Development of Data Processing Methods in Chemical Isotope Labeling Liquid Chromatography-Mass Spectrometry-Based Metabolomics

by

Yunong Li

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

Doctor of Philosophy

Department of Chemistry
University of Alberta

Abstract

Metabolomics is an active research field on the methods development for the analysis of small metabolites in biological systems. It provides powerful approaches that allow us to examine the variations in the metabolic profiles and is capable of detecting complex biological changes using statistical pattern recognition methods. In metabolomics analysis, large amounts of data are produced routinely in order to characterize a sample consisting of hundreds to thousands of metabolites. The conclusions drawn from the metabolomics data rely on the coverage of the detection method, the accuracy of the metabolite concentrations, and the completeness of data to include all the metabolite signals. Therefore, a number of challenges are associated with the data processing methods specific to each experimental platform.

The LC-MS technique has been used widely in the application of metabolomics due to its high sensitivity and high throughput. Traditional LC-MS platforms are limited by the coverage of the detection and the less reproducible quantification results. Thus, the chemical isotope labeling LC-MS method was developed in our group for an improvement of the metabolite separation and a higher detection sensitivity of a broad range of metabolites in a biological sample. In the labeling LC-MS method, each labeled metabolite will generate a peak pair signal in their mass spectra, with the light peak from the individual sample and the heavy peak from the pooled sample. Accordingly, a customized data processing method is required in dealing with the data generated by different chemical isotope labeling LC-MS experiments. My research focused on the development of data processing methods to address the challenges from the growing data processing tasks in the chemical isotope labeling LC-MS. I developed an integrated data

processing workflow that checked the LC-MS raw data in terms of the mass accuracy and retention time reproducibility (Chapters 2 and 3), aligned the peak pair data from individual files, and removed the false peak pairs and redundant peak pairs to improve the confidence of each peak pair in representing a true labeled metabolite (Chapter 4). A missing ratio imputation method was developed to fill all missing ratios and generate a complete metabolite-intensity table for the statistical analyses (Chapter 4). In the peak pair ratio calculations, I developed a data processing method that accounts for natural isotope contributions in the peak intensity of the $^{13}C_2$-labeled peak and improved the quantification accuracy (Chapter 5). An intensity-dependent mass tolerance method was developed to assist the mass-based database search (Chapter 6). All the processing methods were integrated in a program with a graphical user interface that has been implemented in the lab for routine data processing tasks. I used an application of a wine metabolomics study to demonstrate the data processing workflow (Chapter 7). The integrated data processing program can be applied in different chemical isotope labeling LC-MS experiments to facilitate the qualitative and quantitative analyses of different metabolomes.

Preface

A version of Chapter 5 was published as "Yunong Li and Liang Li, Improving Accuracy of Peak-pair Intensity Ratio Measurement in Differential Chemical Isotope Labeling LC–MS for Quantitative Metabolomics, International Journal of Mass Spectrometry, 2018, 434, 202–208." I was responsible for the design of the experiment, data collection, algorithm development, as well as manuscript preparation. Dr. Liang Li supervised the project and edited the manuscript.

Acknowledgments

I received much help in the past five years during my PhD program at University of Alberta. This thesis work would not be possible without the support of many people, to whom I would like to express my thanks here.

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Liang Li, for his continuous support through my PhD study with his invaluable guidance, advice, and encouragement. It has been a great learning experience in Dr. Li's lab with the opportunity to use many state-of-the-art instruments. The research experience I gained there will definitely be an asset in my future career.

Secondly, I would like to extend my appreciation to the members of my supervisory committee, Dr. Michael Serpe and Dr. Guohui Lin, for their comments and suggestions on my annual progress reports and during the candidacy exam. I want to thank to Dr. Monica Li and Dr. Anna Jordan for editing my thesis. I appreciate greatly the help from Anna for her kind help in my thesis writing.

Furthermore, I would like to thank all the Li group members. Importantly, I want to thank Dr. Tao Huan, Dr. Zhendong, Li, Dr. Difei Sun, Dr. Wei Han, Dr. Chiao-Li Tseng, Dr. Yiman Wu, and Dr. Ruokun Zhou for their training, advice, and support at the beginning of my research work. I am especially thankful to Dr. Tao Huan who helped me greatly in learning the skills in program development. I also want to thank to the people that I have worked with: Dr. Helen Wang, Xiaohang Wang, Xingyun Gu, Minglei Zhu, Erik Gomez Cardona, and Kamran Mammadli. I want to thank Xiaohang Wang and Xingyun Gu for providing me data for testing

my programs. I have also benefited from the knowledge of other group members, including but not limit to Kevin Hooton, Jaspaul Tatlay, Dorothea Mung, Shuang Zhao, Xian Luo, Hao Li, Chan Wan, Adriana Zardini Buzatto, and Barinder Bajwa.

In addition, I would like to thank the support from all the staffs in the Department of Chemistry, including the electronic shop, the machine shop, the glass shop, and the IT service team. I'm especially thankful to Dr. Randy Whittal, Dr. Jing Zheng, and Mr. Bela Reiz of the Mass Spectrometry Facility for their professional support.

I want to thank my fellow researchers, Dr. Deying Chen from the First Affiliated Hospital of College of Medicine at Zhejiang University, Dr. Weifeng Sheng from Zhejiang Academy of Agricultural Sciences, and Dr. Chia-Wei Hsu from Chang Gung University. They provided me with many suggestions on my programs, and we have established great friendships through the years.

I want to thank all my friends who are always ready to help in my pursuit of the PhD. Especially, I want to thank Nikki Gao for being such a genuine friend and supporting me in my ups and downs.

Finally and most importantly, I want to thank my family who has been my greatest support of all time. Thank you for giving me the courage to face all the challenges in my life and believing in me for every choice I made. I would never be what I am without you, and I hope I have made you proud.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACN	Acetonitrile
APCI	Atmospheric Pressure Chemical Ionization
API	Atmospheric Pressure Ionization
APPI	Atmospheric Pressure Photoionization Ionization
BPC	Base Peak Chromatogram
BPCA	Bayesian Principal Component Analysis
CE	Capillary Electrophoresis
Da	Dalton
DC	Direct Current
DNA	Deoxyribonucleic acid
DmPA	p-Dimethylaminophenacyl
Dns	Dansyl
EI	Electron Impact Ionization
ESI	Electrospray Ionization
FC	Fold Change
FT-ICR-MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
GC	Gas Chromatography
GC-MS	Gas Chromatography Mass Spectrometry
HMDB	Human Metabolome Database
HILIC	Hydrophilic Interaction Liquid Chromatography
HPLC	High Performance Liquid Chromatography
ICR	Ion Cyclotron Resonance
KNN	k-nearest Neighbors
LC	Liquid Chromatography
LC-MS	Liquid Chromatography Mass Spectrometry
LC-UV	Liquid Chromatography Ultraviolet
MALDI	Matrix-assisted Laser Desorption Ionization
MeOH	Methanol
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
m/z	Mass to Charge
nm	Nano Meter
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDA	Photo Diode Array
PLS-DA	Partial Least Square Discriminant Analysis
PPCA	probabilistic Principal Component Analysis
ppm	part(s) per million
QC	Quality Control
QTOF-MS	Quadrupole Time-Of-Flight Mass Spectrometry
RT	Retention Time

RPLC	Reversed Phase Liquid Chromatography
RSD	Relative Standard Derivation
SCI	Spinal Cord Injury
SDS	Sodium Dodecyl Sulfate
SNR	Signal to Noise Ratio
SPE	Solid Phase Extraction
SVD	Singular-value Decomposition
TOF	Time-Of-Flight
UPLC	Ultra-Performance Liquid Chromatography
UV	Ultra-violet
μM	Micro Molar

Chapter 1 Introduction

1.1 Overview of Metabolomics

Metabolites are defined as the small chemical products in cellular metabolism processes with molecular weight less than 1500 Da in contrast to the relatively large biological polymers, such as DNA, RNA, proteins, and polysaccharides.[1] Similar to the genome, proteome and transcriptome, the complete set of metabolites found in one biological sample is called a metabolome. Metabolomic is the omics science that studies the metabolites within one metabolome, including their identity, quantity, and interactions. Metabolites can include a range of endogenous and exogenous compounds, such as amino acids, small polypeptides, nucleic acids, oligosaccharides, lipids, and many hormones.[2] These small metabolites are related directly to the phenotypic traits of an organism, such as color, shape, pattern, and other physical characteristics.[3] Many functions in a biological system are built on the metabolic processes involving small molecules, including signaling among cells, response to environmental stimuli, and energy transformations.[4]

The collection of metabolites represents both the endpoints of the expressions of the genome and the interaction between one subject and its environment. As an emerging field in omics, metabolomics answers the questions in life science by looking at the pathway networks that involve thousands of metabolites. In contrast to genomics, which can be used to indicate what might potentially happen in an organism, metabolomics data can indicate the current state in a biological process. Specifically, the concentration changes of metabolites can be reflective to the biological processes, such as the development of certain diseases. Therefore, metabolomics

can be complementary to the data generated by genomics, transcriptomics, and proteomics as it is the "omics" approach closest to the phenotype.[5]

In the central molecular biological dogma, genetic information flows directionally from DNA to RNA followed by the translation to proteins. The proteins, such as enzymes, can affect the concentrations of their substrates and products.[6,7] Figure 1.1 shows the different fields of omics studies in the scope of the central dogma of molecular biology. From a biochemical viewpoint, DNA, RNA, and proteins can be described using their smallest building blocks. For example, DNA and RNA are made of four nucleotide monomers, and protein can break down into 20 amino acids. Metabolites, on the other hand, contain a much broader scope of chemicals, including those naturally produced within an organism as well as other exogenous chemicals from the environment. A diverse chemical property and wide concentration distribution make it challenging for detecting these metabolites with high coverage. Thus, research in metabolomics usually focuses on different groups of metabolites for different applications.

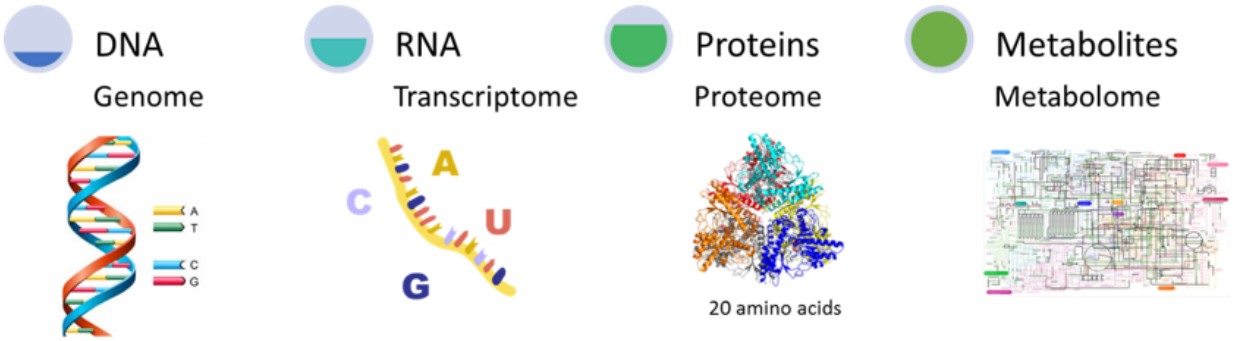


Figure 1.1 Omics studies in the central dogma of molecular biology

Metabolomics can be a powerful tool for characterization of complex phenotypes and the determination of biomarkers for the diagnostics to certain diseases. A biomarker can be defined broadly as a characteristic that can be measured to indicate the physiological state or pathogenic processes for the purpose of differentiating patients with a specific disease from healthy people.[8] This includes a physical readout, such as blood pressure, a molecule concentration level, and even a genome sequencing. The study of disease biomarkers has improved the decision-making process in drug development and disease diagnostics greatly. Although many approved biomarker assays are based on protein or RNA, small molecule biomarkers have attracted increasing interest as they are more sensitive to the biological changes. Small molecular diagnostic biomarkers have been used for risk assessment of certain diseases like cancer before the symptoms become noticeable, enabling an early diagnosis. A prognostic biomarker can indicate the progression and likely outcomes of a disease to assist treatment specifically to each patient. Moreover, a biomarker can be related directly or indirectly to the disease development during its onset or progression. For this reason, biomarkers using small metabolites are studied widely and applied for the diagnosis, assessment of patients, and therapy determination in clinical practice.[9-11]

The understanding of the disease mechanism through studies of metabolic changes also provides a new method for drug development. For years, drug discovery has been focused on the genetic origin through genome sequencing, transcripts profiling.[12] However, the number of targeted disease genes is far fewer than the number of diseases, which is limiting the drug discovery pipeline. It was found that only certain diseases have a clear genetic basis as many of them are the results of exposures to the environment.[13] A study collected the mortality data in the United States over 16 years and showed that the leading contributors to death are related to

tobacco, alcohol, illicit drugs, microbial agents, toxic agents etc.[14] For example, tobacco accounts for over 400,000 deaths each year among Americans. Tobacco exposure contributes substantially to cancer disease in lung, oral cavity, pancreas, kidney, and other organs. A high risk of cardiovascular disease, such as heart disease, stroke, and high blood pressure, was observed in frequent smokers. Metabolites from these exogenous origins can have a great impact on human health.

Metabolomics research provides a new angle in drug development by studying the alteration of metabolic processes. For example, one study has shown several unexpected chemical species that were altered in chronic and complex diseases, including atherosclerosis, cancer, and diabetes.[15] Both exogenous and endogenous metabolites have a far more significant role in cellular signaling, disease development, and physiological homeostasis. For example, cancers are well known as a genetic disease with gene mutations. However, increasing evidence has shown that a metabolic disorder is associated closely with cancer development. A study reported an altered metabolic pattern in tumor cells for aerobic glycolysis and glutaminolysis.[16] Alterations in theses metabolic processes are regulated by oncogenes and tumor suppressor genes. In tumor cells, these metabolic disorders were found to have an elevated concentration of substrates for cell growth. Blocking or restoring some of these key metabolic pathways involved in tumor development potentially can provide a new approach to cancer treatments.

In addition to drug discovery, the development of metabolomics has led to a new research area of precision medicine. This concept takes into account the individual variability and aims at providing a personalized treatment plan targeted to one's own genetic, phenotypic, and psychosocial characteristics.[17,18] With the developments of large-scale databases, powerful analytical platforms, and computational tools, a metabolomics experiment is able to collect a

comprehensive personal metabolic signature to allow sub-classification of diseases and reveal biomarkers for drug response. Certain biomarkers could be used to determine the efficacy of the medication in the treatment of a specific patient. The dose of the drug can be optimized by the biomarkers involved in the metabolism of the drug,[19] and the optimal therapy can be determined further based on the biomarkers profile of an individual. One of the applications of precision medicine is in the cancer therapy. Tumor genomic profiling has been used to classify tumor types, perform assessments, and determine therapeutic decisions.[20] However, the differences in cancer genomes can mask the underlying causes. As mentioned earlier, many cancers can have a number of unique pathway alterations.[16,21] With metabolomics techniques, such as metabolite imaging, one can achieve a better classification of the tumors and enable an informed adjustment to the cancer therapy.

1.2 Strategies of Metabolomics Study

1.2.1 Targeted and untargeted metabolomics

A metabolome may contain a vast variety of metabolites with diverse chemical structures and a wide concentration distribution. There is no analytical method that is capable of the measurement of all metabolites of interest.[22] In a targeted method, the study would focus on a number of predefined metabolites as a subset of the metabolome (most typically, dozens to hundreds of known compounds). This usually requires an a priori knowledge of the sample and the chemical contents. For example, in the study of genetic alteration, the analysis can be constrained to the substrate and direct products of the corresponding encoded protein.[23]

The benefit of knowing the list of metabolites of interest is that the sample preparation can be optimized for those relatively low abundant metabolites, reducing the dominance of metabolite of high concentration in the analyses. Moreover, data processing steps can focus on the selected metabolite signal and filter all other noise signals from entering subsequent analyses.[24] With a targeted metabolomics approach, one can achieve absolute quantification using internal standards (e.g. isotope analogues in mass spectrometry) in the sample preparation for each targeted metabolite. This approach uses the known information in the metabolic pathways and the final metabolite change can be mapped directly onto the biological knowledge and integrated with other omics data.

For the instrumental analysis of a list of targeted compound, predefined metabolite signals are edited in the method so as to separate the metabolites of interest from other signals. For example, in a LC-MS experiment, the retention time of each target analyte is determined first and the m/z of each precursor ion is added into the retention time segments. During the data collection, each targeted analyte would be selected in the quadrupole and further detected in the mass analyzer.

Untargeted metabolomics is used for comprehensive metabolome analysis, that is, the analysis of all the detectable signals in a sample, giving the ability to detect many unknown compounds. This untargeted approach offers the opportunity for the discovery of novel targets for further investigation. Unlike the targeted approach, untargeted analysis requires a sample preparation step to allow a broad range of metabolites for high metabolome coverage. Consequently, the data collected from an untargeted analysis can be large, and it will require an efficient data processing method for metabolite information extraction and identification. To

maximize the detection coverage, one would consider the sensitivity of the relatively low abundant species.[24]

The untargeted LC-MS metabolomics platform involves multiple steps. The first step is to acquire the complete mass spectrometry data in each individual sample. Metabolite information will be extracted from the raw data, and these data will be analyzed using bioinformatics software that performs quantitative analyses to recognize the significant metabolites feature that cause the biological groups separation. Based on the retention time and accuracy mass of each metabolite, one will identify as many as possible metabolites in the data using one or multiple standards databases. To confirm the identifications, tandem mass spectrometry (MS/MS) can be used to collect MS/MS spectra to provide fragments information. A targeted MS/MS analysis is performed typically on one of the pooled samples that usually contain most metabolites present in all samples. Then, the fragmentation pattern of the MS/MS data is compared to the MS/MS database to confirm the identification. The limitation of using MS/MS data matching is that there is a limited number of MS/MS spectra available in a database and not all metabolites in a sample can have a high enough concentration to produce sufficient fragment ions. Therefore, retention time can be used as another identification parameter. The commercial standards will have to be analyzed using the same LC instrument settings and gradient to allow for a comparison of the retention time. Accurate mass, MS/MS data and retention time often work collaboratively for the identification of the unknowns.

Targeted and untargeted approaches are complementary in metabolomics. Most often, untargeted metabolomics is used first for the comprehensive analysis of all detectable analytes in a sample. This list consists of many unknowns, providing the opportunity for discovering novel metabolites and pathways. After significant metabolites are extracted from the global analysis,

targeted analysis can be conducted on these specific metabolites for their identification, absolute quantifications, and pathway analysis. Biological meaning can be explored further for the identified metabolites using their measured characteristics in different biological samples.

1.2.2 Metabolic profiling and fingerprinting

Metabolomic profiling and fingerprinting are both untargeted approach. Metabolic profiling uses a tool to analyze a group of metabolites in a specific metabolic pathway or in a class of metabolites. For example, in a dansylation labeling LC-MS experiment, the amine and phenol containing metabolites are labeled selectively during sample preparation for instrumental analysis. After obtaining the amine and phenol profile in each sample, one can classify different biological groups based further on the concentrations of the labeled metabolites. The differentially expressed metabolites can provide more information for the study of potential alterations in the related metabolic pathways. Compared to the study of a few marker metabolites, the profiling data gives a more detailed description of any metabolic changes.

In metabolic fingerprinting, the metabolites patterns are obtained without identification of each metabolite to enable a quick comparison of metabolic changes in response to stimulants or events such as a disease, exposure of a toxin, environmental stress, or genetic change. Metabolic fingerprinting requires a rapid, high-throughput global analysis technique, such FT-IR, NMR, and MS, in order to be applied to a wide range of metabolites.[25-27] As many diseases are the result of metabolic disorder that involves a stream of metabolites in the related pathway, it can be hard to study the mechanism of a disease development with a few targeted compounds. Often in

disease sample analysis, the real differentiating variables can be buried among many other common features that show no difference among biological groups. To enable a global view for different biological samples, metabolic fingerprinting measures the chemical patterns of metabolites in the whole sample as a way to discriminate samples of different biological origins. Metabolic fingerprinting, without metabolite identification, can be a diagnostic tool by evaluating the metabolic pattern of a sample from a patient in comparison to a healthy or diseased sample. The performance of a treatment strategy can be monitored by looking at the sample fingerprints after treatment and seeing if they fall in the same cluster as a healthy sample.

Although metabolic fingerprinting offers a quick tool for sample classification, without identification of key metabolites that cause the clustering of different experimental groups, one cannot reveal the underlying mechanism of the biological processes. Qualitative and quantitative methods should be developed for investigating the specific metabolites in order to tie metabolic fingerprinting and profiling together.

