectrum should be accounted for in a true PSM. Meanwhile, in a false PSM, the fragment ions are matched by random peaks regardless of their intensities. Thus the feature AnnoPeaks should be able to provide another distinct perspective to distinguish the true and false PSMs. As illustrated in Figure 2.1B, AnnoPeaks is indeed a good feature. Similarly, the same conclusion can be drawn for features including Precursor Mass Error, B, and Y IonFrac (see their definitions in Table 2.1).

In the category of spectral features, it was conventional for Percolator programs (SEQUEST and Mascot Percolator platforms) to include them in the process of differentiation. The rationale was that spectral level information might indicate what types of MS^2 spectra (e.g., precursor charge states and fragment ion intensities) are more likely to lead to correct identifications. Since these features are not a direct measurement of the goodness of a sequence assignment, they might not be powerful discriminators to differentiate true and false positives when used individually. This was exactly what has been observed in this study (see Figure 2.1B). As shown in Figure 2.1B, the distributions of the quasi-molecular ion masses were very similar between the true and decoy PSMs. The median values of their quasi-molecular ion masses were not distinguishable. Similarly, the same conclusion can be drawn for both the total fragment intensity and the maximum fragment ion intensity features. However, even though the spectral features individually were not very indicative in terms of differentiating PSMs, they still might contribute to the task when working collaboratively with each other or with features from the PSM category (see below).

2.3.2 Performance on Validated Dataset

After building a list of useful features based on an experimentally validated dataset, a comparison was carried out on the performance of X!Tandem and X!Tandem Percolator on the same dataset to examine X!Tandem Percolator's accuracy. Unlike most normal shot-gun proteomic

data in which only a part of spectra (30 to 70%) are identifiable, all the un-identifiable spectra were filtered out in this validated dataset. This means that any robust statistical tool should be able to recover close to 100% of all the pre-validated PSMs. In fact, it was found that X!Tandem alone was able to re-identify 98.9% of all the pre-validated PSMs when a lenient E-value threshold (E-value = 1) was applied. Since X!Tandem Percolator was designed to minimize the number of false positives and negatives of the original X!Tandem result, it was reasonable to observe a similarly high recovery rate with better sensitivity and specificity trade-off. As expected, for X!Tandem Percolator the recovery rate was found to be 99.9% when a q-value of 0.4% was applied. Due to the advantage of validated dataset, empirical q-values were calculated by dividing the number of PSMs with the number of incorrect PSMs. When comparing the X!Tandem result with X!Tandem Percolator results at different empirical q-values, the superior sensitivity and specificity that X!Tandem Percolator provided was quite apparent (see Figure 2.2A). In order to assess the accuracy of the statistical evaluation by X!Tandem Percolator, the number of estimated incorrect PSMs was compared to the number of empirical incorrect PSMs. The same analysis was also applied to X!Tandem, Mascot and Mascot Percolator results for comparison. Ideally, the number of incorrect PSMs estimated by a perfect statistical evaluation will always be equal to the number of empirical incorrect PSMs. As shown in Figure 2.2B, all lines closely resemble the 45° diagonal line. This similarity between the number of estimated PSMs and the number of empirical incorrect PSMs indicates that X!Tandem Percolator is able to provide accurate statistical assessment of a search result with similar performance to that of Mascot Percolator.

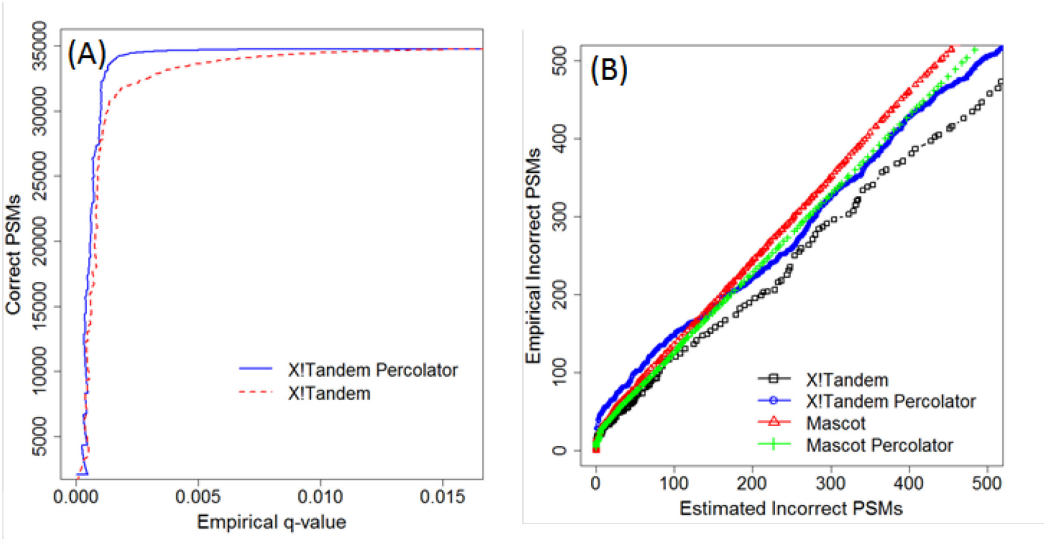


Figure 2.2 (A) Performance comparison between X!Tandem and X!Tandem Percolator at different empirical q-values. (B) Comparison between the number of empirical and estimated incorrect PSMs by X!Tandem Percolator, X!Tandem, Mascot and Mascot Percolator.

The relative contributions of the PSM statistics features and spectral features to the discriminatory power were also investigated. The best way to examine their contributions is through feature removal analysis on real shotgun proteomic data. The *E. coli* dataset was searched with X!Tandem and run on Percolator, eliminating one subset of features at a time. As shown in Figure 2.3, spectral features did make contribution in the process of differentiation, even though individually they did not show strong discriminatory power. The number of estimated correct PSMs at a q-value of 0.01 was summarized in Table 2.2, as well as the percentage decrease in estimated correct PSMs relative to using all the features. As shown in Table 2.2, removing spectral quality features led to a 1% drop in performance, while removing PSM statistics features resulted in a 9% drop. However, comparing to the original X!Tandem result, X!Tandem Percolator equipped with all the features can significantly improve the number of PSMs.

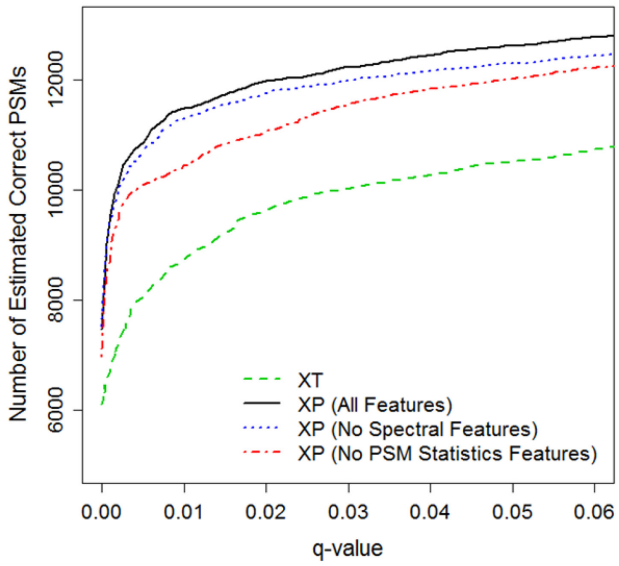


Figure 2.3 Performance of X!Tandem (XT) and X!Tandem Percolator (XP) when fed with different features.

Table 2.2 Performance of X!Tandem Percolator when fed with different features.

	Number of estimated correct PSMs	Drop in performance
All features	11478	-
Spectral quality features removed	11314	1%
PSM statistics features removed	10442	9%
Original X!Tandem	8786	23%

2.3.3 Exemplary Experimental Data

In order to examine X!Tandem Percolator's robustness, Percolator was applied to X!Tandem search results of *E. coli* and human lymphoma cell data, which are representatives of typical proteomic datasets, to demonstrate the superior sensitivity and specificity. As indicated earlier, using our validated dataset, true and false positives were detected and the empirical q-value was calculated accordingly. But in typical shotgun proteomic datasets, there is no validation of sequence assignments. Researchers therefore rely on the X!Tandem and Percolator programs or

other programs to provide statistical assessment for each PSM, e.g., by calculating the q-value (the minimal global FDR at which a PSM is accepted) for each PSM.

First, a performance comparison between Mascot, Mascot Percolator, SEQUEST Percolator, X!Tandem and X!Tandem Percolator was carried out on the shotgun *E. coli* dataset. The *E. coli* system was selected because of its relatively simple proteome complexity (only about 4300 predicted proteins) and its popularity as a model system in biological studies. The dataset was first searched by both X!Tandem and Mascot, and then processed by Percolator programs, respectively. Figure 2.4A shows the number of estimated correct PSMs for X!Tandem, Mascot, Mascot Percolator, SEQUEST Percolator and X!Tandem Percolator at different levels of q-values. A maximum q-value of 0.05 was chosen because performance at this level is the most relevant to the proteomic community, where the usual FDR is controlled to a maximum of 5% (equivalent to q-value 0.05). As indicated in Figure 2.4A, all the Percolator programs offer much better sensitivity and specificity trade-off than Mascot and X!Tandem. To be exact, at q-value of 0.01, X!Tandem Percolator managed to identify 11594 PSMs, corresponding to 1393 proteins. Compared to the X!Tandem result (8875 PSMs and 1209 proteins), this represents 31% and 15% increase in the number of PSMs and proteins, respectively. The same trend was also observed in the comparison between Mascot Percolator and Mascot, which had been reported by Brosh et al. as well.[14] Overall, this result demonstrated the performance advantages of X!Tandem Percolator over the original X!Tandem scoring method when dealing with a simple proteomic system. Moreover, after the statistical analysis by Percolator, X!Tandem and Mascot results became more much agreeable to one another. At q-value of 0.01, the percentage of PSMs appeared in both Mascot and X!Tandem results constituted only 46% of total PSMs that were identified by both search engines. However,

the percentage increased to 82% after Percolator was applied. This can be attributed to the improved sensitivity and more unified statistical assessment of Percolator.

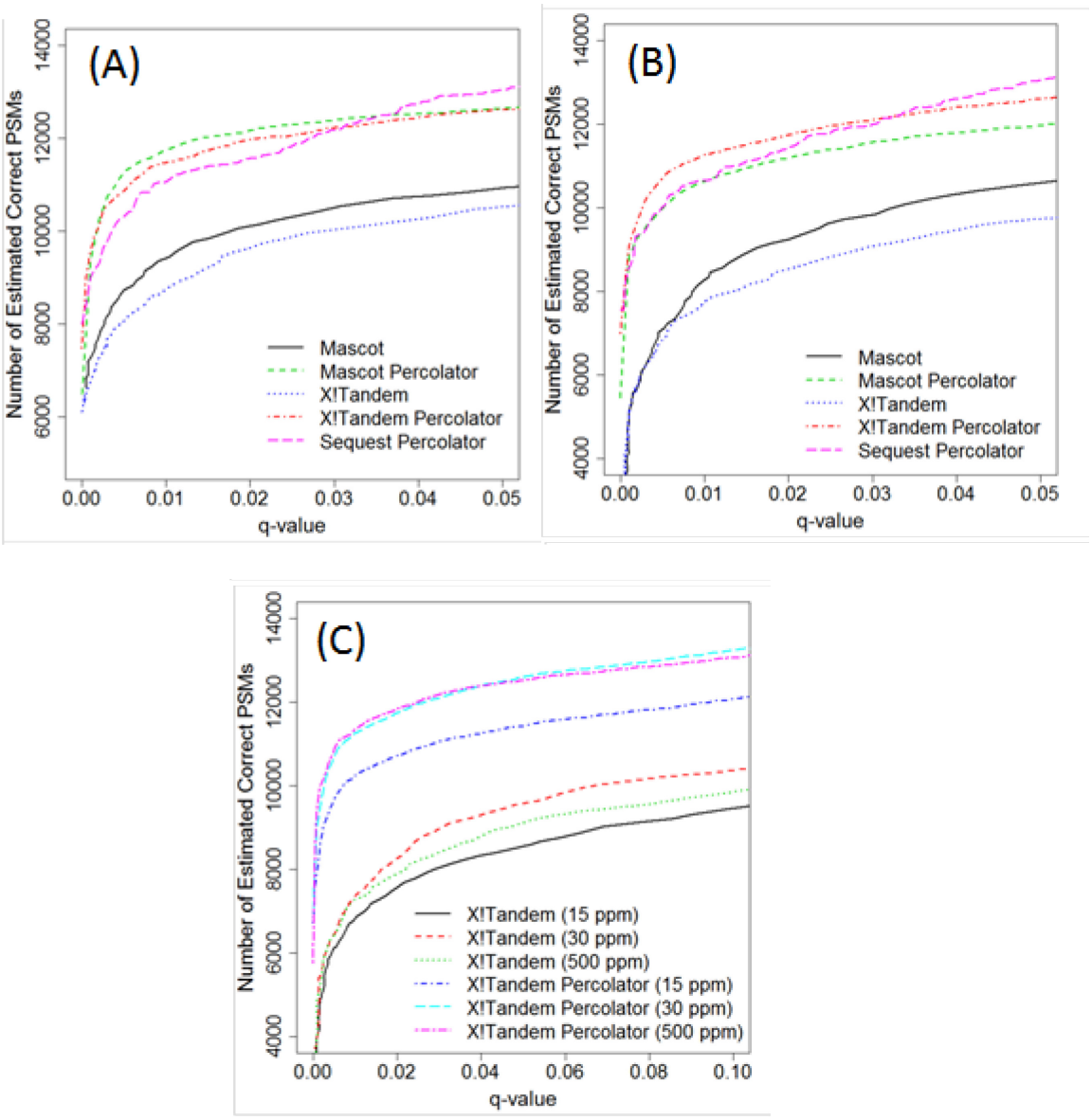


Figure 2.4 Performance comparison between Mascot, Mascot Percolator, SEQUEST Percolator, X!Tandem and X!Tandem Percolator on (A) the shotgun E. coli dataset and (B) the shotgun human dataset. (C) The influence of precursor mass tolerance setting on the performance of X!Tandem and X!Tandem Percolator.

The same analysis was applied to the human dataset to see how X!Tandem Percolator would respond to the searches with a much larger proteome database (87061 protein sequences).

Understandably, a much larger proteome database provides more combinations of amino acids. It is therefore an even bigger challenge to differentiate true and false PSMs. Figure 2.4B shows the number of estimated correct PSMs for X!Tandem, Mascot, Mascot Percolator, SEQUEST Percolator and X!Tandem Percolator at different levels of q-values for the human dataset. As indicated in Figure 2.4B, all the Percolator programs still offer much better sensitivity and specificity trade-off than Mascot and X!Tandem. In fact, at q-value of 0.01, X!Tandem Percolator was able to improve the number of PSMs and protein identifications of the original X!Tandem result by 51% and 29%, from 7477 PSMs and 1419 proteins to 11314 PSMs and 1831 proteins, respectively. The improvement on the human dataset was even greater than the improvement on the *E. coli* dataset, indicating that Percolator was less easily influenced by the complexity of proteome databases.

2.3.4 Sensitivity to Search Space Change.

Sometimes, relaxed searching parameters are chosen by researchers in order to match as many peptide sequences as possible in a proteomic study. For instance, a wide precursor mass tolerance window is often used so as to capture all the potential identifications. Contrary to the original intention, it is often noted that when relaxed searching parameters are set up for a sequence database search engine, a noticeable drop in the number of PSMs is often observed. This is simply due to the fact that increased search space creates more possible random matches. In order to avoid a decrease in accuracy, sensitivity is often sacrificed. In this study, in order to test how well X!Tandem Percolator is able to handle this issue, different precursor mass tolerance settings, including 15, 30 and 500 ppm were used in X!Tandem. After searching the human dataset with these settings, results were processed by Percolator. As shown in Figure 2.4C, as the precursor mass tolerance setting increases from 15 ppm to 30 ppm, an increase in PSM number for both

X!Tandem and X!Tandem Percolator is obvious. In fact, at q-value of 0.01, the improvement in PSM and protein identification were 8% and 1% for the X!Tandem result, and 11% and 4% for X!Tandem Percolator result, respectively. This result indicates that 30 ppm was more appropriate mass accuracy window for this dataset. When increasing the setting from 30 ppm to 500 ppm, little change is observed for either X!Tandem results or X!Tandem Percolator results (see Figure 2.4C). However, at q-value of 0.05, another commonly used identification threshold, a decrease in X!Tandem performance (5% less PSMs, 2% less proteins) can be easily spotted in Figure 2.4C. At the same time, almost no decrease in X!Tandem Percolator is observed. Thus, we can conclude that X!Tandem Percolator is a highly robust statistical tool and less easily influenced by search space increase.

2.4 Conclusions

Percolator was previously shown as a very robust classifier that can dramatically improve sensitivity on various search engines, such as SEQUEST[21] and Mascot.[14] In this study, an interface has been built for Percolator and X!Tandem, a very popular open-source search engine. To successfully integrate Percolator with X!Tandem, a large number of features that define the quality of PSMs were first created. Since an experimentally validated dataset provided the opportunity of isolating the true PSMs from search results, by comparing the features from true and decoy PSMs, the individual discriminatory power of each feature was carefully examined. Moreover, a feature removal analysis was also performed to demonstrate the collective contribution of different subsets of features.

X!Tandem Percolator was then applied to shotgun proteomic data, including the *E. coli* and human datasets. Under various conditions, including different sizes of databases and relaxed search parameters, X!Tandem Percolator was found to substantially outperform the original X!Tandem, showing a similar or even better performance over Mascot Percolator. Overall, it appears that better classification of true and false PSMs can be achieved when multiple factors are working collaboratively instead of just using one or two scoring metrics (e.g., E-values or FDRs). Therefore, we envisage the use of X!Tandem Percolator either as a stand-alone software package or a key part of a sophisticated data analyzing platform in discovery-oriented proteomic studies.

2.5 Literature Cited

(1) Aebersold, R.; Mann, M. *Nature (London, U. K.)* **2003**, *422*, 198-207.

(2) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. *Nat. Methods* **2010**, *7*, 681-685.

(3) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69*, 767-776.

(4) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.

(5) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.

(6) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(7) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787-797.

(8) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.

(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.

(10) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(11) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2007**, *7*, 40-44.

(12) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2007**, *7*, 29-34.

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

(14) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.

(15) Yang, P. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1273-1280.

(16) Xu, M.; Li, L. *J Proteome Res* **2011**, *10*, 3632-3641.

(17) Wu, F.; Wang, P.; Zhang, J.; Young, L. C.; Lai, R.; Li, L. *Mol. Cell. Proteomics* **2010**, *9*, 1616-1632.

(18) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Analytical Chemistry* **2003**, *75*, 4646-4658.

(19) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. *J. Proteome Res.* **2009**, *8*, 3737-3745.

(20) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.

(21) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

Chapter 3

Chemical-vapor-assisted Electrospray Ionization for Increasing Analyte Signals in ESI Mass Spectrometry*

3.1 Introduction

Electrospray ionization mass spectrometry has been widely used for chemical analysis, particularly when it is combined with a solution based separation technique, such as LC. There is always a need to increase the detection sensitivity of the technique in application areas where trace amounts of analytes are analyzed, including nESI MS for proteomic profiling.[1,2] Improvement in detection sensitivity can come from more efficient LC separation[3] or using capillary electrophoresis,[2] increased ion transmission from the source to a mass analyzer,[4] and better ion detection in a mass spectrometer.[5] Another possible means of achieving higher detection sensitivity is to change the ESI conditions in order to generate gas phase ions more efficiently.[6,7] In this work, we report a simple technique to enhance the analyte signals by merely adding an appropriate chemical vapor to the electrospray source without changing any other conditions used in a conventional nESI mass spectrometer. We have investigated several experimental parameters, including the type of chemical vapors used, to achieve the optimal signal enhancement for this chemical-vapor-assisted ESI technique.

Using a chemical additive to improve ESI detection can be done by dissolving it into a solvent or teeing the liquid post-column. For example, addition of 5% dimethyl sulfoxide (DMSO)

*A version of this chapter has been published as"**Chemical-Vapor-Assisted Electrospray Ionization for Increasing Analyte Signals in Electrospray Ionization Mass Spectrometry.**" Li, Z.; Li, L. *Anal. Chem.* **2014**, *86*, 331-335. I was responsible for the design, data collection, analysis, as well as the manuscript writing.

into the eluent has been reported to increase the peptide identification rate from LC-MS experiments, largely due to the collapse of multiple charge states into one easy-to-detect peak.[8] However, a recent study found that the increase in peptide identifications was mainly due to more efficient production of ions in the ESI process, rather than charge state collapse.[9] Post-column incorporation of organic acids such as formic acid has been shown to enhance signals suppressed by trifluoroacetic acid by several fold.[10] Exposing a gas other than the nitrogen or air in the ESI source can also affect the ESI responses of analytes. For examples, gas phase addition of volatile acids,[11] bases,[12] and solvents,[13,14] have been shown to change the charge state distribution in intact proteins or to reduce multiply charge chemical noise in plasma samples.[15] Fenn et al. reported that the addition of water vapor in the counter flow gas could result in an increase in the signal intensity of peptides.[16]

In this work, we systematically exposed nESI—the most widely used ionization technique for proteome analysis—to different classes of gaseous molecules, such as organic acids and alcohols. They were chosen to span a wide range of boiling points, surface tension, and the ability to form azeotropes with water. Each individual chemical was placed in a small container where nitrogen was flown over the headspace and into the sheath around the nESI needle of a Waters QTOF Premier mass spectrometer

3.2 Methods

3.2.1 Reagents and Instrumentation.

Solvents for direct infusion and LC-MS experiments used LC-MS grade acetonitrile, water, and formic acid from Fischer Scientific (Fair Lawn, NJ). Leu-enkphalin and [Glu1]-fibrinopeptide were supplied by Anaspec (Fremont, CA). Alpha casein protein, reserpine, N,N-

dimethyldodecylamine, phenylephrine, syringaldehyde were from Sigma Aldrich (St. Louis, MO). BSA tryptic digest standard was purchased from Waters (Milford, MA). Samples were analyzed by a Waters nanoAcquity UPLC (Milford, MA) and a Waters QTOF Premier mass spectrometer (Milford, MA).

3.2.2 Vapour Introduction

Two 1/16" holes were drilled into the cap of a 3-dram vial. A 1/16" O.D. and 0.020" I.D. tubing coming from the API gas outlet of the mass spectrometer was inserted into one hole, while the same size tubing connected the cap to the sheath around the nESI tip. Liquid was placed into an empty 3-dram glass vial and capped. Gas flow was controlled by setting the API gas pressure. All experiments were done at room temperature (~24°C) and relative humidity of <20%.

3.2.3 Direct Infusion Experiments.

A mixture of 0.2 µg/mL Glu-Fib (m/z 786, 2+) and 0.05 µg/mL Leu-Enk (m/z 556, 1+) was dissolved in 50% acetonitrile in water with 0.1% (v/v) formic acid. The 1+ peak of Glu-Fib and 2+ peak of Leu-Enk were not observed. Electrospray was conducted at 3.3 kV capillary voltage with a 10 µm nESI pico-tip from New Objective (Woburn, MA). Sample was infused at a flow rate of 0.250 µL/min, using the auxiliary pump in the nanoAcquity. The N_2 flow over the headspace of the apparatus was controlled by adjusting the gas pressure to 0.1 bar. The position of the pico-tip was optimized to give the best signal. Data was collected with gas on for ~1 min, then gas off for ~1 min, and then repeated twice. For each section of data collection, 25 scans were summed, and the enhancement ratios were calculated by dividing the enhanced summed intensity and gas off summed intensity, using the monoisotopic peak. For the experiment using different acetonitrile content, the same concentrations of peptides were dissolved in different percentages of acetonitrile

and infused at 0.250 µL/min using the auxiliary pump after it was primed for 3 min and flushed at 10 µL/min for 30 min.

3.2.4 LC-MS for BSA and Alpha Casein Peptide Analysis.

BSA tryptic digest was purchased from Waters. Alpha casein protein was digested in 25% trifluoroacetic acid using microwave-assisted acid hydrolysis. Eluent A was LC-MS water with 0.1% (v/v) formic acid and eluent B was LC-MS acetonitrile with 0.1% (v/v) formic acid. For both experiments, a 75 µm x 150 mm Atlantis dC_{18} column with 3 µm particles (Waters), was used. 50 fmol of BSA tryptic digest and 2 µg alpha casein hydrolysate were injected onto the column, respectively. For the BSA tryptic digest, a 50 min gradient at 0.350 µL/min flow rate was used with a 1 s MS scan time. For alpha casein hydrolysate, a 120 min gradient at 0.300 µL/min flow rate was used; MS survey scans were 1 s, followed by 6 data-dependent fragment scans with 0.5 s scan times. For BSA tryptic digest analysis, MassLynx software from Waters was used to integrate EIC after smoothing and peaks were detected with a peak-to-peak amplitude of 8. GRAVY scores for the peptides were calculated by GRAVY CALCULATOR (http://www.gravy-calculator.de/) and the isoelectric points were calculated by ExPASy's Compute pI/Mw tool (http://web.expasy.org/compute_pi/).

Raw data for the alpha casein hydrolysate was lock-mass corrected, deconvoluted, and deisotoped by ProteinLynx from Waters. Database search was conducted on Mascot version 2.2.06 (http://www.matrixscience.com) with the following settings: Nonspecified cleavage, ±30 ppm precursor tolerance, 0.2 Da fragment tolerance, 1+, 2+, and 3+ peptide charges only, variable modifications include Oxidation (M), deamidation of asparagines and glutamine, phosphorylation (STY), and Gln->pyro-Glu (N-term Q),Glu->pyro-Glu (N-term E). The Mascot threshold score for identity was set at $p < 0.05$.

3.2.5 *E. coli* K12 digestion and LC-MS^2.

The *E. coli* K12 cells (*E. coli*, ATCC 47076) were cultured, collected and disrupted. 1 mL of the frozen lysate was thawed and subjected to reduction with dithiothreitol (Sigma), alkylation with chloroacetamide (Sigma), acetone precipitation, trypsin digestion, then desalting and quantification on a C_{18} column. The desalted samples were then injected into the nanoLC-MS^2 system. Raw data was lock-mass corrected, deconvoluted, and deisotoped by ProteinLynx from Waters. Search was conducted on Mascot 2.2 with an *E. coli* K12 database (4337 sequences, 1372643 residues) with the following parameters: enzyme, trypsin; fixed modifications, carbamidomethylation (C); variable modifications, Acetyl (Protein N-term),Deamidated (NQ),Gln->pyro-Glu (N-term Q),Glu->pyro-Glu (N-term E),Oxidation (M),Phospho (S),Phospho (T),Phospho (Y),Ubiquitination_GG (K),precursor mass error, 30 ppm, fragment mass error, 0.2 Da, maximum missed cleavages, 1, number of ^{13}C, 1. Reversed Decoy (default decoy) search was used for determining the false discovery rate (FDR).

3.3 Results and Discussion

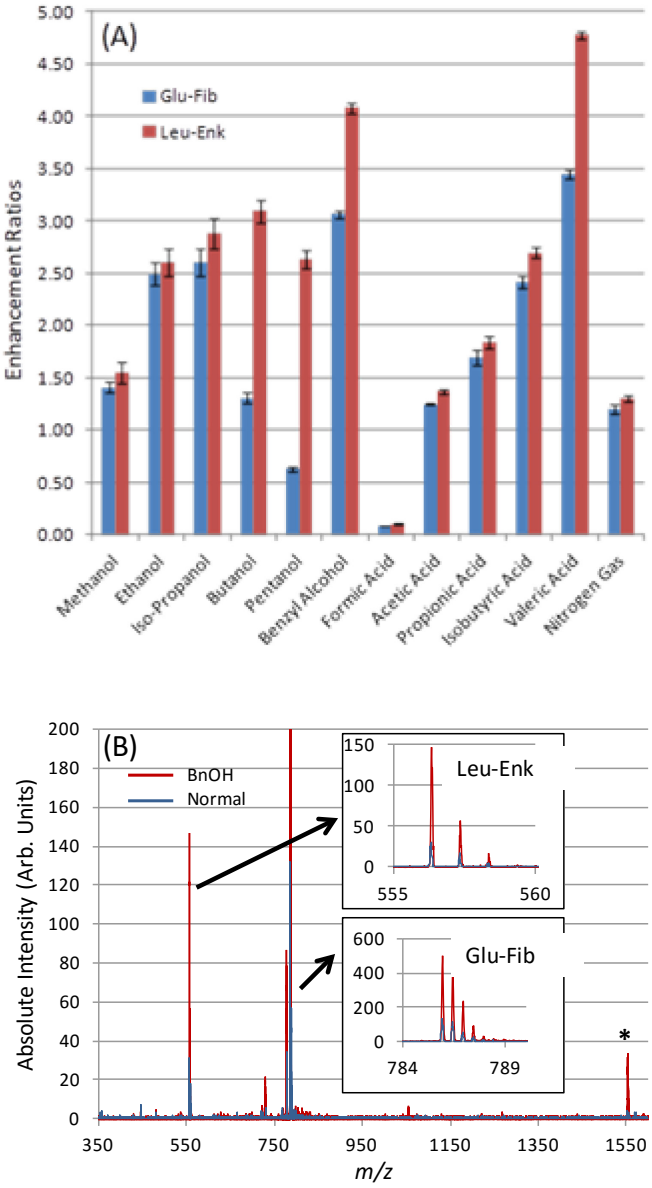


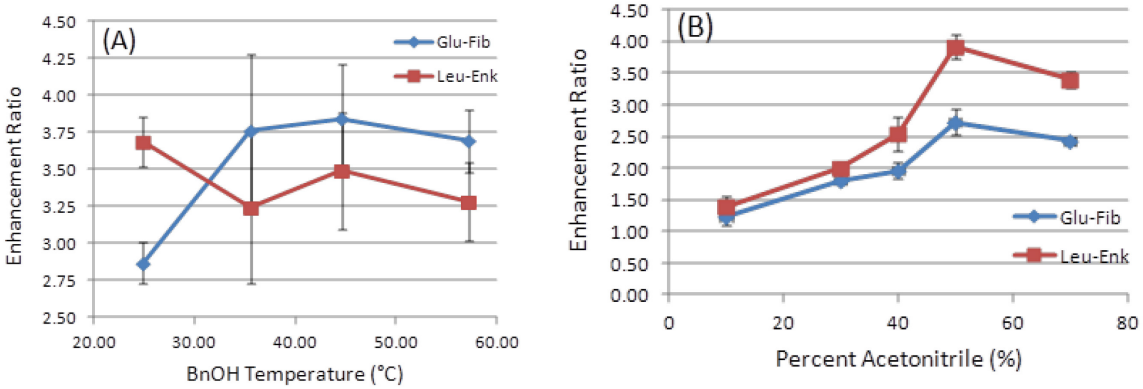
Figure 3.1 (A) Signal enhancement ratios obtained by using different vapors for CAESI. Error bars are 1 standard deviation (n=3). (B) Mass spectra showing the signal enhancement for Glu-Fib and Leu-Enk peaks upon exposure to BnOH. (*) at m/z/ 1552 is the water loss peak of Glu-Fib.