1.2.3 Sub-metabolome analysis

Advances in the sensitivity, dynamic range, and data collection speed in mass spectrometry instrumentation have made it possible for the detection of thousands of metabolites in a biological sample. However, the actual number of metabolites can be much larger than the detectability of one analytical method. A subclass of the whole metabolome usually is selected as the target when developing an analysis method, since certain groups of compound can be the most relevant to one study. With this strategy, metabolomics is finding its applications in various fields, including lipidomics, exposomics.[5]

Lipids are a class of compounds that are present in tissue, cell, and small hormones. They perform a variety of functions in a biological system. Lipids have the unique chemical properties of being relatively non-polar and highly soluble in organic solvents. Lipidomics is the branch of metabolomics focusing on lipid compounds and aims to provide detailed and quantitative information on the construction of cellular lipidome, lipid metabolism, lipid–lipid interaction and lipid protein interaction. Due to these unique chemical properties, sample analysis often begins with extraction of lipids from a tissue or cell with an organic solvent. Then, the complex lipid mixture is analyzed by one or multiple analytical techniques to obtain the lipid profile. The study of lipid metabolism and lipid–protein interactions has been shown to have the potential to reveal the underlying mechanism of many neurological disorder diseases, such Alzheimer's disease.[28]

A metabolome consists of endogenous compounds produced within an organism and a vast variety of exogenous metabolites from the environment. Many diseases can find their causes from different environmental factors. For example, a frequent smoker can have a much higher risk of having lung cancer compared to a non-smoker. All the non-genetic factors that contribute to the development of a disease are considered to be environmental. These include natural or un-natural chemicals, various drugs, infectious agents and stress-related factors. A relatively new research field called "exposomics" attempts to analyze all the environmental stresses in order to obtain the knowledge and understanding in different non-genetic diseases.[29] It studies the effect of exogenous compounds, including their metabolism in different organisms and the potential disturbance on normal pathways. For example, chemical contaminants from air, water, soil, and food were found to be related to inflammation reaction, elevated reactive oxygen species, methylation, and gene expression changes. The study of these groups of exogenous compounds

and their metabolism potentially can reveal the mechanism of many chronic diseases, including cancer and diabetes.[29]

Chemical isotope labeling LC-MS is a strategy for sub-metabolome analysis that is based on the chemical structures of metabolites. The method employs a chemical derivatization method in the sample preparation step that targets a group of analytes with a common functional group; typical examples are the amine and phenol metabolites analysis with dansylation labeling and the acid-containing metabolites analysis with DmPA labeling.[30,31] Compared to protein analysis, in which can break down the structure to the sequence of 20 amino acids, no general analysis method can be applied to all types of metabolites. The divide and conquer strategy in metabolomics focuses on each sub-metabolome and method optimization in each labeling platform to achieve maximum detection. By combining the results of different sub-metabolomes, one can achieve a high coverage of the whole metabolome and obtain the global map of the metabolic pathway networks.

1.3 Analytical Technologies in Metabolomics

1.3.1 Instrumentation

A metabolomics experiment aims at the simultaneous quantification of multiple metabolites using sensitive and specific analytical techniques, such as gas or liquid chromatography combined to mass spectrometry.[32] Unlike genomics, transcriptomics, or proteomics, which have well-developed experimental methods using one or a few instruments, metabolomics usually requires multiple instruments and experimental techniques in analyzing metabolites with diverse chemical properties.[33]

As one of the most established spectrometric techniques, nuclear magnetic resonance spectroscopy (NMR) can identify and quantify a broad range of compounds. The use of NMR in studying metabolism and metabolic processes can be traced back to 1973 when A. L. Burlingame and his colleagues used stable isotope deuterium and carbon-13 to study the ethanol metabolism in rats.[34] NMR has been exceptional in identifying and quantifying most organic chemicals in complex metabolite mixtures, along with a number of unique advantages. In particular, it is non-destructive, highly automated, needs little or no sample separation, provides rich structural information for the identification of novel compounds, and requires no chemical derivatization.[35] It is fast and usually takes around 2–3 min for the analysis of one sample. NMR has been widely used for metabolic profiling, metabolite fingerprinting, and metabolic flux analysis.[36] On the other hand, the major disadvantage of NMR is its relatively low sensitivity (limit of detection at 5 μM) and thus requires a large sample amount.[37] An NMR instrument also has a high start-up cost and a large instrument footprint. With technological advances using higher field magnets and cryogenically cooled probes, the sensitivity of the technique has increased for detecting sub micro-molar analytes with less sample volume required. However, sensitivity remains the number one limitation of the NMR technique for the analysis of large numbers of low-abundance metabolites.

Mass spectrometry (MS) is gaining interest in metabolomics for its superior sensitivity and wide dynamic range suitable for the metabolome profiling analysis, in which the concentration of metabolites can differ by a few magnitudes.[38-40] It can be used as a stand-alone platform to analyze a biological sample by a simple direct injection as a rapid technique for metabolic fingerprinting. However, direct injection of a complex biological sample can introduce strong ion suppression from predominant ions and results in a low ionization efficiency for most

low abundant analytes. Also, usually it requires a sample pre-treatment to remove high concentrations of salts from the sample, which can be harmful to the MS. To avoid these problems, advanced and high throughput separation techniques have been coupled to MS to decrease the complexity of a biological sample so as to increase the number of detected metabolites.

Gas chromatography (GC) is a relatively mature and robust separation technology for volatile and thermally stable compounds. GC-MS has been used extensively in metabolomics for producing efficient and reproducible analysis of a variety of organic molecules and some inorganic compounds.[41] One advantage of GC-MS is that the data generated from different instruments across different labs are consistent with the use of the indexed retention time and the highly reproducible electron impact (EI) ionization technique.[5,42] Most spectral features generated from GC-MS are identifiable with a number of available software packs and GC databases for metabolite identification. For the separation of non-volatile compounds on the GC column, GC-MS requires a derivatization reaction to create volatile compounds. A derivatization method would increase analyte volatility, increase the detector sensitivity, and improve chromatographic behavior of an analyte by decreasing its polarity. The applicability of GC-MS has expanded greatly due to the advances in derivatization techniques, including but not limited to silylation, alkylation, esterification, acylation, and other condensation reactions.[43]

High performance liquid chromatography (HPLC) is suited better for the analysis of a wide range of chemicals, including labile species, nonvolatile chemicals, polar and nonpolar compounds in their native form.[44] In reverse phase LC, the stationary phase is made of non-polar chemical groups, such as C8 and C18, and the mobile phase is a mixture of water with other organic solvents. Several advantages associated with reverse phase LC make it particularly

suitable for metabolomics experiments. Firstly, it requires a very small sample volume and provides high separation efficiency. Usually, a few microliters to sub-microliters of sample are needed as an optimized injection volume. With the development of the UHPLC instrument and a separation column with a particle size less than 2 μm, one can achieve an even higher separation efficiency with less analysis time. Secondly, it is compatible with most non-polar and moderately polar compounds, which covers a wide range of metabolites; since the mobile phase contains polar solvents, it is convenient to analyze most water based biological samples. Thirdly, it is ready to link to common detector techniques, such as UV and mass spectrometer with an Electrospray ionization (ESI) source.[45] ESI is a soft ionization method that does not involve localized heating during the ionization process. Since there is very little internal energy transferred to the molecular ions, most the molecular ions can remain stable during ionization into the gas phase. Therefore, ESI is particularly useful for molecular weight determinations, making LC-ESI-MS a powerful platform for metabolome profiling work.

1.3.2 Chemical isotope labeling LC-MS based metabolomics

With the use of Electrospray ionization (ESI) ionization method, we can obtain the signal from the intact molecule ions. Though suitable for metabolic profile analysis, ESI suffers from several drawbacks, including the most prominent one being the strong ion suppression effect when a complex mixture is ionized together. Analytes will compete for charge during the electrospray as the ionization efficiency is different for different molecules. Therefore, the level of detected ion signal of a particular ion can change, depending on the presence or absence of other co-eluting analytes in the matrix. The mass peak intensity data can have relatively large error for quantification, especially when comparing samples with different matrices.

To improve the quantification accuracy, one can introduce an internal standard in the sample to overcome the ion suppression effect through relative quantification. A stable isotope internal standard often is used with the mass spectrometry platform. The isotope analogue of a targeted compound could have the same retention time as the targeted analyte. As both analyte and isotopic internal standard enter the ESI interface, their ionization will remain the same regardless of ion suppression from other co-eluting compounds. Therefore, the relative intensity ratio can be used to represent the relative concentration of each analyte accurately. However, in untargeted metabolomics, it is not possible to prepare an internal isotope standard for each metabolite. To address this challenge, our group took another approach using a stable chemical isotope labeling LC-MS method. Figure 1.2 shows an example of the labeling reaction with dansylation. The dansyl chloride reagent can react with primary or secondary amine and phenol containing metabolites. Instead of using an isotope analogue of each metabolite, the labeling method introduces a tag molecule to a group of metabolites with either ^{12}C or ^{13}C isotopes. In a typical experimental workflow, each individual sample is labeled with ^{12}C-dansyl chloride, and a pooled sample (mixture of all individual samples) is labeled with ^{13}C-dansyl chloride. Then, the mixture of light and heavy isotope labeled sample is analyzed by LC-MS. For each labeled metabolite, it can be detected in the mass spectrum with a light peak from the ^{12}C-labeled individual sample and a heavy peak from the ^{13}C-labeled pooled sample. The heavy labeled peak here serves as the internal standard for each metabolite. The intensity ratio of light to heavy peak is calculated during data processing for further quantitative analysis.

In addition to a better quantification result, the introduction of a dansyl group also brings other advantages to the metabolome analysis. Firstly, it improves the separation of metabolites in RPLC; the relatively non-polar dansyl group significantly reduces the polarity of some ionic

metabolites. The labeling reaction relatively unifies the chemical diversity of a wide range of metabolites and enables an efficient separation using one LC gradient with RPLC. Secondly, it improves the ionization efficiency for those less ionizable compounds. The tertiary amine group on the dansyl tag can be protonated easily with the addition of formic acid in the mobile phase. This overcomes the bias of ESI-MS for those more ionizable species. Thirdly, the use of an intensity ratio overcomes the quantification bias towards high abundant analytes. The untargeted metabolomics usually favors the high concentration metabolites since they can generate high intensity signals. The intensity ratio value is independent of the absolute peak intensity. Even for compounds of low concentration, the peak pair ratio can be measured accurately. Therefore, a larger quantification range can be achieved with the chemical labeling method.

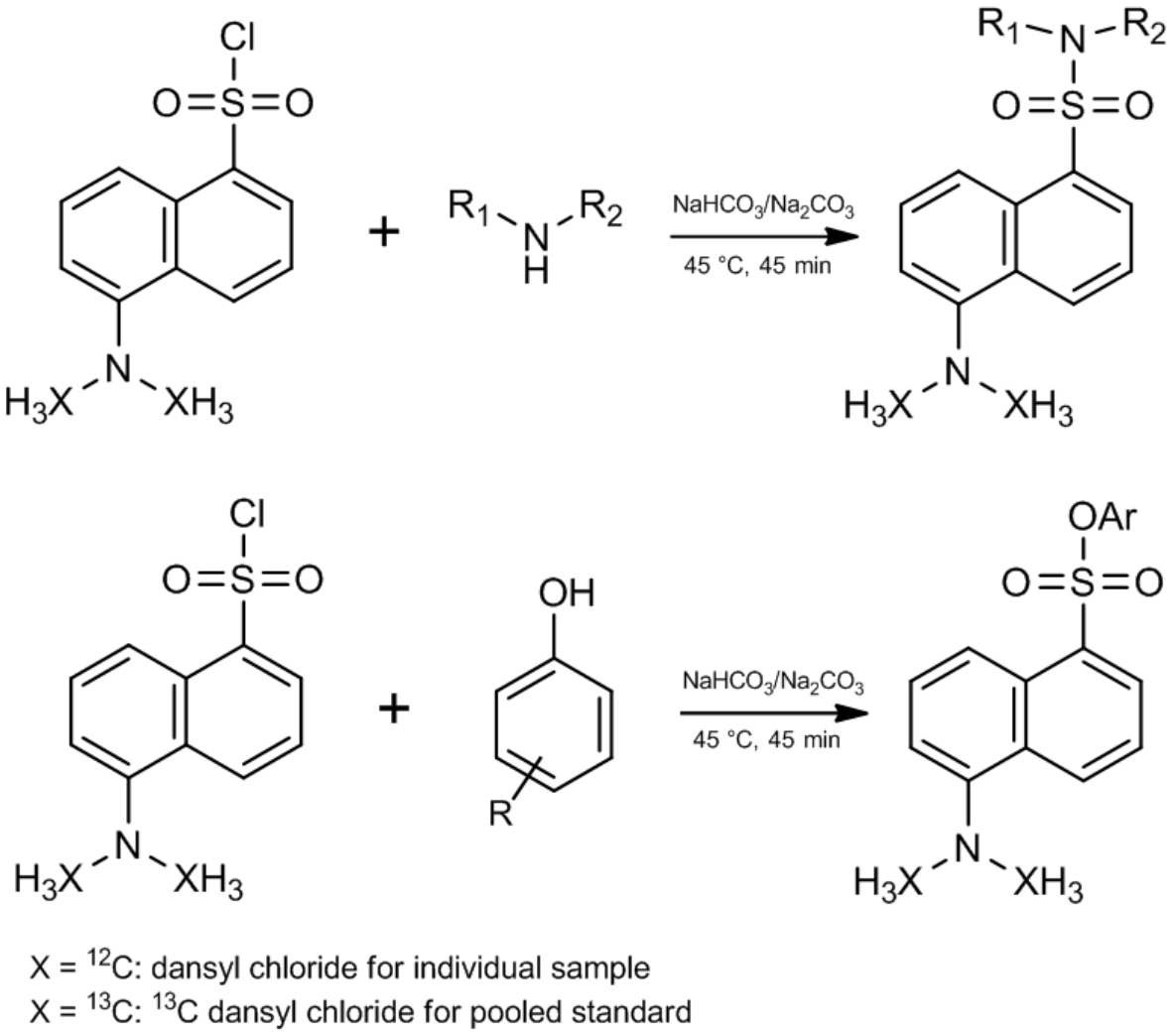


Figure 1.2 Dansylation reaction in the chemical isotope labeling LC-MS platform. The dansyl chloride can react with molecules containing amine and phenol to produce a stable labeled compound. The methyl group in the dansyl chloride can be either ^{12}C or ^{13}C for light and heavy labeling.

The dansylation labeling LC-MS method has been successfully applied to different biological samples, including urine, serum, cerebrospinal fluid, tissues, sweat, and microbial species. Yiman Wu and Xian Luo applies the method in microbial metabolomics and obtained the metabolic profiles from *E. coli* and yeast.[46,47] Kevin Hooton developed the metabolomics

workflow for the human sweat metabolomics using the dansylation labeling method.[48,49] Dorothea Mung used the dansylation labeling method to study the milk metabolomics.[50,51] This work extended the application of the labeling method to nutrition study. Many diseases were also investigated using the dansylation labeling LC-MS method for the discovery of potential biomarkers, including Alzheimer's disease,[52-55] Parkinson's disease,[56] prostate cancer,[57] Osteoarthritis,[58] etc.

1.4 Data Processing and Analysis in Metabolomics

1.4.1 Data processing tools for metabolomics

A large quantity of raw data can be generated from a typical metabolomics experiment with the mass spectrometry platform. The raw data contains both the information of the metabolites of interest and all other detected signals. It is an impossible task to pick out metabolite information by manually examining the spectra, especially in an untargeted analysis. A data processing program usually is developed with an efficient algorithm to extract the useful metabolite information after instrumental analysis. The use of a processing program significantly reduced the time and energy in raw data processing and provided a standard way for processing certain types of data.

Mass spectrometer manufacturers usually have their own data analysis software that comes in a bundle with the instrument. Such software contains most general data preprocessing methods, including peak centroid calculation, mass calibration, data normalization, and feature extraction. In addition to the commercial data processing tools, several web-based data processing platforms, including MetAlign,[59] MZmine,[60] MAVEN,[61] MetaboAnalyst, [62-64] and

XCMS,[65-67] have been developed successfully to facilitate in-depth data processing and analysis. Each platform provides unique capabilities in supporting metabolomic data storage, analysis, annotation, etc. For example, MetaboAnalyst has been developed mainly for targeted metabolomics data analysis, and is used to perform a complete statistical analysis. Users will run data processing on the raw data using their software and upload the aligned data table to MetaboAnalyst for statistical analysis. The combination of different data processing platforms is sufficient for a complete metabolomics data processing pipeline from feature extraction and data alignment to metabolite annotation and exploratory statistical analyses.

In chemical isotope labeling LC-MS based metabolomics, our group has also developed several in-house program to assist the data processing. IsoMS[68] was the first program developed to extract the peak pair data from the raw mass list. It conducts peak pairing, background and adduct peak pair filtering, peak pair ratio calculation, and peak pair grouping. This is a unique processing step in the labeling experiment for feature picking. The zero-fill[69] program was later developed for the peak pair data alignment and missing value retrieval for analyzing multiple samples. As a continuation of the data processing method development, my thesis discusses the new methods in the current workflow that come with unique processing algorithms for a labeling experiment.

1.4.2 Current challenges in metabolomics data analysis

Metabolomics data can have various formats and structures, depending on the experimental method and instrument used. Regardless of the instrument platform, the beginning of the data processing usually starts by exporting the raw data using the commercial software available to

each specific instrument. This converts the instrumental data file to a more accessible format for subsequent processing, which separates the useful metabolite information from other interfering signals. Experimental design is the primary consideration in designing the processing algorithm. For example, in a chemical isotope labeling LC-MS experiment, a metabolite feature is defined by a peak pair consisting of a light and a heavy mass peak. Accordingly, the processing program needs to examine the raw mass data for all possible peak pair features that have the peak pair distance defined by the heavy labeling reagents. Moreover, a data processing program has to be updated frequently to meet the demand from advancements in experimental methods. For example, in different labeling experiments, the labeling reagent and the experimental conditions may be different. The program then needs to adjust its parameters for data generated by different methods in order to generate an accurate and complete metabolite-intensity table. A new function often is needed for a specific task in the data processing, such as the background feature removal from a new labeling reaction. This poses a challenge to each researcher in metabolomics to develop efficient and customized processing programs in dealing with all types of metabolomics data.

The identification of metabolites is important for understanding a biological process. Though one can achieve high detection coverage of a metabolome, the identification of each metabolite signal remains a major challenge in the current metabolomics workflow. These metabolites can have different origins. For example, human biofluids can contain both endogenous metabolites synthesized within human cell and the exogenous compounds from diet, air, drugs, and gut microflora. Many databases have been developed to contain lists of metabolites found in different organisms, including human metabolome database HMDB[2] and MycompoundID,[70] yeast metabolome database YMDB,[71] drug pathway, and metabolites

database DrugBank,[72] etc. Since there is no standard method in the data collection, metabolite identification for data generated in different labs can be challenging. The identification of novel metabolites has to rely on the authentic standards.

1.4.3 R language platform for data processing and statistical analysis

R language is a free and open-source software environment that has been used widely in data processing, statistical analysis, and data visualization. It has good numerical capabilities, flexible visualization capabilities, easy access to databases, and a wide range of statistical and mathematical algorithms available in different R packages. R is accepted widely in the bioinformatics community for its open-source nature and many well developed packages; this makes it convenient for method development on the R platform.

In recent years, our group has developed a series of data processing programs for chemical isotope LC-MS data, including IsoMS,[68] Zero-fill,[73] and IsoMS-Quant.[74] These R based programs have been applied successfully in our daily data processing tasks. In the following Chapters, R language is used as the development language for all processing algorithm designs. Ultimately, the new data processing methods will be integrated with the current programs to provide a complete software package for chemical isotope labeling LC-MS data processing.

1.5 Processing Speed Optimization

1.5.1 Hardware optimization

Data processing speed is another important topic in the development of processing algorithms. The amount of data generated in a lab is increasing with the growing number of mass spectrometers available and the increasing speed in data collection. The processing of a large amount of data can result in a long waiting time and can slow down the research progress. Of all the strategies in improving data processing speed in large data analysis, an upgrade of the hardware infrastructure is often the most feasible and inexpensive one. Nowadays, commercially available computers can handle most data processing tasks with a much improved processing speed thanks to advances in computer hardware technology. When selecting a data processing computer, we usually focus primarily on the CPU performance since it is the center for all calculations. The benchmark score of a CPU can be used as a measure of the processing speed. Figure 1.3 shows the benchmarks of midrange Intel® processors of different generations from 2012 to 2018. The benchmark data were collected by PassMark® software (https://www.passmark.com/), which used a performance test to evaluate the computing power of a CPU. We selected the mainstream i5 CPU from the 2nd to the 8th generation (5th generation data not available) in the comparison. The generation number is indicated by the first digit of the model number (e.g. 8xxx for 8th generation). In the results, we can see that the current 8th generation i5 8400 CPU has double the score of the 2nd generation i5 2400 CPU, indicating a huge improvement of the CPU computing power in recent years.

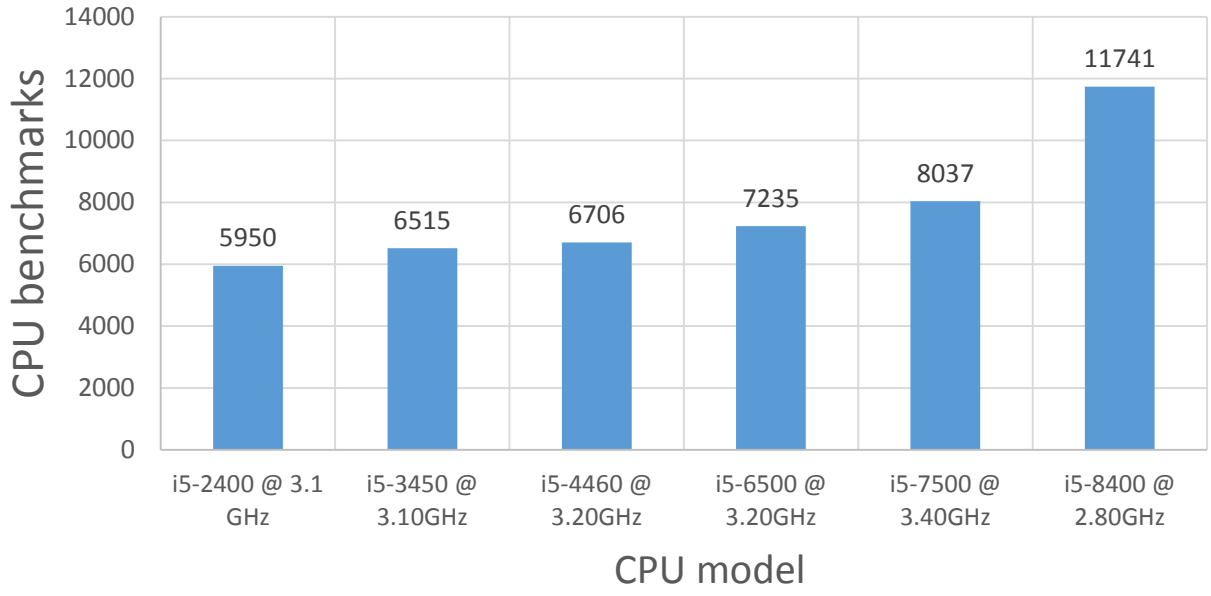


Figure 1.3 CPU benchmarks for midrange Intel CPU in each generation.

A computer is a multi-unit system, in which all components have to work together to achieve an optimal performance; any low performance unit potentially can be a bottleneck in the actual performance. In addition to the CPU unit, other hardware, such as the random-access memory (RAM), hard drive, and motherboard, were carefully selected for building our processing computers.

Data reading and writing are two frequent steps in any data processing. When the program reads the sample data, it stores the data in RAM memory so as to provide a quick access to all the processed information. Additional RAM allows the computer to work with more data at the same time. The hard drives stores the input and output data files of each processing step. The read/write speed of the hard drive must not be too slow when processing a large number of files. Moreover, the hard drive saves the raw files and processed data for each experiment. The robustness and long-term stability of a hard drive are crucial for data safety. Although it rarely

happens, the failure of a hard drive can be disastrous to the precious sample data, especially when the sample volume is limited for re-analysis. A motherboard is the bridge allowing the communication between many crucial electronic components of a system such as CPU, RAM, and hard drive, and provides connectors for other peripherals. Each generation of CPU usually has its compatible motherboard that matches the power requirements and data transfer speed. In the end, a processing computer often has to run at the maximum load for days; this can be demanding for the cooling system and stability of the power supply. The selection of the power supply must match the power requirements of the system, and dust should be vacuumed regularly to prevent heat from accumulating inside.

In conclusion, the upgrade of the computer hardware requires a careful choice of each component to achieve the optimal processing speed and long-term performance. With all these considerations, I built eight data processing computers in the lab as a data processing station to meet the increasing data processing demand from an increasing number of lab users and new instruments. To balance the cost and performance, we selected a midrange Intel i5 CPU and a 16 GB RAM for each computer. We assigned 3–4 computers dedicated to one MS instrument to ensure a minimum waiting time in raw data processing.

1.5.2 Multi-core CPU and parallel processing

In order to run a program more efficiently, i.e., using less time, parallel processing can be used where a set of instructions in a program are given to divide a specific task among multiple processors. In the earliest computers, only one program could run at a time, and each independent program could only be executed when the CPU finished processing the previous

task. Parallel processing at its early form used interleaved execution of two programs. The computer was able to execute the next processor-intensive program while waiting on the current task.[75] The total execution time for the two tasks would be less than the sum of the individual processing times. Multiple tasks are sharing CPU resources in this way.

The next improvement in multiprocessing is the introduction of a multi-core CPU, in which two or more processors are attached in one CPU module for an enhanced performance and a lower power consumption. As the single core CPU rapidly reached its physical limits of complexity and calculation speed, a multi-core CPU has become more and more popular by combining independent processing units into the same socket. Each of the physical cores of a multi-core CPU can read and execute program instructions independently. This is why one can run a number of programs in a computer simultaneously and perform other tasks, such as data transfer and file editing, without significantly slowing down the programs. Efficiency in multitasking has improved much by this multiprocessing technique.

Ideally, the total processing power of a CPU should be the sum of the processing power of all physical cores. However, each core can only run one program task at one time. This means that when executing a program script in a 4-core CPU, the usage of the CPU can only reach up to 25%. In other words, the processing speed is limited most often by a single core speed. Since the multi-core CPU is designed for a better multi-processing efficiency, the structure of the program can be modified to utilize the multi-processing to its full advantage.

Figure 1.4 shows the workflow of raw data processing for a chemical isotope labeling LC-MS dataset. After data collection, the raw data is exported first from the instrument files to mass lists that are easier accessed by customized programs. Then, mass accuracy and retention time shift are analyzed in each raw file. Next, the IsoMS program performs the peak pairs

extraction in each individual files and generates the peak pair list for each LC-MS data.[68] The alignment program aligns the peak pair lists from all individual files into one metabolite intensity table, and the zero-filling program calculates the missing ratios based on the raw mass list file.[76]

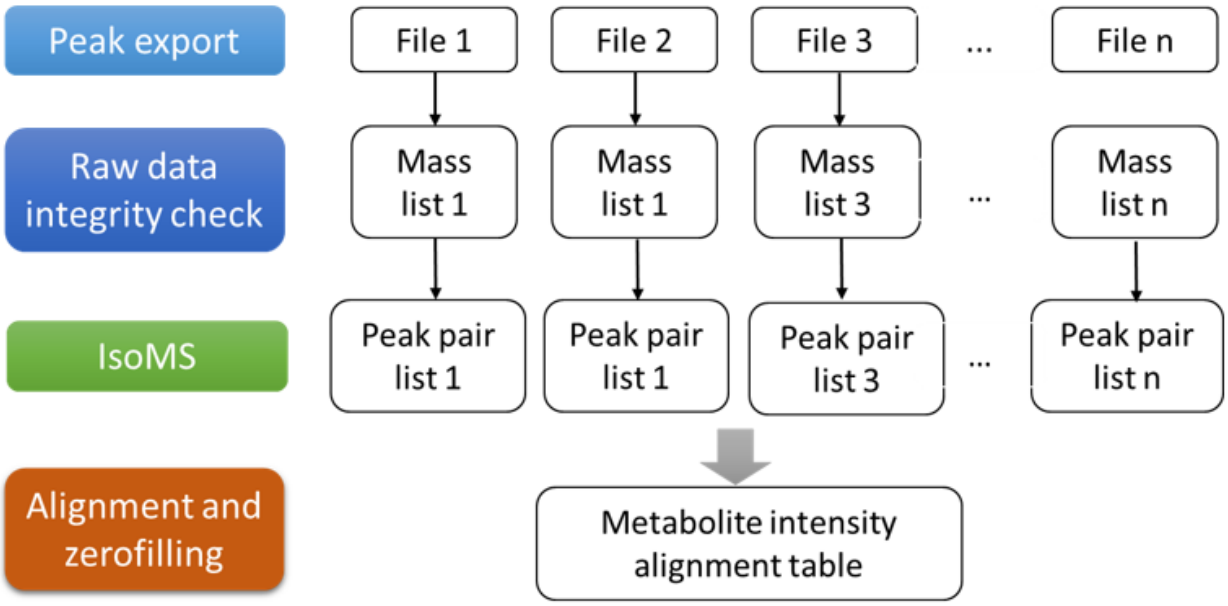


Figure 1.4 Schematic of data processing workflow in chemical isotope labeling LC-MS data.

One of the speed limiting steps in this work flow is IsoMS processing. In the IsoMS program, the raw mass list is processed by a series of functions to generate a final peak pair list. Figure 1.5 shows the functions in the IsoMS program. Each function processes the data generated by the previous step and passes the result to the next function. In the processing pipeline, a function is executed only after the previous function finishes; only one raw file is processed at one time. Since the peak pair lists from different files are independent, it is possible to have multiple IsoMS programs run at the same time.

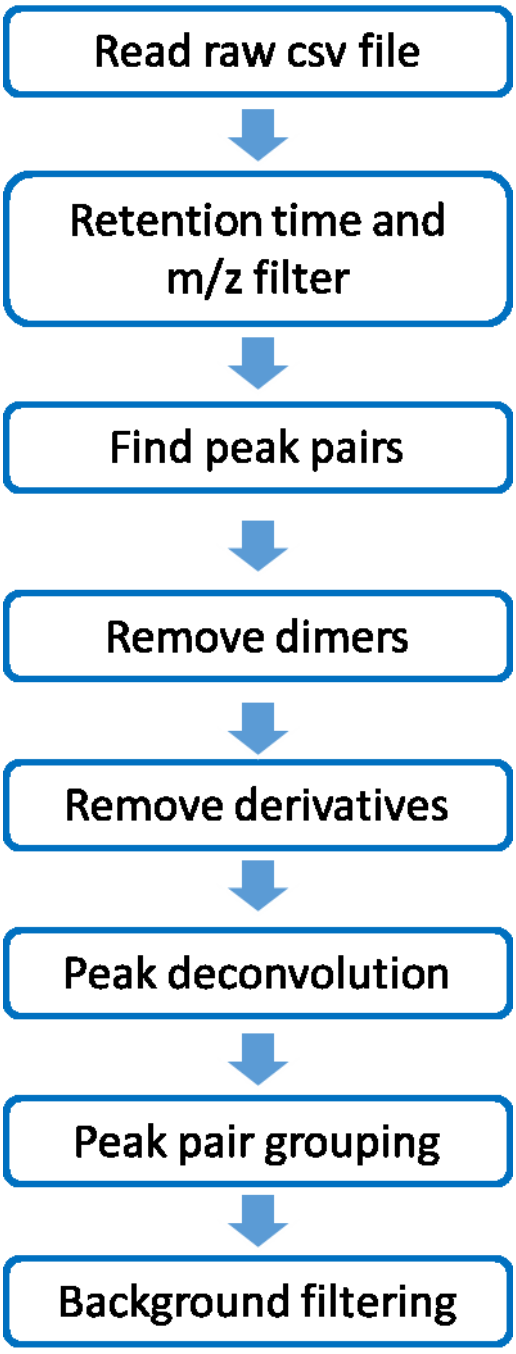


Figure 1.5 Workflow of the functions in the IsoMS program.

To achieve the parallel design, we detect the number of logical cores in the processing CPU at the start of IsoMS program. Then, the raw files are divided into multiple groups and assigned to each CPU core. For example, in a quad-core CPU, the raw files are divided into four groups. Figure 1.6 shows the schematic of file grouping. In a quad-core setting, four independent IsoMS programs will be running in parallel, processing the raw files in each file group. In this way, the processing speed can be 3–4 times faster than in single core processing.

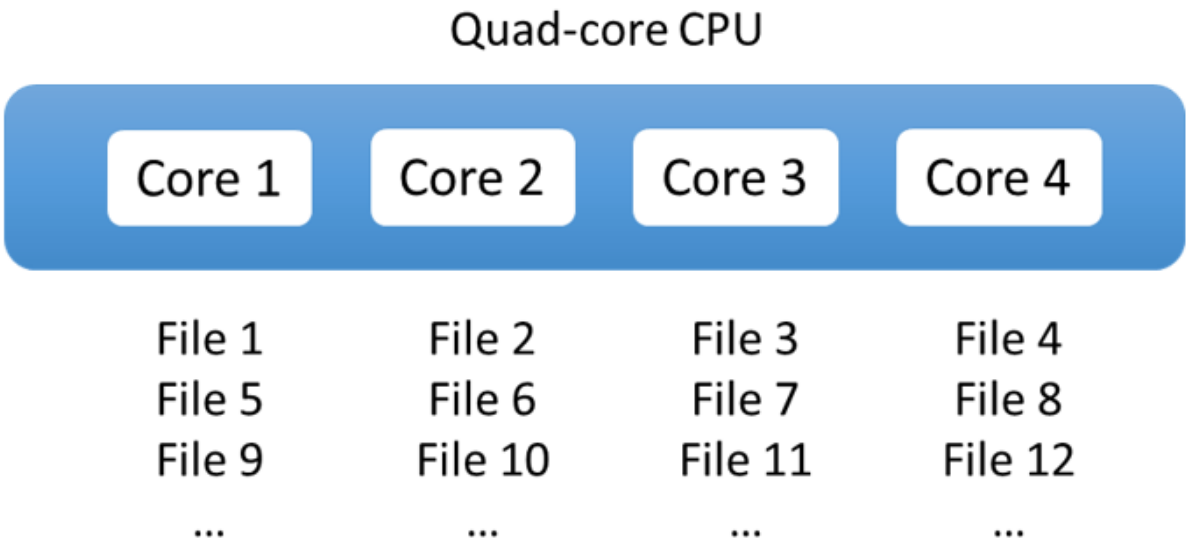


Figure 1.6 File assignment in IsoMS parallel processing. Files are grouped and assigned into each core at the beginning of the processing. 4 Independent IsoMS programs are running in parallel on the sub file list.

Another speed limiting step is the zero-fill program. Figure 1.7 shows an example of a metabolite intensity table after alignment. Each row contains a unique peak pair and their ratio values in each sample. NA indicates a missing value from the IsoMS peak pair extraction. To retrieve the missing value, the zero-filling program searches the peak pair in the raw file by the retention time and mz information (see Chapter 3 for more details). Depending on the number of

missing values in the original alignment table, the zero-filling processing time can vary. When the number of samples or number of peak pairs is large, it can take a long time for peak pair searching for each of the missing values.

Since the ratio values are independent, multiple missing values can be searched simultaneously in the same raw file. To optimize the loop into a parallel structure, we came up with the solution to divide each sample column into multiple segments based on the number of CPU cores. Figure 1.8 shows an example of this method in a quad-core CPU. After the raw data is read into the memory, the corresponding sample column is divided into four segments based on the number of NA. Inside each column segment, an independent zero-fill program is running for the calculation of the missing peak pair ratio using the EIC peak area data. After all individual zero-fill tasks are finished, the resultant segments are pasted to reconstitute the sample column; and the total processing time can be shortened by 3–4 times.

Samples

Variables (peak pair)

				Healthy	Healthy	Healthy	Healthy	Diseased	Diseased	Diseased	Diseased
Peak pair #	RT	mz_light	mz_heavy	1	2	3	4	5	6	7	8
1	124.25	631.8669	635.8801	NA	0.33	NA	NA	NA	NA	NA	NA
2	124.27	432.0074	434.0136	NA	NA	NA	NA	NA	1.51	NA	NA
3	124.39	567.9823	569.9886	NA	0.26	NA	0.3	NA	NA	1.26	NA
4	124.77	364.0200	366.0261	NA	2.87	NA	NA	3.4	NA	2.22	3.41
5	124.94	389.1272	391.1338	0.92	0.95	0.92	0.93	1.04	0.94	0.96	0.91
6	124.97	692.8343	694.8346	0.85	0.96	0.98	0.95	0.81	0.83	1.03	0.88
7	125.06	499.9947	502.0011	0.94	0.93	1.02	0.97	0.89	0.97	1	1.02
8	125.24	627.9443	629.9471	NA	NA	NA	NA	0.92	NA	1.03	NA
9	125.26	465.9684	467.9712	2.94	2.67	NA	2.33	1.92	NA	2.25	2.11
10	125.28	635.9690	637.9761	0.9	0.87	0.85	0.93	0.94	0.92	0.94	0.96
11	125.31	286.0300	288.0341	NA	NA	NA	1.82	NA	NA	NA	NA
12	125.33	546.8322	550.8416	NA	NA	0.59	NA	NA	NA	1.44	NA

Figure 1.7 Example of an aligned data table from a CIL LC-MS experiment. Each row contains the information of a peak pair and the peak pair ratio in each sample. NA represents a missing value.

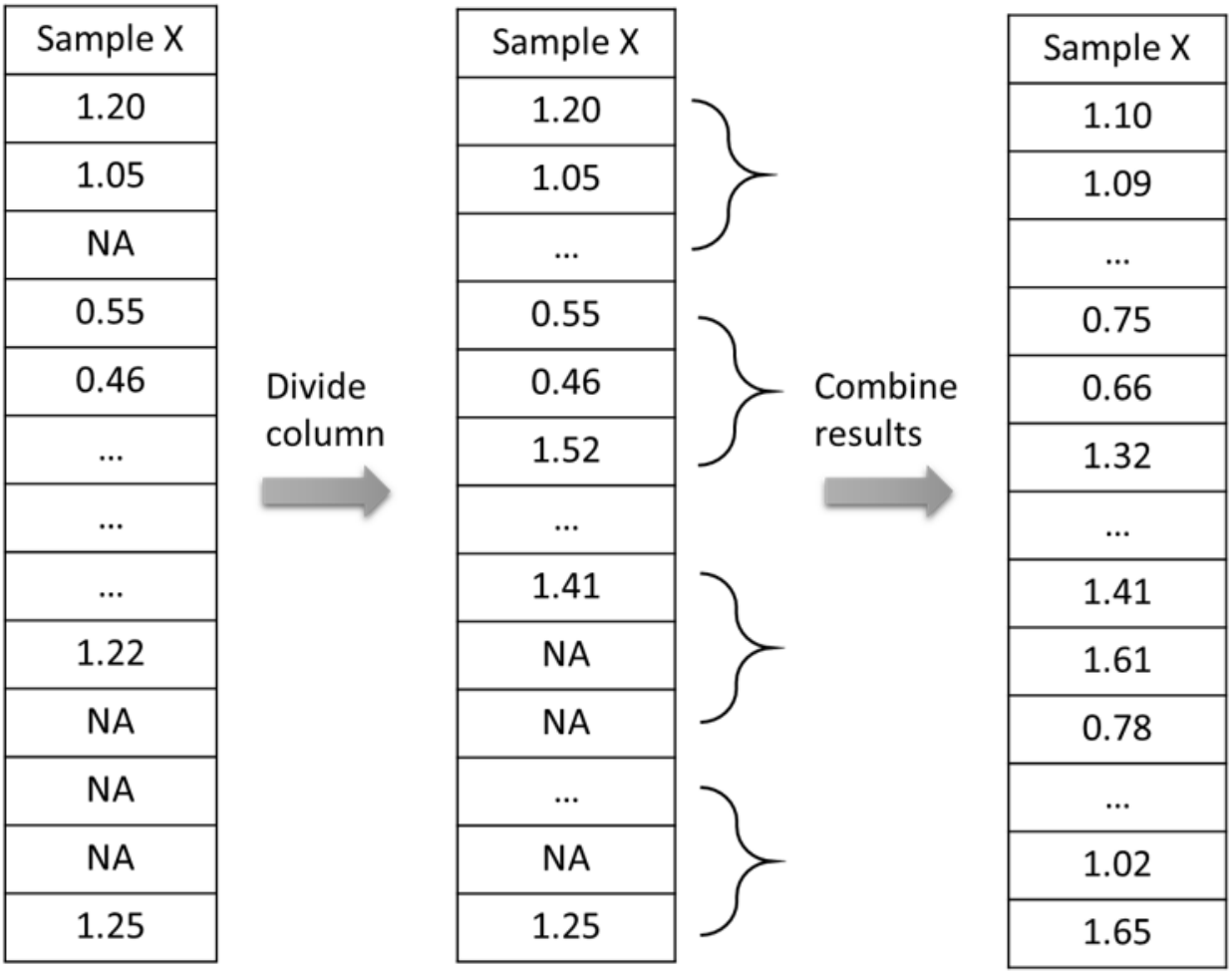


Figure 1.8 Parallel processing in the zero-fill program. The column shows the peak pair ratio values in one sample.

One major concern in parallel processing is the potential conflict in combining the outputs from parallel tasks. This means that the data structure must be kept the same during the parallel process and there should be no dependency from one result on another. For this reason, not all processing steps can be modified into a parallel structure. However, after the optimization

of the two major speed limiting processing steps, a speed improvement of at least three-fold was observed for processing the same number of files.

With parallel processing, the processing speed can keep increasing with more cores in a CPU. Currently, both Intel and AMD have their high-end CPU with up to 32 cores. Meantime, the cost of the CPU plus their compatible peripherals (such as RAM, power supply, and cooling) also increases significantly, leading to a much higher cost for each desktop. Therefore, we picked the newest generation of Intel i5 CPU as a balance between performance and cost. To ensure a sufficient access to the computer resources for each lab user, a multi-computer data processing system was built to meet the challenges of the growing data collection speed and increasing data size.