Figure 3.1A shows the signal intensities relative to those with no vapor added (i.e., enhancement ratios) of two peptide ions, singly charged Leu-enkphalin (Leu-Enk^{1+}) (m/z 556) and

doubly charged [Glu1]-fibrinopeptide (Glu-Fib^{2+}) (m/z 786), detected when different chemical vapors were used. The singly charged Glu-Fib, and doubly charged Leu-Enk were not detected under the settings used. Prior to exposing the nESI spray with the vapors, blank runs with nitrogen flow alone showed a minor signal increase, compared to no gas flow (1.20±0.05-fold increase for Glu-Fib and 1.30±0.03-fold increase for Leu-Enk from triplicate results). In contrast, alcohols and carboxylic acids enhance the signals of the two peptide ions by up to 4-fold, compared to no gas flow. Benzyl alcohol (BnOH) increased the signal by 4.08±0.05 times (n=3) for Leu-Enk, and 3.06±0.04 (n=3) times for Glu-Fib. Figure 3.1B shows the comparison of the mass spectra obtained with and without the addition of BnOH. There was no increase in noise level when BnOH was added and thus the use of this chemical vapor resulted in a net increase in analyte signals. It should be noted that, in the initial discovery of the phenomenon that certain chemical vapors could enhance the ESI signals, chemical vapor was introduced directly to the area near the spray tip by placing a vapor inlet tubing orthogonal to the spray needle. However, it was found that adding the vapor to the sheath gas flow is a simpler solution and generated more reproducible results while achieving similar signal enhancement.

At this stage, we cannot determine any strong correlation of chemical properties with the extent of signal enhancement. There appears to be some correlation with chain length and boiling point, especially for the carboxylic acids; however, there is only minor correlation observed in alcohols and a decrease of ionization efficiency was observed for one of the standard peptides as the boiling point increased. There is also little correlation of the proton affinity of the chemical vapors used and the extent of signal enhancement. For example, in Figure 3.1A, as the chain length of alcohols increases from methanol to ethanol and isopropanol, both Glu-Fib^{2+} and Leu-Enk^{1+} peak intensities increase. Using 1-butanol, the Leu-Enk^{1+} peak keeps increase, while there is a

decrease in Glu-Fib^{2+}. For 1-pentanol, the Glu-Fib^{2+} signal becomes suppressed, compared to no vapor added, while the Leu-Enk^{1+} signal is more than doubled. Using carboxylic acids as the chemical vapors, similar trend can be seen in Figure 3.1A except that, in the case of Glu-Fib^{2+}, there is no dramatic drop in signal intensity for longer chain acids as for alcohol series. Although the peptides were in a solvent that contained 0.1% formic acid, the addition of gaseous formic acid greatly decreased the peptide signals. This was due to the disruption of the Taylor cone in the presence of formic acid vapour. The spray quickly resumed after the removal of the formic acid vapours and restarting the spray voltage. Valeric acid gives 3.44±0.04 (n=3) and 4.77±0.04 (n=3) times signal increase for Glu-Fib^{2+} and Leu-Enk^{1+}, respectively, which are greater than BnOH. However, the downside to valeric acid is that it produces extra peaks in the m/z 300-400 region of the spectrum. Isobutyric acid also produces extra peaks in the low mass region. In addition, these carboxylic acids smell unpleasant and are toxic. On the other hand, BnOH is known for its low toxicity[17] and lack of strong odour. Thus, we chose BnOH for the subsequent studies.



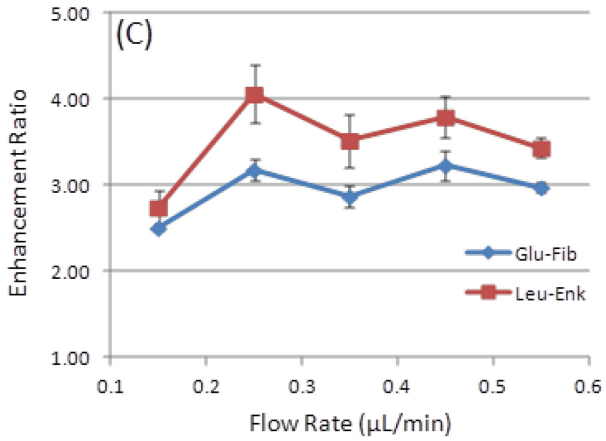


Figure 3.2 Signal enhancement ratios (n=3) obtained at different (A) BnOH liquid temperature, (B) percent organic concentration, and (C) flow rate.

Various conditions that may affect the extent of signal increase were optimized for BnOH-vapor-assisted ESI. The first condition was the vapor concentration and Figure 3.2A shows the effect of heating the reservoir containing benzyl alcohol, thereby increasing the vapour concentration at the nESI spray tip. Between room temperature (23°C) and 60°C, there were only small enhancement of the Glu-Fib peaks and a decrease in the Leu-Enk enhancement. In addition, the enhancement ratio is not as reproducible as that at room temperature. Thus, room temperature operation was determined to be adequate to achieve optimal signal enhancement. Next, the effect of acetonitrile percentage in the mobile phase used to introduce the peptides into ESI was examined. Figure 3.2B shows the signal enhancement ratios of the peptide ions at different acetonitrile percentages. Adding BnOH vapors increases the signals in all cases and more signal gain is attained as the acetonitrile percentage increases. This result suggests that analyte signal enhancement can be achieved across the entire peptide elution time window in a typical LC-MS experiment. Lastly, the effect of LC flow rate was investigated using the solution containing 50% acetonitrile. Figure 3.2C shows that the enhancement increases until 0.250 µL/min where it plateaus off. To summarize, it was found that BnOH produced the greatest signal enhancement when the BnOH reservoir was

set at room temperature and the mobile phase used to introduce the analyte into nanospray contained higher acetonitrile percentages with an optimal flow rate of 0.250 to 0.550 µL/min.

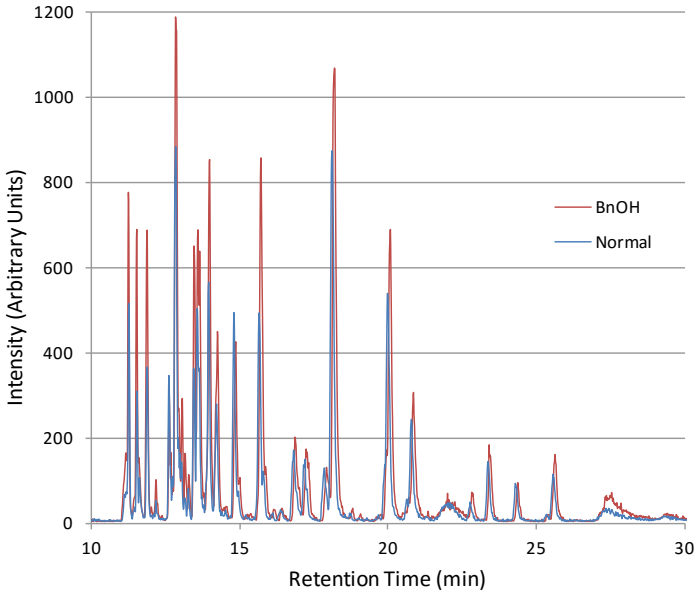


Figure 3.3 Comparison of base peak ion chromatograms between the control run and run with BnOH vapour obtained from 50 fmol injections of BSA trypsin digest.

After finding the optimal conditions, we first applied our technique to LC-MS analysis of a trypsin digest of bovine serum albumin (BSA) to see how peptides of different properties eluted out at different mobile phase compositions respond to BnOH. Figure 3.3 shows the base peak chromatogram with and without adding BnOH vapor. Qualitatively, peak intensities were increased across the entire chromatogram. We then measured the peak areas of extracted ion chromatograms of the BSA peptides, and determined the enhancement ratios. The enhancement ratio ranges from 0.5 to 2.2 across all charge states, with an average increase of 45%. Nine out of 41 peptides show a peak area decrease (between ratios of 0.5 and 0.8) in one of its charge states, but with an increase in the other charge states. Plotting the enhancement ratio as a function of GRAVY score and isoelectric point of each peptide revealed that there was weak correlation ($r < 0.5$) between these factors with the signal increase (Figure 3.4). From these data, we conclude that signal enhancement

was observed for most peptides with no apparent property correlation, and was not a result of charge state shift.

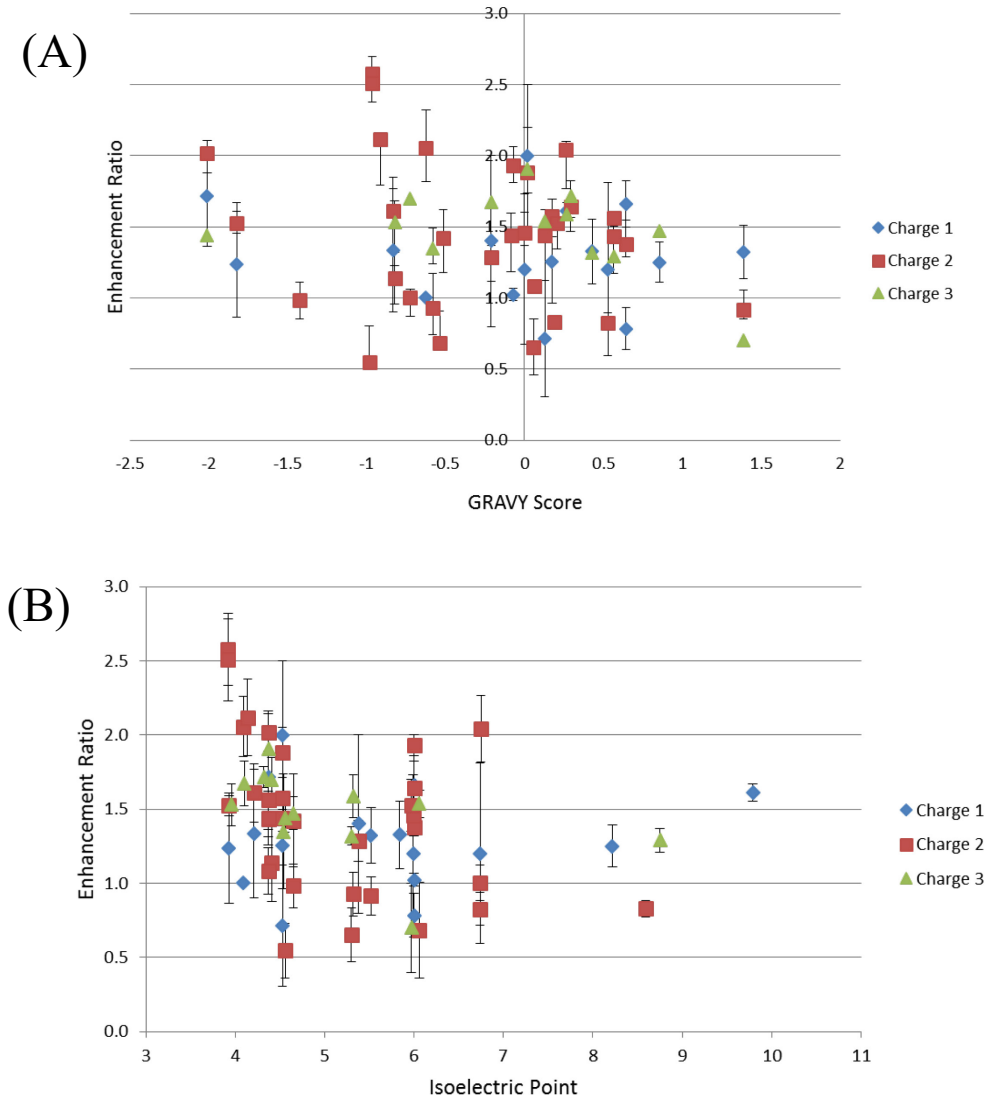


Figure 3.4 Effect of peptide propertie on the enhancement by BnOH vapour. (A) Effect of hydrophobicity (GRAVY score) on signal enhancement of BSA peptides. (B) Effect of pI on signal enhancement of BSA peptides.

The mechanism of chemical-vapor-assisted ESI for increasing analyte signals is very likely related to the enhancement of the ionization efficiency of the analytes during the ESI process. Enhanced ionization, not merely charge-state shifting, by using DMSO has been proposed by

Hahne et al. in their work of using DMSO as a co-solvent added to the analyte solution.[9] The major difference between our method and the DMSO-addition method is that in our method a proper chemical vapor is added to the spray area, not in the liquid stream, thereby there is no need to change any conditions related to sample injection (e.g., trap column operation may be affected by a chemical additive) and LC conditions. We tested DMSO as a chemical vapor and found that it only provided a small signal enhancement (1.4±0.4, n=3, for Leu-Enk and 1.6±0.4, n=3, for Glu-Fib). However, when we added BnOH, instead of DMSO, to the solution of the peptide mixture, we observed a large signal enhancement. The average enhancement factors from triplicate experiments were found to be 4.5±0.1 for Leu-Enk and 3.7±0.1 for Glu-Fib when 1.0% BnOH (v/v) was present in the solution. The enhancement factors were 3.6±0.1 for Leu-Enk and 3.9±0.3 for Glu-Fib with 3.0% BnOH, and 3.49±0.06 for Leu-Enk and 1.62±0.03 for Glu-Fib with 5% BnOH. However, adding BnOH to the solution is not desirable as it would affect the LC separation.

On the mechanism of enhancing ESI efficiency by BnOH vapors, if we assume that the vapors are condensed to the surfaces of the Taylor cone, filament and droplets during the ESI process, higher boiling point and lower surface tension of BnOH, compared to acetonitrile and water, would increase the sequestration of analyte molecules into nano-droplets and hence increase the efficiency of ion production[9]. However, this explanation would not readily apply to the case of isopropyl alcohol, which has a similar boiling point as acetonitrile and would have evaporated before water. On the other hand, if the vapors are not condensed to the surfaces in an appreciable amount, one plausible explanation may be related to the change of the energy barrier[7] that a solution ion needs to overcome to escape the droplet to form a gaseous ion. The added vapor molecules might stabilize the intermediate species, thereby lowering the energy barrier (or destabilizing and increasing it in the case of 1-pentanol or for some peptides). In any case, we feel that the mechanism

of enhancing analyte signals using a chemical vapor may be different from that encountered in the case of adding a chemical modifier to the analyte solution.

Table 3.1 Effect of BnOH on the detectability of peptides generated from microwave-assisted acid hydrolysis of alpha casein.

	# PSM	# Unique peptides	Average match score
Normal (n = 3)	2063±73	492±21	33.14±0.07
BnOH (n = 3)	2752±42	714±10	33.4±0.2
Change	689±84	222±23	0.3±0.2

Table 3.2 Effect of BnOH on the detectability of peptides and proteins from a trypsin digest of *E. coli* K12 cell lysate.

	# Spectra	# PSM	#Unique peptides	# Unique proteins	Average match score (peptide)	FDR (peptide)
Normal (n = 3)	4372±36	1941±22	1257±3	384±2	61.6±0.6	3.3%±0.4%
BnOH (n = 3)	4958±10	2154±35	1422±18	439±7	66.2±0.1	2.8%±0.2%
Change	585±37	212±41	165±18	55±7	4.6±0.6	0.5%±0.4%

A microwave-assisted acid digest of alpha casein sample, using a procedure previously optimized by our group,[18] was analyzed to see if BnOH-vapor-assisted ESI could substantially increase the number of identified peptides in the analysis of more complex protein digest samples. The results from the Mascot searches of the MS^2 spectra generated are summarized in Table 3.1. The total number of peptide-spectrum-matches (PSMs) above the Mascot identity threshold increases from 2063±73 (n=3) to 2752±42 (n=3). The increased PSMs translates to an increase of 222 unique peptides, or 45%. The average score for these non-trypic peptides remains unchanged,

suggesting that the increase in number of PSMs did not result from increase in spurious matches and the data quality remains similar. A trypsin digest of *E. coli* K12 cell lysate was also performed to examine whether peptide and protein identifications could be improved by using BnOH-vapor-assisted ESI. Table 3.2 shows the results from triplicate LC-MS^2 analyses of the digest. Application of BnOH vapors results in an increase of 165 unique peptides (13%) and 55 proteins (14%). The average score of these tryptic peptides increases by 4.6, showing that the stronger signals are giving more confident matches.

Beyond peptides and protein digest samples, several small molecules were tested. N,N-dimethyldodecylamine, reserpine, and syringaldehyde showed little increase in signal, while phenylephrine showed a 3-fold increase. More investigations on the use of the BnOH-vapor-assisted ESI MS technique for possible signal enhancement in analyzing a variety of other molecules will be carried out in the future. In addition, the use of a combination of different chemical vapors to explore their complementarities for enhancing the signals of a wide variety of analytes will be explored.

3.4 Conclusion

In conclusion, we have developed a simple technique, chemical-vapor-assisted ESI, for increasing detection sensitivity of peptides and other molecules in nESI MS. This technique can be safely and readily implemented at minimum cost, i.e., using a solvent bottle, BnOH and a tubing connected to a sheath gas line in nESI MS or a tubing with its outlet placed close to the spray tip.

3.5 Literature Cited

(1) Gatlin, C. L.; Kleemann, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. *Anal. Biochem.* **1998**, *263*, 93-101.

(2) Zhu, G. J.; Sun, L. L.; Yan, X. J.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 2569-2573.

(3) Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M. *Mol. Cell. Proteomics* **2011**, *10*, 1-9.

(4) Page, J. S.; Marginean, I.; Baker, E. S.; Kelly, R. T.; Tang, K.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2265-2272.

(5) Contino, N. C.; Jarrold, M. F. *Int. J. Mass Spectrom.* **2013**, *345*, 153-159.

(6) Shuford, C. M.; Muddiman, D. C. *Expert Rev. Proteomics* **2011**, *8*, 317-323.

(7) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. *Anal. Chem.* **2013**, *85*, 2-9.

(8) Meyer, J. G.; A. Komives, E. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1390-1399.

(9) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Medard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat. Methods* **2013**, *10*, 989.

(10) Apffel, A.; Fischer, S.; Goldberg, G.; Goodley, P. C.; Kuhlmann, F. E. *J. Chromatogr. A* **1995**, *712*, 177-190.

(11) Kharlamova, A.; Prentice, B. M.; Huang, T.-Y.; McLuckey, S. A. *Anal. Chem.* **2010**, *82*, 7422-7429.

(12) Kharlamova, A.; McLuckey, S. A. *Anal. Chem.* **2011**, *83*, 431-437.

(13) Winger, B. E.; Light-Wahl, K. J.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 624-630.

(14) Hopper, J. T. S.; Sokratous, K.; Oldham, N. J. *Anal. Biochem.* **2012**, *421*, 788-790.

(15) Hassell, K. M.; LeBlanc, Y. C.; McLuckey, S. A. *Anal. Chem.* **2011**, *83*, 3252-3255.

(16) Nguyen, S.; Fenn, J. B. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1111-1117.

(17) Nair, B. *Int. J. Toxicol.* **2001**, *20 Suppl. 3*, 23-50.

(18) Wang, N.; Li, L. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1573-1587.

Chapter 4

Construction of a High Resolution Human Metabolite MS^2 Library*

4.1 **Introduction**

Liquid chromatography coupled mass spectrometry has started a revolution in the field of metabolomics. Thousands of metabolite mass features can be detected from a single sample, and hundreds of samples can be run in a week. The high-throughput analysis of metabolite features facilitates the discovery of biomarkers during metabolomics study of diseases.[1-3]

After data acquisition of samples and multivariate analysis, the final results are metabolite features that can discriminate between healthy controls and patients. Due to the large number of metabolite features profiled in every study, the final list could be between 50 to 100 metabolite features long. In order to understand the important biological processes that is at work in the studied disease, these metabolite features must be identified and placed in their respective metabolic pathways.

Using LC-MS, this identification process is usually done by searching the accurate mass of these unknowns against the measured or predicted accurate mass of metabolites in a repository, such as the Human Metabolite Database (HMDB).[4-6] However, even with low error in the measured

*Mingguo Xu, Jaspaul Tatlay, Tran Tran Ngoc, Tao Huan, Dr. Yiman Wu, Dr. Ruokun Zhou, Dr. Chiao-Li Tseng, and Wei Han contributed to the collection of data in this chapter. HMDB standards were supplied by Professor Wishart and Edison Dong.

mass of 1 ppm, this approach often cannot give unambiguous identification due to a large number of metabolites that have nearly identical masses.[7,8]

CID fragmentation can be used to improve the identification power of the LC-MS platform. In this process, unknown metabolite ions are isolated and accelerated through a cloud of gas molecules. Collision with the gas molecules imparts energy and causes bond breakage, generating many fragments. The fragment spectrum of each metabolite feature is unique, and their identity can be elucidated with manual interpretation, computer assisted interpretation, or matching with fragmentation spectra of standards.

Due to a wide chemical variety of metabolites, manual interpretation is often too laborious, and sometimes unreliable for unambiguous identification of large numbers of unknowns. Computer assisted interpretation can process large number of unknowns; however, at the time of writing, they are also unreliable due to poor understanding of the fragmentation mechanism for many metabolites. There has been recent work in creating algorithms that aim to predict fragmentation using statistical methods,[9-11] but they are still far from being useful as standalone tools for identification.

Matching CID spectra with spectra from purified standard remains the method of choice. In this way, each fragment can be verified to be from the metabolite of interest and a perfect match usually guarantees the identity of an unknown. With batch searching mode, library programs can process hundreds of unknown spectra in a few seconds. There are large MS^2 libraries already in existence, covering many different classes of chemicals from pharmaceuticals to natural products, and some are freely available on the internet.[6,12-14]

Due to the age of these libraries, many were built using low resolution ion-traps and triple quadrupoles mass spectrometers. Newer high-resolution mass spectrometers based on QTOF designs, such as the Bruker Impact HD, can generate spectra with better mass accuracy and higher resolution than older instruments. Better quality data increases identification confidence, because fragment peaks can be matched with less error tolerance.

In order to take advantage of this QTOF platform for future human metabolome analysis, a new high resolution HMDB MS^2 library, containing only endogenous human metabolites, was built. The library was built using rigorous data acquisition and data processing procedures that ensured the library was high quality and free from chemical noise.

4.2 Methods

4.2.1 Chemicals and Reagents

All chemicals and reagents were purchased from Sigma-Aldrich Canada (Markham, ON, Canada) except those otherwise noted. LC–MS grade water and acetonitrile were purchased from Thermo Fisher Scientific (Edmonton, AB, Canada).

4.2.2 Instrumentation

All data was generated with an impact HD QTOF mass spectrometer (Bruker Daltonics, Germany). Flow injection analysis was performed with an 1100 HPLC (Agilent, USA) with no column attached.

4.2.3 QC Samples

The purpose of QC samples were to monitor the fragmentation reproducibility of the instrument and the sensitivity. Therefore, they were made of only a single compound at a low

concentration. Positive mode MS^2 QC: 1 μM of melatonin in 1:1 (v/v) H_2O:ACN 0.1% formic acid. Negative mode MS^2 QC: 1 μM citric acid in 1:1 (v/v) H_2O:ACN 0.1% formic acid.

4.2.4 HMDB Standards

Metabolite standards were obtained from Professor David Wishart at the University of Alberta, Canada. Roughly 1 mg of solid or liquid were used to prepare a stock solution of every standard in 1:1 (v/v) H_2O:ACN. Stock solutions were then diluted approximately 500 fold with 1:1 (v/v) H_2O:ACN 0.1% formic acid. Diluted samples were transferred to 2 mL HPLC injection vials prior to analysis.

4.2.5 Library Data Acquisition

The 1100 HPLC autosampler was programmed as follows:

1. 4 μL of 5 mM lithium formate 1:1 (v/v) H_2O:ACN 0.1% formic acid (v/v)
2. 3 μL of air
3. 25 μL of sample
4. Inject

Flow injection conditions were:

Mobile Phase A: H_2O with 0.1% formic acid (v/v)

Mobile Phase B: Acetonitrile with 0.1% formic acid (v/v)

Entire run was 50% A, 15 μL/min for 1.4 minutes, followed by 110 μL/min until 2.2 minutes. Total time 2.2 minutes.

Mass spectrometer source settings were: Nebulizer gas 2.0 bar, dry gas 8.0 L/min, 200 °C, capillary voltage 4.5 kV, capillary offset 0.5 kV, acquisition rate 8 Hz, mass range 20 – 1000 m/z, CID gas was nitrogen.

There were 12 individual sections in the 2.2 minutes MS method. The first section was the MS precursor measurement, followed by 5 sections of MS^2 with 6 m/z isolation window, and repeated with 5 more sections of MS^2 with 1 m/z isolation window. For the 5 sections, different collision energies were used. Sections 1 to 5 used 10, 20, 30, 20-50, 40 eV collision energies, respectively. The last section measured the lithium formate signal in MS mode. The scan rate was 0.25 Hz.

A QC sample was run for every 20 HMDB standards. The fragment spectrum of the QC sample was searched against an initial measurement of the QC sample. The QC was passed as long as the fit score for the match was above 900.

All samples were first measured in positive mode. Only those samples that were not ionized in the positive mode were then measured in negative mode.

4.2.6 Library Data Processing

Raw data was automatically processed by a script, written in house using the data processing program DataAnalysis (Bruker, Germany). This script automatically averages all of the scans within one section, and then labels each averaged scan with the name and collision energy. It then performs automated mass calibration based on the last section containing lithium formate signals.

Each scan was then inspected manually for signal intensity and mass accuracy. Mass error must be less than 2 ppm, and intensity must be less than 3×10^6 counts. The 6 m/z window MS^2

scans were compared with 1 m/z window MS^2 scans, to remove noise; this process is detailed in the discussion. 2 fragment structures were assigned to the 2 most intense peaks in the 20-50 eV MS^2 spectrum. Based on the chemical formula of the metabolite, DataAnalysis' SmartFormula3D tool was then used to assign chemical formulas to all of the fragment peaks based on their accurate masses. Simulated fragment spectra for all collision energies were generated using the calculated formulas for each energy level.

4.2.7 Apple juice experiment

Human urine was collected from a healthy volunteer. Prior to drinking juice, the volunteer drank one cup of water in the morning and collected urine. Immediately following urine collection, one cup of apple juice was consumed. The first urine after drinking juice was collected. This urine sample was then filtered by 0.22 μm PVDF syringe filter (Milipore) and stored at -80 °C. Before injection, the samples were thawed, then diluted by half with 1:1 (v/v) H_2O:ACN 0.1% formic acid.

An Ultimate 3000 RS-LC system from Thermo Scientific (California, USA) was coupled to the impact HD QTOF mass spectrometer. Column was a Waters 1.7um 100x2.1 BEH C_{18} w/ VanGuard BEH C_{18} guard column (Waters, USA). Column was kept at 30 °C with a column oven. The final gradient chosen was: 0 min (1% B), 0-2.0 min (1% B), 2.0-17 min (1-99% B), 17 min (99% B), 17-20 min (99% B). The flow rate was 250 μL/min and the injection volume was 5 μL. A wash and equilibration injection was run between samples; the gradient was: 0-4.5 min (100% B), 4.5-10 min (1% B) at 350 μL/min. Mobile phase A was 0.1 % LC-MS formic acid in LC-MS water and mobile phase B was 0.1 % LC-MS formic acid in LC-MS acetonitrile.