1.5.3 Optimizing data processing workflow

Data processing involves multiple steps, and the total processing time for a batch of data files is the sum of the processing time of all steps. Now that we have optimized the processing speed in individual functions, the next step is to establish a data processing workflow that optimizes the total processing time from the raw data to the final data tables. Figure 1.9 shows the design of the optimized workflow. We first attach the post-acquisition mass calibration and mass exportation methods to each LC-MS method. Then, each LC-MS data is exported to a mass list file at the end of the acquisition. While the next LC-MS analysis is running, the data from the previous run is exported to a csv file. In this way, all mass list csv files are available at the end of the LC-MS analysis sequence.

As an example, the analysis of a total number of 120 samples, each with a 47 min (32 min gradient and 15 min equilibrium) LC running time, we can analyze around 30 samples per day. On Day 2, the 30 LC-MS data files are checked for their mass accuracy and retention time shift, followed by the IsoMS processing. Peak pair lists are generated on Day 2 for data collected in Day 1. Since data alignment requires all the individual peak pair lists, after all peak pair lists are generated, we align all data files and run the zero-fill program to generate the final data tables for further statistical analysis. The total data alignment and zero-fill processing will take about one day.

In this workflow, the total data processing time is only 1 to 2 extra days on top of four days' instrumental analysis time. Instead of waiting for 1–2 weeks on the computer for data processing before the optimization, now one can move quickly on to statistical analysis workflow after data collection.

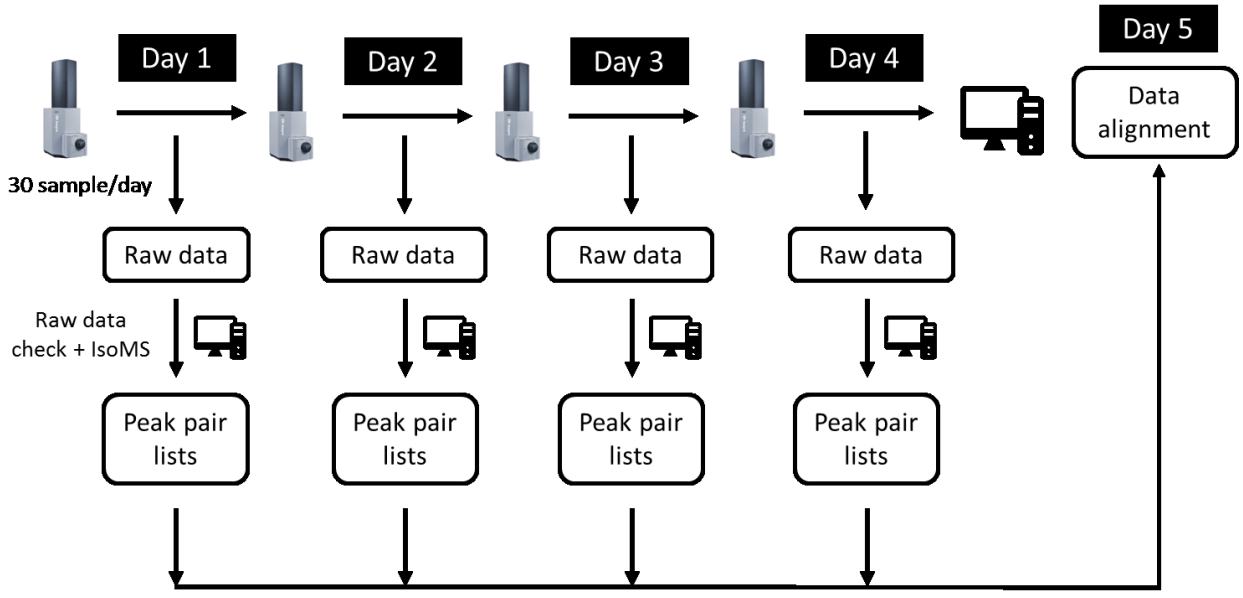


Figure 1.9 Optimized workflow of data processing in dealing with a batch of data files.

1.6 Overview of the Thesis

My research focused on the development of data processing methods to meet the challenges in metabolomics data processing in chemical isotope labeling LC-MS. Chapters 2 and 3 discussed the LC-MS raw data quality control in terms of mass accuracy and retention time shift. In LC-MS based metabolomics, m/z and retention time are the two most important measurements of one metabolite. All the downstream data processing steps rely on these two parameters for feature extraction, data alignment, and metabolite identification, etc. Any large error with mass and retention time should be corrected at the beginning of the data processing.

Chapter 4 discussed the chemical isotope labeling LC-MS data processing workflow and explained the algorithms in each function. Unique peak pair features were extracted first from each LC-MS analysis. Then, background peak, adducts, and repeatedly detected signals were removed from the peak pair list. Next, peak pairs of the same metabolite from different samples were aligned into a metabolite intensity table. We calculated each peak pair ratio using the average value from the whole peak area to minimize random errors. Each peak pair was evaluated on the dependence of the light and heavy peak to ensure it represents a labeled metabolite truly. Redundant peak pairs were evaluated by the distance of mass, retention time, and similarity of within-sample peak pair ratios. False positive and redundant peak pairs were removed from the data table. Missing peak pair ratios were searched in the raw data for possible peak pair signals. Different searching algorithms were applied for a multi-layer missing value calculation. For peak pairs with the intensity information available in the raw file, we calculated

the ratio based on the original data. For peak pairs missing one or both peak intensities, the peak pair ratio was predicted based on any existing intensity and peak pair information in other samples. In the end, we generated a complete metabolite intensity table with each peak pair uniquely representing a labeled metabolite.

Chapter 5 studied the natural isotope peak intensity in dansylation labeling LC-MS data. It discussed the challenge of quantification in any labeling method with the presence of natural isotope peaks. A new method was proposed to remove the peak intensity contribution from natural isotope peak to improve the accuracy of the peak pair ratio calculation further.

Chapter 6 discussed the metabolite identification in metabolomics study. Metabolites in a biological sample can have a wide concentration distribution. Accordingly, a wide distribution of peak intensities can be observed in the resultant data, and mass error can be affected by the peak intensity. We investigated the relationship between mass peak intensity and mass error, and an intensity dependent mass tolerance was calculated for each query mass in the library search to improve the accuracy and efficiency in the metabolite identification.

Lastly, we designed a program that incorporates all processing functions with a graphical user interface. Figure 1.10 shows a program window of mass accuracy check function. On the left are the parameters used in the function, including the reference mass and mass search window. At the end of the processing of a total of 280 sample files, a plot was generated in the program window with the average and standard deviation of the measured mass in each sample. The data in the plot is saved in the local folder for further examination.

As an objective of this thesis work, an all-around data processing program was designed to facilitate the metabolomics research. The integrated data processing program has been

implemented as a standard in the lab for data generated with different experimental methods. Moreover, feedback was collected from different users to help us update the software further with more functions. Metabolite identification and statistical analysis modules could be added to the current platform to provide a complete solution to metabolomics data analysis in the future.

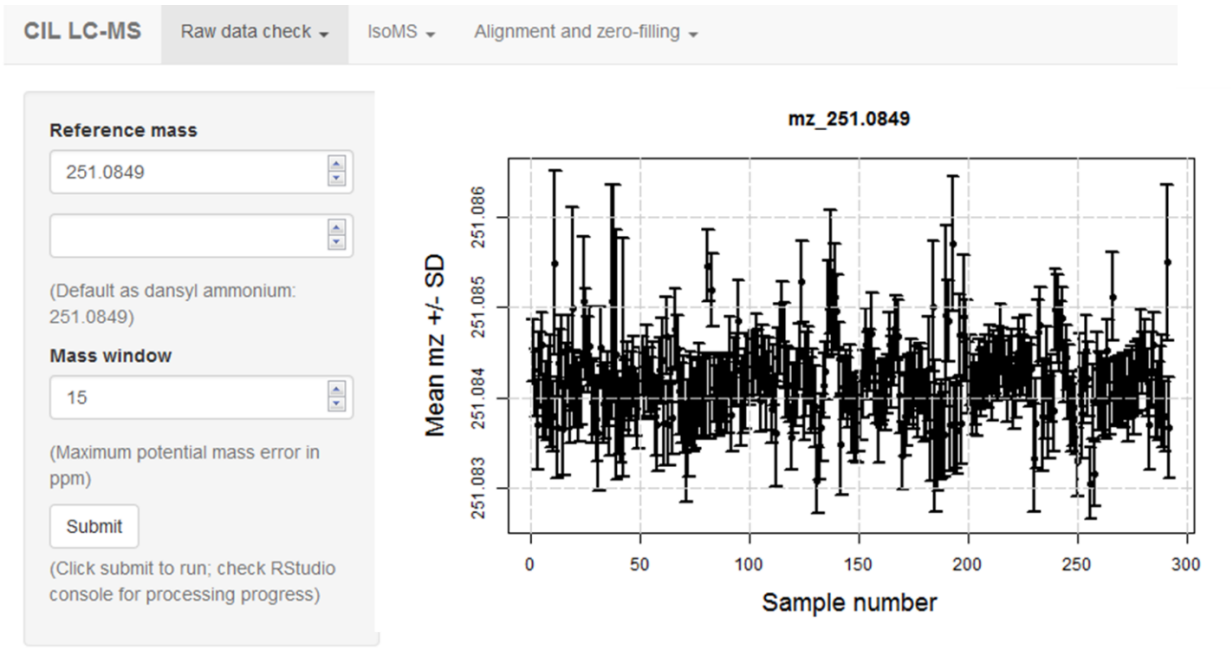


Figure 1.10 Example of graphical user interface in the CIL LC-MS data processing program. Parameters can be modified in each of the text boxes. An instant plot is generated at the end of the processing to provide a quick view of the results.

Chapter 2 Mass Accuracy and Precision Check for LC-MS Raw Data Using Background Mass Peaks

2.2 Introduction

The data produced by a mass spectrometer are the masses and intensities of analytes and their fragments. The accuracy of the mass measurement determines directly the usefulness of the data in the identification of metabolites and other statistical analyses. Instruments introduced in the past decades have improved the mass resolution and the mass accuracy greatly.[77] For example, a TOF mass spectrometer with an energy correcting reflectron can attain a mass accuracy at low ppm values for detecting low mass ions (<1000 Da).[78] Other ion-trap-based mass spectrometers, such as a FT-ICR mass spectrometer, provide a resolving power potentially over 1,000,000 with an average mass error generally less than 1 ppm.[77,79]

Although high in theoretical resolution, a TOF mass spectrometer requires a regularly conducted mass calibration to correct mass shift caused by ambient temperature changes or voltage fluctuations in order to maintain the best possible mass accuracy.[78] External calibration methods have been developed with calibration solutions covering different mass ranges, and they are used routinely for mass calibration in different mass spectrometers.[80-82] However, compared to an internal calibration, an external calibration cannot account for mass shift during data acquisition, especially for data collection over a long period of time. To address this issue, a lock mass calibration method was introduced with one or multiple mass standards that were injected to the mass spectrometer with the sample. The lock mass standards generated signals in every mass spectrum to be used as mass references in post-acquisition internal mass calibration.[83] The

mass standard can be either from an external compound introduced in the ESI source or from a known background in the LC-MS analysis. For example, polydimethylcyclosiloxanes, a group of ubiquitous contaminants of the laboratory air, were found to be the source of extreme background signals in nano-electrospray mass spectrometry.[84] Mann et al. later used this known background mass peak produced by electrospray for lock mass calibration and achieved sub-ppm mass accuracy.[85]

Despite a much improved mass accuracy with lock mass calibration, the method often is not compatible with metabolomics experiments, in which the concentration of metabolites can vary by a few magnitudes. The introduction of high concentrations of internal mass standards can suppress the signals of metabolites of relatively low concentration easily and thus decreases the metabolome coverage. To obtain the best mass accuracy possible, our lab employed another strategy by implementing a mass calibration segment during the LC dead time for each LC-MS run. Figure 2.1 (A) shows a LC chromatogram with a sodium formate calibration solution injected at the start of the LC run. In the first two minutes, the eluent from the separation column went to waste, and another line of sodium formate solution was connect to the ESI source of the mass spectrometer. Figure 2.1 (B) shows a mass spectrum of sodium formate adducts in positive mode for mass calibration (see adducts formulas and exact masses in Table 2.1). After data collection, each LC-MS data file is calibrated using its calibration segment. Since mass accuracy is relatively stable over the course of one sample analysis, the method is able to correct mass shift in each individual sample more accurately than traditional external mass calibration.

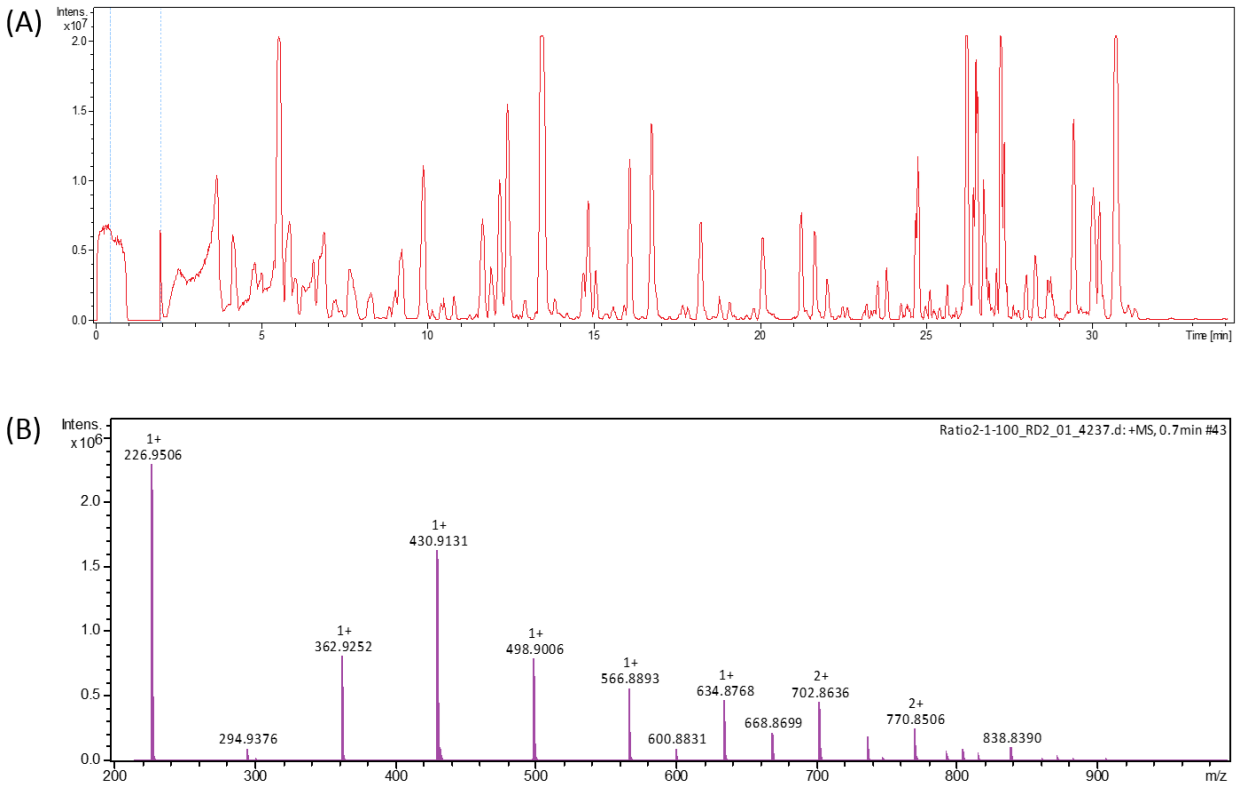


Figure 2.1 (A) LC chromatogram with calibration segment at the beginning two minutes and (B) mass spectrum of sodium formate adducts peaks used for mass calibration.

Table 2.1 List of sodium formate adducts and their exact mass for mass calibration.

Formula	m/z
$Na(NaCOOH)_3$	226.951493
$Na(NaCOOH)_4$	294.938917
$Na(NaCOOH)_5$	362.926341
$Na(NaCOOH)_6$	430.913765
$Na(NaCOOH)_7$	498.901189
$Na(NaCOOH)_8$	566.888613
$Na(NaCOOH)_9$	634.876037

$Na(NaCOOH)_{10}$	702.863461
$Na(NaCOOH)_{11}$	770.850884
$Na(NaCOOH)_{12}$	838.838308
$Na(NaCOOH)_{13}$	906.825732
$Na(NaCOOH)_{14}$	974.813156

With either external or internal mass calibration, one can achieve the optimal mass accuracy available in one instrument. However, mass shift sometimes can occur in the middle of the analysis, causing an excessively large mass shift in some of the spectra within a sample. In such a situation, the mass calibration method may not correct the mass shift back to within the range of the instrument tolerance. Since there is no evaluation method to show the calibrated mass accuracy and precision after mass calibration, it is difficult for users to pick out the mass spectrum with a mass accuracy issue.

Inspired by the work of Mann et al[77] who used a naturally occurring background mass peak in the nano-LC-MS system for lock mass calibration, I studied the background mass peaks in the chemical isotope labeling LC-MS data and found a list of common background peaks associated with different labeling methods. The signals of these background peaks may vary by a few magnitudes at different retention times, making it less suitable for lock mass calibration even with a known exact mass. However, the consistent presence of a background peak can be used as a mass reference to evaluate the mass accuracy and precision of the mass measurement in each calibrated spectrum. In this work, I developed a program to search the background masses that are present in all sample files. Mass errors were calculated in each mass spectrum using the reference background mass peaks to show the mass accuracy and precision after mass calibration.

2.3 Materials and Methods

2.3.1 Chemicals and reagents

All the chemicals and reagents, unless otherwise stated, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). In a dansylation labeling reaction, the ^{12}C-labeling reagent (dansyl chloride) was purchased from Sigma-Aldrich, and the ^{13}C-labeling reagent was synthesized and purified in our lab using the procedure published previously.[30] LC-MS grade water, methanol, and acetonitrile (ACN) were purchased from ThermoFisher Scientific.

2.3.2 Dansylation labeling

Mouse serum samples were collected, and a pooled mouse serum sample was prepared by mixing equal aliquots of each individual sample. In a microcentrifuge tube, 30 µL of pooled serum were mixed with 90 µL of methanol. The mixture was then incubated at -20 °C for 2 h before centrifuging at 15,000 g for 15 min to precipitate the proteins. 90 µL of clear supernatant was taken and dried in a SpeedVac vacuum concentrator. The sample was re-dissolved to 75 µL with 2:1 H_2O/ACN. After that, 25 µL of 250 mM sodium carbonate/sodium bicarbonate buffer were added to the sample to introduce a basic environment for the labeling reaction. The solution was vortexed, spun down, and mixed with 50 µL of freshly prepared ^{12}C-DnsCl solution (18 mg/mL) (for light labeling) or ^{13}C-DnsCl solution (18 mg/mL) (for heavy labeling). After the sample was incubated at 40 °C for 45 min, 10 µL of 250 mM NaOH were added to quench the excess dansyl chloride. The solution was incubated further at 40 °C for another 10 min to allow the unreacted dansyl chloride to be hydrolyzed fully. Finally, 50 µL of formic acid (425 mM) in 1:1 ACN/H_2O were used to acidify the solution. A quality control (QC) sample was prepared by

mixing a 1:1 volume of light labeling and heavy labeling pooled samples. The QC sample was injected between individual samples to monitor the instrument stability. A total of 35 QC injections were conducted during the LC-MS analysis of individual mouse serum samples.

2.3.3 LC-MS analysis

The ^{12}C- and ^{13}C-labeled samples were mixed and centrifuged at 20,800 g for 10 min before injecting into a Bruker Maxis Impact QTOF mass spectrometer (Billerica, MA, USA) linked to a Dionex UltiMate 3000 UHPLC system. A Zorbax Eclipse Plus C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1% (v/v) LC-MS grade formic acid in 5% (v/v) grade CAN, and solvent B was 0.1% (v/v) LC-MS grade formic acid in LC-MS grade ACN. The gradient elution profile was as follows: t=0.0 min 20% B, t=3.5 min, 35% B, t=18.0 min, 65%B, t=24 min, 99%B, t=28 min, 99% B. The flow rate was 180 μL/min. The QC sample injection volume was 2 μL.

2.4 Results and Discussion

2.4.1 Background mass peak search

Background mass peaks are not uncommon in LC-MS data. Impurities in reagents, and contaminations in sample vials, the mobile phase, or the column itself can introduce background signals. The actual type of background peaks can be different from one experiment to another. To find potential method-specific background peaks, we designed the first module of the program for searching the most frequent background peaks in the raw mass data.

Figure 2.2 shows the workflow for the function. We assumed that the background mass peaks were present in more than 80% all spectra within one sample. In searching for background signals, the program picked 10 spectra randomly from the sample file and combined all the masses into one mass list. From this mass list, the program removed any redundant mass to create a background mass candidates list. Based on the threshold (80%) set for each background peak, the probability of each qualified background peak present in a random spectrum is 0.8. Thus, the probability that the background peak is not present in the candidate mass list is 0.2^{10}, which is equal to 1.024×10^{-7}. Thus, we can say with confidence that the candidate list should include all potential background mass peaks. This random sampling strategy reduced the processing time significantly compared to scanning all masses in the raw data.