The instrument was operated in the MS^1 mode, with no fragmentation. Each sample was injected in triplicate and in the positive MS mode. Raw data was calibrated and molecular features

were extracted with ProfileAnalysis 1.3 (Bruker Daltonics, Germany) software. The software then aligned each of the molecular features together, across all samples, and tabulated the molecular features' peak areas for multivariate analysis. Each molecular feature's peak area in one sample was normalized to the sum of the peak area of that feature across all samples. PLS-DA was performed with the ProfileAnalysis with 3 principle components. A clear separation was observed between the before and after drinking apple juice classes. The list of important metabolite features were then taken from the loadings plot generated by the software.

An LC- MS^2 experiment was performed on one of the urine sample. An MS^2 inclusion list was used to ensure that CID spectra were obtained for the metabolite features. From the CID spectra, a library search was performed using DataAnalysis with a 5 ppm precursor mass tolerance, and results greater than >700 fit score were reported.

4.3 Results and Discussion

4.3.1 Library Construction Considerations

In order to facilitate the MS acquisition of the 803 HMDB standards, it was decided to use flow injection analysis MS (FIA-MS) instead of performing LC-MS analysis. Each dissolved sample was picked up by the autosampler followed by 4 µL of lithium formate; this plug of sample and calibrant was then pushed into the ESI source of the QTOF-MS where data were acquired in positive or negative polarity. LC-MS analysis would have provided useful retention time information and removed interfering contaminants; however, performing gradient elution for all the standards would have taken too much time. An LC-MS analysis would have needed 30 minutes per sample, whereas FIA-MS required only 2 minutes. An advantage of FIA-MS was that the standard's signal could be observed for the entire 2 minutes analysis (Figure 4.1). During this time,

10 different CID experiments can be performed; for each experiment, many spectra were averaged to give spectra of high intensity and signal-to-noise ratio. This would not have been possible in an LC-MS analysis where analytes elute over a time peaks widths are less than 6 seconds.

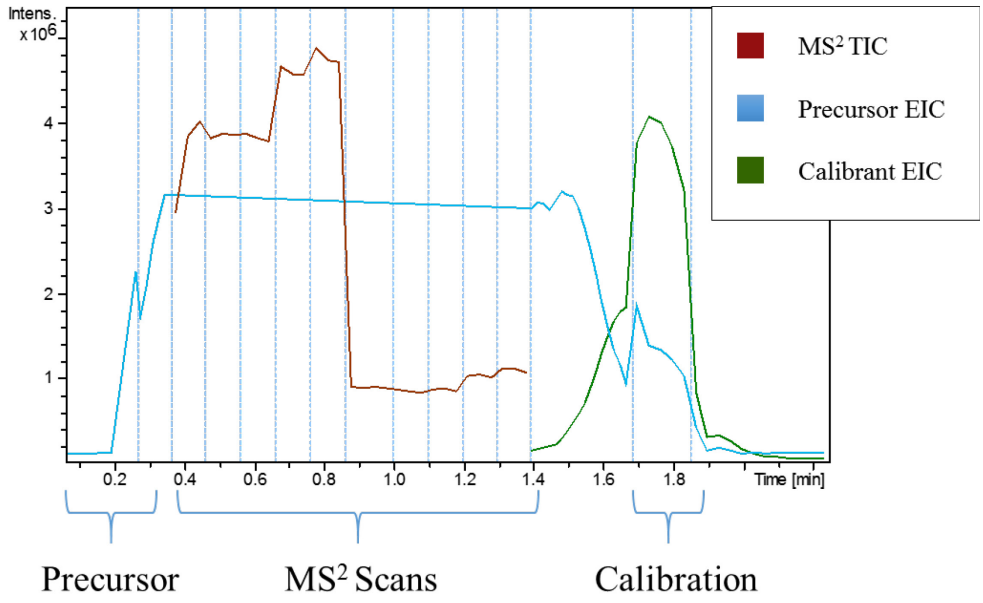


Figure 4.1 An example of the flow injection analysis performed on a HMDB standard. Important sections of the analysis are labeled below.

Five different collision energy conditions were collected for all of the HMDB standards: 10 eV, 20 eV, 30 eV, 40 eV, and a combined spectrum of 20 eV and 50 eV. These collision energies ensured that compounds of different robustness towards CID fragmentation were able to generate at least one or more spectra that contained unique fragmentation patterns. Compounds that contained fragile bonds gave rich fragmentation patterns in the lower energy, but at higher collision energies they were completely fragmented and gave few signals that can be used for identification. Compounds that contained rigid structures, such as porphyrins, were found to be highly resistant to fragmentation. At the lower collision energy there were no observable fragment signals, and only higher energy spectra contained signals. These observations suggest that for metabolomic

analysis, different collision energies are needed in order to produce useful CID spectra for identification.

The identities of metabolites are unknown in untargeted metabolomics, therefore it is difficult to select an optimal energy during the analysis. The combined 20 eV and 50 eV spectrum was collected as a generic setting that can be used in any untargeted metabolomic experiment so that there would be useful fragments for all compounds. During each MS^2 scan, the isolated molecule was fragmented at 20 eV for half of the scan time and then fragmented for 50 eV for the other half of the scan time. The two spectra were then averaged together to give the 20-50 eV spectrum. By using this energy setting both fragile and robust compounds can produce unique MS^2 spectra. The trade-off for using this method was that only half the scan time was used for each collision energy. Compounds that fragment with only one of the energy level would have fragmentation signal for half the time, while no signal for the other half. When these two scans are summed together, it results in a lower quality fragment spectrum than if the optimal energy was used for the entire scan time. This was also the reason that only 2 collision energies were selected for the 20-50 eV CID experiment. Adding more collision energy steps to the combined spectrum might give a richer fragmentation pattern, because fragment signals can be collected at the intermediate energies; but more scans would have meant that less scans can be summed together for each collision energy, resulting in lower signals. In the end, it was found that the 20-50 eV was the most generic collision energy, and the most useful for untargeted metabolomic studies.

The long acquisition time of FIA-MS also permitted the use of two different isolation windows for all 5 different collision energies. One wide isolation window of 6 m/z and one narrow isolation window of 1 m/z were used. During LC-MS^2 analysis, there are many metabolite eluting from the column at any one time. The instrument isolates the compound ion of interest and then

adjusts the AC current in the first quadrupole to only allow that specific mass through to the collision cell to be fragmented. The isolation window has a finite width around the mass of interest, and is user adjustable. The choice of window width is a balance between selectivity and sensitivity. Due to the parabolic transmission profile of the quadrupole fields, the optimal mass windows are usually selected to be several m/z around the ion of interest. However, this usually means that other ions around the m/z of the ion of interest are co-isolated. The resulting fragment spectrum contains signal from many different ion species, which convolutes the identification process. Using smaller windows ensure that only the compound ion of interest enters the collision cell and the resulting fragments are specific to that compound. The sacrifices for increased specificity are: lower fragment intensity due to transmission losses with smaller transmission window; and loss of isotopic peaks, which can be used for chemical formula determination of the fragment peaks (Figure 4.2).

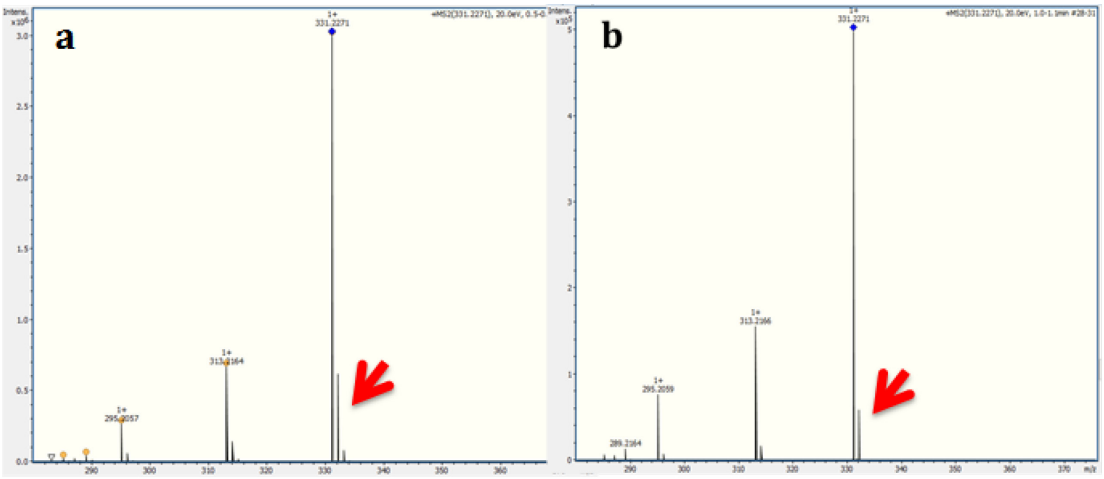


Figure 4.2 MS^2 spectra of 17-Hydroxyprogesterone, acquired with (a) 6 m/z isolation window and (b) 1 m/z isolation window. Red arrows show the loss of isotopic peaks at the two different window settings.

In order to develop a high quality MS^2 library, the advantages of both isolation window widths were combined. Some of the HMDB standards were not pure and contained side products

or contaminants that have masses close to the $[M+H]^+$ of the standard and were co-isolated by the wide 6 m/z isolation window. The final data for the library cannot contain any co-isolated chemical noise peaks, and it cannot miss crucial isotopic peaks, which are used for formula calculation. To realize this goal, manual data processing was done by comparing the spectra from 6 m/z with the 1 m/z isolation window. Spectra acquired with the 1 m/z isolation window were assumed to be pure, because interfering species were not co-isolated. Any fragment peaks in the 6 m/z window spectrum that did not appear in the 1 m/z window were due to co-isolated interferences and were deleted manually (Figure 4.3). The processed 6 m/z window spectra, which contained complete isotope pattern for all the fragment peaks without interferences, were imported into the final library. This labour intensive data processing yielded clean final library spectra that reduced the chance of spurious matches to chemical noise peaks in the library.

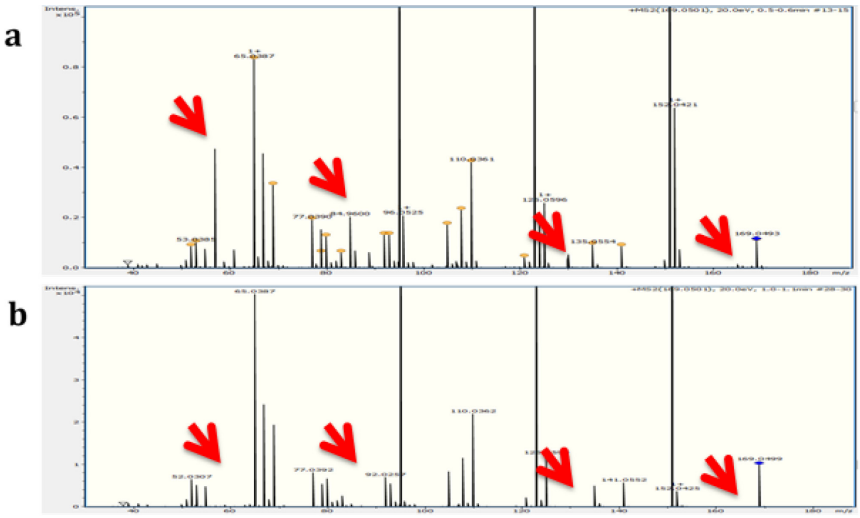


Figure 4.3 MS^2 spectra of 5-Methoxysalicylic acid, acquired with 6 m/z isolation window (a) and 1 m/z isolation window (b). Red arrows show the contaminant peaks that were present in the 6 m/z isolation window spectrum, that were not present in the 1 m/z isolation window.

One further data quality improvement were implemented through the simulation of fragment spectra based on the calculated molecular formulas for fragment peaks. This

"SmartFormula3D" algorithm was included in the DataAnalysis data processing software, and it calculated the chemical formulas of fragments using the accurate mass and isotopic distribution in the MS^2 spectra. Only fragments with calculated formulas that were subsets of the parent molecule's molecular formula were considered to be true fragments. The true fragments were simulated with the theoretical accurate mass and theoretical isotopic distribution, to create a new virtual copy of the MS^2 spectrum. It was found that this algorithm did as good a job at identifying co-isolated interferences as the above two window method, however, it was often not able to identify low intensity fragment peaks that were experimentally confirmed to be real fragments due to bugs in the algorithm. Therefore, the utility of the final virtual $MS^2$is limited, due to the missing peaks.

Besides the protocols, the high resolution QTOF helped to generate spectra with high resolution and high mass accuracy. The peak FWHM from this instrument was in the range of 0.01 m/z which converted to a resolving power of 60,000 at a mass of 622 m/z. The typical mass error of a calibrated spectrum is usually around 3 ppm. A previous version of the HMDB MS^2 library, available on the HMDB website, was measured using an ion trap mass spectrometer capable producing data with a peak FWHM of 0.8 m/z mass error of greater than 100 ppm. The high resolution data was also able to be acquired on a time scale that was compatible with ultrahigh performance liquid chromatography (UHPLC) flow rates, commonly used in rapid metabolomic analysis. The acquisition rate can be increased to 8 Hz without the loss of resolution, unlike with other high resolution MS such as the orbitrap.[15] At 8 Hz, peak of less than 6 seconds wide can be accurately reproduced, improving data quantification.

Both the high resolution and high mass accuracy of the QTOF are essential for metabolite identification using CID MS^2. Figure 4.4 shows the improved MS^2 data quality acquired with the

QTOF when compared with ion trap MS (IT-MS) MS^2 data. The high resolution spectrum showed greater detail in the busy low-mass region of the spectrum, and this detail will be crucial for identifying compounds from spectra that contain co-isolated fragment peaks. On a low resolution instruments co-isolated fragment peaks will be un-resolved, causing reduced matching scores with the MS^2 library.

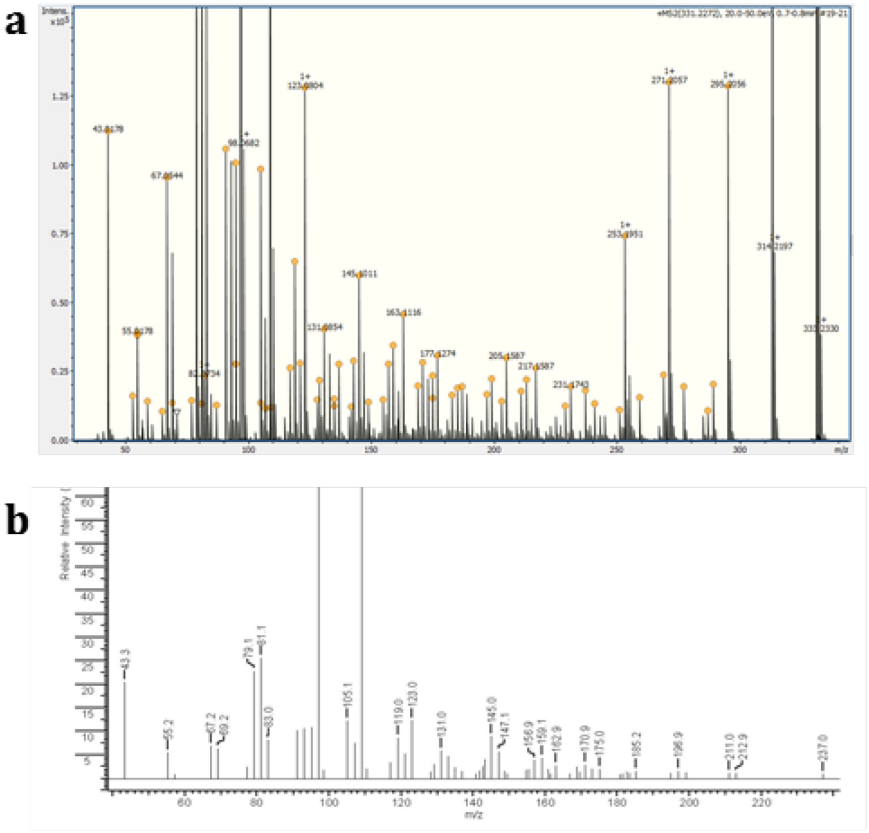


Figure 4.4 MS2 spectra of 17-Hydroxyprogesterone, acquired with the (a) impact HD and (b) QTRAP 2000.

High mass accuracy is used to narrow down the identity of unknowns with their measured $[M+H]^+$ masses. For example, the accurate $[M+H]^+$ mass of glutamine (147.076419 m/z) was searched in the HMDB database with a specified mass tolerance of 3 ppm, the same accuracy obtained on the QTOF, 5 matches were returned by the query. All 5 matches were isomers of

glutamine, and therefore indistinguishable from the correct identification. When searched with the IT-MS mass accuracy of 100 ppm there were 25 matches or 5 times more possible identities. The larger number of incorrect matches complicates the metabolite identification process, and requires more resources to validate.

The performance contrast between the current generation QTOF and the older IT-MS showed that the QTOF would be the future platform for generating rapid LC-MS data for metabolomic analysis. Therefore, new MS^2 data need to be generated using this platform in order to facilitate metabolite identification in metabolomic workflows using the new QTOF instrument.

4.3.2 Final Library Analysis

Table 4.1 Chemical Composition of the HMDB standards used to create the library.

HMDB Classes	
Organic acids	55
Lipids	151
Aliphatic	83
Carbohydrates and Carbohydrate Conjugates	69
Amino Acids, Peptides, and Analogues	137
Aromatic	230
Organophosphorus Compounds	5
Homogeneous Non-metal Compounds	10
Nucleosides, Nucleotides, and Analogues	44
Unspecified	20

In total, 803 separate HMDB standards were measured on the QTOF platform, encompassing a wide range of classes of human metabolites that can be found in blood, urine, and cerebrospinal fluid. Table 4.1 summarizes the chemical variety of the HMDB standards. Out of the 803, 101 compounds did not produce observable $[M+H]^+$ or $[M-H]^-$ signal, and therefore yielded

no usable MS^2 spectra. A closer investigation of the 101 compounds was conducted, and it was found that 62% were from the lipid and aromatic compound classes (Table 4.2). These two classes are difficult to ionize using ESI because they lack basic sites that facilitate protonation. In order to measure these compounds using MS, other ionization techniques such as atmospheric pressure photoionization (APPI) and atmospheric pressure chemical ionization (APCI) are required.[16] However, the majority of compounds ionized efficiently with ESI and ESI has been the ionization method of choice for a large number of MS labs conducting metabolomic studies. Thus, it was decided to complete the first version of the library with an ESI source only.

Table 4.2 Composition of the standards that were not ionized.

Non ionizable compounds	
Organic acids	5
Homogeneous Non-metal Compounds	8
Lipids	32
Amino Acids, Peptides, and Analogues	3
Aromatic	9
Aliphatic	30
Carbohydrates and Carbohydrate Conjugates	9
Nucleosides, Nucleotides, and Analogues	3
Organophosphorus Compounds	1

For some organic acids and carbohydrates that did not ionize in positive mode, negative mode was used to obtain an $[M-H]^-$ signal and the negative mode MS^2 spectra were acquired. 74 compounds were detected in negative mode only. It was found that negative mode MS^2 produced poor fragmentation, except for polymers such as polysaccharides. Out of 74 negative mode only compounds, 32 had no detected fragment ions at any collision energies, and 18 had measureable

fragment ions only in the lower collision energies. With compounds that had measurable fragment ions, they were generally of much lower intensity with respect to the parent ion intensity (Figure 4.5); this behaviour was not observed in the positive mode data. Poor fragmentation of C-C bonds in negative mode was previously reported for lipids,[17] and simple carboxylic acids.[18] From these observations, it was concluded that negative mode was not suitable for untargeted metabolite profiling work because of poor fragmentation. The exceptions are cases where the compound contains negatively charged head groups or ester bonds that fragment in negative mode to give useful information, such as in the analysis of oligosaccharides[19,20].

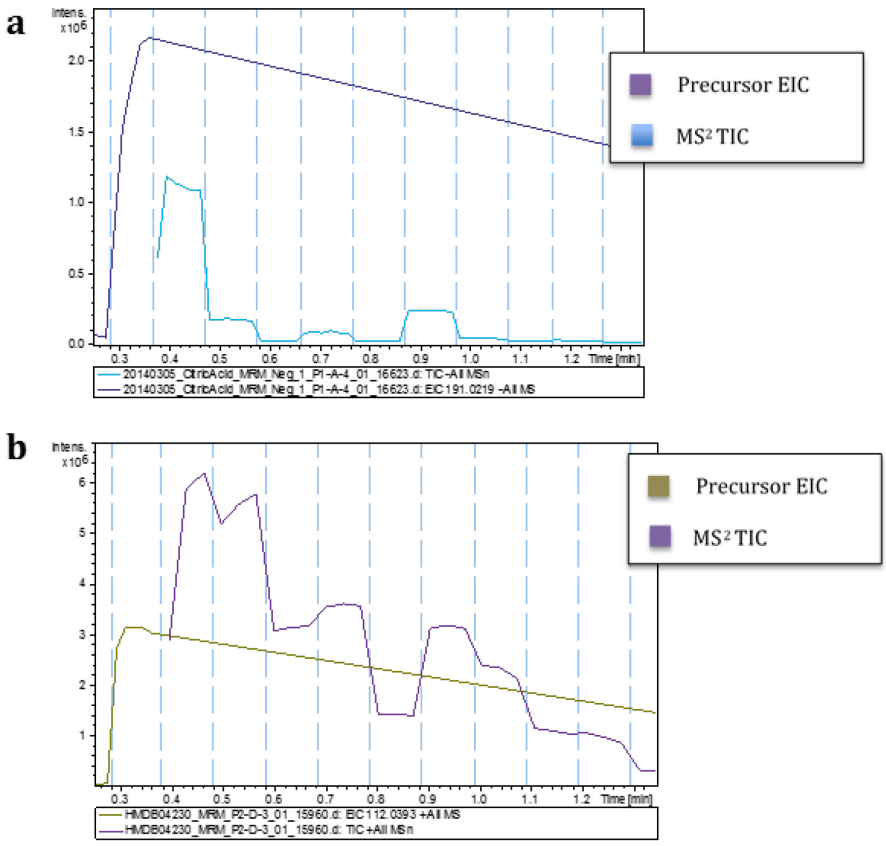


Figure 4.5 Comparison of ion intensities of negative mode CID (a) versus positive mode CID (b). Compared to positive mode, fragment intensities in negative mode is substantially lower with respect to its $[M-H]^-$ precursor.

When the library was being built, metadata were kept on a variety of parameters such as standard concentration, solubility, parent ion intensity, amount of in-source fragmentation, and adduct formation. These information were tabulated in an excel file that was included with the final library package. Information such as in-source fragmentation, parent ion intensity, and adduct formation aid in the optimization of targeted metabolite analysis. When a compound is expected to have strong in-source fragmentation according to the metadata, the $[M+H]^+$ ion would be of low intensity. Monitoring the $[M+H]^+$ in such a situation would reduce the sensitivity of the method. Sensitivity can be improved by monitoring the in-source fragment, instead. If a compound was expected to form sodium adducts more readily than forming $[M+H]^+$, sodium could be spiked into the sample to increase sensitivity.

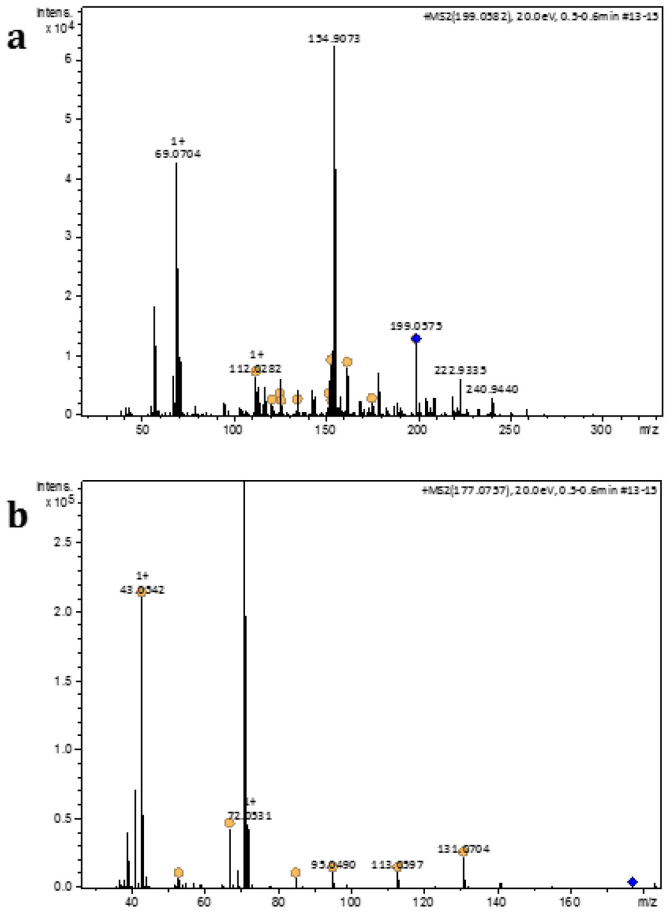


Figure 4.6 CID MS2 spectra of sodiated (a) and protonated (b) 2-Isopropylmalic acid. Yellow dots marks peaks that can be explained by the SmartFormula3D algorithm. All peaks in the protonated MS2 spectra can be explained, whereas a large number of peaks are unexplainable in the sodiated MS^2 spectrum.

Due to the frequency of observing $[M+Na]^+$ ions with higher intensity than the $[M+H]^+$ ions, the fragmentation efficiency of these sodium adducts was investigated to see if they could be used to generate unique MS^2 spectra for metabolite identification. Figure 4.6 shows sodiated ions of small molecules rarely provide useable fragmentation and require more energy to fragment. In this figure, 2-Isopropylmalic acid's $[M+H]^+$ ion generated peaks that were identified as fragments of the parent ion by calculating their molecular formula. The $[M+Na]^+$ spectrum was filled with unidentified masses and masses above the $[M+Na]^+$ mass, therefore the majority of the observed

signal were chemical noise, making the data unsuitable for spectral identification. Furthermore, the parent $[M+Na]^+$ ion was more resistant towards fragmentation and was still present in the MS^2 spectrum, while the parent $[M+H]^+$ ion was completely fragmented in the MS^2 spectrum.

The situation improves when fragmenting $[M+Na]^+$ of polymers, such as oligosaccharides. These polymers produced clean MS^2 spectra that were completely different from their $[M+H]^+$ MS^2 spectra and the fragments were explainable by molecular formula calculation and can be used for identification. However, these large sodiated polymers did not make up a large portion of the library.

From these observations, fragmentation of $[M+Na]^+$ ions will not give useful experimental data and should be avoided for small molecule analysis. For this reason, sodiated MS^2 spectra were not included in this version of the library.

4.3.3 Searching Strategies

The spectral matching algorithm for identifying experimental spectra with the library was built into the DataAnalysis. The algorithm employed was a classic fit and purity score calculation. The fit score reflected the presence of fragment peaks from the library standard in the experimental spectrum. Extra peaks in the experimental spectrum did not penalize the scoring, as long as all library standard peaks were found. The purity score looked for exact matching between library spectrum and experimental spectrum; extra unmatched peaks will reduce the score. Fit score was found to be the most useful metric for identification of unknowns with real complex samples such as urine. The reason was that due to the large number of metabolites in the body, each mass spectrum in the chromatogram contains many metabolites within the same MS^2 isolation window of 2-3 m/z. These metabolites were co-isolated and contributed to extra fragment peaks in addition

to the fragments from the parent ion of interest. It was found that using the fit score was more resistant to this issue, because it did not take into account extra peaks in the experimental spectrum. As long as all of the library spectrum peaks were found, the fit score was high. This problem is also commonly encountered in LC-MS based proteomic analysis, and creative solutions were proposed.[21,22] However, for metabolomics there are no predictable fragmentation rules that can be applied, so those solutions cannot be used.

Unambiguous identification of a particular ion in a mass spectrum requires two parameters: a good match to the fragmentation pattern of a standard, and the precursor mass that gave rise to the fragmentation pattern must also match to the mass of the standard within an instrument specific mass tolerance. In untargeted metabolomics, matching both criteria limits the amount information that can be derived from the experimental data. In a typical biological sample there is a vast number of metabolites, certainly many times more than possible in any standards library. Most of these metabolites are built on similar chemical scaffolds and share parts of their chemical structures; while their intact precursor masses are different they will share certain fragment ions because the core structures are similar. In these cases, by finding similar fragmentation patterns without specifying the precursor mass can help elucidate the chemical class and provide some information of biological function. This expands the power of the library for metabolomic applications, without the need for additional standards.

Table 4.3 Table of identified metabolites in human urine.

Biofluid	MS^2 Spectra	Putative ID	Similar Fragments

Plasma	4544	45	151
Urine	2558	49	163
Saliva	3392	18	101

The spectral matching algorithm comes with the ability to toggle off the requirement to match the precursor mass and perform only fragment pattern matching. These functions were tested with human urine samples. A short LC-MS^2 experiment yielded 45 putatively identified metabolites (Table 4.3) with a 3 ppm precursor mass tolerance, and a fit score of above 700 which represents a good match. By removing precursor matching, the metabolite number was tripled to 151 matches, with a fit score of above 700. Figure 4.7 shows an example of this type of match. Precursor ion 369.1710 m/z was selected for fragmentation and the resulting pattern matched closely with the pattern from the dehydroepiandrosterone standard. From this information it was concluded that the unknown was closely related to the steroid, but it was 80 m/z heavier. Upon literature search it was found that dehydroepiandrosterone was usually found to be sulfated to induce a negative charge and modify the transport behavior in the body.[23] By using this strategy, identification power of the library can be expanded beyond the standards that were measured.

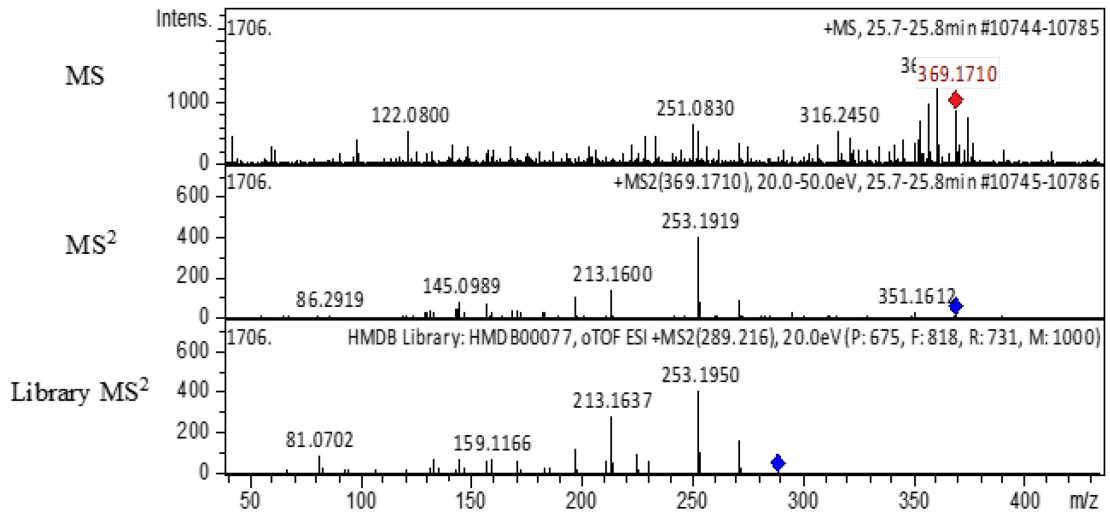


Figure 4.7 Experimental MS^2 spectrum of ion 369 m/z compared with the library MS^2 spectrum of dehydroepiandrosterone.

4.3.4 Proof-of-Concept

A major advantage of the Bruker suite of post-processing software was its tight integration for the entire metabolomic workflow. Large batches of raw data from an untargeted metabolomic study can be seamlessly calibrated, processed, and then passed on for multivariate analysis, such as PCA or PLS-DA, without complicated file format conversions required when using popular open-source software.[24-27] Important masses can be determined from the multivariate analysis, and the library annotates the identity of the unknown mass features. The entire workflow was demonstrated by a proof-of-concept study in which LC-MS^2 was used to discover apple juice metabolites from a healthy volunteer's urine.

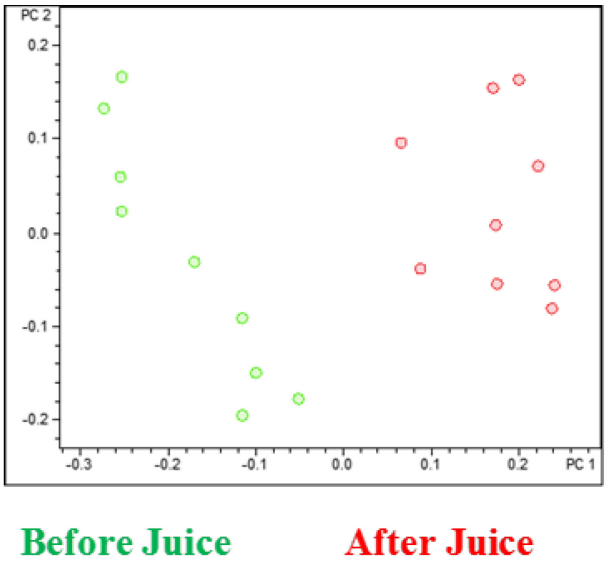


Figure 4.8 PLS-DA scores plot for the apple juice metabolite data set. Clear separation of urine samples from before and after drinking apple juice was observed.

Figure 4.8 shows that the PLS-DA was able to fully separate the before and after drinking juice samples. The Bruker multivariate analysis software also produced a list of masses that was ranked based on how much it contributed to the separation. Higher rank means the mass changed abundance the most between urine samples before and after drinking apple juice. The library was used to identify the compounds that were on the list, and the results are shown in Table 4.4. When exact mass and fragmentation pattern were used for the identification, 4 unknowns were identified, including the top 3 metabolites that contributed to the difference between before and after juice samples. With fragment pattern matching only, 3 additional unknowns were identified to be all carnitine related compounds. All identified metabolites were found to have metabolic relevance to apple juice consumption as reported in previous studies. Uric acid was reported to be the major metabolite of fructose,[28] which is in high abundance in apple juice. Hippuric acid was linked to the metabolism of polyphenols that were found in apple juice.[29] L-carnitine and acetyl-L-carnitine were found to be involved in the regulation of glucose metabolism,[30] and could be affected by

glucose contained in the apple juice. This proof-of-concept study showed that the new library allowed rapid features identification and facilitated biologically relevant hypothesis from a simple urine study.

Table 4.4 PLS-DA results. Metabolites Up-Regulated After Drinking Juice (ordered by PLS loading score ranking)

	Purative Identification	Measured m/z
1	Uric Acid	169.0347
2	Hippuric Acid	180.0642
3	L-Acetylcarnitine	204.1229
4	2-Octenoylcarnitine (partial match to acetylcarnitine, manually interpreted)	286.2010
5	No ID	193.0370
6	Alpha-N-phenylacetyl-L-glutamine	265.1180
7	No ID	84.9600
8	Isobutyryl-L-Carnitine (Partial match to carnitine, manually interpreted)	232.1540
9	No ID	337.0650
10	Carnitine Derivative (partial match to carnitine, no accurate ID can be proposed)	310.2010

4.4 Conclusion

The high resolution HMDB MS^2 library was demonstrated to be capable of confidently identifying important metabolites in human urine samples. The identification process is easily integrated into the metabolomic workflow, and will facilitate future metabolomic studies using the Bruker platform.

From this work, several guidelines were developed for future metabolomic studies. It was found that negative mode MS^2 was not as efficient as positive mode in generating useable spectra for identification. Therefore, negative mode should not be the first method to use in exploratory

work unless negative mode friendly chemical classes are targeted. Sodiated adducts also reduced MS^2 quality, but their abundance is often difficult to control due to the prevalence of sodium in the lab. Lastly, more unknowns can have their chemical structures narrowed down by performing a fragmentation pattern match without specifying the precursor mass.

MS^2 library is still only one part of the identification workflow. MS^2 alone is not sufficient at distinguishing the differences between some structural isomers and the majority of stereoisomers. Their masses are exactly the same and fragmentation patterns are similar. LC separation is already performed prior to MS^2 analysis, mostly for reducing ion suppression due to sample complexity. However, retention time information is also another crucial physical property that can be used for unambiguous identification. The current MS^2 library can be supplemented with retention times for each standard in the future.

Standards are still crucial for the building of the library. As new metabolite standards become available, the library can be expanded with new data. However, this process is expensive and time consuming. It will be infeasible to synthesize enough standards to cover the entire metabolome. The future for metabolomics and small molecule identification lies in the precise understanding of the CID process, so that chemical structures can be derived *de novo* from fragmentation spectra. Progress is being made in these methods, however more accurate algorithms are required before metabolite MS^2 library can be retired for good.

4.5 Literature Cited

(1) Núñez, O.; Gallart-Ayala, H.; Martins, C. P. B.; Lucci, P.; Busquets, R. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2013**, *927*, 3-21.

(2) Wang, G.; Zhou, Y.; Huang, F. J.; Tang, H. D.; Xu, X. H.; Liu, J. J.; Wang, Y.; Deng, Y. L.; Ren, R. J.; Xu, W.; Ma, J. F.; Zhang, Y. N.; Zhao, A. H.; Chen, S. D.; Jia, W. *J. Proteome Res.* **2014**, *13*, 2649-2658.

(3) Trivedi, D. K.; Iles, R. K. *Biomed. Chromatogr.* **2014**, *28*, 1491-1501.

(4) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annu. Rep. Comput. Chem.*, Ralph, A. W.; David, C. S., Eds.; Elsevier, 2008, pp 217-241.

(5) Cho, K.; Mahieu, N. G.; Johnson, S. L.; Patti, G. J. *Curr. Opin. Biotechnol.* **2014**, *28*, 143-148.

(6) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801-807.

(7) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2012**, *9*, 44-66.

(8) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.

(9) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2013**, 1-8.

(10) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics (Oxford, England)* **2012**, *28*, 2333-2341.

(11) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148.

(12) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*,

703-714.

(13) Stein, S. *Anal. Chem.* **2012**, *84*, 7274-7282.

(14) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826-828.

(15) Zubarev, R. a.; Makarov, A. *Anal. Chem.* **2013**, *85*, 5288-5296.

(16) Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P. *J. Chromatogr. A* **2013**, *1292*, 51-65.

(17) Bao, J.; Gao, X.; Jones, A. D. *Rapid Commun. Mass Spectrom.* **2014**, *28*, 457-464.

(18) Bandu, M. L.; Watkins, K. R.; Bretthauer, M. L.; Moore, C. A.; Desaire, H. *Anal. Chem.* **2004**, *76*, 1746-1753.

(19) Doohan, R. A.; Hayes, C. A.; Harhen, B.; Karlsson, N. G. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1052-1062.

(20) Harvey, D. J. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 622-630.

(21) Alves, G.; Ogurtsov, A. Y.; Kwok, S.; Wu, W. W.; Wang, G.; Shen, R.-f.; Yu, Y.-K. *Biol. Direct* **2008**, *3*, 27.

(22) Chapman, J. D.; Goodlett, D. R.; Masselon, C. D. *Mass Spectrom. Rev.* **2013**, 1-19.

(23) Pizzagalli, F.; Varga, Z.; Huber, R. D.; Folkers, G.; Meier, P. J.; St-Pierre, M. V. *J. Clin. Endocrinol. Metab.* **2003**, *88*, 3902-3912.

(24) Allwood, J. W.; AlRabiah, H.; Correa, E.; Vaughan, A.; Xu, Y.; Upton, M.; Goodacre, R. *Metabolomics* **2014**, 438-453.

(25) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *86*, 6931-6939.

(26) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(27) Zhu, Z.-J.; Schultz, A. W.; Wang, J.; Johnson, C. H.; Yannone, S. M.; Patti, G. J.; Siuzdak, G. *Nat. Protoc.* **2013**, *8*, 451-460.

(28) Cirillo, P.; Sato, W.; Reungjui, S.; Heinig, M.; Gersch, M.; Sautin, Y.; Nakagawa, T.; Johnson, R. J. *J. Am. Soc. Nephrol.* **2006**, *17*, S165-S168.

(29) Kahle, K.; Kempf, M.; Schreier, P.; Scheppach, W.; Schrenk, D.; Kautenburger, T.; Hecker, D.; Huemmer, W.; Ackermann, M.; Richling, E. *Eur. J. Nutr.* **2011**, *50*, 507-522.

(30) Giancaterini, a.; De Gaetano, a.; Mingrone, G.; Gniuli, D.; Liverani, E.; Capristo, E.; Greco, a. V. *Metabolism.* **2000**, *49*, 704-708.

Chapter 5

Construction of a Metabolite Retention Time Library

5.1 Introduction

CID MS^2 libraries have been instrumental in facilitating the high-throughput identification of unknown metabolites in LC-MS metabolomics. As mass spectrometers become more advanced—with higher resolutions and higher mass accuracy—these identifications have also increased in confidence. From our experience, CID MS^2 spectra contained in these libraries are reproducible between instruments of the same make and model, making them portable to labs that use the same make of MS. The reproducibility is adequate between different instrument vendors, especially after fragmentation calibration procedures.[1,2] Major fragment peaks will be common between instruments, but differ in their intensities. While different instruments might not produce exact matches, partial fragment matches can still be useful for compound identification.

Most of metabolomic studies using a mass spectrometry platform utilize LC separation prior to ionization—mostly for reducing sample complexity.[3,4] The retention times reported for each unknown metabolite ion are used for aligning different samples of the same study, and not for compound identification. Despite the identification powers of CID MS^2, there are situations in which it will be challenged; an orthogonal method of narrowing down the identity is needed, such as retention time.

These challenging situations include when MS^2 is performed on a low intensity precursor ion.[5] The generated fragment spectrum will be of poor quality, with many peaks missing. There may be enough for a partial match to a library entry but that is not very confident on its own. If

there were a partial MS^2 match, accurate precursor match, and retention time match, then the confidence of identification would increase.

Another situation is if the MS^2 spectrum were not unique to a standard. This occurs in a few chemical classes: First, this occurs when the unknown is a stereoisomer, which generates identical fragmentation patterns.[6] Secondly, structural isomers with only a small difference to other isomers will not produce unique spectra. Lastly, some compounds fragment in a way to only give dominant ions that are not unique to that compound,[5] such as acylcarnitines.[7] In all of these situations, MS^2 data are not conclusive and require retention time information for identifying metabolites.

The challenge for metabolite identification with retention time is that metabolites have diverse chemical properties,[8] and their behaviors on a LC column cannot be accurately predicted *de novo*. In limited situations, retention time can be predicted if the analyte of interest is part of a well characterized chemical class, such as peptides. Peptide sequence was used to predict retention time on reversed phase columns, and the correlation of prediction to experimental data had R^2 of greater than 0.9.[9] Currently, there has been no useable model for predicting metabolite retention; therefore, standards must be obtained and retention time must be empirically measured. Retention times also depend on many other factors, which will cause variability in measured data. These factors include: Stationary phase, mobile phase composition, tubing length, leaks, injector needles, gradient programming, etc. These variabilities cause retention time to shift tens of seconds between labs and even between runs; such shifts prevent accurate identification. For these reasons, LC retention time libraries have seen little use in the metabolomics community compared to GC libraries.[10-12]

With these challenges in mind, the purpose of this work is to generate a usable retention time library for HMDB metabolites. To reduce the variabilities described above, we provide a very specific, optimized LC-MS condition—one that is applicable to most metabolomic work flows. The end user is expected to follow these exact conditions, and only then can the library be used. However, there are still minor variables causing retention time shifts that cannot be accounted for. A LC calibration method was developed to counter these variables. With the final library, we demonstrate that metabolite identification from human urine was more robust using retention time information. Also, in-source fragmentation was revealed to be prevalent in metabolomic analysis and convoluted identification greatly. This final retention time library composed of 792 measured metabolites measured on a reversed phase column.

5.2 Materials and Procedure

5.2.1 Chemicals and Reagents

All chemicals and reagents were purchased from Sigma-Aldrich Canada (Markham, ON, Canada) except those otherwise noted. LC–MS grade water and acetonitrile were purchased from Thermo Fisher Scientific (Edmonton, AB, Canada).

5.2.2 Instrumentation

All data were generated with an Ultimate 3000 RSLC system from Thermo Scientific (California, USA) and an impact HD high resolution QTOF mass spectrometer from Bruker Daltonics (Bremen, Germany). Mass spectrometer source settings were: Nebulizer gas 2.0 bar, dry gas 8.0 L/min, 200 °C, capillary voltage 4.5 kV, capillary offset 0.5 kV, acquisition rate 8 Hz, mass range 20 – 1000 m/z, CID gas is nitrogen, mass calibration with sodium formate.

5.2.3 Chromatography

An Acclaim 120 C_{18} 100 x 2.1 mm 2.2 µm column (Thermo Scientific) with a Vanguard BEH C_{18} 1.7 µm guard column (Waters) was used for all experiments. Column was kept at 30 °C with a column oven. The final gradient chosen was: 0 min (1% B), 0-2.0 min (1% B), 2.0-17 min (1-99% B), 17 min (99% B), 17-20 min (99% B). The flow rate was 250 µL/min and the injection volume was 5 µL. A wash and equilibration injection was run between samples; the gradient was: 0-4.5 min (100% B), 4.5-10 min (1% B) at 350 µL/min. Mobile phase A was 0.1 % LC-MS formic acid in LC-MS water and mobile phase B was 0.1 % LC-MS formic acid in LC-MS acetonitrile.

5.2.4 Quality Control (QC) Mixture

The following chemicals were used: Acetaminophen, Aminoacridine, Aminoanthracene, Caffeine, Dansyl Alanine, Dansyl Arginine, Dansyl Asparagine, Dansyl Aspartic acid, Dansyl Glutamic acid, Dansyl Glycine, Dansyl Histidine, Dansyl Isoleucine, Dansyl Leucine, Dansyl Lysine 2Tag (two dansyl groups were attached), Dansyl Methionine, Dansyl Phenylalanine, Dansyl Proline, Dansyl Serine, Dansyl Threonine, Dansyl Tryptophan, Dansyl Tyrosine 2tag, Dansyl Valine, N-decyl-N-N-Dimethyl-3-amino-1-propane-sulfonate, Nicotinic acid, P-coumaric acid, Reserpine, Sinapinic Acid, Sulfanilamide. These compounds were dissolved in 1:1 (v/v) H_2O:ACN. 100 µL from each stock solution were mixed together to make the QC stock mixture. 10 µL of stock mixture was diluted by 800 µL of 1:1 (v/v) H_2O:ACN for the final QC mixture.

5.2.5 Column Efficiency Calculation

Equation 5.1 was used to calculate the column efficiency of the Thermo Acclaim column. N is the number of theoretical plates for a peak. t_r is the retention time of the peak in minutes or seconds. $W_{0.5}$ is the full width at half maximum of the peak in minutes or seconds.

$$N = 5.54 \left(\frac{t_r}{W_{0.5}}\right)^2 \quad (5.1)$$

5.2.6 HMDB Standards

Pure standards were obtained from Professor David Wishart at the University of Alberta, Canada. Roughly 1 mg of solid or liquid was used to prepare a stock solution of every standard in 1:1 (v/v) H_2O:ACN. 10 µL from between 10-24 stock solution were mixed together, and 10 µL of this mixture was then diluted by 800 µL of 1:1 (v/v) H_2O:ACN 0.1% formic acid. 10 µL of QC stock solution was added to the final solution.

5.2.7 Library Data Acquisition

HMDB standards mixtures were injected in triplicate, with QC injection every 10 samples. QC standards were monitored for retention time deviation of less than 0.1 minute. The retention time of the HMDB standards were extracted with TargetAnalysis 1.3 (Bruker Daltonics), and tabulated in an Excel (Microsoft) spreadsheet. Using a Visual Basic script (Microsoft), the retention times were imported into the existing Bruker HMDB MS^2 library.

5.2.8 Urine Samples

One human urine sample was collected from a healthy individual. This urine sample was then filtered by 0.22 µm PVDF syringe filter (Milipore) and stored at -80 °C. Before injection, the samples was thawed then diluted by half 1:1 (v/v) H_2O:ACN 0.1% formic acid. The same LC-MS method as described above was used for the analysis of urine. Data dependent MS^2 acquisition was enabled in the mass spectrometer. The settings for this mode were: Active exclusion 3 scans, intensity threshold 1500, isolation window width 4 m/z, collision energy 20-50 eV.

DataAnalysis 4.2 (Bruker Daltonics) was used to automatically calibrate the samples, extract metabolite features, group MS^2 spectra to their respective metabolite feature, and perform library search.

5.2.9 Retention Time Calibration

The QC mixture and urine were injected onto the column with guard column, then the injections were repeated after removing the guard column. The retention times of all QC standards in the mixture were extracted for both with and without guard column data using TargetAnalysis 1.3 (Bruker). Metabolites with retention times were extracted from the urine data by the DataAnalysis 4.2 (Bruker Daltonics) molecular feature algorithm.

The retention time correction factor (CF) was calculated for each QC standard using Equation 5.2, where $RT_{no\ guard}$ is the retention time of the standard using the column without guard column, and RT_{guard} is the retention time of the standard using the column with guard column:

$$CF = RT_{no\ guard} - RT_{guard} \tag{5.2}$$

The algorithm then looks through all of the metabolite from the urine analyzed on the column without guard column, and correct the retention time to be the same as the retention time in the data measured on the column with guard column. Each metabolite is bracketed by two QC standard in terms of retention time. Using both CF, Equation 5.3 is then applied to correct the retention time of the metabolite:

$$Corrected\ RT = RT_{exp} + \frac{CF_{i+1} - CF_i}{RT_{i+1} - RT_i} \times \left(RT_{exp} - RT_i\right) \tag{5.3}$$

RT_{exp} is the uncorrected retention time of the metabolite. RT_i and CF_i are the retention time and correction factor for the standard that is before the metabolite. RT_{i+1} and CF_{i+1} are the retention time and correction factor for the standard that is after the metabolite.

For metabolites that elute before the first standard, or elute after the last standard, a single CF was applied (Equation 5.4).

$$Corrected\ RT = RT_{exp} + CF_i \tag{5.4}$$

Where RT_{exp} is the metabolite retention time, and CF_i is the correction factor of the first or last standard, depending on which end of the chromatogram the metabolite is on.

5.3 Results and Discussion

5.3.1 LC Conditions

Conventionally, metabolomics is conducted with reversed-phase columns. These columns retain metabolites of weak to medium polarity in a wide variety of biological samples, but they show poor retention of high polarity metabolites. Reversed phase columns are widely available from a variety of vendors, and are popular within the pharmaceutical industry and academia. Reversed phase LC has been used for the analysis of hydrophobic metabolites for studies involving many different biological systems.[13,14]

In this work, a C_{18} column with 2.2 µm porous particles was used for data acquisition, and we have accepted the fact that a number of polar metabolites will not be retained and are possibly lost due to ion suppression at the beginning of the analysis. The long-term plan was to continue to

build on this reversed phase data with other phases that will patch up the missing polar metabolite data.

The elution system composed of water with 0.1% (v/v) formic acid in channel A, and acetonitrile with 0.1% (v/v) formic acid in channel B. This system was chosen to be compatible with electrospray ionization in mass spectrometer sources and also because the mobile phases are widely available in metabolomics laboratories. By choosing to prepare pure solvents with formic acid, instead of premixed solvents (i.e. 95% water and 5% acetonitrile in channel A), complexity and variation due to mobile phase preparation will be reduced.

The large number of metabolites contained in biological samples requires the implementation of gradient elution to detect and quantify as large a number of metabolites as possible. A linear gradient of 99% A to 1% A over 15 minutes was chosen to allow a broad coverage of different polarities of metabolites. The starting composition of 99% A retained moderately polar metabolites, and the final composition of 1% A ensured that strongly retained non-polar compounds will be eluted from the column to prevent carryover.

The gradient time of 15 minutes balanced the metabolome coverage and the sample cycle time. Due to the large number of samples required when conducting metabolomic studies, the entire analysis time of each sample must be as low as possible while offering enough separation power for the given sample. The analysis included an additional 2-minute hold at 99% A prior to the gradient, and a 3-minute hold at 1% A following the gradient; combined with a wash and equilibration of 10 minutes, the entire analysis was 30 minutes long. This meant that 44 samples can be ran during a 24-hour period, with 4 QC injection every 10 samples. A typical sample size of 200 can be analyzed in one week. This is in line with the recommendation of the optimal 180 samples per week for reproducible LC-MS metabolomic data, proposed by Zelena et al.[15] This

gradient profile and cycle time is sufficient in handling metabolomic studies in a reasonable amount of time, and was selected as the library method.

Finally, an ESI based MS method optimized for the detection of metabolites between 50 – 1000 m/z was selected. This was the same MS method previously optimized for the building of the Bruker HMDB MS^2 library. Data acquired with this method was found to be compatible with the MS^2 entries contained in that library. The only change was the increase in spectral acquisition rate from 1 Hz to 8 Hz, in order to produce better peaks on the UHPLC system.

5.3.2 Column Performance

The goal for this project was to build a library of retention times for the accurate identification of metabolites in addition to MS^2 spectral matching. In order for the library retention times to be useful, the UHPLC-MS system chosen in this study must have sufficient retention time reproducibility. Based on previous experience, the chosen column from Thermo Scientific was known to have reproducible batch-to-batch retention times, and each column comes validated with a certificate of analysis. The UHPLC was recently installed prior to this work and passed installation and performance qualification. In order to assess the reproducibility of the entire UHPLC system with the method to be used in the library, a set of standards that elutes evenly spaced was selected to complete an operating qualification. Table 5.1 shows the complete list of standards and their retention time %RSD (n=14). The standards showed reproducible retention times with %RSD between 0.00% - 1.65%. Two early eluting compounds, sulfanilamide and nicotinic acid, showed the highest variation of 0.96% and 1.65%. The remaining standards that eluted after 3 minutes all showed good reproducibility between 0.00% to 0.48%, which was close to the theoretical predicted LC system variability of 0.29% proposed by Boswell et al.[16]

Table 5.1 Retention time of quality control standards and their standard deviations (n = 14).

Name	Average (min)	% RSD	Standard Deviation (min, sec)
Nicotinic acid	1.44	0.96	0.01(0.8 s)
Sulfanilamide	2.98	1.65	0.05(3.0 s)
Acetaminophen	6.46	0.32	0.02(1.2 s)
Caffeine	7.33	0.26	0.02(1.1 s)
Aminoacridine	7.67	0.44	0.03(2.0 s)
Dansyl Histidine	7.85	0.34	0.03(1.6 s)
Dansyl Arginine	8.33	0.36	0.03(1.8 s)
P-coumaric acid	8.54	0.15	0.01(0.8 s)
Dansyl Asparagine	8.66	0.14	0.01(0.7 s)
Sinapinic Acid	8.74	0.14	0.01(0.7 s)
Dansyl Serine	9.32	0.10	0.01(0.5 s)
Dansyl Glutamic acid	9.52	0.08	0.01(0.5 s)
Dansyl Aspartic acid	9.57	0.08	0.01(0.5 s)
Dansyl Threonine	9.90	0.10	0.01(0.6 s)
Dansyl Glycine	10.28	0.06	0.01(0.4 s)
Dansyl Alanine	10.80	0.05	0.01(0.3 s)
Reserpine	10.85	0.48	0.05(3.1 s)
N-decyl-N-N-Dimethyl-3-amino-1-propane-sulfonate	11.05	0.03	0.00(0.2 s)
Dansyl Proline	11.99	0.04	0.00(0.3 s)
Dansyl Methionine	12.23	0.03	0.00(0.2 s)
Dansyl Valine	12.25	0.06	0.01(0.4 s)
Dansyl Tryptophan	12.39	0.04	0.00(0.3 s)
Aminoanthracene	12.41	0.05	0.01(0.4 s)
Dansyl Phenylalanine	12.95	0.03	0.00(0.3 s)
Dansyl Isoleucine	13.05	0.24	0.03(1.9 s)
Dansyl Leucine	13.06	0.27	0.04(2.1 s)
Dansyl Lysine 2Tag	14.31	0.03	0.00(0.3 s)
Dansyl Tyrosine 2tag	15.98	0.02	0.00(0.2 s)

The average peak width (FWHM) of all the standards was found to be 0.085 min (n=3). These peak widths showed that this column was able to provide separation of complex mixture of metabolites, while allowing adequate time for data-dependent acquisition of 60-80 MS^2 spectra per peak.

With a particle size of 2.2 µm, this column can be considered as a borderline UHPLC-class column, where true UHPLC columns have particles below 2 µm. In exchange for slightly lower column efficiency than the sub-2 µm columns, this 2.2 µm was found to have substantially lower back pressure. During a run at 250 µL/min, the maximum pressure was 215 bar at the beginning of the gradient. This low pressure allows the column to be used on a variety of HPLC and UHPLC instruments, and there is a large margin until the column overpressures a typical HPLC at 400 bar. Meanwhile, 1.7 µm Agilent UHPLC columns of the same dimension typically gave maximum pressures of 280 to 300 bar at a slower flow rate of 180 µL/min. The column selected was robust for metabolomics studies and can be portable to many different LC instruments.

5.3.3 Retention time Results

To complete the acquisition of all 792 metabolites in a reasonable amount of time, between 10 to 24 compounds were mixed together according to their chemical classification in order to prevent any unwanted reactions. Each retention time sample was also spiked with the previously mentioned QC mixture. Isobaric compounds were separated into different vials to prevent convoluting the measured retention times. Overall, there were 53 individual HMDB standard mixtures and they were all measured in triplicates, in both positive and negative modes.

Table 5.2 Retention time distribution for the 800 compounds measured.