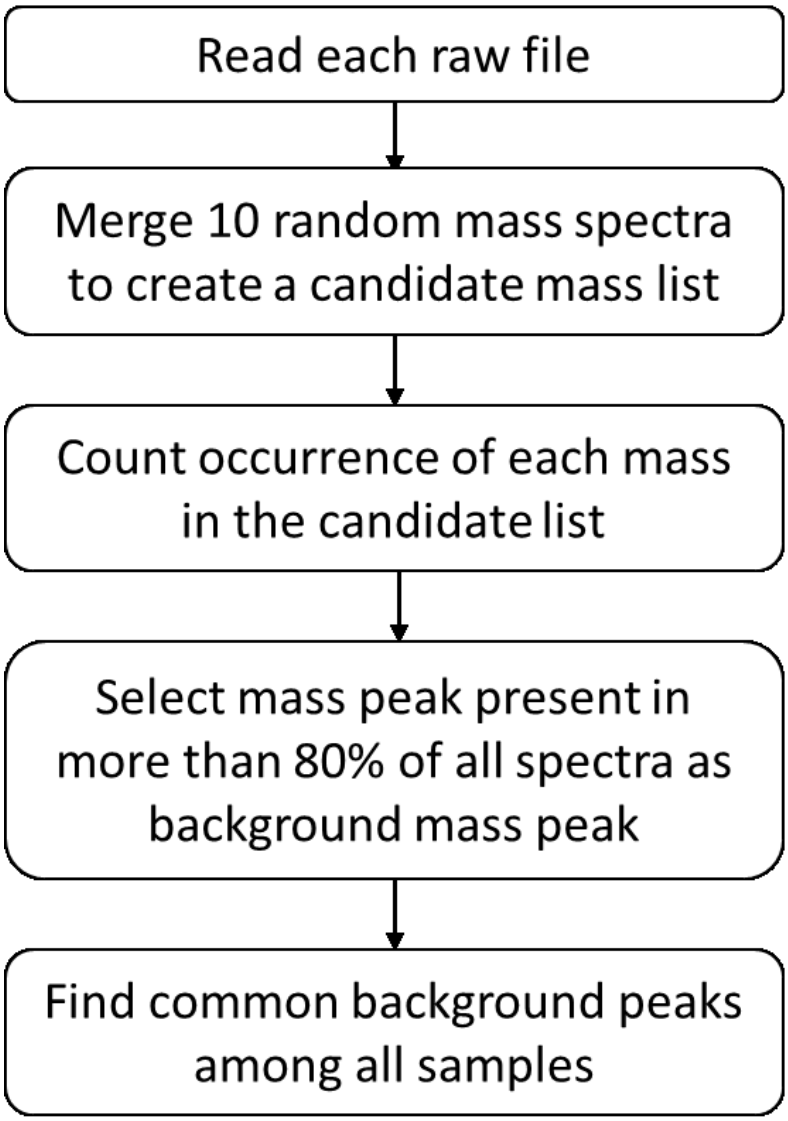


Figure 2.2 Workflow for finding background mass peak.

Based on the candidate mass list, the program next searched each candidate mass against the whole sample data by the user-defined mass tolerance and calculated a total number of occurrences of each mass. Table 2.2 shows an example of the resultant table of background mass candidates found in one QC sample. Each mass in the "average.mz" column was calculated as the averaged measured mass from this file. The standard deviation of measured masses, peak intensity average (average.int), and signal to noise ratio average (average.sn) were calculated for

each mass candidate. The Frequency column shows how many times each mass candidate appears in all spectra, and the Occurrence column gives the percent of each mass candidate showing up in all spectra.

Table 2.2 Example of candidate background peak list (only peaks with over 50% occurrence is shown here).

No	average.mz	Standard deviation	average.int	average.sn	Frequency	Occurrence
1	252.0695	0.0010	21483.33	119.79	1753	0.922
2	251.0846	0.0009	121745.94	442.06	1746	0.918
3	279.1152	0.0009	163049.60	706.43	1669	0.877
4	254.0756	0.0011	23302.88	128.72	1665	0.875
5	283.1286	0.0010	163577.33	713.37	1659	0.872
6	253.0899	0.0012	135623.77	495.52	1572	0.826
7	274.0505	0.0010	11315.31	60.62	1548	0.814
8	217.1040	0.0009	14775.07	58.87	1449	0.762
9	276.0570	0.0011	10739.00	57.34	1444	0.759
10	236.0653	0.0011	10421.51	57.45	1405	0.739
11	261.1304	0.0012	10527.87	43.21	1363	0.717
12	319.1376	0.0014	7201.40	40.01	1353	0.711
13	250.1777	0.0010	16419.76	75.99	1339	0.704
14	228.1958	0.0009	13613.21	59.51	1332	0.700
15	234.0585	0.0011	9326.33	52.63	1312	0.69
16	305.1559	0.0014	8062.06	34.76	1300	0.683
17	284.1314	0.0014	34018.62	147.78	1253	0.659
18	255.0805	0.0016	3408.49	18.81	1245	0.655
19	226.1802	0.0011	22843.41	89.32	1231	0.647
20	256.0730	0.0019	42746.62	216.29	1182	0.621
21	241.0292	0.0010	2486.41	12.12	1116	0.587
22	245.0989	0.0013	3817.82	16.22	1087	0.572
23	259.1163	0.0015	4045.83	16.57	1075	0.565
24	289.1260	0.0016	4069.66	17.48	1064	0.559
25	391.2831	0.0024	43409.93	219.93	1044	0.549
26	215.0888	0.0012	2659.68	11.51	1012	0.532
27	282.2780	0.0012	163544.99	526.87	1002	0.527
28	305.2464	0.0020	2408.09	11.07	997	0.524
29	282.1071	0.0013	117137.28	682.07	994	0.523
30	265.1026	0.0014	8876.73	52.54	993	0.522
31	413.2664	0.0022	103592.30	557.88	978	0.514

Figure 2.3 is a snapshot of the graphical user interface for the function. One can choose the number of raw files to be analyzed and the number of scans in each raw data for creating the mass candidate list. The “Initial scan number” is used to skip the data during the LC dead time, and the “Minimum background peak occurrence” is used to define the occurrence frequency of a background peak. After processing all individual files, the program compared the background peaks in each sample and combined the common peaks into a final background peaks table in the program window (see in Figure 2.3).

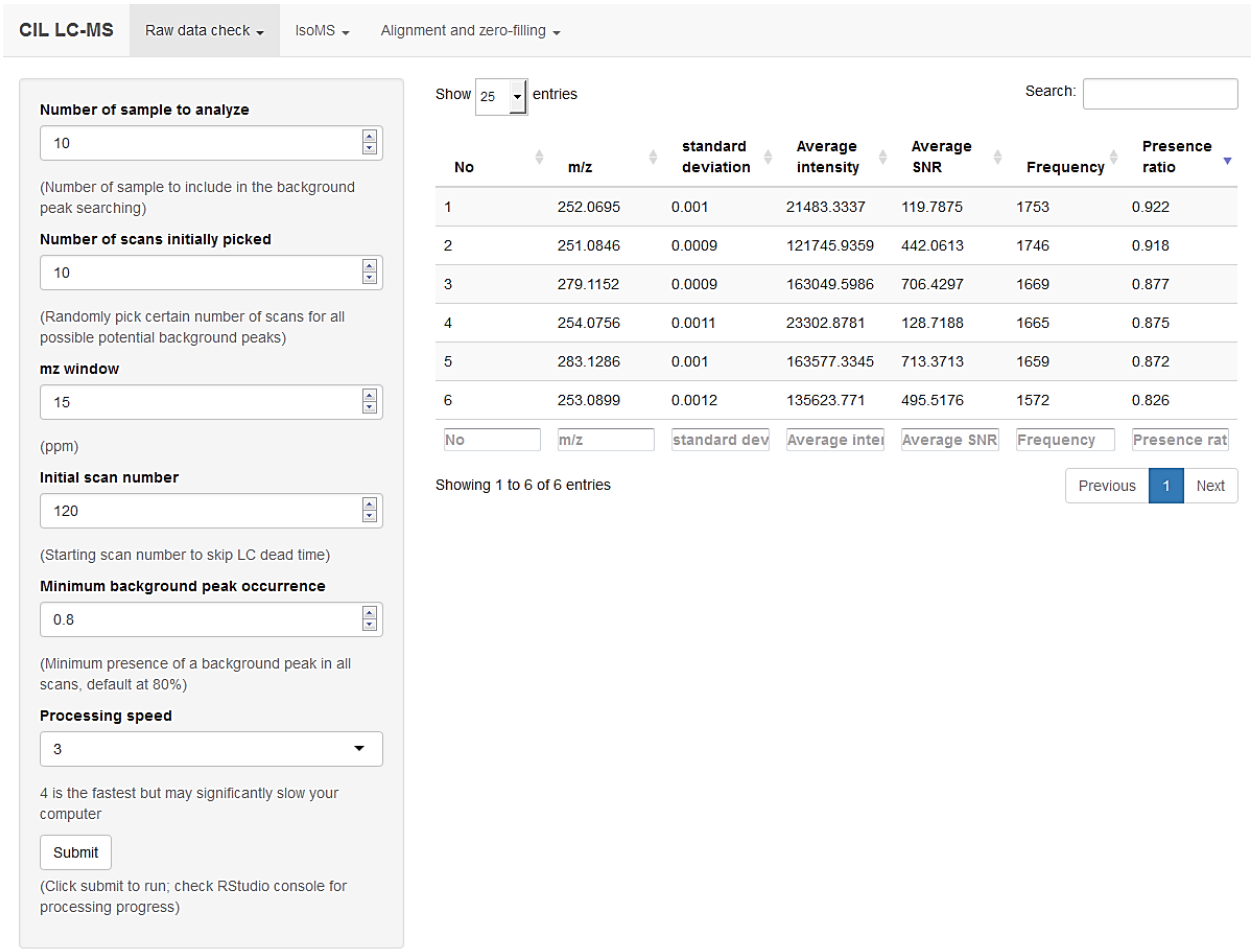


Figure 2.3 Graphical user interface of searching background mass peaks. The column on the left shows the parameters used in the processing. The right side of the window shows the resultant background mass peak list. The table was generated automatically at the end of the program.

2.4.2 Mass accuracy and precision check

The background peak list in Figure 2.3 summarized the mass information of the background peaks found in the whole sample data. Some of these masses can be identified further by searching in the dansyl library[86] using both the accurate mass and the retention time. For identified compounds, their exact masses can be used as internal references to check the mass accuracy and precision. For other unidentified masses, the averaged measured mass can be used to evaluate mass precision.

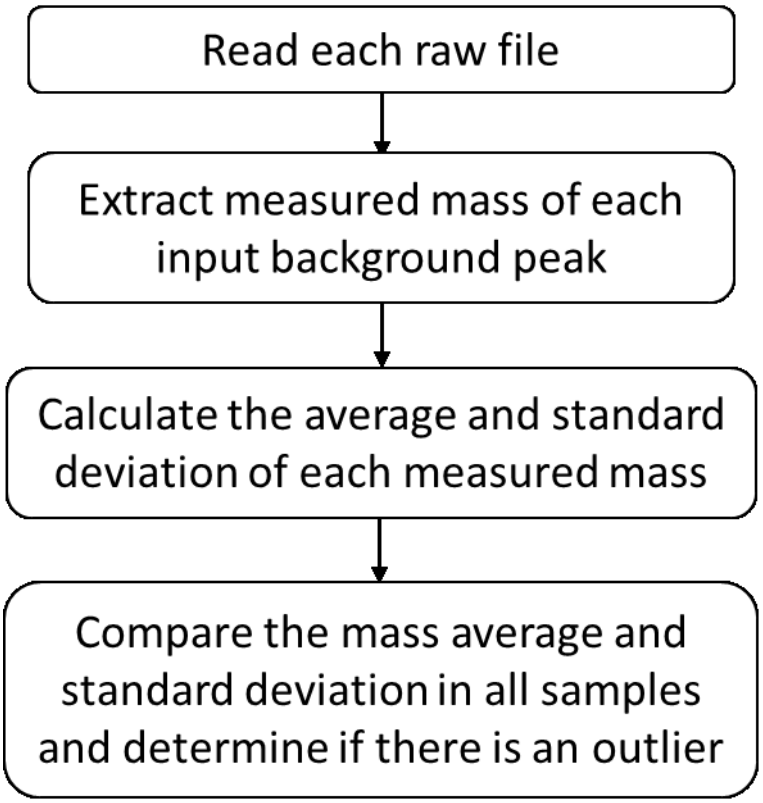


Figure 2.4 Workflow for mass accuracy and precision check.

Figure 2.4 shows the workflow for the mass accuracy and precision check function. I will demonstrate the processing algorithm using dansyl ammonia, which is a common background in the dansylation labeling method. In the user interface shown in Figure 2.5, we input the theoretical m/z of dansyl ammonia, 251.0849, in the reference mass text box; a second optional reference mass also is available. One or two reference masses are sufficient to detect any mass accuracy issue since mass shift usually occurs in the whole mass range.[78] With each input mass, the program extracts its measured mass in each spectrum using the user-defined mass tolerance and generates an extracted mass list. Mass average and standard deviation are calculated for each input mass.

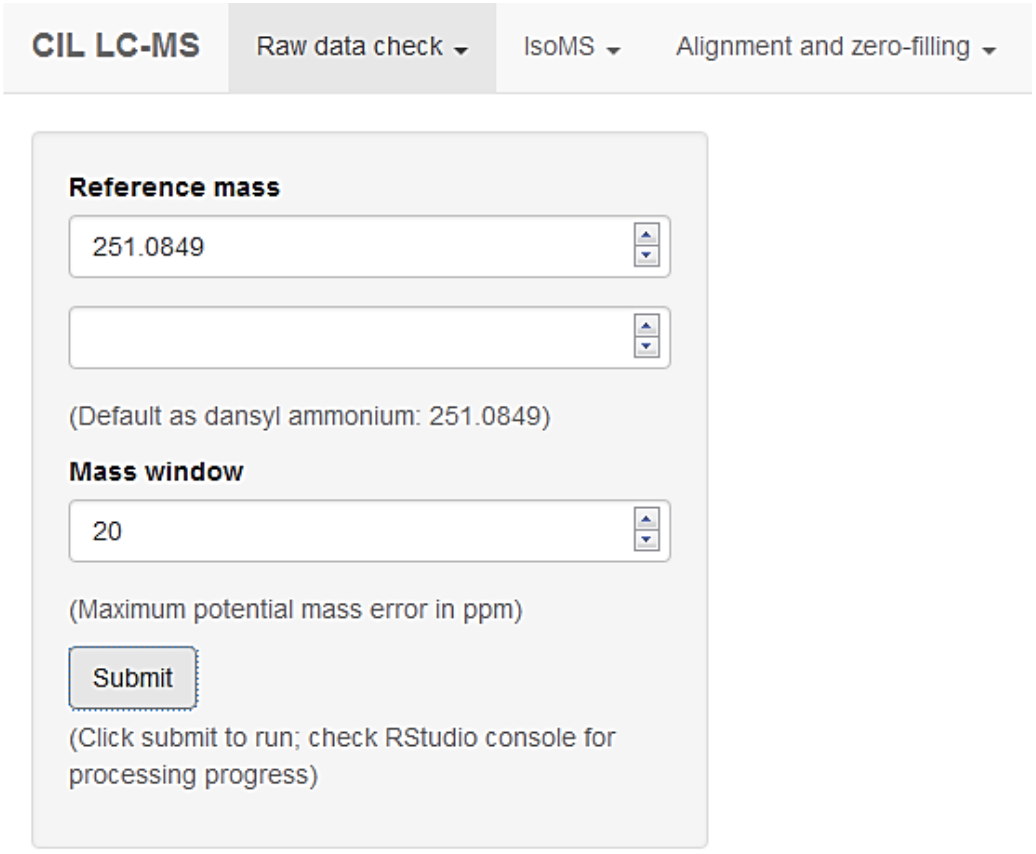
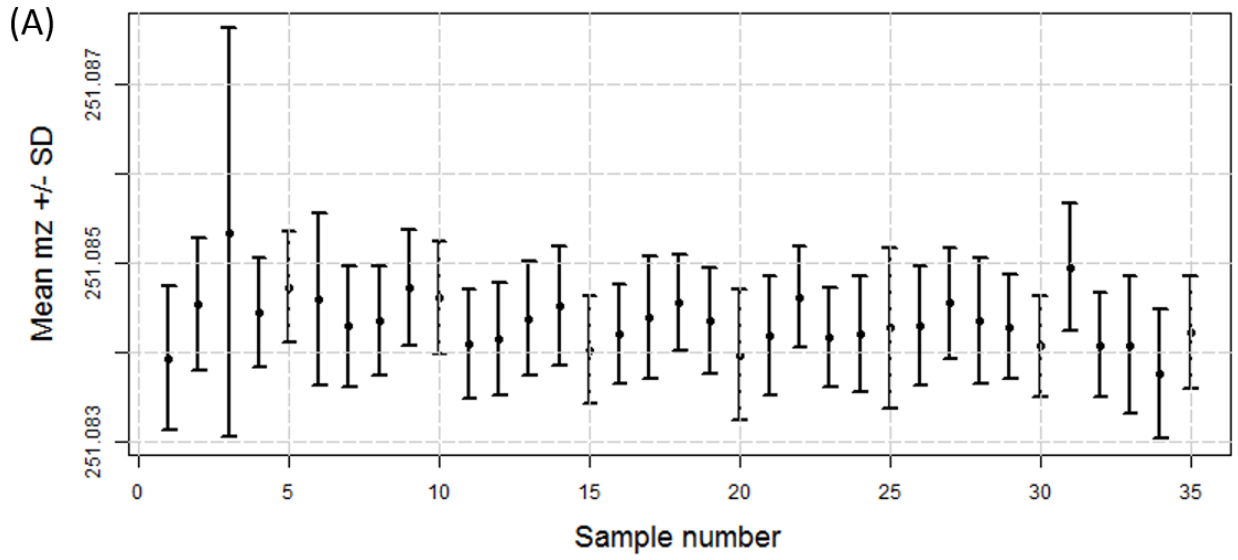


Figure 2.5 Graphical user interface for mass accuracy and precision check.

Figure 2.6 (A) shows the results for the 35 QC samples. The plots were generated in the user interface at the end of the processing. Each data point represented the mass average and standard deviation of the dansyl ammonia peak in one sample. The mass average showed a mass error within 5 ppm (0.0012 Da) compared to the theoretical value of 251.0849. The standard deviation within a sample gave a consistent value in most samples except the 3rd sample. A relatively large standard deviation may indicate a large mass shift within the sample. Figure 2.6 (B) shows the number of dansyl ammonia peaks found in each sample. We can observe that the 5th sample has a significantly smaller value compared to the other samples. This indicates a major mass shift in this sample that caused the measured mass of dansyl ammonia to be outside the mass tolerance in some of the mass scans.



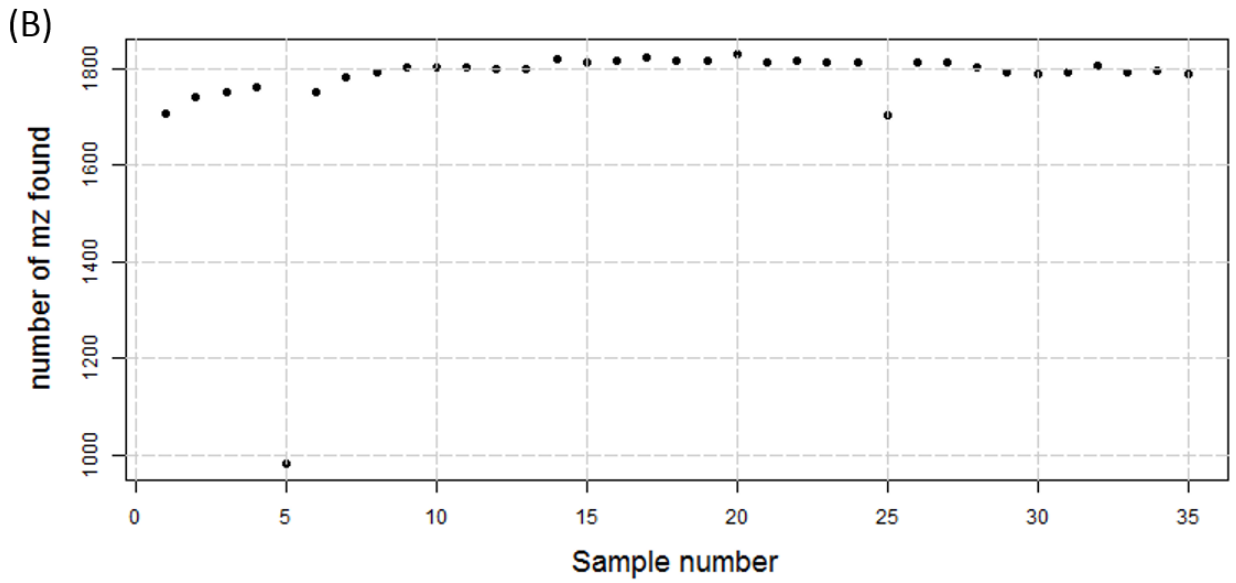
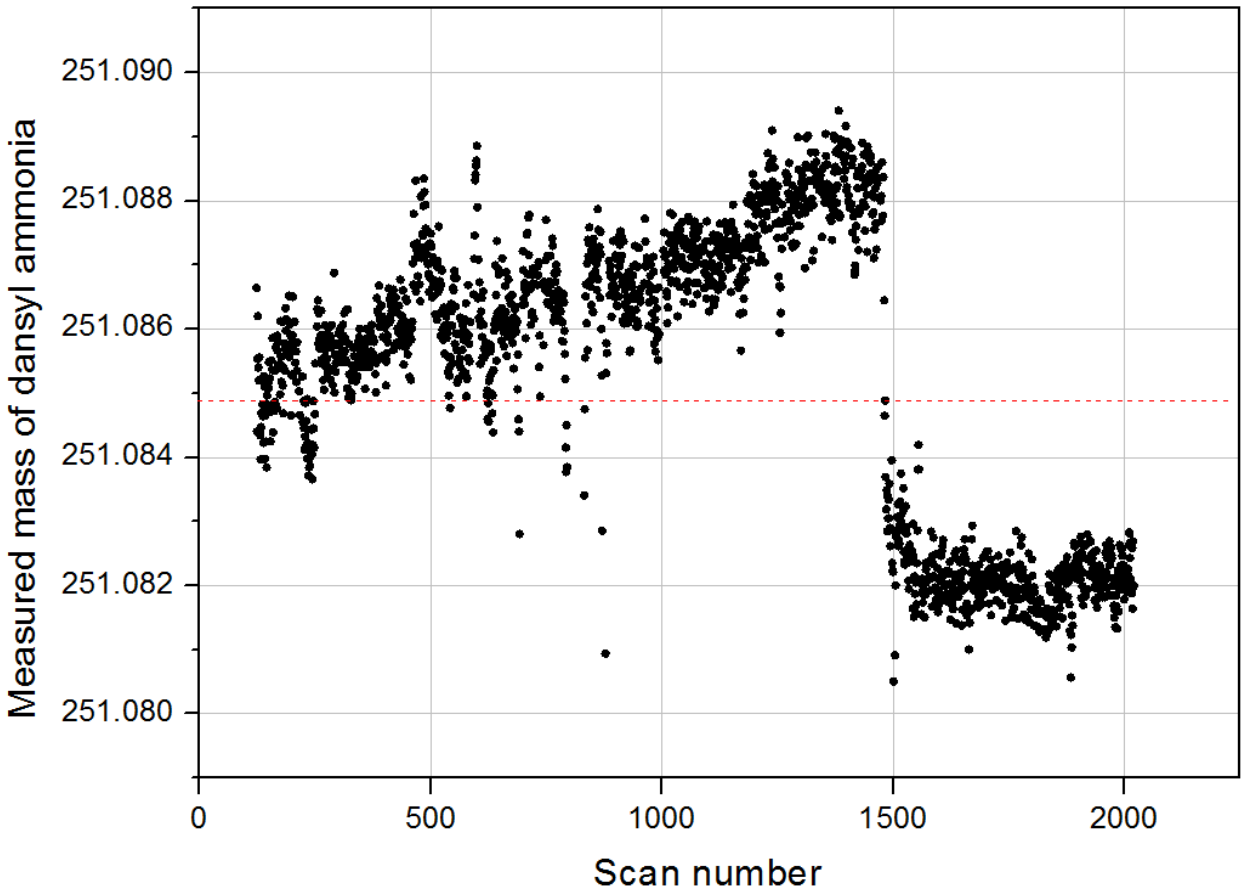


Figure 2.6 (A) Average and standard deviation of the dansyl ammonia peak in 35 samples and (B) number of dansyl ammonia peaks found in each sample data.

From the initial plots, we found two suspicious files that may have mass accuracy issues, therefore, we looked into the extracted peak list of the two specific files. Figure 2.7 (A) and (B) show the measured mass of dansyl ammonia in each spectrum for Files 3 and 5. We can see that the mass shift occurred at scan 1500 in File 3, and the dansyl ammonia peak was missing in File 5 after scan 1200. By checking the original data in File 5, we found that the mass shift in File 5 after scan 1200 exceeded the mass tolerance used in the processing. The large mass shift in these two files was due to the instability of the power supply of the mass spectrometer. Data affected by the mass shift was excluded from further data processing since they are QC samples. For individual sample data, one can re-analyze the sample to replace the data with a mass accuracy issue.

(A)


Measured mass of dansyl ammonia
251.090
251.088
251.086
251.084
251.082
251.080
0
500
1000
1500
2000
Scan number

(B)

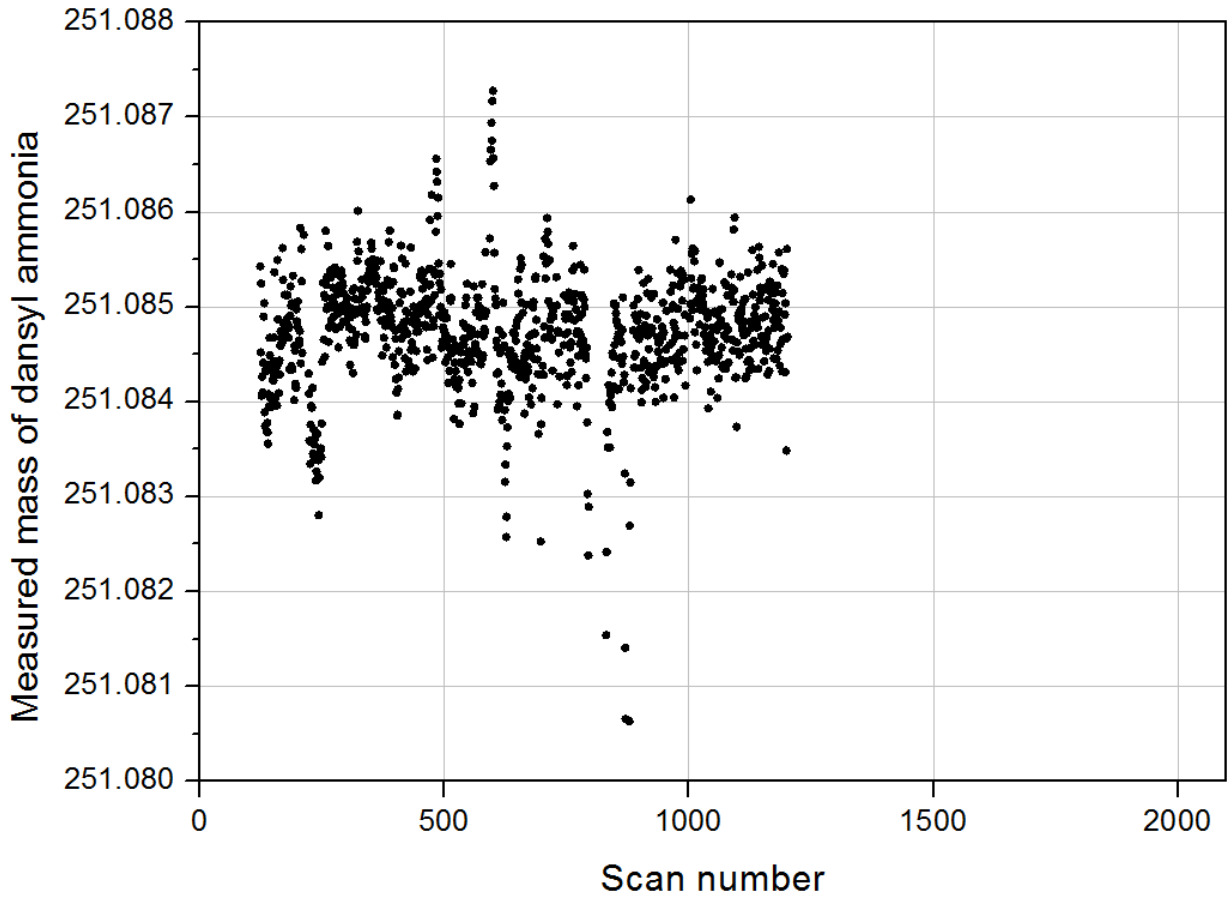


Figure 2.7 Dansyl ammonia peaks in each spectrum in File 3 (A) and File 5 (B).

2.5 Conclusions

I have developed a data processing program that is able to evaluate mass accuracy and precision comprehensively for any type of LC-MS raw data. The program detects the potential background mass peaks in one data set and uses one of the two background masses to check the accuracy and precision of mass measurement in the whole sample data. A graphical user interface was designed to facilitate the use of the program with adjustable parameters. The results of the processing are shown in the program window at the end of the processing with the average and

standard deviation of the background peak and the number of background peaks found in each individual sample. One can pick out any sample file with a mass accuracy issue quickly. A list of detailed extracted mass data is created in the local user folder for further investigation of the issue files. Due to the nature of the background peak of its instable peak intensity, an internal mass calibration cannot be conducted to the issue file. To fix the potential mass accuracy issue, one would re-analyze the same sample and replace the issue file with the new data.

The program has been implemented in the data processing pipeline for CIL LC-MS data processing. The processing speed was optimized at around 10 sec in processing one file (~1800 mass spectra/file). An application of the program was used to show the ability of the processing method to pick out any mass spectrum with a major mass shift. Processing results were shown in plots in the program window to reduce the manual work of the user in dealing with a large number of samples. The program is powerful and efficient for keeping mass data quality in the beginning of the data processing pipeline.

Chapter 3 Retention Time Shift Analysis and Retention Time Correction for Chemical Isotope Labeling LC-MS Raw Data

3.1 Introduction

Retention time and accurate mass are the most important measurements in LC-MS based metabolomics. In Chapter 2, we discussed the methods for mass accuracy and precision check as a control for mass data quality. This Chapter will focus on the retention time shift analysis and the correction of retention time in raw LC-MS data.

As a characteristic parameter of a compound under a specific LC separation method, retention time has been used widely as a criterion in metabolite identification and structure confirmation. In the development of different chemical isotope labeling LC-MS methods and the standards libraries, an optimized HPLC separation method was developed in each labeling method to achieve the separation of a variety of small metabolites, and a high confidence metabolite identification was enabled using both the retention time and the accurate mass.[30,31,47,87,88] The same LC method and separation column have to be used for the same labeling method so that retention time data can be used readily for data alignment and peak identification.

Although unique to the molecular structure, the retention time in an actual application can vary greatly depending on a number of factors, including instrumental setup, column type, elution conditions, etc. For example, tubing volume before and after column, differences of C18 columns due to different manufacturing processes, or even the same mobile phase prepared by different individuals all can introduce retention time errors from a few seconds to over 10 sec.

Other events such, as leaks in the LC system, degradation of the separation column stationary phase over long a period of time, and other instrumental errors potentially can cause a major retention time shift, which can be hard to notice until one manually checks the data files. Therefore, a retention time correction always is needed when comparing two batches of LC-MS data and identifying compounds in the standards library using both accurate mass and retention time.

In addition to metabolite identification, retention time also plays an important role in data alignment. One challenge of data processing for comparative analysis is to match peak pairs that represent the same metabolite from different samples. To overcome the retention time drift from sample to sample, an often-used method is to spike a small number of internal standards during sample preparation. After data collection, signals of the internal standards are identified in the sample data for correcting the retention time shift in the whole dataset.[89] However, this method requires additional steps in sample preparation, and the spiked standards can suppress signals of low concentrations of metabolites easily. A different correction algorithm was developed using data dependent internal peaks for retention time alignment in the XCMS online processing tool.[90] In this method, the "internal standards" were selected from hundreds of peaks that are repeatedly detected in most of the sample files. These peaks had a high probability of being matched in each individual sample and could be used as temporary standards for retention time correction.

In the chemical isotope labeling LC-MS data, we usually can find a list of commonly detected metabolites for the same type of sample. For example, certain amino acids are always present in some biological samples, such as urine or serum. These common metabolites usually have relatively high peak intensities and can be captured accurately in each sample data. The peak pair pattern of these sample-specific compounds also can help exclude many interfering

singular peaks within the mass window of the nominal mass during the search. With the retention time and accurate mass of these common metabolites, we can extract their retention times in raw LC-MS data and check the retention time deviation in all sample files. As a result, a retention time shift correction can be conducted without using extra standards during sample preparation. In this Chapter, we will introduce a program that evaluates the retention time distributions in sample data using the pre-identified internal standards within a biological sample. A retention time correction method was developed for normalizing the retention time in raw LC-MS data files. The correction achieved a better accuracy using the identified internal compounds.

3.2 Materials and Methods

3.2.1 Chemicals and reagents

All the chemicals and reagents, unless otherwise stated, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). In the dansylation labeling reaction, the ^{12}C-labeling reagent (dansyl chloride) was from Sigma-Aldrich, and the ^{13}C-labeling reagent was synthesized and purified in our lab using the procedure published previously.[30] LC-MS grade water, methanol, and acetonitrile (ACN) were purchased from ThermoFisher Scientific.

3.2.2 Human urine collection

Human urine samples were collected from six healthy individuals under the Ethics Approval from the University of Alberta. Each urine sample was centrifuged at 14,000 rpm for 10 min, and the supernatant was filtered twice through a 0.22 μm filter. A pooled urine sample was prepared

by mixing all the individual samples by equal volume. A total of 276 human urine samples were collected. The filtered urine was aliquoted and stored at -80 °C until further use.

3.2.3 Dansylation labeling

The frozen urine samples were thawed in an ice-bath and then centrifuged at 14,000 rpm for 15 min, and 25 μL of supernatant were transferred into an Eppendorf for the labeling reaction. Next, 25 μL of 250 mM sodium carbonate/sodium bicarbonate buffer and 25 μL of ACN were added. The solution was vortexed, spun down, and mixed with 50 μL of freshly prepared ^{12}C-dansyl chloride solution (18 mg/mL, for light labeling) or ^{13}C-dansyl chloride solution (18 mg/mL, for heavy labeling). After 45 min incubation at 40 °C, 10 μL of 250 mM NaOH were added to the reaction mixture to quench the excess dansyl chloride. Then, the solution was incubated at 40 °C for another 10 min. Finally, 25 μL of formic acid (425 mM) in 50/50 ACN/H_2O were added to consume excess NaOH and to make the solution acidic. The ^{12}C- or ^{13}C-labeled sample was centrifuged at 14,000 rpm for 10 min before injecting onto LC-UV for quantification.[91] For LC-MS analysis, the ^{12}C- and ^{13}C-labeled samples were mixed in equal amounts based on the quantification results.

3.2.4 LC-MS analysis

The ^{12}C- and ^{13}C-labeled samples were mixed and centrifuged at 20,800 g for 10 min before injecting into a Bruker Maxis Impact QTOF mass spectrometer (Billerica, MA, USA) linked to an Agilent 1100 HPLC system (Palo Alto, CA, USA). A Zorbax Eclipse Plus C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1%

(v/v) LC-MS grade formic acid in 5% (v/v) LC-MS grade CAN, and solvent B was 0.1% (v/v) LC-MS grade formic acid in LC-MS grade ACN. The gradient elution profile was as follows: t=0.0 min 20% B, t=3.5 min, 35% B, t=18.0 min, 65% B, t=24 min, 99% B, t=28 min, 99% B. The flow rate was 180 μL/min.

3.3 Results and Discussion

3.3.1 Retention time shift analysis

Peak pair matching in data alignment usually allows a retention time tolerance from a few seconds to tens of seconds to account for retention time shift from sample to sample. If the overall retention time shift is well within the tolerance value over the chromatographic profile for all samples, a retention time correction may not be necessary for data alignment. At the beginning of the processing, we checked the retention time in all sample data to determine if any major retention time deviation occurs.

For the same type of sample, certain compounds always will show up in the data with relatively high peak intensities. Such “well-behaved” peaks can be used as the temporary internal standards for checking and correcting the retention time. In this experiment, we generated data from the human urine samples in which a list of commonly detected amino acid standards are present. These compounds were identified previously as common metabolites in a human urine sample. Table 3.1 shows 10 such compounds that were selected for checking the retention time in each individual sample. Their retention times are distributed evenly over the significant portions of the chromatographic profile. We tried to avoid isomer compounds in the internal

standards selection, such as leucine and isoleucine, as they are close in retention time and peak intensity (see Figure 3.1).

Table 3.1 Ten amino acids used as retention time internal standards for checking retention time in each sample file.

Name	mz_light	mz_heavy	RT	#Charge	#Tag
Arginine	408.1700	410.1767	2.51	1	1
Glutamine	380.1275	382.1342	3.32	1	1
Threonine	353.1166	355.1233	5.37	1	1
Alanine	323.1060	325.1127	6.90	1	1
Proline	349.1217	351.1284	9.23	1	1
Methionine	383.1094	385.1161	9.86	1	1
Phenylalanine	399.1373	401.1440	11.67	1	1
Cystine	354.0703	356.0770	13.04	2	2
Lysine	307.1111	309.1178	16.09	2	2
Tyrosine	324.5953	326.6020	21.23	2	2

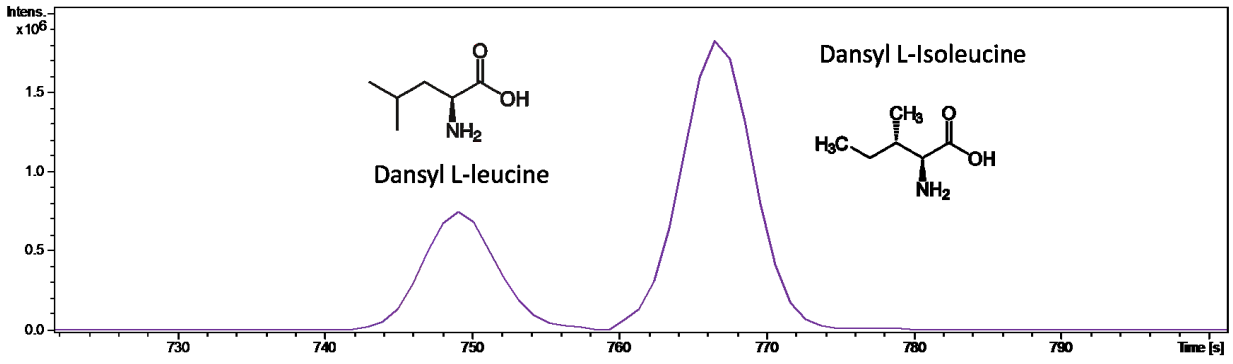


Figure 3.1 LC chromatographic peaks of dansyl leucine and dansyl isoleucine. The dansyl group is omitted in the structures.

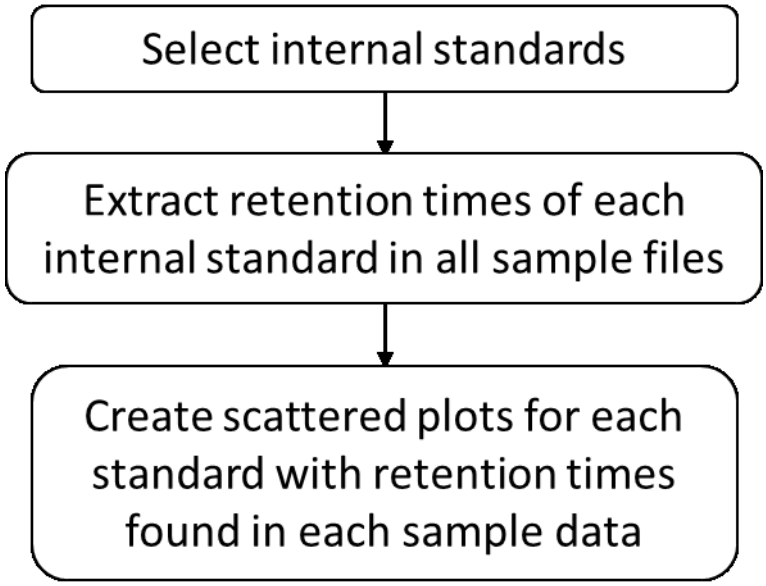
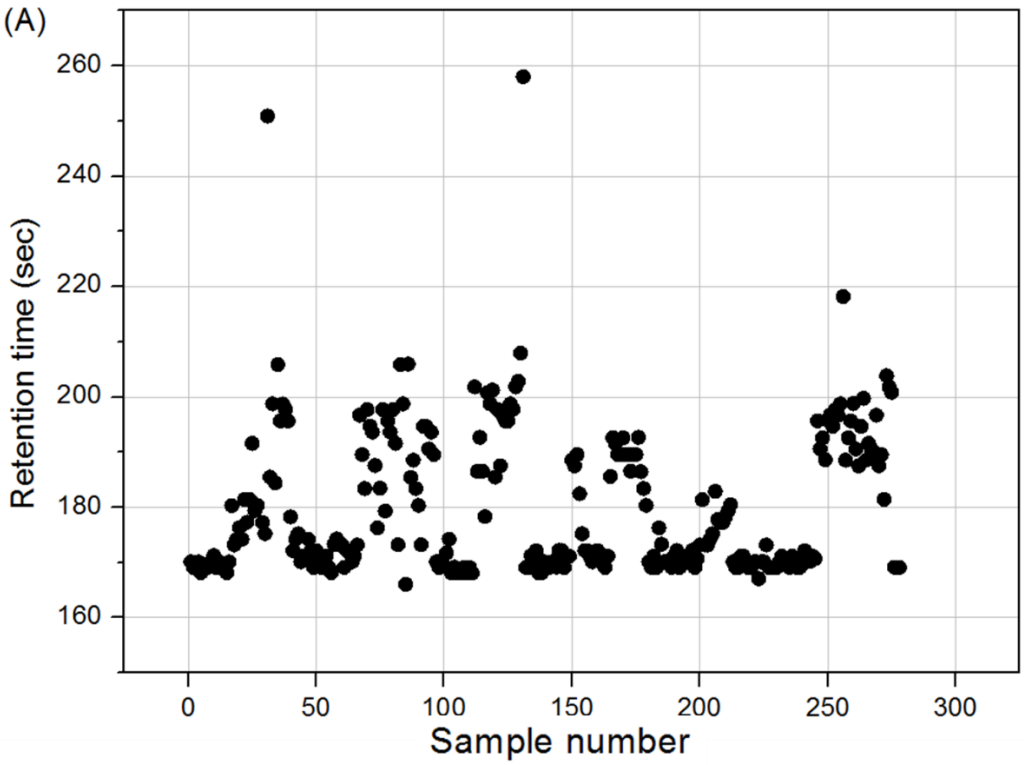


Figure 3.2 Workflow of retention time shift analysis.