Retention time (s)	Frequency

100	296
200	2
300	2
400	55
500	79
600	50
700	48
800	42
900	18
1000	6
1100	4
1200	8
1300	0

Table 5.2 shows the summary of the retention times for all 610 standards that were detected in the LC-MS method. With a total gradient delay volume of approximately 390 µL (190 µL autosampler needle + 200 µL pump tubing), the actual gradient started at 214 seconds. 49% of the metabolites eluted before 100 seconds, which showed that the majority of the compounds in the library were highly polar and did not retain on the column. For these unretained compounds, the reversed phase column cannot separate them, and they will most likely be ion-suppressed in real samples because they elute with the complex unretained matrix plug. The rest of the metabolites, 48% of total, eluted between 400 seconds and 900 seconds.

In order to analyze the reason why half of the metabolites were unretained, the chemical classes of retained and unretained compounds are summarized in Table 5.3. The majority of the metabolites that were unretained were polar amino acids or their analogues (31%), carbohydrates (18%), and polar aromatic compounds (18%). It was well known that these compound classes are too polar to be retained on C_{18} columns,[17] so this result was expected. For the metabolites that were retained, majority was composed of aromatic (49%), lipids (25%), and polypeptides (12%), which are hydrophobic and retained more on the C_{18} column.

Table 5.3 Chemical classes for unretained and retained compounds, as percentages of total unretained or total retained compounds. Unretained compounds had retention times <= 100 seconds; retained compounds had retention times > 100 seconds and <= 900 seconds.

	% Unretained	% Retained
Aliphatic*	13%	3%
Amino Acids, Peptides, and Analogues	31%	12%
Aromatic	18%	49%
Lipids	2%	25%
Carbohydrates and Carbohydrate Conjugates	18%	1%
Alkaloids and Derivatives	1%	1%
Nucleosides, Nucleotides, and Analogues	9%	2%
Organic acids	5%	5%
Organophosphorus Compounds	1%	0%
Unspecified	2%	2%
Polyketides	0%	0%
Homogeneous Non-metal Compounds	1%	0%

*Small compounds with 3 or more carbons, and may contain other heteroatoms.

The 1:1 ratio of retained to unretained metabolites highlighted the need for a polar separation methods, such as HILIC, along with reversed phase C_{18} methods to study the metabolome. Studies have already shown that analyzing the same samples with both methods improves the metabolome coverage.[18,19] For analysis that use reversed phase columns, this current library data provides an invaluable source of data to aid in the identification of metabolites.

5.3.4 Analysis of Human Urine

The retention time information was imported into the Bruker Library program, alongside the existing HMDB MS^2 library. Searching with the new retention time information was straightforward and handled in the DataAnalysis program. The simple searching algorithm only requires an input of retention time tolerance windows in minutes. First, metabolites are identified

and filtered based on their MS^2 spectrum match to the library. Then, experimental retention time of the valid matches from this step are compared with the library retention time; if the experimental retention time is within the tolerance window of the library data, then that match will show up in the final results. If the experimental retention time is outside of the window, they are removed from the final list no matter how perfectly the MS^2 spectrum matched to the library.

Human urine was used to test the number of matches possible in a typical sample. The urine was analyzed on the same column, with the same LC and MS methods used during the library construction. 1042 unknown metabolite features, each with a MS^2 spectrum, were identified with the compound finding algorithm built into the data processing software, DataAnalysis. Searching those metabolite features with the library, matching both accurate precursor mass and MS^2 spectrum but without specifying the retention time, 89 features were matched with fit scores above 500. Of that 89, 74 were unique compounds. The duplicates were mostly creatinine that appeared at different times on the chromatogram.

Using the same data file, the search was repeated to include retention time information. Based on the QC sample retention time reproducibility, the retention time tolerance window was set to 0.1 min, and a broader tolerance window of 0.3 min. This meant that experimental retention time within ±0.1 min was considered to be a perfect match; between ±0.1 min and ±0.3 min there was a penalty to the Fit matching score; and beyond ±0.3 min, the results were eliminated from the final list.

After specifying retention time matching, the number of matches reduced to 23 metabolites. 20 of these 23 metabolites (87%) were very confident matches with fit scores above 900, whereas only 56 out of 74 (76%) were confident matches when no retention time information was used. By

using retention time information, the proportion of confident results were improved—albeit with a reduction in metabolite detection sensitivity.

Table 5.4 High scoring MS^2 library matches for metabolites human urine data, without using retention time data, compared with the result from searching with retention time information (RT Result).

Compound Name	Fit'	Found in RT Result	Library RT (s)	Experimental RT (s)	In-Source Fragmentation
5-Methylcytidine - HMDB00982	1000	Not found	78.18	132	Did not observe in-source fragmentation
1-Methyladenosine - HMDB03331	1000	Not found	78.44	141	Did not observe in-source fragmentation
Adenosine - HMDB00050	1000	Not found	78.18	309	Did not observe in-source fragmentation
Cyclic AMP - HMDB00058	1000	Not found	78.44	331.2	Did not observe in-source fragmentation
Salicyluric acid - HMDB00840	1000	Found - Wrong RT	518.32	388.8	Did not observe in-source fragmentation
Salicyluric acid - HMDB00840	1000	Found - Wrong RT	518.32	409.8	Did not observe in-source fragmentation
5-Hydroxyindoleacetic acid - HMDB00763	1000	Not found	447.78	492	Did not observe in source fragmentation
Salicyluric acid - HMDB00840	1000	Found	518.32	518.4	Correct Assignment
Indoleacetic acid - HMDB00197	1000	Found	586.74	586.8	Correct Assignment
Trimethylamine N-oxide - HMDB00925	999	Found	59.74	64.2	Correct Assignment
L-Acetylcarnitine - HMDB00201	999	Found	81	99	Correct Assignment
2-Furoylglycine - HMDB00439	999	Found	367.8	375	Correct Assignment
Isovalerylcarnitine - HMDB00688	999	Found	430.32	429	Correct Assignment
Indoleacetic acid - HMDB00197	999	Found - Wrong RT	586.74	517.2	352.1029, 176.0704*
Creatinine - HMDB00562	999	Found - Wrong RT	61.18	1180.8	Carryover

Ribothymidine - HMDB00884	998	Not found	336.96	81	Did not observe in source fragmentation
Glycylproline - HMDB00721	998	Not found	79.08	98.4	Did not observe in source fragmentation
Symmetric dimethylarginine - HMDB03334	997	Found	65.08	70.2	Correct Assignment
Betaine - HMDB00043	996	Found	61.88	62.4	Correct Assignment
1,3-Dimethyluric acid - HMDB01857	995	Found	371.4	371.4	Correct Assignment
D-Glutamine - HMDB03423	995	Not found	58.56	454.2	265.1179, 147.0768*
Creatinine - HMDB00562	994	Found - Wrong RT	61.18	83.4	Did not observe in source fragmentation
4-Aminophenol - HMDB01169	994	Not found	67.38	199.2	190.0166, 110.0596
Creatinine - HMDB00562	993	Found	61.18	67.2	Correct Assignment
Creatinine - HMDB00562	993	Found - Wrong RT	61.18	309.6	Did not observe in source fragmentation

In theory, high scoring matches to both the precursor and the fragmentation pattern should have been very high confident identifications; therefore, they should have also matched the retention times in the library. The numbers of confident matches utilizing retention time should be the same as that without using retention time. In order to find out why this was not the case, 25 highest scoring matches from the searches without and with retention time information were closely examined. Table 5.4 shows the comparison of these high scoring matches. All of the 25 highest scoring matches were very confident and their MS^2 spectra perfectly matched the library's with scores above 900. 15 out of the 25 had experimental retention times that differed from the library retention time by more than 18 seconds, and so they were not found in the RT result. Of the 15 not-found, 9 had library retention times of less than 2 minutes, meaning that they were unretained polar compounds. However, in the experimental urine run, these metabolites were retained more. They eluted later but still near the gradient dead volume, around 2-5 minutes. One possible hypothesis is that these polar compounds were more susceptible to the matrix effect of the urine, causing greater retention time shift early in the gradient than later. Several previous research have shown biological matrix can affect retention time of columns. Another 3 out of the 15 were retained compounds, however their deviations were unexplained. The remaining 3 not-found metabolites had an interesting reason for their shifted experimental retention times: In-source fragmentation.

5.3.5 In-Source Fragmentation

After investigating the data, it was found that these 3 misidentified metabolites were artefacts from in-source fragmentation of larger metabolite conjugates. In a living organism, there are many metabolic processes that add chemical groups to small molecules to make metabolite conjugates, in order to change their properties. For example, addition of sulfate to cholesterol

increases polarity, and allows cholesterol to be excreted in the urine or feces.[20] What was observed was that some of these conjugates are prone to cleavage during collision with gas molecules in the source region. Figure 5.1 shows how the in-source fragment is generated in the mass spectrometer. This type of conjugate in-source fragmentation generated a fragment with the exact same mass and MS^2 spectrum as the unconjugated metabolite. This caused the conjugated metabolite to be misidentified, and because conjugation usually changes the chemical properties of the metabolite, the retention time was shifted from the unconjugated species. For this reason, high scoring exact mass and MS^2 matches were made at retention times far away from the library retention times.

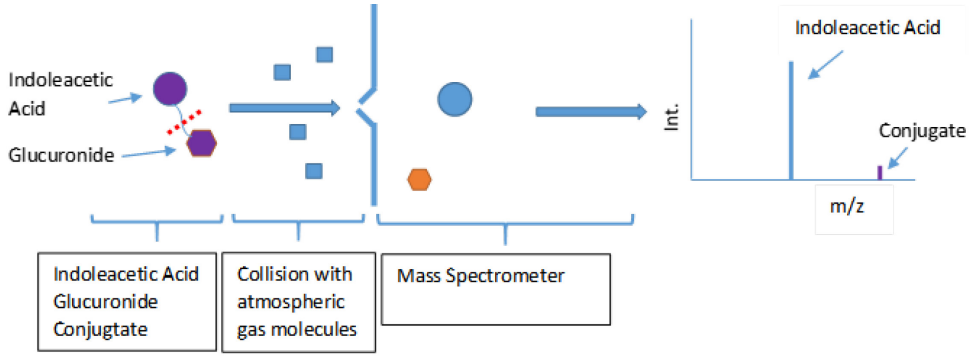


Figure 5.1 Schematic of metabolite conjugates generating in-source fragmentation. The ionized conjugate is first accelerated through atmospheric gas molecules resulting in bond cleavage, indicated by the red dotted line. After breaking apart, the ionized indoleacetic acid fragment enters the mass spectrometer while the glucuronide is lost as a neutral. The resulting mass spectrum predominately shows indoleacetic acid, and not the conjugate.

Due to the fragment and the metabolite being identical, it was impossible to determine when in-source fragmentation was occurring by examining the precursor spectrum. To find out if these 4 misidentified metabolites were in-source fragments of larger conjugates, MS^2 spectra of higher masses in the same precursor spectrum were inspected. In these MS^2 spectra of larger precursors, if an exact mass corresponding to the suspicious metabolite was found, then that gives sufficient

evidence to putatively identify the larger mass as the conjugate that gave rise to the in-source fragmentation metabolite.

The first conjugate was identified as indoleacetic acid with a fit score of 999, which indicated an exact match of the MS^2 spectrum. Applying the procedure for conjugate identification, the exact mass of 176.0704 (indoleacetic acid) was found in the MS^2 spectrum of 352.1029 m/z (Appendix Figure A5.1). The mass difference between the two precursor ions was 176.0325 m/z, which was determined to be glucuronidation. The theoretical mass of glucoronic indoleacetic acid was 1.2 ppm of the measured mass. While glucuronidation of indoleacetic acid was not reported in literature to the authors' knowledge, a drug with a similar indoleacetic acid structure was found to be glucuronidated for excretion by urine;[21] therefore this observed conjugate was also likely to form in the body. It was found that the glutamine ion was generated by the in-source fragmentation of 265.1179 m/z, a mass increase of 118.0411 m/z. In literature, glutamine was found to conjugate with phenylacetic acid,[22] and the theoretical mass of phenyacetyl glutamine was 1.5 ppm off the measured mass. The conjugation of phenylacetic acid would also explain the increase of retention time from less than one minute for glutamine standard to the experimental 7.5 minutes. Lastly, 4-aminophenol was found to be an in-source fragment of 190.0166, a mass increase of 79.957, which was most likely due to the addition of a sulfate (mass error of 1.4 ppm). Out of 25 results, 12% were misidentified due to in-source fragmentation.

In-source fragmentation has recently been recognized as prevalent source of error for untargeted metabolomics, where a large portion of identified metabolite features in a study might be artefacts generated by in-source fragments.[23] In-source fragments are commonly thought of as extra noise that does not contribute real metabolite identification when using MS^2; in this work, it was shown that in-source fragmentation of conjugated metabolites can produce the exact same

accurate mass and MS^2 spectrum as the native metabolites. This phenomenon misidentified metabolites, which are difficult to distinguish without retention time information. Due to the prevalence of conjugates in biology, this problem is expected to occur often in LC-MS metabolomics and highlights the need for retention time libraries.

5.3.6 Retention Time Calibration

A requirement of retention time libraries is that it must be portable to other labs, which use different LC configurations. In order to ensure the reproducibility of the retention times, this library is to be used only with narrowly defined LC parameter, such as columns that are the same part number as the column used in this work, and the gradient conditions must be the same as the ones specified in the method section. Even with these precautions, there will be small differences in the instrument configuration that result in different dead volume and gradient delay. Such differences might still occur even if the same model of LC was used but the tubings have been changed to different dimensions. However, these changes are usually small, and retention time calibration can solve these problems.

For this work, a pseudo retention time indexing strategy that involved calibrating the experimental data with the retention times of the QC standards was used—similar to GC retention time indexing.[10] However, the aim is to correct for small retention time shifts using the exact same gradient conditions, rather than predicting retention time with different gradient profiles.[16] The standards elute roughly every minute from the column, and forms a ladder of retention times. Retention time of this experimental QC data would then be compared to the QC data used to generate the library. Subsequent experimental retention times are then changed according to the calculated retention time differences. For the ease of calculation, retention time shifts between two adjacent standards are assumed to be linear. The calibrated experimental data would then be

compatible with the library. This method has been previously implemented in a similar manner for proteomic retention time correction.[24]

In order to test if this method would be able to account for small volume changes in the flow path, the QC standard mixture and human urine were ran on the C_{18} column with and without a guard column. Retention times of several metabolites in the urine throughout the gradient were compared between the column configurations. By removing the guard column, the retention times were shifted by an average of -6.3 seconds, which reflected the reduction of dead volume. Figure 5.3 shows that the shift was not linear throughout the entire analysis. The later eluting metabolites had larger retention time shifts than earlier peaks. Therefore, it would not have been possible to simply figure out the dead time difference and apply a linear correction to all of the data. After applying the retention time calibration method, the average retention time shift was greatly reduced to +0.25 second. This showed that after calibration, retention times of the experimental data should match closely with the data in the library and increase the identification confidence.

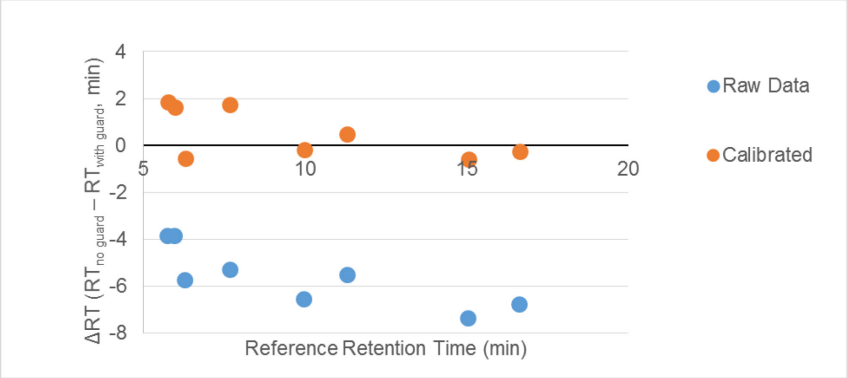


Figure 5.2 Retention time shifts of selected metabolites throughout the gradient during urine analysis. Blue data points indicate the shift in the uncorrected data, and orange points indicate the shift after retention time calibration.

The QC standard mixture and the retention time calibration algorithm would be supplied with the library. At the time of writing, the retention time calibration algorithm was not

implemented in the Bruker data processing software. Therefore, it was not possible to calibrate experimental data and test whether the calibration improved identification with the retention time library.

5.4 Conclusion

Following the completion of the Bruker HMDB MS^2 library, work was under way to fill in retention time information to boost the library's identification power. In the first part of this work a robust reverse phase gradient method was developed and validated. This method was applied to the collection of human metabolite retention time information, and the compiled library tested with real human urine data. Retention time information increased the confidence of the metabolite identifications by removing duplicates or incorrect identifications. As one of the largest retention time library for metabolites, this work will be a valuable resource to the metabolomic community around the world.

It is clear from the findings in this study that more work is needed to address the issue of in-source fragmentation, which contributes toward misidentifying metabolites if retention time information were unavailable. For unknown metabolites not in the library, in-source fragmentation must be checked manually. Even with manual checking, it is sometimes impossible to determine if a peak is an in-source fragment or an intact metabolite, which obscures the identification process.

Currently, data processing programs have the ability to reduce the chances of picking an in-source fragment as a metabolite. This is done by extracting all ions in a spectrum and plotting their EIC, and ions with chromatographic peaks that overlap are then considered to be all from the same compound. The problem with this approach is that in a complex sample with inadequate data acquisition rate, many chromatographic peaks from different compounds overlap perfectly. If these

peaks were to be grouped as one metabolite, then the detected number of metabolites would decrease. One way to counter this issue is to implement the manual inspection procedure described in this work as an algorithm to automatically identify in-source fragments. The MS^2 spectrum of large ions are searched for the lower masses that are observed in the precursor spectrum. If a fragment matched an ion in the precursor spectrum, then that ion can be identified as an in-source fragment and grouped together with the larger parent ion. This is a more robust way of identifying the in-source fragments, and would lead to less sensitivity loss.

Completing the reversed phase data was only the start of the project. Of the 792 compounds measured, nearly 50% were too polar to be retained on the reversed phase column. The next step would be to develop a generally applicable HILIC method that can separate these polar metabolites. With both reversed phase and HILIC data, this library would be able to coverage a large portion of the endogenous human metabolome.

5.5 Literature Cited

(1) Bristow, A. W. T.; Nichols, W. F.; Webb, K. S.; Conway, B. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2374-2386.

(2) Hopley, C.; Bristow, T.; Lubben, A.; Simpson, A.; Bull, E.; Klagkou, K.; Herniman, J.; Langley, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1779-1786.

(3) Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 443-458.

(4) Patti, G. J. *J. Sep. Sci.* **2011**, *34*, 3460-3469.

(5) Stein, S. *Anal. Chem.* **2012**, *84*, 7274-7282.

(6) Awad, H.; El-Aneed, A. *Mass Spectrom. Rev.* **2013**, *32*, 466–483.

(7) Zuniga, A.; Li, L. *Anal. Chim. Acta* **2011**, *689*, 77-84.

(8) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; Dame, Z. T.; Poelzer, J.; Huynh, J.; Yallou, F. S.; Psychogios, N.; Dong, E.; Bogumil, R.; Roehring, C.; Wishart, D. S. *PLoS One* **2013**, *8*, e73076.

(9) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785-7795.

(10) Babushok, V. I.; Linstrom, P. J.; Reed, J. J.; Zenkevich, I. G.; Brown, R. L.; Mallard, W. G.; Stein, S. E. *J. Chromatogr. A* **2007**, *1157*, 414-421.

(11) Erban, A.; Schauer, N.; Fernie, A.; Kopka, J. In *Metabolomics*, Weckwerth, W., Ed.; Humana Press, 2007, pp 19-38.

(12) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, *81*, 10038-10048.

(13) Wikoff, W. R.; Pendyala, G.; Siuzdak, G.; Fox, H. S. *J. Clin. Invest.* **2008**, *118*, 2661-2669.

(14) Yanes, O.; Clark, J.; Wong, D. M.; Patti, G. J.; Sánchez-Ruiz, A.; Benton, H. P.; Trauger, S. A.; Desponts, C.; Ding, S.; Siuzdak, G. *Nat. Chem. Biol.* **2010**, *6*, 411-417.

(15) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*,

1357-1364.

(16) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218*, 6732-6741.

(17) Spagou, K.; Tsoukali, H.; Raikos, N.; Gika, H.; Wilson, I. D.; Theodoridis, G. *J. Sep. Sci.* **2010**, *33*, 716-727.

(18) Contrepois, K.; Jiang, L.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14*, 1684-1695.

(19) Ivanisevic, J.; Zhu, Z.-J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876-6884.

(20) Marinkovic-Ilsen, A.; van den Ende, A.; Wolthers, B. G. *Arch. Dermatol. Res.* **1984**, *276*, 364-369.

(21) Dixon, C. M.; Saynor, D. A.; Andrew, P. D.; Oxford, J.; Bradbury, A.; Tarbit, M. H. *Drug Metab. Dispos.* **1993**, *21*, 761-769.

(22) Mokhtarani, M.; Diaz, G. A.; Rhead, W.; Lichter-Konecki, U.; Bartley, J.; Feigenbaum, A.; Longo, N.; Berquist, W.; Berry, S. A.; Gallagher, R.; Bartholomew, D.; Harding, C. O.; Korson, M. S.; McCandless, S. E.; Smith, W.; Vockley, J.; Bart, S.; Kronn, D.; Zori, R.; Cederbaum, S.; Dorrani, N.; Merritt, J. L.; Sreenath-Nagamani, S.; Summar, M.; LeMons, C.; Dickinson, K.; Coakley, D. F.; Moors, T. L.; Lee, B.; Scharschmidt, B. F. *Mol. Genet. Metab.* **2012**, *107*, 308-314.

(23) Xu, Y.-F.; Lu, W.; Rabinowitz, J. D. *Anal. Chem.* **2015**, *87*, 2273–2281.

(24) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. *Proteomics* **2012**, *12*, 1111-1121.

Chapter 6

Nanoflow LC-MS for Chemical Isotope Labeling Quantitative Metabolomics*

6.1 Introduction

The growth of metabolomics in the past decade has been directly linked to the development of modern analytical techniques that are able to quantitatively profile a wide range of metabolites in a sample. Liquid chromatography mass spectrometry (LC-MS) has become a powerful tool for metabolomic profiling.[1,2] To increase the sensitivity of the LC-MS platform, researchers are continually developing more sensitive mass spectrometers, new LC techniques and improving ionization efficiency of metabolites. The latter can be done using chemical labeling such as isotope encoded chemical derivatization or chemical isotope labeling (CIL).[3-10] In CIL, one isotopic form of a reagent is used to target a broad submetabolome (e.g., all amines and phenols when using dansyl chloride,[4] or all carboxylic acids using DmPA[5]). In parallel, a reference sample of very similar composition but distinct from the sample which is most commonly made by pooling all available samples is labeled with another isotopic form of the reagent.[11,12] The derivatized sample and reference are then mixed together and injected into LC-MS for analysis. Peak pairs detected from differentially labeled metabolites are used for metabolite quantification and identification. By using a proper labeling reagent,[3-5] CIL LC-MS allows concomitant improvement in LC separation

* A version of this chapter has been submitted for publishing as Li, Z.; Tatlay, J.; and Li, L. "Nanoflow LC-MS for Chemical Isotope Labeling Quantitative Metabolomics."

and MS detection. Accurate relative and absolute quantification of thousands of metabolites can be obtained from a single experiment.[12]

Further sensitivity increase in LC-MS is still highly desirable in handling samples of limited amounts, particularly those requiring multiple analyses. For example, in CIL LC-MS, each labeling reagent covers a selected submetabolome. Therefore, multiple labeling of the same sample using different aliquots needs to be carried out in order to increase the coverage of the overall metabolome. If multidimensional separation of a metabolome or submetabolome is used, the amount of metabolites in individual pre-fractionated aliquots for LC-MS analysis may be very limited,[13-15] requiring a sensitive detection technique. In this regard, there exists a high sensitivity platform that is already widely used in proteomics,[16] but less common in metabolomics: the nanoflow-LC MS. Only a few studies were reported using nLC-MS for metabolomic analysis.[17-20] This can be attributed to several reasons including technical challenges. In untargeted metabolomic profiling, four modes of LC-MS experiments using two different stationary columns (e.g., reversed phase (RP) and hydrophilic interaction (HILIC) columns) with each operated at positive and negative ion MS detection are often performed on a sample to detect both polar and nonpolar metabolites.[21-25] In nLC-MS, it is a relatively time-consuming process to switch different capillary columns and then optimize their performances thereafter. In addition, injecting a large volume of sample to increase sample loading to nLC is a major challenge.[18] nLC-MS systems used for shotgun proteome analysis is often equipped with a trap column to capture peptides in several microliters of volume prior to nLC separation. However, high efficiency trapping of all metabolites with wide variations in chemical and physical properties is very difficult in metabolome analysis.

CIL metabolomic profiling can overcome these technical challenges, because chemical labeling increases hydrophobicity and allows polar metabolites to retain on RP columns.[4,5] Both

RP trap column and analytical column can be used. There is no need to switch columns to handle different classes of metabolites. CIL also reduces the impact of larger retention time shifting in nLC than conventional LC, because quantification is not reliant on accurate chromatographic alignment between different samples and each metabolite is quantified with its own isotopic counterpart as a peak pair in a mass spectrum.[11,12]

In this work, we report a workflow based on nLC-MS for routine analysis of chemical isotope labeled metabolomic samples and describe its performance, particularly in comparison with microbore LC-MS (mLC-MS) commonly used in metabolomics. Dansylation labeling was used for analyzing metabolite standards and the amine/phenol submetabolome of human urine and sweat to demonstrate the improvement of detection sensitivity and metabolome coverage by using nLC-MS.

6.2 Experimental Section

6.2.1 Chemicals and Reagents

All chemicals and reagents were purchased from Sigma-Aldrich Canada (Markham, ON, Canada) except those otherwise noted. The synthesis of 13C2-dansyl chloride has been reported (reference 4). LC-MS grade water and acetonitrile were purchased from Thermo Fisher Scientific (Edmonton, AB, Canada).

6.2.2 Dansyl Labeling

The labeling protocol has been reported (reference 4). Briefly, 25 µL of sample was diluted with 25 µL of water and 25 µL of sodium bicarbonate buffer (250 mM) was added, and the solution vortexed. 75 µL of 13 mg/mL $^{12}C_2$ or $^{13}C_2$-dansyl chloride was added and vortexed. Sample was incubated for 40 minutes at 45°C and then quenched with 10 µL of 250 mM sodium hydroxide.

Sample was heated for 10 minutes at 45°C for complete quenching. Finally, the sample was acidified with 50 µL of 425 mM formic acid.

6.2.3 LC-UV Quantification

The method for LC-UV quantification of total labeled metabolites was reported previously (reference 23). Briefly, a dansyl labeled sample was injected into Waters LC-UV instrument with a reversed phased column, and eluted with a step gradient. The peak area of the UV absorbance at 338 nm was used to quantify the total labeled metabolites in the sample.

6.2.4 nLC-MS

All nLC-MS experiments were performed on a Waters nanoACQUITY UPLC (Milford, MA, USA) connected to a Waters Q-TOF Premier quadrupole time-of-flight (QTOF) mass spectrometer (Milford, MA, USA) equipped with a nano-ESI source. Mass spectrometer settings were: capillary voltage 3.5 kV, sampling cone 30 V, extraction cone 3.0 V, source temperature 110°C, and collision gas 0.45 mL/min. A 5 µm I.D. PicoTip by New Objective (Woburn, MA, USA) was used with the nano-ESI source. Chromatographic separations were performed on an Acclaim PepMap RSLC C18 (75 µm x 150 mm, 2 µm) and Acclaim PepMap 100 trap column (75 µm x 20 mm, 3 µm). A Waters nanoAcquity C18 (75 µm x 200 mm, 1.7 µm) column and nanoAcquity Atlantis trap column (180 µm x 20 mm, 5.0 µm) was also evaluated. Mobile phase A was 0.1 % LC-MS formic acid in LC-MS water and mobile phase B was 0.1 % LC-MS formic acid in LC-MS acetonitrile. The 45 minute gradient conditions were; 0 min (15% B), 0-2.0 min (15% B), 2.0-4.0 min (15-25% B), 4.0-24 min (25-60% B), 24-28 min (60-90% B), and 28-45 min (90% B). A wash and equilibration injection was run between samples; the gradient was: 0-10 min (90% B), 10-25 min (15% B). The flow rate was 350 nL/min and the injection volume was 5 µL (the

maximum volume of the sample loop used) in most cases except that of studying the trapping efficiency.

6.2.5 LC-MS

All LC-MS experiments were performed on an Agilent 1100 Series Binary LC System (Santa Clara, CA, USA) connected to the same Q-TOF Premier mass spectrometer used in the nLC-MS experiment, with the nESI source swapped out for an ESI source. Mass spectrometer settings were: capillary voltage 3.5 kV, sampling cone 30, extraction cone 3.0, source temperature 110°C, desolvation temperature 220°C, desolvation gas 800 L/hr, and collision gas 0.45 mL/min. Chromatographic separations were performed on a Waters Acquity BEH C18 column (2.1 mm x 100 mm, 1.7 μm) with the same mobile phases as the nano-LC. The 45 minute gradient conditions were; 0 mins (20% B), 0-3.5 min (20-35% B), 3.5-18 min (35-65% B), 18-24 min (65-99% B), 24-37 min (99% B), and 37.1-45 min (20% B). The flow rate was 180 μL/min.

6.2.6 nLC-MS Trapping Efficiency

A mixture of amino acids at a concentration of 1 mM each was dansylated[4] (see Supplemental Note N1). $^{12}C_2$ and $^{13}C_2$-dansyl labeled amino acids were mixed 1:1 by volume and diluted to 1000, 2000, 4000, 6000, 8000, and 10000 fold using serial dilution. Injection volume was varied for each diluted sample to ensure 120 fmol of dansylated amino acids are loaded onto the column for each injection. Data was de-noised, smoothed, centered and peak areas extracted using Waters QuanLynx software.

6.2.7 Dynamic Range of Peak Pair Detection

$^{12}C_2$-dansylated amino acids were diluted by half and mixed with undiluted $^{13}C_2$-dansylated amino acids in a 1:1 volume ratio. The theoretical peak ratio of $^{12}C_2$- to $^{13}C_2$-labeled amino acid

should be 1:2. The sample was then diluted using serial dilution and increasing sample amounts were injected into the nLC-MS and mLC-MS. Ratios were calculated by dividing the $^{12}C_2$-labeled amino acid peak area by the $^{13}C_2$-labeled amino acid peak area.

6.2.8 Urine and Sweat Analysis

A human urine sample was split into two vials; one was $^{12}C_2$-dansyl labeled and the other was $^{13}C_2$-dansyl labeled. The $^{12}C_2$-dansyl urine was quantified to be 48.2 mM using the LC-UV method[26] (see Supplemental Note N1). The $^{12}C_2$-dansyl urine and $^{13}C_2$-dansyl urine were mixed 1:1 by volume then diluted using serial dilution. These diluted samples were injected at increasing concentrations into the nLC-MS and LC-MS. Peak pairs were then extracted from the processed data using IsoMS.[11]

A human sweat sample was treated the same way as the urine sample. The concentration of the sweat was determined to be 8.4 mM using the LC-UV method. The $^{12}C_2$-dansyl sweat and $^{13}C_2$-dansyl sweat were mixed 1:1 (v/v) for injection into nLC-MS and mLC-MS for analysis.

6.3 Results and Discussion

6.3.1 Column Selection

We recognize that some users may re-purpose an existing nLC-MS system used for shotgun proteomic analysis to analyze metabolomic samples for metabolomics. Various factors need to be considered to make such a switch including column selection. We initially used a set of Waters trap column and analytical column used for proteomic analysis to analyze the dansyl labeled urine samples. The resulting chromatogram showed wide peak widths of around 0.8 min with tailing (an example is shown in Appendix Figure A6.1A), compared to widths of ~0.06 min for peptides. This problem was found to be caused by the Waters Symmetry C18 trap which uses high purity silica

with end capping. Peak broadening was not observed when the sample was directly loaded onto the analytical column which uses a polymeric bonded phase. It is very likely that a substantial amount of residual silanol activity existed in the trap column that caused broadening for the basic dansylated metabolites but not the peptides. We then switched the trap and column set to the Thermo Scientific PepMap 100 C_{18} set which also used high purity silica. However, the PepMap 100 set was found to give adequate peak widths of 0.2 minutes with reduced tailing (see Appendix Figure A6.1B). This set was thus selected for the subsequent experiments. This example illustrates the importance of selecting a proper column and trap combination for profiling labeled metabolites.

6.3.2 Separation Parameters

Several nLC parameters were optimized to achieve an optimal coverage of the amine/phenol submetabolome profile within the shortest run time. First, analytical flow rates of 500, 350, 150 nL/min were tested. With the three different flow rates, we found no significant differences in the number of peak pairs or metabolites detected. However, while using 500 nL/min gave the shortest run time, it increased backpressure, resulting in popping the fused-silica capillaries out of their fittings. At the lowest flow rate of 150 nL/min, the analysis time was increased by 10 min. Thus the flow rate of 350 nL/min was chosen as a compromise for the work.

Next we optimized the gradient separation condition. It was found that the majority of the labeled metabolites eluted between 15% and 60% mobile phase B (acetonitrile 0.1% formic acid). Thus, a shallow gradient from 25% to 60% over 20 min was used to improve separation.

The solvent composition of the diluent used to prepare the dansylated samples was also optimized. Initially, the samples were diluted using the same solvent composition as that used for the dansylation labeling reaction, i.e., 1:1 acetonitrile:water (v/v) 0.1% formic acid, to prevent any

potential precipitation of highly non-polar dansylated metabolites. This sample plug with a high organic composition greatly reduced the retention of metabolites on the trap column, causing metabolites to be flushed out of the trap column and into the waste. As a result, a portion of early eluting peaks were reduced in intensity (see Appendix Figure A6.2A). After testing a number of different diluents, a diluent composed of 1:9 acetonitrile:water (v/v) 0.1% formic acid was found to give no sample loss or precipitation for the urine samples studied (see Appendix Figure A6.2B).

6.3.3 Trapping Optimization and Efficiency

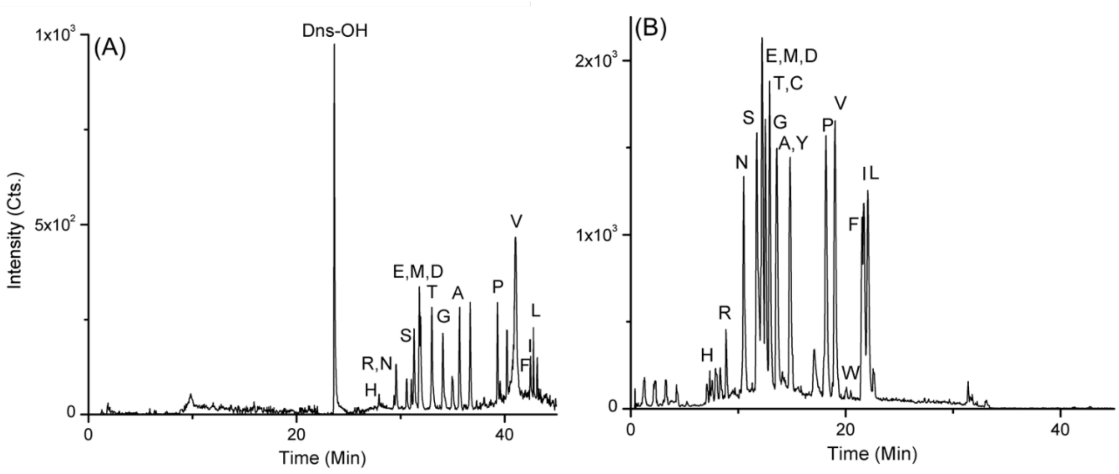


Figure 6.1 nLC-MS chromatograms of a mixture of 18 dansylated amino acids obtained (A) without using a trap column and (B) with the use of a trap column. The peak at 23.63 min was from dansyl-OH, a product of dansyl reagent after quenching with NaOH. This product did not retain on the RP trap column and thus did not show up in (B).

A trapping column is an integral part of nLC-MS for injecting a relatively large volume of samples. It is not commonly used in mLC-MS, as injection of several microliters of sample is compatible with the high flow rate. In nLC-MS, prior to separating on the analytical column, the sample is first pushed through a short trap column, usually at a higher flow rate compared to the analytical flow rate. Analytes are retained on the trap while extra diluent and other non-retaining matrix components are flushed into the waste. This serves two functions: the first is to remove salts

and other interfering chemicals and the second is to reduce the time it takes for samples to reach the column. As a result, a large volume of sample can be loaded onto the column in a short time. Figure 6.1 shows the chromatograms of the separation of a mixture of dansylated amino acids using a 5 μL injection loop at a flow rate of 350 nL/min. Without the use of the trap column, there was a dead time of 16.67 min and the first retained analyte eluted in 23.63 min (see Figure 6.1A). With the trap column, at a trapping flow rate of 7.0 μL/min, the dead time of the sample loop was reduced to 0.71 min leaving only the dead time of the gradient delay which was 7.10 min (see Figure 6.1B). Overall, there was a reduction of 16.53 min in run time when the trap was used. Therefore, analyte trapping is essential for reducing the dead time of nLC-MS operating at nanoliter flow rates.

In using the trap, the goal is to have metabolites completely retained on the trap while mobile phase is pushed through at the highest flow rate possible to wash out salts and other non-analytes. The concern is that with higher flow rate there will be more metabolites that are flushed into the waste. Therefore, the trapping flow rate, trapping mobile phase composition and trapping time need to be carefully balanced. Several trapping flow rates ranging from 1 μL/min to 20 μL/min was tested; 20 μL/min was the highest flow rate possible without over pressuring the trap column. By increasing the trap flow rate, the number of peak pairs detected was reduced due to sample loss. Decreasing the flow rate caused a longer dead time and longer overall run time with no significant increase in peak pair number. The optimal flow rate was found to be 7 μL/min which was the highest flow rate without substantial sample loss. The trapping mobile phase composition was optimized to be 2% acetonitrile in water. Increasing the organic composition washed away the sample. Finally, the shortest trapping time to wash the entire sample out of the 5 μL sample loop and onto the trap, in addition to washing out salts, was set at 1 min.

While trapping can increase the detection sensitivity by allowing for the injection of a large volume of dilute sample, there is a greater chance that the analytes might be washed out with larger loading volume. After optimizing the trapping conditions, we investigated the trapping efficiency by injecting a series of diluted dansylated amino acid mixtures where the injection amount was kept constant at 120 fmol by adjusting the injection volume and concentration. No substantial sample loss was observed from 1000 to 10000 fold diluted samples when looking at the measured peak area for selected dansylated amino acids as shown in Figure 6.2. This shows that the nLC-MS trapping condition used was efficient at trapping low-concentration and high-injection-volume dansylated samples without affecting chromatographic separation or incurring large amounts of sample loss.

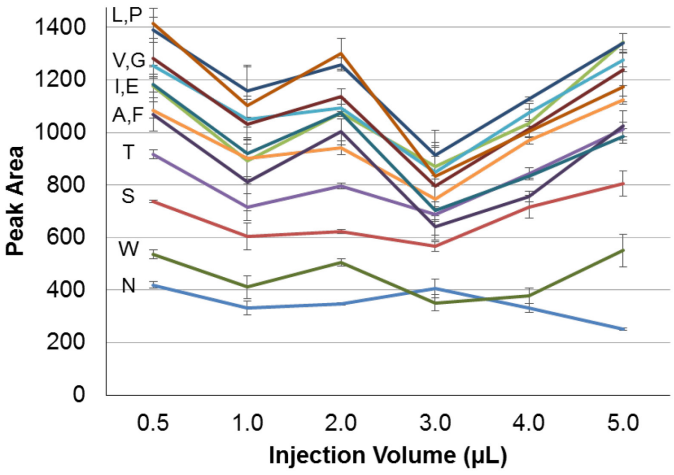


Figure 6.2 Chromatographic peak areas of dansylated amino acids with the same sample injection amount (120 fmol).

6.3.4 Chromatographic Reproducibility

In untargeted metabolomic studies, reproducible retention time is required for data file alignment to generate accurate abundance information across hundreds of samples that are run on different days or even different weeks. Appendix Table A6.1 shows the intraday retention time

reproducibility of dansylated amino acids measured using the nLC and mLC. The average relative standard deviation (%RSD) of the nLC retention times was 0.48%, which was worse than the %RSD of the mLC at 0.06%. This confirms reports by other groups that nLC retention time is not as stable as mLC.[27] The lower retention time reproducibility may be due to the reduced quality in stationary phase packing in preparing the nLC columns and a larger flow rate variation with nLC pumps vs. mLC pumps. Retention time stability has a negative effect on the quantification of unlabeled metabolites between different samples and several peak alignment methods have been reported to reduce the effect.[28] However, with CIL, each $^{12}C_2$ dansylated metabolite in a sample is quantified relative to the $^{13}C_2$ dansylated metabolite in a control and thus precise alignment is not required for relative quantification.

In addition to retention time, the intensity of the metabolites needs to be stable between sample runs for accurate quantification. Appendix Table A6.2 shows the intensity %RSDs of the dansylated amino acids. The average %RSD of nLC intensities was found to be similar to that of mLC (3.6% vs. 3.3%). Thus, the quantitative precision of the two systems is similar.

6.3.5 Sensitivity Improvement

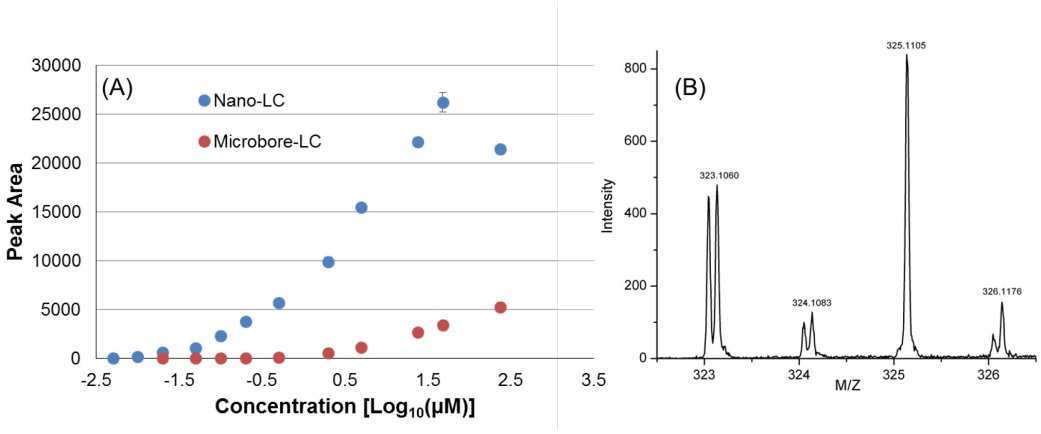


Figure 6.3 (A) Chromatographic peak area as a function of sample solution concentration for dansyl alanine analysis. Error bar represents one standard deviation (n=3). (B)

Molecular ion region of the mass spectrum obtained from 1:2 mixture of 12C-dansyl alanine and ^{13}C-dansyl alanine at 5 nM with an injection of 5 µL solution (i.e., 25 fmol). The extra peak next to the ^{12}C-dansyl alanine was from a background species.

Figure 6.3A shows the plots of peak areas as a function of sample injection amount for nLC- and mLC-MS using dansyl alanine as an example. Signal saturation was observed for nLC-MS when the analyte concentration was over 48 µM, corresponding to 240 pmol with 5 µL injection. In contrast, even at 238 µM, the peak area obtained by mLC-MS was not very high. In fact, it was slightly lower than that obtained using the solution of 0.5 µM in nLC-MS. This result demonstrates a more than 476-fold increase in detection sensitivity. At the low limit, as Figure 6.3B shows, mass spectral signals were still detectable at S/N 25 with the injection of the 0.005 µM or 5 nM solution, corresponding to 0.025 pmol or 25 fmol amount. For other labeled amino acids tested (see Figure 4 and Appendix Figure A6.3), injections of 5 nM of dansylated glycine, glutamic acid, asparagine, phenylalanine, leucine and tryptophan gave signals with S/N 130, 120, 11, 30, 150 and 7, respectively. These results illustrate that with a 5-µL loop injection we can now analyze metabolites at <5 nM concentrations with an analyte amount of <25 fmol.

6.3.6 Dynamic Range for Relative Quantification

Quantitative metabolomics relies on relative quantification of all the metabolites in comparative samples, not just one or a few metabolites. In CIL LC-MS, relative quantification of each metabolite is achieved by calculating the peak ratio of the ^{12}C-labeled metabolite in a sample and the ^{13}C-labeled same metabolite in a control. It is always desirable to detect as many peak pairs as possible in a mass spectrum to quantify the low and high abundance metabolites. However, if a peak pair becomes saturated in a mass spectrum, the highest peak in the pair will become compressed, distorting the measured peak ratio. In our work, the dynamic range for detecting peak pairs was evaluated by analyzing a series of diluted solutions of a 1:2 mixture of a ^{12}C-labeled

amino acid and its ^{13}C-labeled counterpart for several amino acids. The theoretical peak intensity ratio should be 0.5. However, the ^{13}C-labeled peak should be saturated first when the concentration of the mixture increases. Thus, the measured peak ratio will be greater than 0.5 when the ^{13}C-peak is saturated.

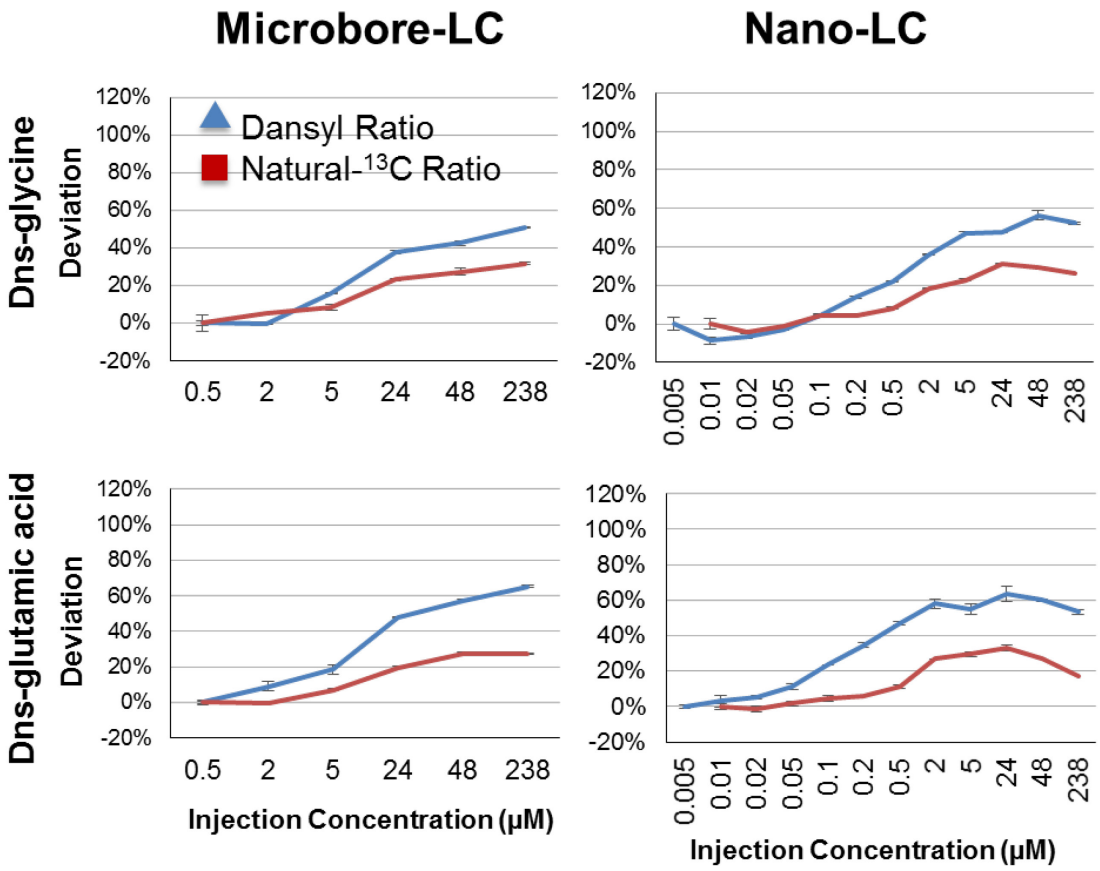


Figure 6.4 Effect of detector saturation on the calculated peak pair ratio in mLC-MS and nLC-MS. Derivation from the expected 1:2 ratio is plotted as a function of the solution concentration of 1:2 mixture of ^{12}C-dansyl amino acid and ^{13}C-dansyl amino acid.

Figure 6.4 shows the deviation of the peak ratios of glutamic acid and glycine at different mixture concentrations. Both mLC-MS and nLC-MS showed deviations from the theoretical ratio of 0.5 when the sample concentration increased, showing the effect of detector saturation on the quantification of metabolites. The trends were similar with the other dansyl amino acids (see

Appendix Figure S6.3). The mLC-MS dansyl peak ratios for glycine and glutamic acid deviated more than 20% at concentrations of > 5 μM, and below 0.5 μM the amino acids were not observed. This means that an accurate peak ratio can only be obtained within 1 order of magnitude in concentration for this mLC-MS setup. The nLC-MS deviated above 20% at 0.5 μM for glycine and 0.1 μM for glutamic acid. The higher sensitivity allowed the lower end concentration to be reduced down to 0.005 μM, giving a concentration range increase of 2 and 1.3 folds for glycine and glutamic acid, respectively.

To extend the dynamic range when the peaks become saturated, the natural-^{13}C peaks can be used to recover the accurate peak ratio because they are of lower intensity and still reflect the sample ratios.[29] Figure 6.4 shows that the natural-^{13}C peak ratios were more resistant to deviation caused by detector saturation. In mLC-MS, the natural-^{13}C peak ratios of glycine deviated above 20% for glycine and glutamic acid at 24 μM, instead of 5 μM, when measuring dansyl peak ratios. The nLC-MS deviated past 20% at 2 μM for both amino acids which was between 4 and 20 times higher concentration than using the dansyl peak ratio. Combining the concentrations that deviated less than 20% using both dansyl and natural-^{13}C peak ratios, the nLC-MS had a range of 2.6 orders of magnitude, while mLC-MS had 1.7, for both amino acids.

The above results demonstrate that nLC-MS offers a greater dynamic range for detecting peak pairs with accurate quantification, compared to mLC-MS. If we relax the deviation to ~30%, instead of 20%, the quantitative dynamic range becomes 476-fold (i.e., 0.5 to 238 μM) for mLC-MS and 47600-fold (i.e., 0.005 to 238 μM) for nLC-MS for a 1:2 mixture of ^{12}C-/^{13}C-dansyl glycine or ^{12}C-/^{13}C-dansyl glutamic acid.

6.3.7 Urine Submetabolome Profiling

Dansylated human urine was used for the direct comparison of metabolomic analysis sensitivity between nLC-MS and mLC-MS. A urine sample was split and labeled with ^{12}C- and ^{13}C-dansyl chloride, followed by mixing together in a 1:1 ratio. The total concentration of all dansylated metabolites in urine was quantified to be 48.2 mM. The ^{12}C-/^{13}C-labeled urine mixture was diluted up to 10000-fold and injected in triplicate in decreasing metabolite amounts.

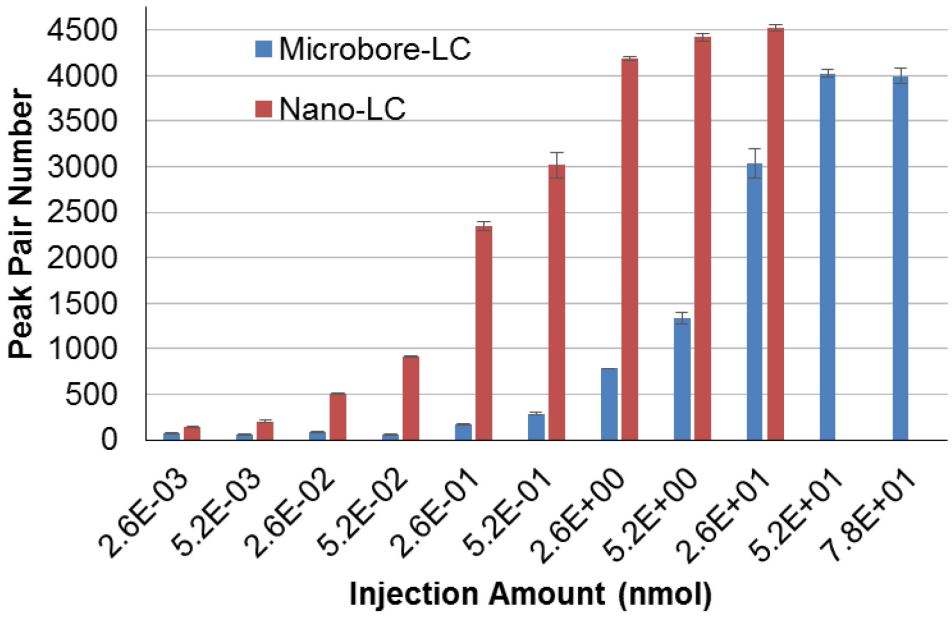


Figure 6.5 Number of peak pairs detected as a function of the sample injection amount from mLC-MS and nLC-MS analysis of 12C-/13C-labeled human urine sample.

Figure 6.5 shows that the maximum number of detected metabolites in urine using nLC-MS was 4524±37 (n=3) at 26.076 nmol of metabolites injected, while mLC-MS gave a maximum number of 4019±40 at 52.151 nmol injection. Thus, nLC-MS detected about 13% more metabolites than mLC-MS, likely due to improved dynamic range of detecting peak pairs. At the optimal injection amount for nLC-MS of 26.076 nmol, mLC-MS detected only 67% of the metabolites (i.e.,

3034±161). This means that the optimal injection amount for nLC-MS was 2 times lower than using mLC-MS. The improved sensitivity of nLC-MS was more apparent at lower sample loading amounts; below 0.522 nmol loading, nLC-MS detected at least 8 times more metabolites than mLC-MS. It is clear that if sample amount is not limited, mLC-MS can still be used for metabolomic profiling without incurring too large drop in the number of metabolites detected. However, as Figure 6.5 shows, sample dilution has a much greater effect on mLC-MS than nLC-MS. For example, injecting 2.6 nmol detected less than 1/3 of the peak pairs found in the 26 nmol injection by mLC-MS, while injecting 0.26 nmol in nLC-MS detected more than half of the peak pairs found in the 2.6 nmol injection. Thus, nLC-MS would have a clear advantage in handling samples of limited amounts or diluted samples.

6.3.8 Sweat Submetabolome Profiling

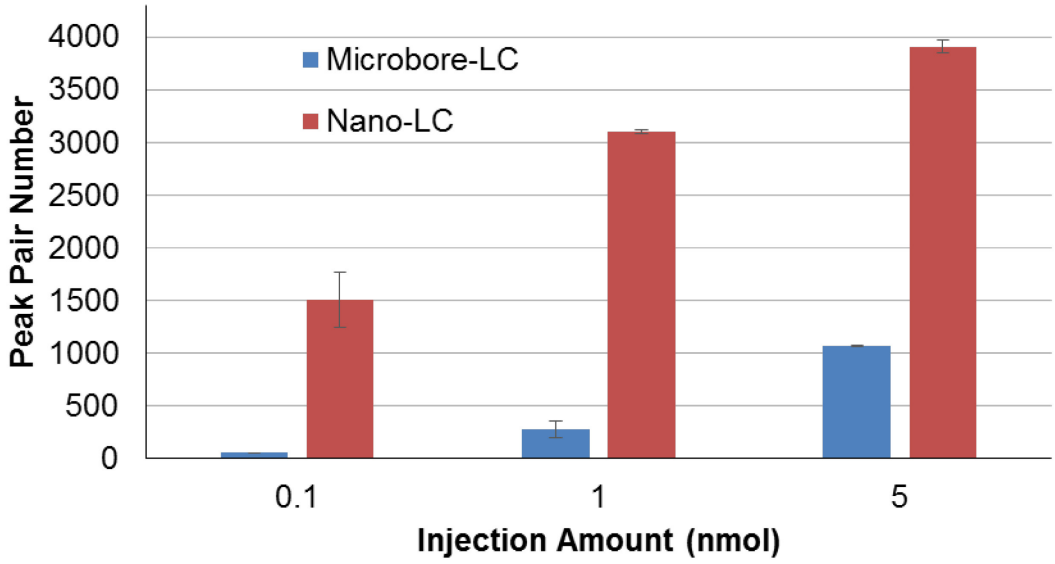


Figure 6.6 Number of peak pairs detected as a function of the sample injection amount from mLC-MS and nLC-MS analysis of 12C-/13C-labeled human sweat sample.

The advantage of nLC-MS for analyzing a limited amount of sample can be demonstrated in profiling the human sweat metabolome. Typically, only several microliters of sweat can be collected from a subject without needing a prolonged collection and using a very large collection area. For this study, about 10 μL of human sweat was collected after exercise from an arm of a healthy individual. The total concentration of the dansyl labeled metabolites in the sweat was determined to be 8.4 mM using LC-UV, which was 6 times lower than the total concentration of metabolites in urine. For sweat analysis, the lower total concentration and the lower volume require the extra sensitivity offered by the nLC-MS. Figure 6.6 shows that at the maximum injection amount of 5 nmol of dansylated sweat, 3908±62 peak pairs were detected for nLC-MS and 1064±6 peak pairs were detected for mLC-MS, or a 4-fold increase in the number of metabolites detected. Due to the limited amount of sample, it was not possible to inject an optimal amount for mLC-MS like with the urine sample. The higher sensitivity was again observed at lower sample loading amounts of 1 nmol where nLC-MS has 11-fold higher peak pair values of 3098±16 compared to 275±78 from mLC-MS. We envisage the use of nLC-MS for analyzing many types of metabolomic samples where the sample amount is limited, such as a microliter of sweat collected naturally, a droplet of blood from a finger prick, etc.