Figure 3.2 shows the overall workflow of the retention time check function. With the selected internal standards in Table 3.1, the program extracts the retention times of each standard in each sample file. In different individual samples, the concentration of the standards may vary, and it is possible that the signal of one internal standard is relatively low in some of the samples. Fortunately, the heavy peak of each standard will provide a more stable intensity as it is from the pooled sample. The program searches the heavy peak of each standard in the sample data and checks the light peak in the same mass spectrum to confirm the peak pair. Once the peak pairs of the standard are found, the program records the retention time of the standard using the highest intensity heavy peak. At the end of the processing for all sample files, a plot is generated in the program window with the retention times extracted for each input standard. Figure 3.3 shows the results of the retention time analysis from three selected standards. Each data point in the plot represents a retention time of the standard in one sample. The three standards show the retention

time shift at different regions of the chromatographic profile. From the results, we can observe a significant retention time delay in some of the samples. The retention time shift is up to nearly two min, which is extreme for the LC system. It was found later that due to the aging of the injection needle and needle seat in the LC, a minor leak randomly occurred during the data collection and resulted in a decreased flow rate. Although retention time increased for some of the samples, peak intensities in these samples did not decreased significantly, and the peak pair ratio can still be useful with the heavy peaks from the pooled sample as internal standards. To make use of the collected data, we will correct the retention time shift before further data processing.



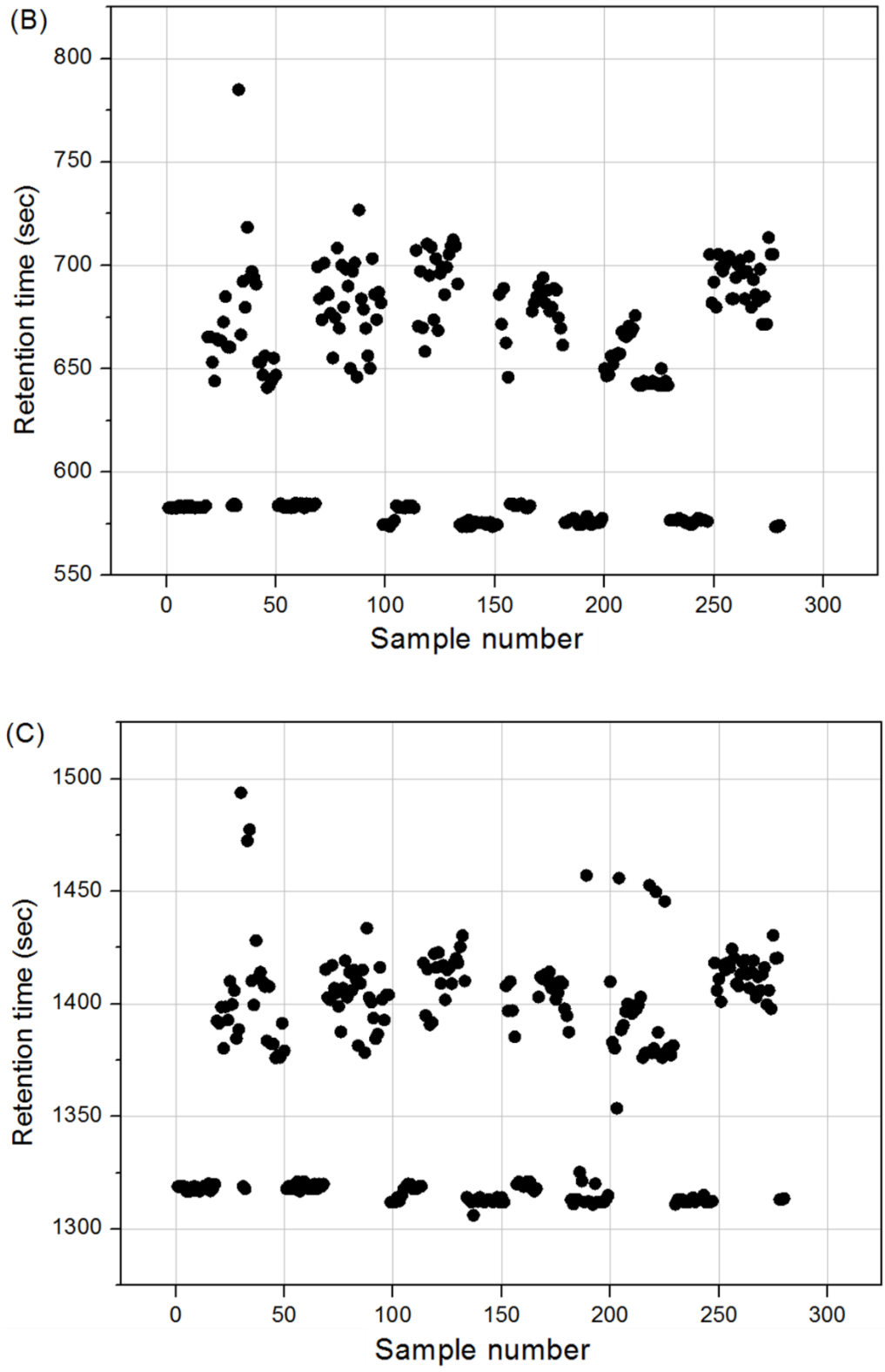


Figure 3.3 Retention times of (A) dansyl arginine, (B) dansyl proline and (C) dansyl lysine in in each sample.

3.3.2 Retention time correction using segmented linear calibration

The retention time shift analysis results showed a large retention shift in some of the human urine data, however, the relationship between the retention time deviation and the retention time points is still unknown. To study the retention time shift pattern within a sample, we selected one sample file as the reference file and calculated the retention time differences of other sample files against the reference file. A list of 13 dansyl labeled standards (see Table 3.2) were selected for the retention time difference calculation over the chromatographic profile. Figure 3.4 shows the retention time difference of one sample compared to the reference file. We can observe that the relative retention time shift is not constant at different retention time points but has an increasing trend over the profile. Many studies[90,92-94] also have pointed out that deviation of retention time is not a linear relationship with retention time since the shift is a result of multiple independent factors, as discussed in the introduction of this Chapter. Other studies also demonstrated that the shift of retention time cannot be approximated well by a quadratic function or other higher-order polynomials.[90] Therefore, it is not possible to use one fitting function and accurately predict the retention time shift over the whole chromatogram. To correct the retention time shift across samples better, our group developed a segmented correction strategy that divided the retention time into multiple segments and used a local linear calibration method for correcting retention time data.[86] The method was designed for metabolite identification in the dansyl library by correcting retention differences between aligned sample data and library standard data. In this work, we updated the algorithm and applied it for retention time correction in the LC-MS raw data for the first time.

Table 3.2 Retention time reference table. Thirteen dansyl labeled standards were selected for retention time correction. Their retention times were extracted from the reference file as the reference retention time for the correction of retention time in other samples.

No	mz_light	mz_heavy	RT	nCharge	nTag
1	408.17	410.1767	2.82	1	1
2	339.1009	341.1076	4.86	1	1
3	353.1166	355.1233	6.29	1	1
4	323.106	325.1127	7.99	1	1
5	349.1217	351.1284	10.7	1	1
6	383.1094	385.1161	11.43	1	1
7	399.1373	401.144	13.25	1	1
8	354.0703	356.077	14.613	2	2
9	307.1111	309.1178	17.864	2	2
10	389.1281	391.1343	18.48	1	1
11	324.5953	326.602	22.97	2	2
12	354.1159	356.1222	25.3	1	1
13	611.1222	615.135	28.23	1	2

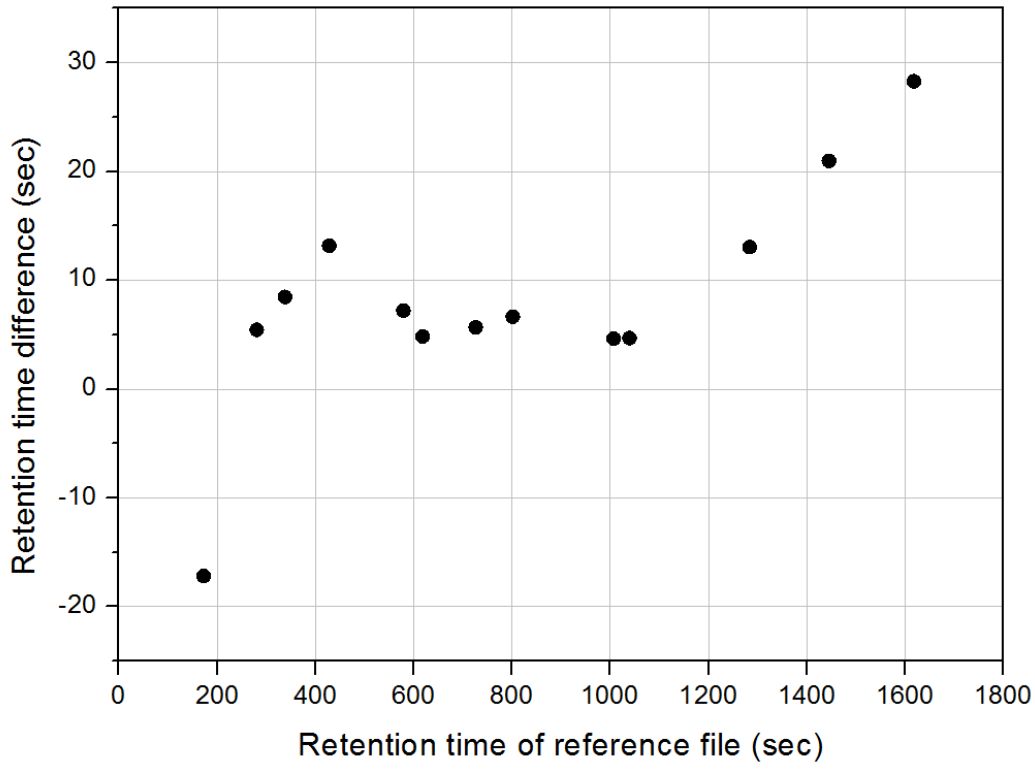


Figure 3.4 Retention time shift in a sample file compared to a reference file.

Figure 3.5 shows the overall workflow of the retention time correction function. To correct the retention time in the 278 human urine data files, we selected one data file as the reference file to correct the retention times in all other files. The retention time of the reference standards were extracted from the reference file first; Table 3.2 shows the details of the reference table. Figure 3.6 shows the extracted ion chromatograms of these standards in the reference file, and all have a relatively high peak intensity in the file. The reference retention times were then passed to the program for retention time correction in sample files.

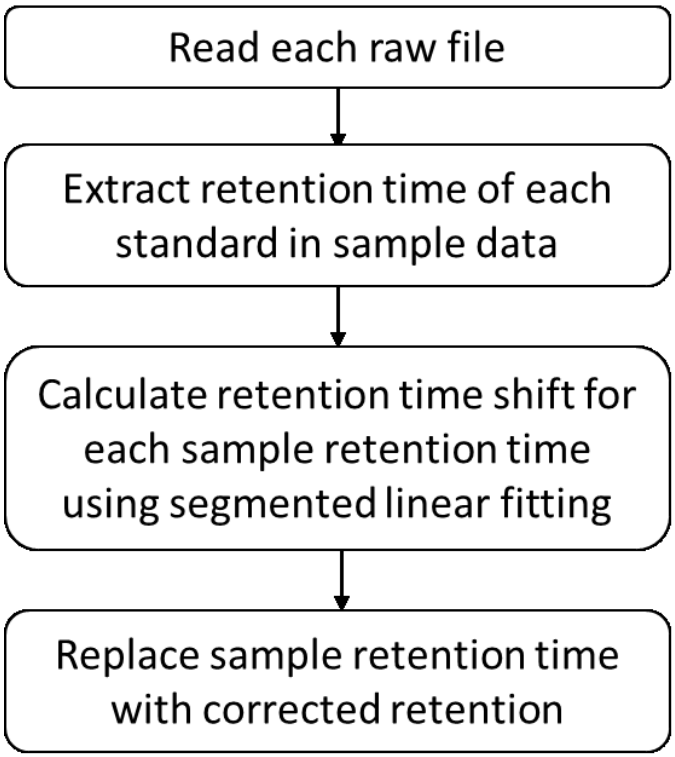


Figure 3.5 Workflow of retention time correction in raw LC-MS data.

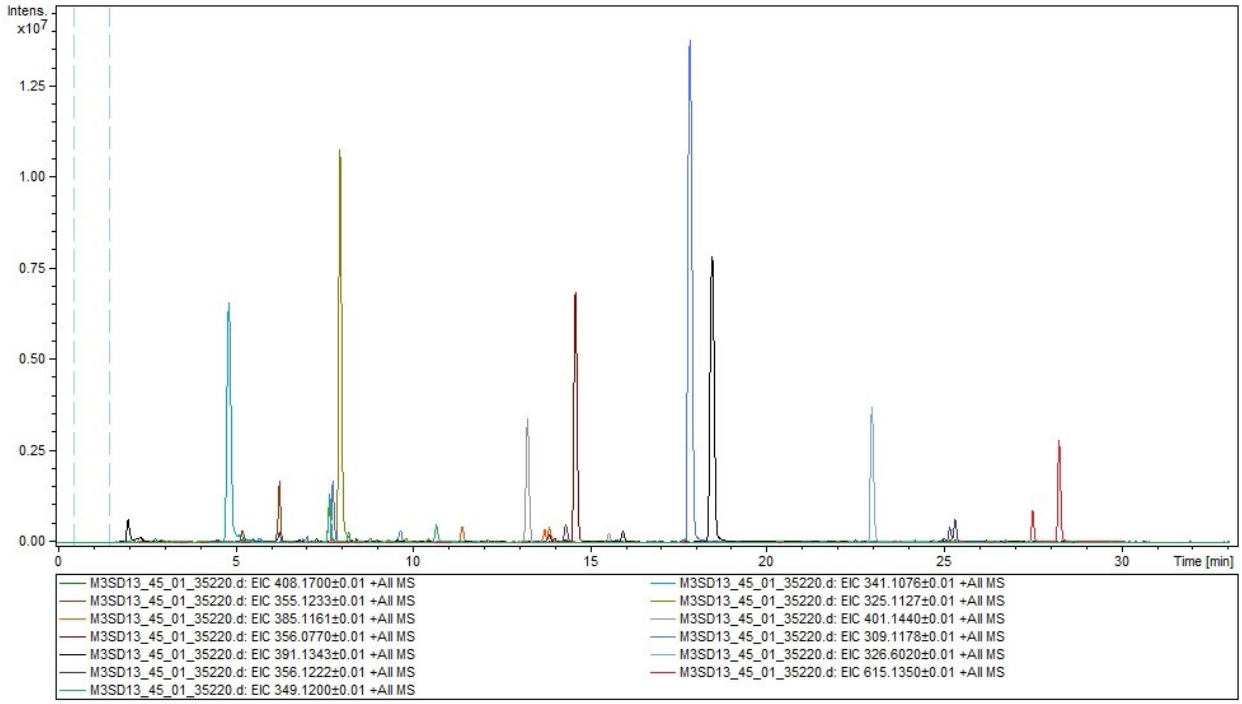


Figure 3.6 Extracted ion chromatogram (EIC) of 13 internal standards in reference file. Retention time of each standard is used as reference in retention time correction.

Figure 3.7 shows a schematic of the retention time correction algorithm. The correction is conducted on one sample file at a time. The program reads each sample data file and extracts the retention time of each reference standard. As a result, each reference standard has one reference retention time t_{ref} from the reference file and one sample retention time t_{smp} from one sample file. At the i^{th} standard (i is any integer between 1 and 13), the retention time difference of sample to reference is calculated as $\Delta t_i = t_{smp,\,i} - t_{ref,\,i.}$ For the retention time between standard i and standard i+1 (t_i to t_{i+1}), we applied a linear interpolation to calculate the retention time difference Δt_x by,

$$\Delta t_x = \Delta t_i + [(\Delta t_{i+1} - \Delta t_i) / (t_{smp,\,i+1} - t_{smp,\,i})] * (t_{smp,x} - t_{smp,i}) \qquad (3.1)$$

where Δt_x is the retention time shift at any retention time point in the sample data between t_i to t_{i+1}; $t_{smp,\,i}$ is the retention time of i^{th} standard in the sample data and $t_{smp,\,x}$ is the retention time in the sample where Δt_x is measured.

For retention time segments before the first standard and after the last standard, only one standard is available, so the equation above cannot be applied directly. For the retention time before the first reference standard, we assumed a zero retention time shift at time zero and a hypothetical standard at t=0 was added in each standard table. For the retention time after the last reference standard, we assumed that the retention time shift is relatively constant in this range as the mobile phase composition is fixed. The retention time shift at the last reference standard point was applied to the data after the last reference standard. In conclusion, the retention time shift is calculated as a piecewise linear function of sample retention time by,

$$RT_{cor} = RT_{smp} - \mathrm{f}_{cor}\,(RT_{smp}) \qquad (3.2)$$

where $\mathrm{f}_{cor}(\mathrm{x})$ is the piecewise function described in Equation (3.1) for the calculation of the retention time shift at each segment. RT_{smp} and RT_{cor} are the retention time before and after the correction, respectively.

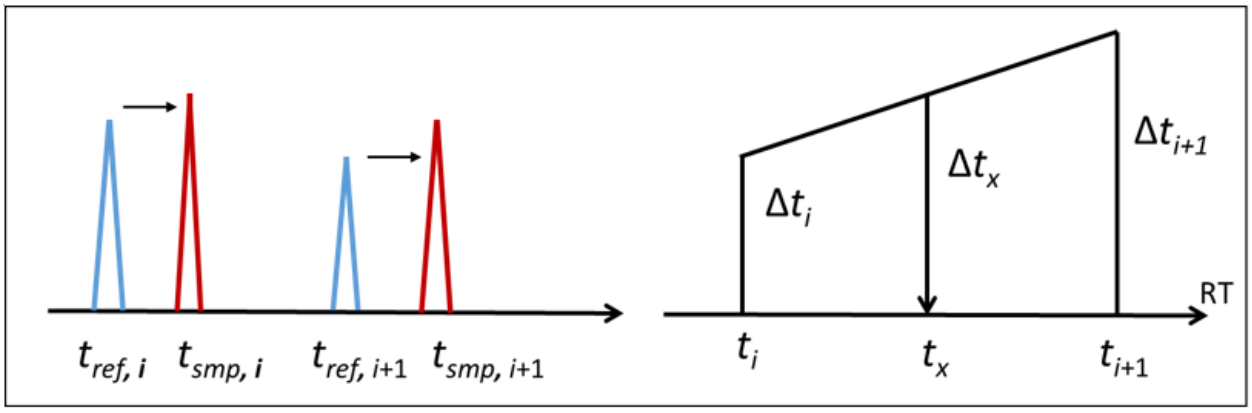


Figure 3.7 Schematic of the retention time calibration method. $t_{ref,\,i}$ and $t_{smp,\,i}$ refer to the retention time of the i^{th} standard in the reference file and sample file, respectively. Δt_i and Δt_{i+1} refer to the retention time shift at the i^{th} and $i+1^{th}$ standard from the sample data to the reference data. Red color peaks are the sample peaks and blue color peaks are the reference peaks.