6.3.9 Robustness

For routine metabolomic analysis, an analytical tool needs to be robust in dealing with a large number of samples. Due to the small inner diameter of fused-silica capillaries, columns, and nESI emitters used in nLC-MS, the entire system is more finicky to maintain, compared to mLC-MS. Firstly, all of the fused-silica components are much more fragile than the polymer and stainless steel components used in mLC-MS and must be handled with care. The small internal diameter also means that the capillaries are more prone to clogging from particles in the samples and silica

particles from poorly cut and ragged tubing edges. The small nESI emitters are more prone to clogging from sample matrix precipitation and silica particles from the fused-silica, and backpressure must be regularly monitored for clogging.

There are a few precautions that can be taken to reduce the frequency of catastrophic clogging of nLC-MS. In our laboratory, a typical capillary internal diameter of 20 μm and outer diameter of 360 μm was found to be the optimal balance between robustness and chromatographic performance by reducing dead volume. For cutting fused-silica capillaries to the necessary lengths, a rotating diamond cutter is expensive but highly recommended for its ability to reproducibly give clean cuts that are free of capillary clogging particles. Following cutting, new fused-silica columns and capillaries must be flushed and their ends washed with clean solvent to remove any particles. Although the nLC-MS platform can be less robust than mLC-MS platform, by following these precautions and being careful, the system can be operated for months with little downtime. Recent advancements in nLC technology, such as the use of an integrated microfluidic column or capillary cartridge that can be conveniently connected to an MS interface,[30] are expected to make nLC-MS more robust for routine metabolomic analysis.

6.4 Conclusions

We report a nanoflow LC-MS system combined with chemical isotope labeling of metabolites for metabolomic profiling with high coverage. A reversed phase trap column is used to capture the labeled metabolites at a flow rate of 7 μL/min, followed by separation on a capillary RPLC column at a flow rate of 350 nL/min. The sample injection volume is typically at 5 μL, allowing the analysis of a diluted sample solution. Dansylation CIL was demonstrated for sensitive profiling of the amine/phenol submetabolome in human urine and sweat; however, the technique should be applicable to other labeling chemistries where labeled metabolites can be retained on

RPLC. Because the configuration of the nLC-MS system described herein is similar to those widely used for shotgun proteome analysis, this metabolomic profiling platform should be readily adapted.

6.5 Literature Cited

(1) Yin, P. Y.; Xu, G. W. *J. Chromatogr. A* **2015**, *1374*, 1-13.

(2) Rainville, P. D.; Theodoridis, G.; Plumb, R. S.; Wilson, I. D. *Trends Anal. Chem.* **2014**, *61*, 181-191.

(3) Zhou, R.; Huan, T.; Li, L. *Anal. Chim. Acta* **2015**, *881*, 107-116.

(4) Guo, K.; Li, L. *Anal. Chem.* **2009**, *81*, 3919-3932.

(5) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82*, 8789-8793.

(6) Liu, P.; Huang, Y. Q.; Cai, W. J.; Yuan, B. F.; Feng, Y. Q. *Anal. Chem.* **2014**, *86*, 9765-9773.

(7) Tayyari, F.; Gowda, G. A. N.; Gu, H. W.; Raftery, D. *Anal. Chem.* **2013**, *85*, 8715-8721.

(8) Dai, W. D.; Huang, Q.; Yin, P. Y.; Li, J.; Zhou, J.; Kong, H. W.; Zhao, C. X.; Lu, X.; Xu, G. W. *Anal. Chem.* **2012**, *84*, 10245-10251.

(9) Yuan, W.; Anderson, K. W.; Li, S. W.; Edwards, J. L. *Anal. Chem.* **2012**, *84*, 2892-2899.

(10) Song, P.; Mabrouk, O. S.; Hershey, N. D.; Kennedy, R. T. *Anal. Chem.* **2012**, *84*, 412-419.

(11) Zhou, R.; Tseng, C. L.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86*, 4675-4679.

(12) Huan, T.; Li, L. *Anal. Chem.* **2015**, *87*, 7011-7016.

(13) Guo, K.; Peng, J.; Zhou, R. K.; Li, L. *J. Chromatogr. A* **2011**, *1218*, 3689-3694.

(14) Mirnaghi, F. S.; Caudy, A. A. *Bioanalysis* **2014**, *6*, 3393-3416.

(15) Vuckovic, D. *Anal. Bioanal. Chem.* **2012**, *403*, 1523-1548.

(16) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. *Nat. Methods* **2010**, *7*, 681-685.

(17) Jones, D. R.; Wu, Z.; Chauhan, D.; Anderson, K. C.; Peng, J. *Anal. Chem.* **2014**, *86*, 3667-3675.

(18) Chetwynd, A. J.; Abdul-Sada, A.; Hill, E. M. *Anal. Chem.* **2015**, *87*, 1158-1165.

(19) David, A.; Abdul-Sada, A.; Lange, A.; Tyler, C. R.; Hill, E. M. *J. Chromatogr. A* **2014**, *1365*,

72-85.

(20) Uehara, T.; Yokoi, A.; Aoshima, K.; Tanaka, S.; Kadowaki, T.; Tanaka, M.; Oda, Y. *Anal. Chem.* **2009**, *81*, 3836-3842.

(21) Ivanisevic, J.; Zhu, Z. J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876-6884.

(22) Contrepois, K.; Jiang, L. H.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14*, 1684-1695.

(23) Vorkas, P. A.; Isaac, G.; Anwar, M. A.; Davies, A. H.; Want, E. J.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2015**, *87*, 4184-4193.

(24) Tulipani, S.; Mora-Cubillos, X.; Jauregui, O.; Llorach, R.; Garcia-Fuentes, E.; Tinahones, F. J.; Andres-Lacueva, C. *Anal. Chem.* **2015**, *87*, 2639-2647.

(25) Mahieu, N. G.; Huang, X. J.; Chen, Y. J.; Patti, G. J. *Anal. Chem.* **2014**, *86*, 9583-9589.

(26) Wu, Y. M.; Li, L. *Anal. Chem.* **2013**, *85*, 5755-5763.

(27) Percy, A. J.; Chambers, A. G.; Yang, J. C.; Domanski, D.; Borchers, C. H. *Anal. Bioanal. Chem.* **2012**, *404*, 1089-1101.

(28) Smith, R.; Ventura, D.; Prince, J. T. *Brief. Bioinform.* **2015**, *16*, 104-117.

(29) Zhou, R.; Li, L. *J Proteomics* **2015**, *118*, 130-139.

(30) Rainville, P. D.; Langridge, J. I.; Wrona, M. D.; Wilson, I. D.; Plumb, R. S. *Bioanalysis* **2015**, *7*, 1397-1411.

Chapter 7

Conclusion and Future Work

7.1 Thesis Summary

The use of LC-MS in the field of biology, and analytical chemistry, has started a new era of protein and metabolite analysis. The ease-of-use and the sheer number of proteins and metabolites that can be analyzed on this platform is unrivaled by any other technique. Continued fundamental research and method development in the field is needed in order to produce reliable and impactful results.

Chapter 1 outlined the fields of proteomics and metabolomics, and highlighted the instrumentation used in these fields. The chapter introduced areas that require further development and are of interest to the research community as a whole. These areas include metabolite quantification, sensitivity, peptide spectrum matching accuracy, and reference spectra for identification. The focus of the thesis was to improve these areas, so more metabolites and peptides can be accurately quantified.

The thesis began with a projected aimed at improving the sensitivity and accuracy of peptide identification for proteomic database search engines. In Chapter 2, the Percolator machine learning algorithm was coupled with X!Tandem, a popular open-source search engine. An experimentally validated dataset was used to find the best combination of scores and other matching statistics as features to distinguish between true positives from false positives. After demonstrating that the list of features was optimized, *E. coli* and human LC-MS proteomic data were fed through the X!Tandem Percolator algorithm to show real sample performance. The final

number of correct PSMs for X!Tandem Percolator was higher than the original X!Tandem and as high as the commercially available Mascot Percolator. The use of X!Tandem Percolator improved the sensitivity of peptide identification, and it was an easy to implement software available to the general proteomic community.

Chapter 3 showed that by pushing chemical vapours into the source region of nESI source can improve the sensitivity of peptide and small molecule detection. A variety of volatile compounds were screened for the largest sensitivity increase for two standard peptides, and it was found that benzyl alcohol gave the best performance and was non-toxic. This technique increased the number of unique peptides identified in the acid hydrolysate of α-casein increased by 45%; the number of peptides and proteins identified in a tryptic digest of *E. coli* cell lysate by 13% and 14%, respectively. The average PSM scores were also increased for both sample types. Again, this project showed another easy-to-implement method of improving protein detection for LC-MS proteomics.

Chapters 4-6 marks a shift from proteomics to metabolomics research, another area that is dominated by LC-MS. Chapters 4 and 5 improved the way that unknowns are identified in metabolomics studies, and Chapter 6 improved the sensitivity of metabolites in CIL metabolomics.

In Chapter 4, a high resolution QTOF instrument was used to create a new generation of MS^2 spectral library using HMDB standards. The completed library was used to demonstrate the identification of biologically relevant apple juice metabolites from urine.

Beyond the use of the library in identification, several crucial observations were made to improve future metabolomics studies. By observing the ionization behavior of all the standards, it was found that negative mode MS^2 was not as useful as positive mode for identification due to low

fragmentation efficiency. Sodiated adducts were also commonly observed, and their levels were not easily controlled.

Chapter 5 expanded on the work completed in Chapter 4, by acquiring RT information on RP-LC columns. It was found that in-source fragmentation occurred in metabolomics studies, and was a source of error; because an in-source fragment of a metabolite conjugate can be misidentified as the unconjugated metabolite. The additional RT information was shown to help correct this type of misidentification.

Retention time shifts, due to instrumental or environmental effects when the same column and elution conditions are used, hinders the usefulness of the RT library. The measured retention times in the library can no longer be matched with the retention times measured for the experiment. A RT calibration scheme developed in our lab was used to demonstrate that small non-linear shifts in retention time could be corrected by a calibration mixture and a computer algorithm. After simulating a change in LC tubing, the average shift before calibration was found to be 6.3 seconds; after the calibration process, the average shift was found to be 0.3 seconds. After implementing with the RT library it should be able to improve RT matching accuracy by accounting for small shifts in retention time.

Lastly, Chapter 6 described the optimization and implementation of nLC-MS in CIL metabolomics. The use of the more sensitivity technique detected 1000% more metabolites when the sample concentration was low, and 106% more metabolites when injecting the maximum amount of low-volume sweat, when compared with the traditional LC-MS methods.

7.2 Future Work

Peptide identification using LC-MS is a field of ongoing software developments, and there are new versions of database search engines coming out continuously. One of the future work required for the X!Tandem Percolator project is to update the program code, so that the software is kept up-to-date with the latest versions of X!Tandem.

As high resolution data is becoming more wide spread in the proteomic community, the features—those selected for the best discrimination between true and false positives—might need to be re-optimized for the new data.

For the chemical vapour study, it and other research has pointed to a more nuanced process in the ESI mechanism. While the process of ion ejecting from ESI droplets are now well understood, there is still a gap of knowledge of what happens after ions are ejected and before entering the mass spectrometer. And, it is at this critical step that the vapours might be causing the increases in ion intensity. Further mechanistic studies are needed to fully understand this step.

To make this vapour enhancement method widely available, a vapour introduction device needs to be devised for the variety of sources on mass spectrometers. This way, users in different labs with different instruments can take advantage of the higher sensitivity.

Since MS^2 spectral libraries are still the primary method for metabolite identification, the immediate future work is to expand the HMDB library with new compounds as soon as they are available. Long-term goal of metabolite identification would be to abandon the time consuming and costly library building process, and predict small molecule fragmentation to a level in which spectra can be identified with high confidence. This has been a long-term goal for mass

spectrometrists, and new work has improved on the topic; however, that stage is not likely to be reached for some time.

Much like the MS^2 spectral library, new standards are needed to expand the HMDB RT library so it will be even more useful. Again, this is a time consuming and costly endeavour and some sort of computer prediction of RT would be ideal. However, the likelihood of being able to accurate predict retention times on the variety of columns available to researchers, is even more unlikely than creating an accurate fragmentation prediction software.

In the short term, the RT library need to be bolstered with additional HILIC RT information. The 50% of metabolites that were studied in Chapter 5 did not retain on the reversed phase column, and HILIC would be the perfect separation method to analyze these compounds.

The RT normalization algorithm that was demonstrated in Chapter 5 was not implemented in the library searching program at the time of writing. Without the algorithm, small changes in the LC plumbing will have a large effect on the retention time. Currently, users are limited to using the exact same plumbing as the LC used to generate the RT; having the normalization algorithm would make the library more useful and flexible.

Lastly, the nLC CIL analysis showed great promise of using metabolomics with the small nanoflow columns. Major drawbacks of current generation nLC are the practical issues. The columns and nESI sources are all small diameter that gets blocked very easily. Constructed from fused silica, anytime fluidic connections were made there is a likely chance of dislodging a piece of silica from the tubing that will plug the nESI source. The future development for nLC, to combat improve practicality, is to use integrated microfluidic chips. On these chips, the trap, column, and nESI emitter are all integrated and no manual fluidic connections need to be made. This reduces

clogging and reduces the time needed for connecting columns. Such devices are already available from Waters and New Objective, but they have not seen wide spread use in the community. The future work would be to test and optimize these new products for routine CIL metabolomics analysis. To further enhance sensitivity, the combination of the vapour enhancement with benzyl alcohol and nLC is worth investigating for CIL metabolomics.

The promise of both metabolomics and proteomics are endless in the field of medicine and biology. With these technologies, we might be able to discover the mechanism of diseases and improve treatments or diagnosis. It is hoped that the work described in this thesis can play a part bringing that future to reality, and further the field.

Bibliography

(1) Aebersold, R.; Mann, M. *Nature (London, U. K.)* **2003**, *422*, 198-207.
(2) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2013**, 1-8.
(3) Allwood, J. W.; AlRabiah, H.; Correa, E.; Vaughan, A.; Xu, Y.; Upton, M.; Goodacre, R. *Metabolomics* **2014**, 438-453.
(4) Almeida, R.; Pauling, J. K.; Sokol, E.; Hannibal-Bach, H. K.; Ejsing, C. S. *J. Am. Soc. Mass Spectrom.* **2014**, *26*, 133-148.
(5) Alves, G.; Ogurtsov, A. Y.; Kwok, S.; Wu, W. W.; Wang, G.; Shen, R.-f.; Yu, Y.-K. *Biol. Direct* **2008**, *3*, 27.
(6) Anderson, D. M.; Biemann, K.; Orgel, L. E.; Oro, J.; Owen, T.; Shulman, G. P.; Toulmin, P.; Urey, H. C. *Icarus* **1972**, *16*, 111-138.
(7) Antti, H.; Ebbels, T. M. D.; Keun, H. C.; Bollard, M. E.; Beckonert, O.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 139-149.
(8) Apffel, A.; Fischer, S.; Goldberg, G.; Goodley, P. C.; Kuhlmann, F. E. *J. Chromatogr. A* **1995**, *712*, 177-190.
(9) Asara, J. M.; Christofk, H. R.; Freimark, L. M.; Cantley, L. C. *Proteomics* **2008**, *8*, 994-999.
(10) Awad, H.; El-Aneed, A. *Mass Spectrom. Rev.* **2013**, *32*, 466–483.
(11) Babushok, V. I.; Linstrom, P. J.; Reed, J. J.; Zenkevich, I. G.; Brown, R. L.; Mallard, W. G.; Stein, S. E. *J. Chromatogr. A* **2007**, *1157*, 414-421.
(12) Baker, M. *Nature* **2012**, *484*, 271-275.
(13) Bandu, M. L.; Watkins, K. R.; Bretthauer, M. L.; Moore, C. A.; Desaire, H. *Anal. Chem.* **2004**, *76*, 1746-1753.
(14) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. *Anal. Bioanal. Chem.* **2007**, *389*, 1017-1031.
(15) Bao, J.; Gao, X.; Jones, A. D. *Rapid Commun. Mass Spectrom.* **2014**, *28*, 457-464.
(16) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protoc.* **2007**, *2*, 2692-2703.
(17) Bjornson, R. D.; Carriero, N. J.; Colangelo, C.; Shifman, M.; Cheung, K.-H.; Miller, P. L.; Williams, K. *J. Proteome Res.* **2008**, *7*, 293-299.
(18) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annu. Rep. Comput. Chem.*, Ralph, A. W.; David, C. S., Eds.; Elsevier, 2008, pp 217-241.
(19) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218*, 6732-6741.
(20) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; Dame, Z. T.; Poelzer, J.; Huynh, J.; Yallou, F. S.; Psychogios, N.; Dong, E.; Bogumil, R.; Roehring, C.; Wishart, D. S. *PLoS One* **2013**, *8*, e73076.
(21) Bristow, A. W. T.; Nichols, W. F.; Webb, K. S.; Conway, B. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2374-2386.
(22) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. *Mol. Cell. Proteomics* **2008**, *7*, 962-970.
(23) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J. Proteome Res.* **2009**, *8*, 3176-3181.
(24) Buszewski, B.; Jaroniec, M.; Gilpin, R. K. *J. Chromatogr. A* **1994**, *673*, 11-19.
(25) Buszewski, B.; Noga, S. *Anal. Bioanal. Chem.* **2011**, *402*, 231-247.
(26) Chapman, J. D.; Goodlett, D. R.; Masselon, C. D. *Mass Spectrom. Rev.* **2013**, 1-19.

(27) Chen, A.; Lynch, K. B.; Wang, X.; Lu, J. J.; Gu, C.; Liu, S. *Anal. Chim. Acta* **2014**, *844*, 90-98.
(28) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849-865.
(29) Chetwynd, A. J.; Abdul-Sada, A.; Hill, E. M. *Anal. Chem.* **2015**, *87*, 1158-1165.
(30) Cho, K.; Mahieu, N. G.; Johnson, S. L.; Patti, G. J. *Curr. Opin. Biotechnol.* **2014**, *28*, 143-148.
(31) Choi, H.; Fermin, D.; Nesvizhskii, A. I. *Mol. Cell. Proteomics* **2008**, *7*, 2373-2385.
(32) Chubukov, V.; Gerosa, L.; Kochanowski, K.; Sauer, U. *Nat. Rev. Microbiol.* **2014**, *12*, 327-340.
(33) Cirillo, P.; Sato, W.; Reungjui, S.; Heinig, M.; Gersch, M.; Sautin, Y.; Nakagawa, T.; Johnson, R. J. *J. Am. Soc. Nephrol.* **2006**, *17*, S165-S168.
(34) Contino, N. C.; Jarrold, M. F. *Int. J. Mass Spectrom.* **2013**, *345*, 153-159.
(35) Contrepois, K.; Jiang, L.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14*, 1684-1695.
(36) Contrepois, K.; Jiang, L. H.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14*, 1684-1695.
(37) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.
(38) Crick, F. *Nature* **1970**, *227*, 561-563.
(39) Cutillas, P. R. *Current Nanoscience* **2005**, *1*, 65-71.
(40) Dai, W. D.; Huang, Q.; Yin, P. Y.; Li, J.; Zhou, J.; Kong, H. W.; Zhao, C. X.; Lu, X.; Xu, G. W. *Anal. Chem.* **2012**, *84*, 10245-10251.
(41) David, A.; Abdul-Sada, A.; Lange, A.; Tyler, C. R.; Hill, E. M. *J. Chromatogr. A* **2014**, *1365*, 72-85.
(42) Dixon, C. M.; Saynor, D. A.; Andrew, P. D.; Oxford, J.; Bradbury, A.; Tarbit, M. H. *Drug Metab. Dispos.* **1993**, *21*, 761-769.
(43) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365-8374.
(44) Doohan, R. A.; Hayes, C. A.; Harhen, B.; Karlsson, N. G. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1052-1062.
(45) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2012**, *9*, 44-66.
(46) Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; Yeates, T. O. *Nature* **2000**, *405*, 823-826.
(47) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207-214.
(48) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-989.
(49) Erban, A.; Schauer, N.; Fernie, A.; Kopka, J. In *Metabolomics*, Weckwerth, W., Ed.; Humana Press, 2007, pp 19-38.
(50) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. *Proteomics* **2012**, *12*, 1111-1121.
(51) Fenyö, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.
(52) Fenyoe, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768-774.
(53) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301-305.
(54) Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P. *J. Chromatogr. A* **2013**, *1292*, 51-65.
(55) Gangl, E. T.; Annan, M.; Spooner, N.; Vouros, P. *Anal. Chem.* **2001**, *73*, 5635-5644.
(56) Garibyan, L.; Avashia, N. *J. Invest. Dermatol.* **2013**, *133*, e6.

(57) Gartland, K. P.; Sanins, S. M.; Nicholson, J. K.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C. *NMR Biomed.* **1990**, *3*, 166-172.
(58) Gatlin, C. L.; Kleemann, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. *Anal. Biochem.* **1998**, *263*, 93-101.
(59) Geiger, T.; Madden, S. F.; Gallagher, W. M.; Cox, J.; Mann, M. *Cancer Res.* **2012**, *72*, 2428-2439.
(60) Giancaterini, a.; De Gaetano, a.; Mingrone, G.; Gniuli, D.; Liverani, E.; Capristo, E.; Greco, a. V. *Metabolism.* **2000**, *49*, 704-708.
(61) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, O111.016717.
(62) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *86*, 6931-6939.
(63) Grada, A.; Weinbrecht, K. *J. Invest. Dermatol.* **2013**, *133*, e11.
(64) Griffiths, J. *Anal. Chem.* **2008**, *80*, 5678-5683.
(65) Gritti, F.; Kazakevich, Y. V.; Guiochon, G. *J. Chromatogr. A* **2007**, *1169*, 111-124.
(66) Guo, K.; Li, L. *Anal. Chem.* **2009**, *81*, 3919-3932.
(67) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82*, 8789-8793.
(68) Guo, K.; Peng, J.; Zhou, R. K.; Li, L. *J. Chromatogr. A* **2011**, *1218*, 3689-3694.
(69) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.
(70) Guthals, A.; Clauser, K. R.; Frank, A. M.; Bandeira, N. *J. Proteome Res.* **2013**, *12*, 2846-2857.
(71) Gygi, S. P. *Nat. Biotechnol.* **1999**, *17*, 994-999.
(72) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Medard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat. Methods* **2013**, *10*, 989.
(73) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Médard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat. Methods* **2013**, *10*, 989-991.
(74) Harvey, D. J. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 622-630.
(75) Hassell, K. M.; LeBlanc, Y. C.; McLuckey, S. A. *Anal. Chem.* **2011**, *83*, 3252-3255.
(76) Heinemann, M.; Sauer, U. In *Curr. Opin. Microbiol.*, 2010, pp 337-343.
(77) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics (Oxford, England)* **2012**, *28*, 2333-2341.
(78) Heiskanen, L. a.; Suoniemi, M.; Ta, H. X.; Tarasov, K.; Ekroos, K. *Anal. Chem.* **2013**, *85*, 8757-8763.
(79) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111-3118.
(80) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. *Anal. Chem.* **2008**, *80*, 5574-5582.
(81) Hopley, C.; Bristow, T.; Lubben, A.; Simpson, A.; Bull, E.; Klagkou, K.; Herniman, J.; Langley, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1779-1786.
(82) Hopper, J. T. S.; Sokratous, K.; Oldham, N. J. *Anal. Biochem.* **2012**, *421*, 788-790.
(83) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703-714.

(84) Huan, T.; Li, L. *Anal. Chem.* **2015**, *87*, 7011-7016.
(85) Hutschenreuther, A.; Kiontke, A.; Birkenmeier, G.; Birkemeyer, C. *Analytical Methods* **2012**, *4*, 1953.
(86) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2007**, *6*, 392-398.
(87) Ikonomou, M. G.; Blades, A. T.; Kebarle, P. *Anal. Chem.* **1991**, *63*, 1989-1998.
(88) Issaq, H.; Veenstra, T. *BioTechniques* **2008**, *44*, 697-698.
(89) Ivanisevic, J.; Zhu, Z. J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876-6884.
(90) Ivanisevic, J.; Zhu, Z.-J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876-6884.
(91) Jarussophon, S.; Acoca, S.; Gao, J.-M.; Deprez, C.; Kiyota, T.; Draghici, C.; Purisima, E.; Konishi, Y. *The Analyst* **2009**, *134*, 690-700.
(92) Jones, D. R.; Wu, Z.; Chauhan, D.; Anderson, K. C.; Peng, J. *Anal. Chem.* **2014**, *86*, 3667-3675.
(93) Juraschek, R.; Dülcks, T.; Karas, M. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 300-308.
(94) Kaell, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.
(95) Kahle, K.; Kempf, M.; Schreier, P.; Scheppach, W.; Schrenk, D.; Kautenburger, T.; Hecker, D.; Huemmer, W.; Ackermann, M.; Richling, E. *Eur. J. Nutr.* **2011**, *50*, 507-522.
(96) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.
(97) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2007**, *7*, 40-44.
(98) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2007**, *7*, 29-34.
(99) Kamleh, A.; Barrett, M. P.; Wildridge, D.; Burchmore, R. J. S.; Scheltema, R. A.; Watson, D. G. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1912-1918.
(100) Karr, J. R.; Sanghvi, J. C.; Macklin, D. N.; Gutschow, M. V.; Jacobs, J. M.; Bolival, B.; Assad-Garcia, N.; Glass, J. I.; Covert, M. W. *Cell* **2012**, *150*, 389-401.
(101) Kebarle, P.; Verkerk, U. H. *Mass Spectrom. Rev.* **2009**, *28*, 898-917.
(102) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383-5392.
(103) Kharlamova, A.; McLuckey, S. A. *Anal. Chem.* **2011**, *83*, 431-437.
(104) Kharlamova, A.; Prentice, B. M.; Huang, T.-Y.; McLuckey, S. A. *Anal. Chem.* **2010**, *82*, 7422-7429.
(105) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.
(106) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, *81*, 10038-10048.
(107) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. *Anal. Chem.* **2013**, *85*, 2-9.
(108) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785-7795.
(109) Lemeer, S.; Heck, A. J. *Curr. Opin. Chem. Biol.* **2009**, *13*, 414-420.
(110) Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 443-458.
(111) Liu, P.; Huang, Y. Q.; Cai, W. J.; Yuan, B. F.; Feng, Y. Q. *Anal. Chem.* **2014**, *86*, 9765-9773.
(112) Lundgren, D. H.; Hwang, S.-I.; Wu, L.; Han, D. K. *Expert Rev. Proteomics* **2010**, *7*, 39-53.
(113) Mahieu, N. G.; Huang, X. J.; Chen, Y. J.; Patti, G. J. *Anal. Chem.* **2014**, *86*, 9583-9589.
(114) Mak, C. M.; Lee, H.-C. H.; Chan, A. Y.-W.; Lam, C.-W. *Crit. Rev. Clin. Lab. Sci.* **2013**, *50*, 142-162.

(115) Mamyrin, B. a. *Int. J. Mass Spectrom.* **2001**, *206*, 251-266.
(116) Marinkovic-Ilsen, A.; van den Ende, A.; Wolthers, B. G. *Arch. Dermatol. Res.* **1984**, *276*, 364-369.
(117) Marshall, A. G.; Hendrickson, C. L. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* **2008**, *1*, 579-599.
(118) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1-35.
(119) Mayr, B. M.; Kohlbacher, O.; Reinert, K.; Sturm, M.; Gröpl, C.; Lange, E.; Klein, C.; Huber, C. G. *J. Proteome Res.* **2006**, *5*, 414-421.
(120) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69*, 767-776.
(121) Meyer, J. G.; A. Komives, E. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1390-1399.
(122) Miller, P. E.; Denton, M. B. *J. Chem. Educ.* **1986**, *63*, 617.
(123) Mirnaghi, F. S.; Caudy, A. A. *Bioanalysis* **2014**, *6*, 3393-3416.
(124) Mokhtarani, M.; Diaz, G. A.; Rhead, W.; Lichter-Konecki, U.; Bartley, J.; Feigenbaum, A.; Longo, N.; Berquist, W.; Berry, S. A.; Gallagher, R.; Bartholomew, D.; Harding, C. O.; Korson, M. S.; McCandless, S. E.; Smith, W.; Vockley, J.; Bart, S.; Kronn, D.; Zori, R.; Cederbaum, S.; Dorrani, N.; Merritt, J. L.; Sreenath-Nagamani, S.; Summar, M.; LeMons, C.; Dickinson, K.; Coakley, D. F.; Moors, T. L.; Lee, B.; Scharschmidt, B. F. *Mol. Genet. Metab.* **2012**, *107*, 308-314.
(125) Musunuri, S.; Wetterhall, M.; Ingelsson, M.; Lannfelt, L.; Artemenko, K.; Bergquist, J.; Kultima, K.; Shevchenko, G. *J. Proteome Res.* **2014**, *13*, 2056-2068.
(126) Nair, B. *Int. J. Toxicol.* **2001**, *20 Suppl. 3*, 23-50.
(127) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646-4658.
(128) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787-797.
(129) Nguyen, S.; Fenn, J. B. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1111-1117.
(130) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. *Nat. Methods* **2010**, *7*, 681-685.
(131) Núñez, O.; Gallart-Ayala, H.; Martins, C. P. B.; Lucci, P.; Busquets, R. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2013**, *927*, 3-21.
(132) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nat. Methods* **2007**, *4*, 709-712.
(133) Orth, J. D.; Conrad, T. M.; Na, J.; Lerman, J. A.; Nam, H.; Feist, A. M.; Palsson, B. Ø. In *Mol. Syst. Biol.*, 2011.
(134) Page, J. S.; Marginean, I.; Baker, E. S.; Kelly, R. T.; Tang, K.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2265-2272.
(135) Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2005**, *24*, 508-548.
(136) Patterson, S. D.; Aebersold, R. H. *Nat. Genet.* **2003**, *33*, 311-323.
(137) Patti, G. J. *J. Sep. Sci.* **2011**, *34*, 3460-3469.
(138) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43-50.
(139) Percy, A. J.; Chambers, A. G.; Yang, J. C.; Domanski, D.; Borchers, C. H. *Anal. Bioanal. Chem.* **2012**, *404*, 1089-1101.
(140) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.
(141) Pizzagalli, F.; Varga, Z.; Huber, R. D.; Folkers, G.; Meier, P. J.; St-Pierre, M. V. *J. Clin. Endocrinol. Metab.* **2003**, *88*, 3902-3912.

(142) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.
(143) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J. *Bioinformatics (Oxford, England)* **2009**, *25*, 758-764.
(144) Polson, C.; Sarkar, P.; Incledon, B.; Raguvaran, V.; Grant, R. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **2003**, *785*, 263-275.
(145) Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2012**, *4*, 565-583.
(146) Qendro, V.; Lundgren, D. H.; Rezaul, K.; Mahony, F.; Ferrell, N.; Bi, A.; Latifi, A.; Chowdhury, D.; Gygi, S.; Haas, W.; Wilson, L.; Murphy, M.; Han, D. K. *J. Proteome Res.* **2014**, *13*, 5031-5040.
(147) Rainville, P. D.; Langridge, J. I.; Wrona, M. D.; Wilson, I. D.; Plumb, R. S. *Bioanalysis* **2015**, *7*, 1397-1411.
(148) Rainville, P. D.; Theodoridis, G.; Plumb, R. S.; Wilson, I. D. *Trends Anal. Chem.* **2014**, *61*, 181-191.
(149) Rauniyar, N.; Yates, J. R. *J. Proteome Res.* **2014**, *13*, 5293-5309.
(150) Righetti, P. G.; Castagna, A.; Antonucci, F.; Piubelli, C.; Cecconi, D.; Campostrini, N.; Antonioli, P.; Astner, H.; Hamdan, M. In *J. Chromatogr. A*, 2004, pp 3-17.
(151) Ross, P. L. *Mol. Cell. Proteomics* **2004**, *3*, 1154-1169.
(152) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Mol. Cell. Proteomics* **2004**, *3*, 1154-1169.
(153) Shiio, Y.; Aebersold, R. *Nat. Protoc.* **2006**, *1*, 139-145.
(154) Shuford, C. M.; Muddiman, D. C. *Expert Rev. Proteomics* **2011**, *8*, 317-323.
(155) Smith, R.; Ventura, D.; Prince, J. T. *Brief. Bioinform.* **2015**, *16*, 104-117.
(156) Snyder, L. R.; Kirkland, J. J.; Dolan, J. W. *Introduction to modern liquid chromatography*, 3rd ed.; John Wiley & Sons: Hoboken, 2010.
(157) Song, P.; Mabrouk, O. S.; Hershey, N. D.; Kennedy, R. T. *Anal. Chem.* **2012**, *84*, 412-419.
(158) Spagou, K.; Tsoukali, H.; Raikos, N.; Gika, H.; Wilson, I. D.; Theodoridis, G. *J. Sep. Sci.* **2010**, *33*, 716-727.
(159) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. *J. Proteome Res.* **2009**, *8*, 3737-3745.
(160) Stein, S. *Anal. Chem.* **2012**, *84*, 7274-7282.
(161) Sterling, H. J.; Prell, J. S.; Cassou, C. A.; Williams, E. R. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1178-1186.
(162) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9*, 504.
(163) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826-828.
(164) Tayyari, F.; Gowda, G. A. N.; Gu, H. W.; Raftery, D. *Anal. Chem.* **2013**, *85*, 8715-8721.
(165) Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M. *Mol. Cell. Proteomics* **2011**, *10*, 1-9.
(166) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. *Anal. Chem.* **2003**, *75*, 1895-1904.
(167) Trivedi, D. K.; Iles, R. K. *Biomed. Chromatogr.* **2014**, *28*, 1491-1501.
(168) Trygg, J.; Holmes, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469-479.

(169) Tulipani, S.; Mora-Cubillos, X.; Jauregui, O.; Llorach, R.; Garcia-Fuentes, E.; Tinahones, F. J.; Andres-Lacueva, C. *Anal. Chem.* **2015**, *87*, 2639-2647.
(170) Uehara, T.; Yokoi, A.; Aoshima, K.; Tanaka, S.; Kadowaki, T.; Tanaka, M.; Oda, Y. *Anal. Chem.* **2009**, *81*, 3836-3842.
(171) Van Berkel, G. J.; Zhou, F. *Anal. Chem.* **1995**, *67*, 3958-3964.
(172) Venta, R. *Clin. Chem.* **2001**, *47*, 575-583.
(173) Vorkas, P. A.; Isaac, G.; Anwar, M. A.; Davies, A. H.; Want, E. J.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2015**, *87*, 4184-4193.
(174) Vuckovic, D. *Anal. Bioanal. Chem.* **2012**, *403*, 1523-1548.
(175) Wang, G.; Zhou, Y.; Huang, F. J.; Tang, H. D.; Xu, X. H.; Liu, J. J.; Wang, Y.; Deng, Y. L.; Ren, R. J.; Xu, W.; Ma, J. F.; Zhang, Y. N.; Zhao, A. H.; Chen, S. D.; Jia, W. *J. Proteome Res.* **2014**, *13*, 2649-2658.
(176) Wang, N.; Li, L. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1573-1587.
(177) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001**, *19*, 242-247.
(178) Wenner, P. G.; Bell, R. J.; van Amerom, F. H. W.; Toler, S. K.; Edkins, J. E.; Hall, M. L.; Koehn, K.; Short, R. T.; Byrne, R. H. *Trends Anal. Chem.* **2004**, *23*, 288-295.
(179) Wikoff, W. R.; Pendyala, G.; Siuzdak, G.; Fox, H. S. *J. Clin. Invest.* **2008**, *118*, 2661-2669.
(180) Winger, B. E.; Light-Wahl, K. J.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 624-630.
(181) Wishart, D. S. *Trends Anal. Chem.* **2008**, *27*, 228-237.
(182) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801-807.
(183) Wiśniewski, J. R.; Friedrich, A.; Keller, T.; Mann, M.; Koepsell, H. *J. Proteome Res.* **2015**, *14*, 353-365.
(184) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148.
(185) Wu, F.; Wang, P.; Zhang, J.; Young, L. C.; Lai, R.; Li, L. *Mol. Cell. Proteomics* **2010**, *9*, 1616-1632.
(186) Wu, W. W.; Wang, G.; Baek, S. J.; Shen, R.-F. *J. Proteome Res.* **2006**, *5*, 651-658.
(187) Wu, Y.; Li, L. *Anal. Chem.* **2014**, *86*, 9428-9433.
(188) Wu, Y. M.; Li, L. *Anal. Chem.* **2013**, *85*, 5755-5763.
(189) Xu, M.; Li, L. *J. Proteome Res.* **2011**, *10*, 3632-3641.
(190) Xu, Y.-F.; Lu, W.; Rabinowitz, J. D. *Anal. Chem.* **2015**, *87*, 2273–2281.
(191) Yamashita, M.; Fenn, J. B. *The Journal of Physical Chemistry* **1984**, *88*, 4451-4459.
(192) Yanes, O.; Clark, J.; Wong, D. M.; Patti, G. J.; Sánchez-Ruiz, A.; Benton, H. P.; Trauger, S. A.; Desponts, C.; Ding, S.; Siuzdak, G. *Nat. Chem. Biol.* **2010**, *6*, 411-417.
(193) Yang, P. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1273-1280.
(194) Yin, P. Y.; Xu, G. W. *J. Chromatogr. A* **2015**, *1374*, 1-13.
(195) Yuan, W.; Anderson, K. W.; Li, S. W.; Edwards, J. L. *Anal. Chem.* **2012**, *84*, 2892-2899.
(196) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357-1364.
(197) Zhou, R.; Huan, T.; Li, L. *Anal. Chim. Acta* **2015**, *881*, 107-116.
(198) Zhou, R.; Li, L. *J Proteomics* **2015**, *118*, 130-139.

(199) Zhou, R.; Tseng, C.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86*, 4675-4679.
(200) Zhou, R.; Tseng, C. L.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86*, 4675-4679.
(201) Zhu, G. J.; Sun, L. L.; Yan, X. J.; Dovichi, N. J. *Anal. Chem.* **2013**, *85*, 2569-2573.
(202) Zhu, Z.-J.; Schultz, A. W.; Wang, J.; Johnson, C. H.; Yannone, S. M.; Patti, G. J.; Siuzdak, G. *Nat. Protoc.* **2013**, *8*, 451-460.
(203) Zubarev, R. a.; Makarov, A. *Anal. Chem.* **2013**, *85*, 5288-5296.
(204) Zuniga, A.; Li, L. *Anal. Chim. Acta* **2011**, *689*, 77-84.

Appendix

Chapter 5 Construction of a Metabolite Retention Time Library

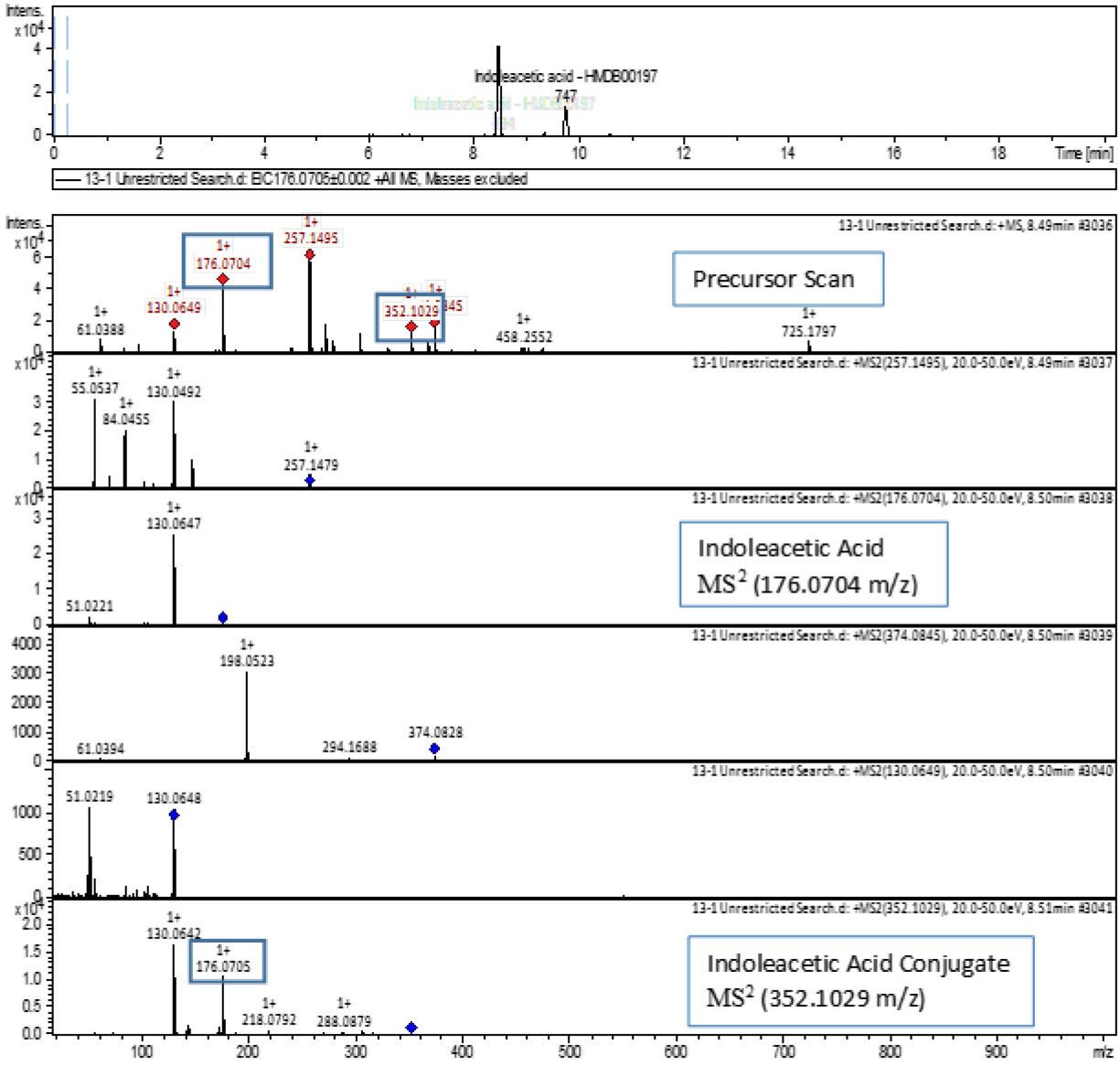


Figure A5.1 In-source fragmentation of indoleacetic acid conjugate, and its fragmentation spectra. In the MS^2 spectrum of the indoleacetic acid conjugate, the exact mass of indoleacetic acid can be seen. This is evidence that the 176 m/z, seen in the precursor scan, is an in-source fragmentation of the conjugate.

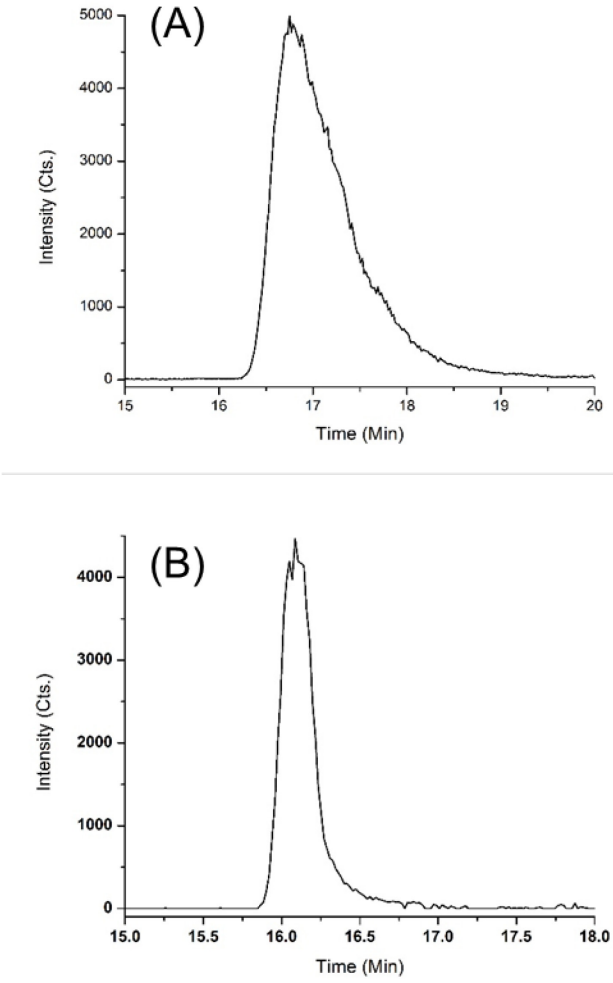


Figure A6.1 Chromatographic peaks of a dansyl analyte in labeled urine obtained by using (A) Waters nanoACQUITY column and (B) Thermo Acclaim PepMap column.

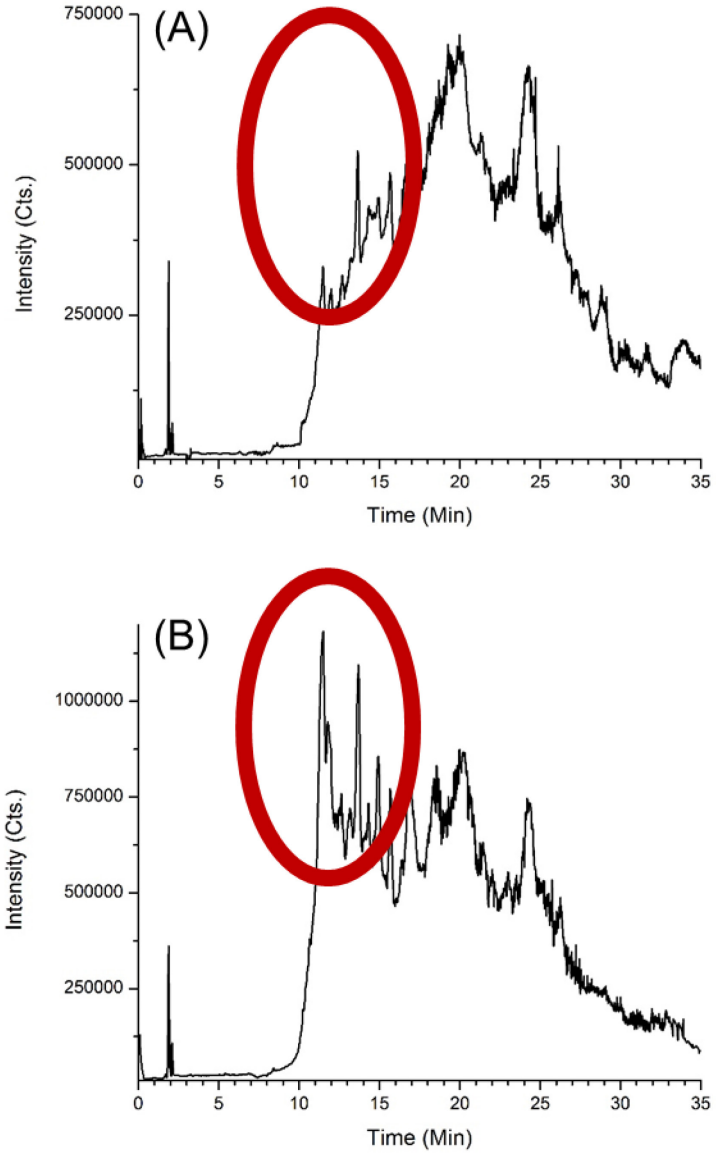
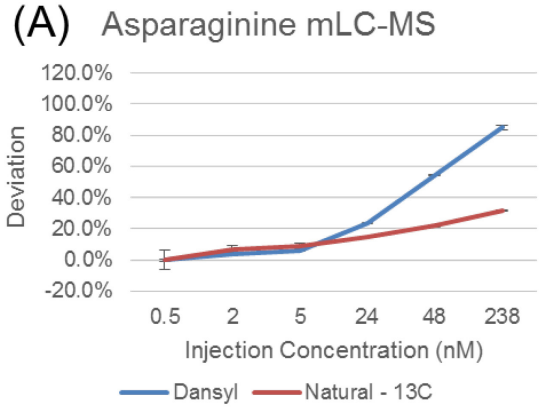
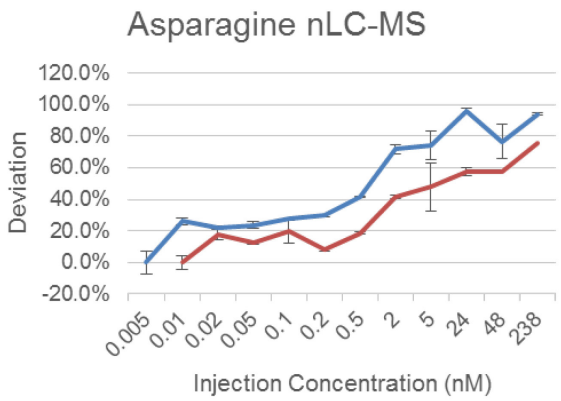


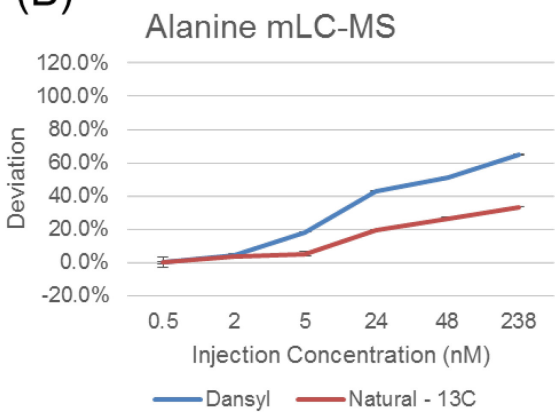
Figure A6.2 TIC comparison of (A) 1:1 ACN:H_2O diluent and (B) 1:9 ACN:H_2O diluent. Large portion of the early eluting peaks are reduced in intensity when using the 1:1 ACN:H_2O diluent.

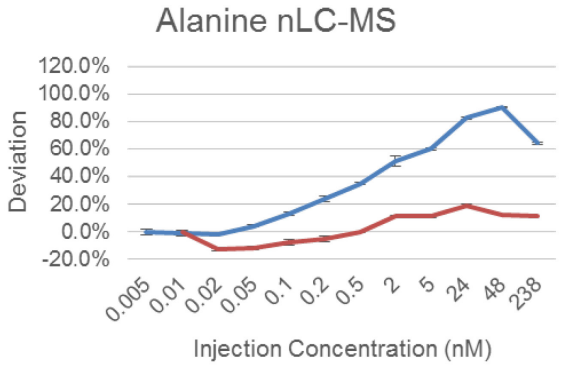
(A)


Asparaginine mLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.5
2
5
24
48
238
Injection Concentration (nM)
Dansyl
Natural - 13C

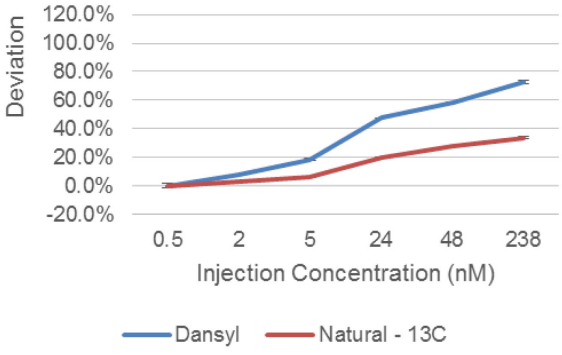

Asparagine nLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.005
0.01
0.02
0.05
0.1
0.2
0.5
2
5
24
48
238
Injection Concentration (nM)

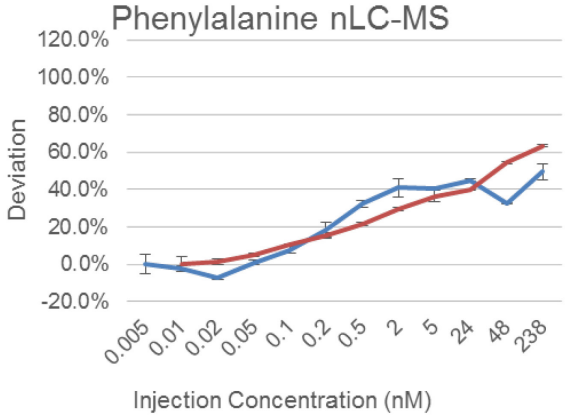
(B)


Alanine mLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.5
2
5
24
48
238
Injection Concentration (nM)
Dansyl
Natural - 13C


Alanine nLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.005
0.01
0.02
0.05
0.1
0.2
0.5
2
5
24
48
238
Injection Concentration (nM)

(C)


Phenylalanine mLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.5
2
5
24
48
238
Injection Concentration (nM)
Dansyl
Natural - 13C


Phenylalanine nLC-MS
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
Deviation
0.005
0.01
0.02
0.05
0.1
0.2
0.5
2
5
24
48
238
Injection Concentration (nM)

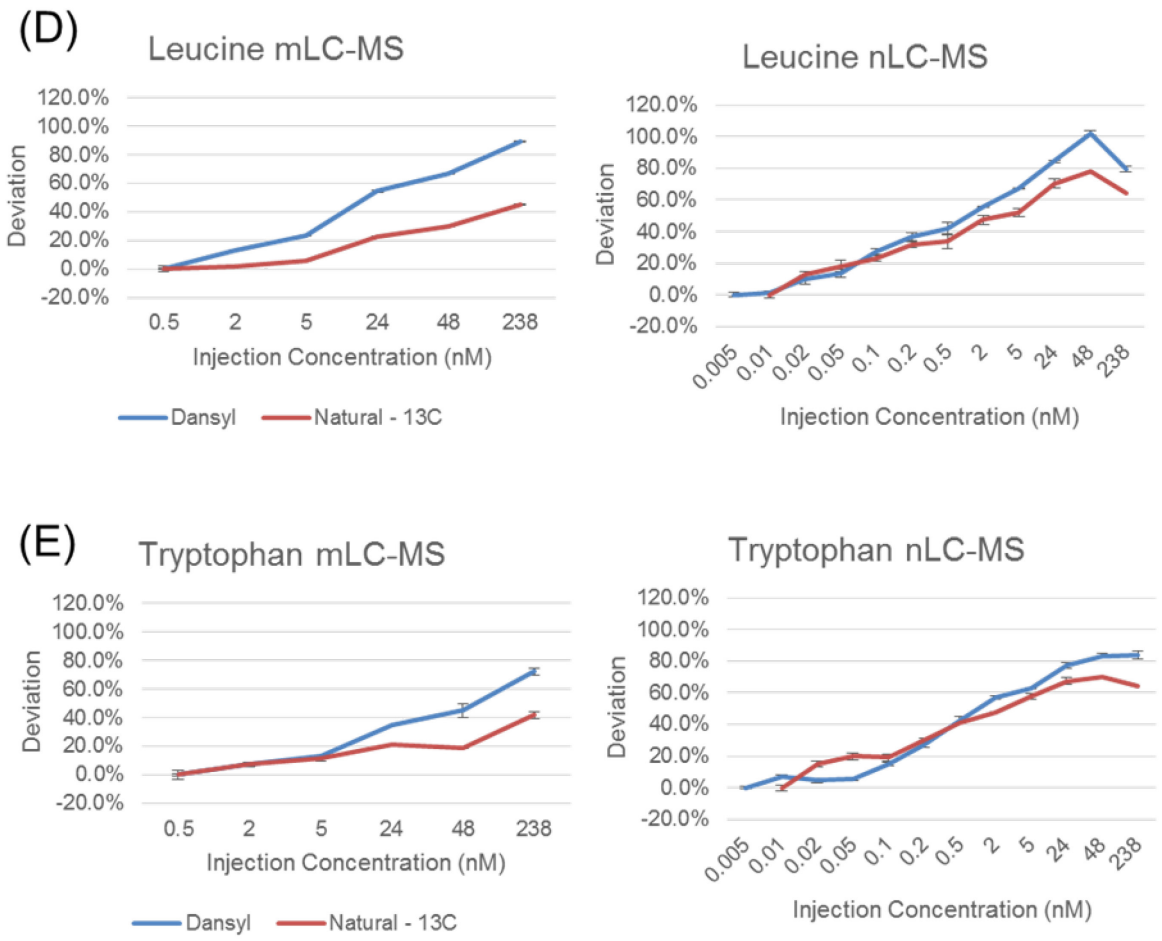


Figure A6.3 Effect of detector saturation on the calculated peak pair ratio in mLC-MS and nLC-MS. Derivation from the expected 1:2 ratio is plotted as a function of the solution concentration of 1:2 mixture of ^{12}C-dansyl amino acid and ^{13}C-dansyl amino acid.

Table A6.1 List of relative standard deviations of retention times of dansylated amino acids measured by nLC-MS and mLC-MS (n=3).

	nLC		mLC	
	RSD (%)	σ (s)	RSD (%)	σ (s)
Asparagine	1.07%	7.30	0.00%	0.00
Glutamic Acid	0.84%	6.61	0.18%	0.69
Glycine	0.40%	3.52	0.00%	0.00
Alanine	0.65%	6.26	0.00%	0.00
Proline	0.34%	3.99	0.10%	0.69
Tryptophan	0.21%	2.71	0.09%	0.69
Phenylalanine	0.16%	2.16	0.00%	0.00
Leucine	0.13%	1.83	0.11%	0.92

Table A6.2 List of relative standard deviations of peak areas of dansylated amino acids measured by nLC-MS and mLC-MS (n=3).

	nLC			mLC		
	Average Area	SD	RSD	Average Area	SD	RSD
Asparagine	2238.0	153.5	6.9%	49.3	1.3	2.7%
Glutamic Acid	6353.0	98.6	1.6%	121.7	1.4	1.2%
Glycine	6553.0	138.4	2.1%	173.8	9.4	5.4%
Alanine	5684.3	129.8	2.3%	126.7	5.6	4.4%
Proline	5553.6	166.8	3.0%	156.2	1.4	0.9%
Tryptophan	3560.9	99.6	2.8%	49.8	0.7	1.4%
Phenylalanine	4483.3	268.8	6.0%	124.0	3.3	2.7%
Leucine	5514.4	209.5	3.8%	184.0	14.4	7.8%
Average	4992.6	158.1	3.5%	123.2	4.7	3.3%