The selection of the reference standards is critical to the correction accuracy. Each retention time reference standard must be a peak pair feature and have a relatively high intensity in most sample data. The high peak intensity of the standard ensures the accuracy of the retention time extraction as there might be other interfering signals within the mass window. One also should check each of the selected standards in the chromatogram and see if there are isomers with a similar peak intensity in the nearby retention time. An isomer compound can cause miss-picking of the retention time extraction. In the example shown in Figure 3.1, leucine and isoleucine have the same accurate mass and are close in retention time. One can hardly differentiate the isomers using solely the accurate mass and an input retention time window.

The linear interpolation is an approximation for the data points between two standards. The list of reference standards should be distributed evenly over the chromatographic profile of data to achieve better accuracy in the correction. A closer distance between two adjacent standards will increase the accuracy of the predicted retention time shift.

In searching for the retention time of each standard in the sample data, we sometimes found that one of the standards' peak pairs is missing from some of the sample data. This will cause an error in Equation (3.1) due a data point for the linear function. To address this issue, whenever there is a missing standard, the program deletes the standard both in the reference and sample standards list; the retention time segment is then extended to the next available standard. In this way, even with one or two samples with one standard missing, the retention time correction still can calculate the retention time shift to complete the correction process. Each missing standard will be recorded in the processing log. One can adjust the reference standards list to re-correct the retention time in a specific file.

3.3.3 Calibrated retention time evaluation

We applied the retention time correction for a total of 278 human urine data. The processing speed is optimized at 20 sec/file (~1800 spectra/file). To evaluate the retention time after correction, the program automatically generated the overlay plots of base peak ion chromatogram (BPC) for each individual sample before and after the correction. Figure 3.8 shows one example of the 278 overlay plots. The black colored chromatogram is the BPC of the reference file, and the red and blue colored chromatogram are the data before and after the retention time correction, respectively. We clearly can see that the retention time shift was corrected to a much smaller deviation after applying the correction algorithm for the high intensity peaks. The corrected retention time difference is now within a 30 sec window compared to a two min shift before correction. One can go through the 278 overlay plots quickly to check the correction results of each individual sample.

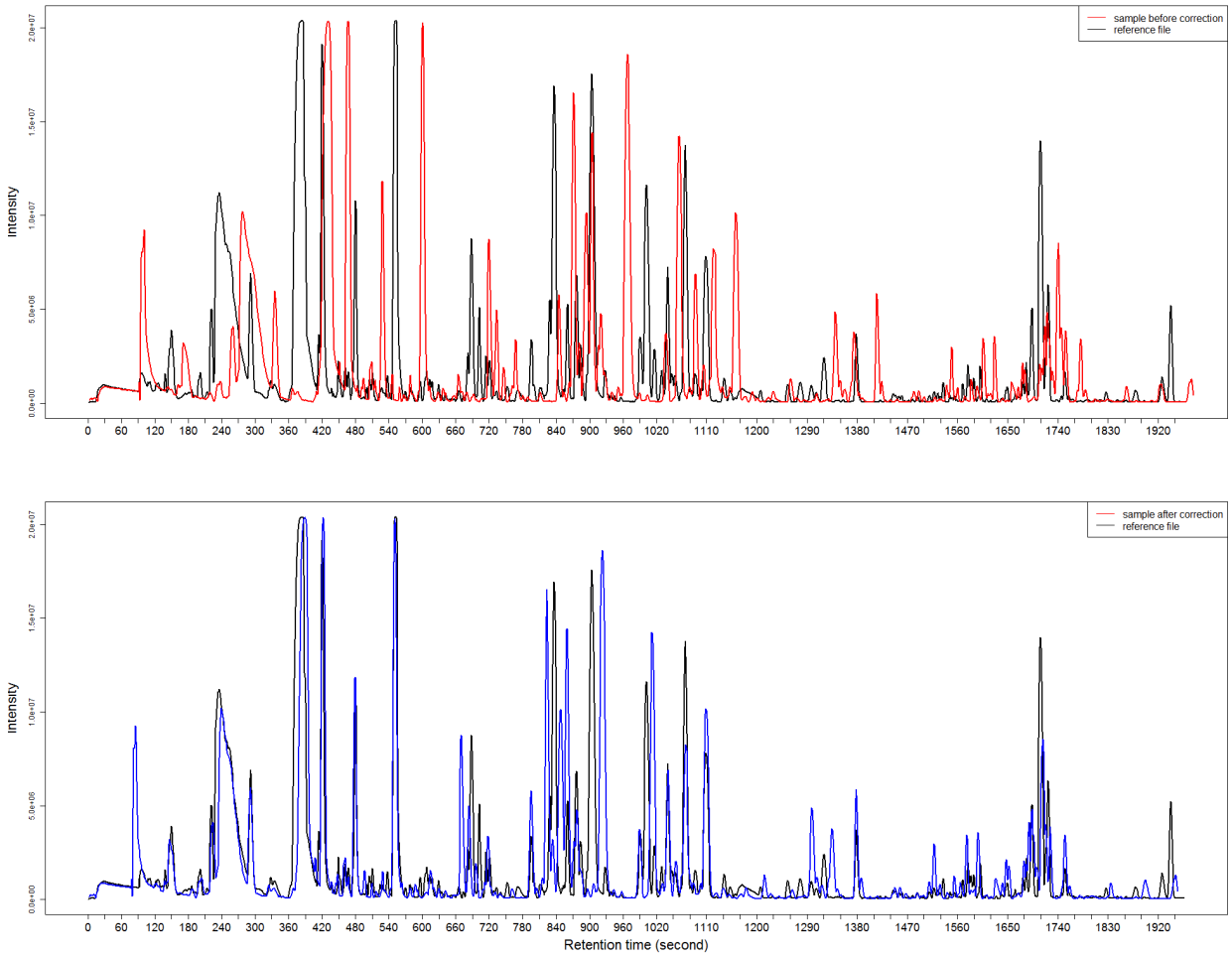


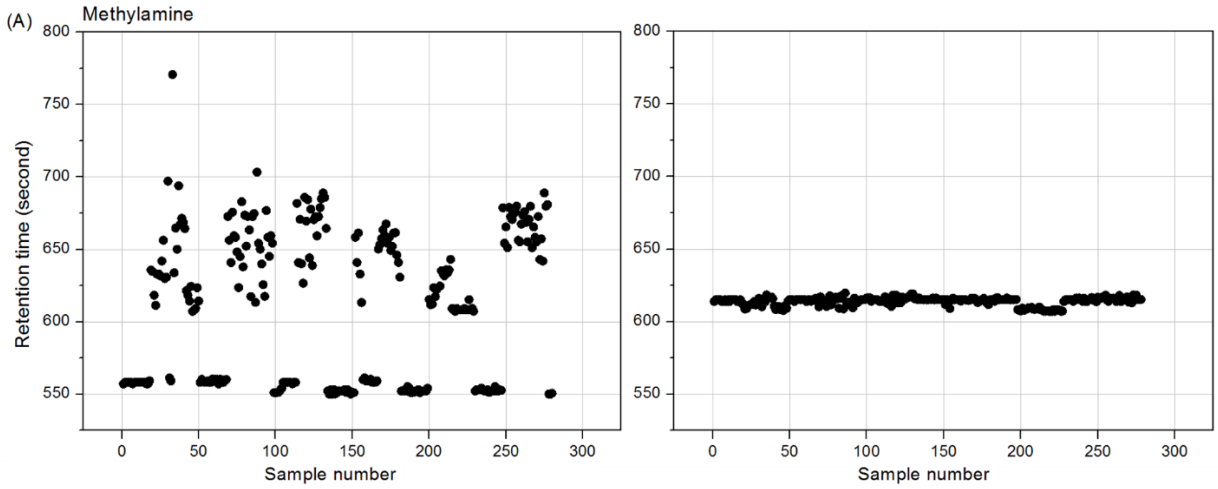
Figure 3.8 Overlay plots of BPC of one file before (black) and after (red and blue) retention time correction against the reference file.

To test the performance of the correction algorithm on the 278 sample files further, we picked another three commonly detected metabolites (see Table 3.3) as testing standards that are not in the correction standard list. We used the retention time and accurate mass of the testing standards and extracted their measured retention time in each data file before and after the retention time correction. The testing results are summarized in Figure 3.9. On the left are the retention times of the testing standards in the original LC-MS data, and on the right are the

corrected retention times. We can observe that after correction, the retention time of the testing standards fell into a ± 30 sec window compared to a two min window before the correction.

Table 3.3 Testing standards used to evaluate the retention time shift after applying retention time correction.

No	Name	mz_light	mz_heavy	RT	nCharge	nTag
1	Methylamine	265.1011	267.1074	11.2	1	1
2	Dimethylamine	279.1168	281.1223	16.2	1	1
3	Cystine isomer	354.0654	356.0720	22.9	2	2



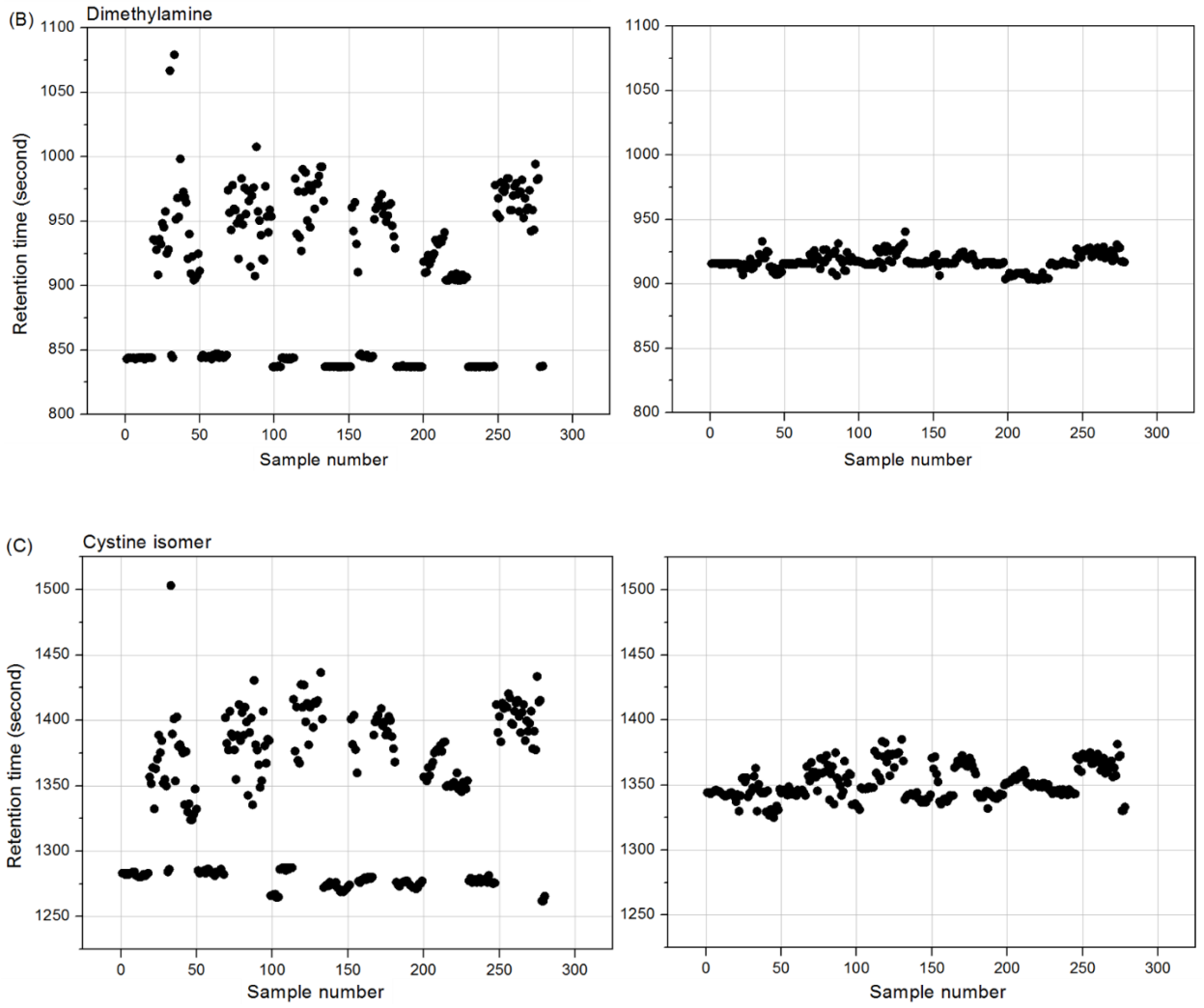


Figure 3.9 Retention time of three testing standards before and after retention time correction.

3.4 Conclusions

In this work, we developed a program that checks and corrects the retention time in the LC-MS raw data. A list of commonly detected metabolites was used as internal standards to extract the retention time over the chromatographic profile in each sample data. The analysis results were shown in the scattered plots to enable a quick view of the retention time distributions at different retention time points. The method provides a fast and convenient way for one to

comprehensively evaluate the retention time shift over the LC profile in all sample data. Any major retention time shift can be picked out easily with the results. The retention time distribution plots also can assist researchers in selecting an appropriate retention window for data alignment. The retention time check program in combination with the mass accuracy check program ensure the raw data quality before the downstream data processing.

In addition to the retention time check, a retention time correction program was developed for the correction of retention time shift in LC-MS raw data. The correction method used the internal standards picked from the commonly detected metabolites in a biological sample and enabled an internal calibration without adding extra standards in the sample preparation. We applied the correction method to a human urine dataset with a large retention time shift. The performance of the correction was evaluated using external testing standards to show the accuracy and efficiency of the correction program.

Chapter 4 Integrated Data Processing Workflow for Generating a Complete Metabolite Intensity Table in Differential Chemical Isotope Labeling LC-MS for Quantitative Metabolomics

4.1 Introduction

Chemical isotope labeling (CIL) LC-MS has become a powerful platform for quantification of metabolites in metabolomics. With the growing number of labeling methods to analyze chemical-group based submetabolomes for increasing metabolome coverage,[30,87,88,95] it becomes more demanding for rapid data processing workflow to generate a complete quantitative dataset with better data integrity and accuracy. Chapter 2 and 3 dealt with the methods for LC-MS raw data quality check, focusing on mass accuracy and retention time shift. After examination, the raw data require further processing for peak pair extraction and data alignment for comparative analysis. In recent years, our group has developed a number of data processing programs[68,73,74] that worked for different chemical isotope labeling LC-MS experiments. With continuing advances in metabolomics, along with constant user feedback, it became necessary for a substantial upgrade to the current processing workflow with both algorithm updates and new function integrations.

The ultimate goal of CIL LC-MS raw data processing is to generate a complete data table with each labeled metabolite accurately picked from LC-MS raw data and with each peak pair in the final data table representing one unique metabolite. The data processing relies on a set of criteria in the program for finding and matching peak pairs in multiple sample data. Due to the

complexity of signals from a real sample, some of the peak pair patterns may not be truly from a labeled metabolite and one labeled metabolite could be repeatedly picked in the data table. We collected sample data from different users and looked into each of these issues. Two new functions, "peak pair validation" and "redundant peak pair merging", were proposed and integrated in the data processing workflow to deal with false peak pairs and redundant peak pairs. These methods further increased the specificity of each peak pair feature as a linkage to a unique labeled metabolite.

The peak pair intensity ratio is the foundation of quantification in a chemical isotope labeling experiment as it is used to reflect the relative concentration change of one metabolite in different samples. To account for the concentration variation in different metabolome samples, the light labeled individual sample and the heavy labeled pooled sample are normalized and mixed in the same total amount prior to injection to LC-MS.[91] In the data processing step, the peak pair ratio has been calculated simply by using the peak intensity of the light and heavy peak within a mass spectrum. With the use of a heavy labeled pooled sample, the intensity ratio of one peak pair in the same sample can be very reproducible regardless of absolute peak intensity changes. However, a measurement error can be associated with the ratio data due to the uncertainty of each individual peak intensity. Instead of using the peak pair ratio in one spectrum, we updated the ratio calculation by using all peak pair signals detected in multiple spectra for one metabolite and used the average of the peak pair ratios to fill the alignment data table to improve the ratio accuracy further.

Although most of the peak pair ratios can be calculated after a thorough inspection of LC-MS raw data, a missing value still can occur in the resultant data table. These missing data could have a great influence on the conclusions drawn from different data analysis methods.[96]

Currently, there have been many missing value imputation methods developed for metabolomics studies that were designed for metabolite peak intensity data.[97-99] In CIL LC-MS, however, the abundance of one metabolite is presented as a peak pair ratio value calculated from two peak intensities. The mathematical meaning of the ratio value is different from the peak intensity value. To generate a complete metabolite intensity table, we investigated the origin of the missing ratio in CIL LC-MS data and developed a missing value prediction method based on the intensity of light- and heavy-labeled peaks in the LC-MS data.

Finally, an integrated data processing workflow is presented here in Figure 4.1, with functions of raw data integrity check, peak pair extraction, data alignment, false peak pair and redundant peak pair exclusion, missing data imputation, and metabolite identification. This workflow comprehensively evaluates the raw data in terms of retention time and mass accuracy and generates a well-aligned dataset with each peak pair uniquely linked to a labeled metabolite. It provides a new way to examine each metabolite feature to exclude the false and redundant one, and the missing data imputation method provide an accurate estimation of the missing ratio

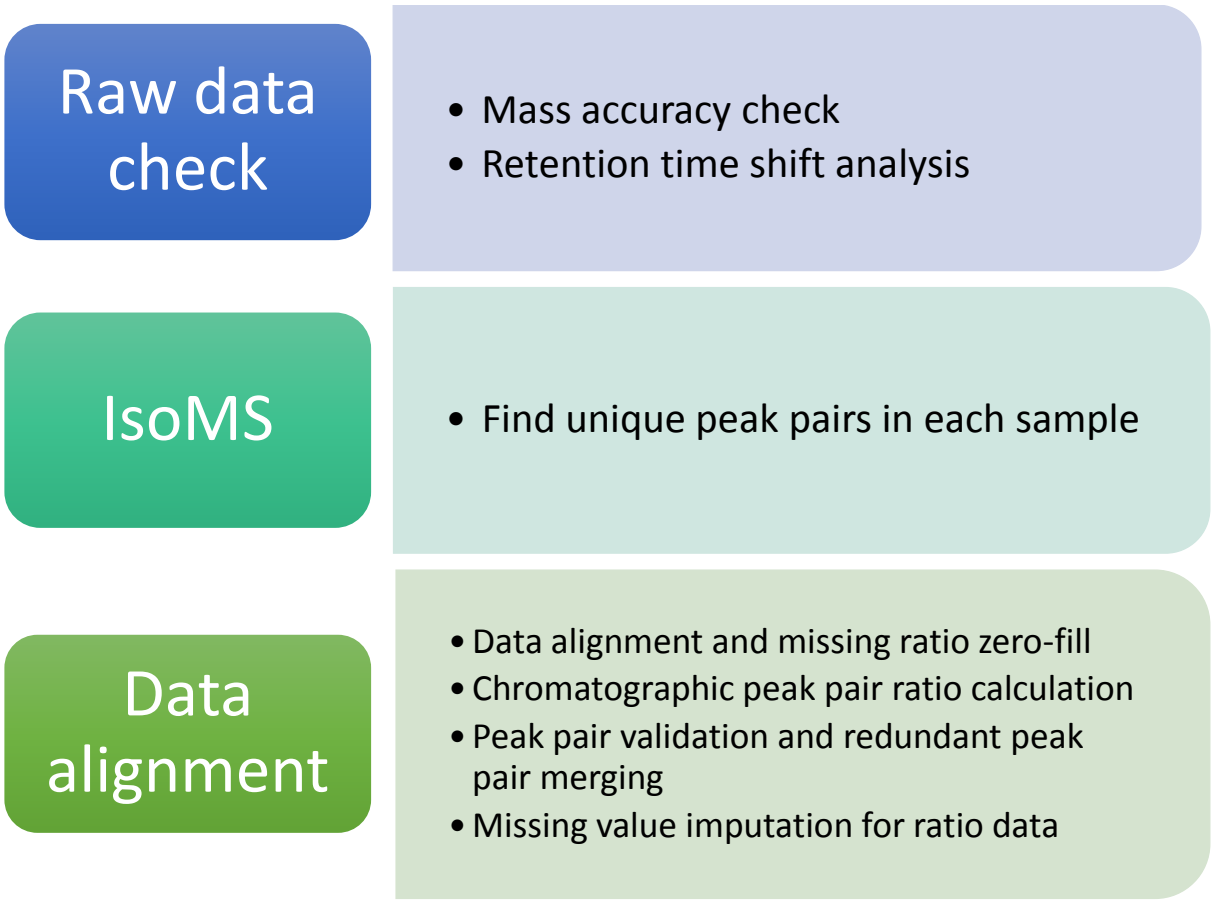


Figure 4.1 Overview of functions in the data processing workflow.

4.2 Materials and Methods

4.2.1 Chemicals and reagents

All the chemicals and reagents, unless otherwise stated, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). In a dansylation labeling reaction, the ^{12}C-labeling reagent (dansyl chloride) was purchased from Sigma-Aldrich, and the ^{13}C-labeling reagent was synthesized and purified in our lab using the procedure published previously.[30] LC-MS grade water and acetonitrile (ACN) were purchased from ThermoFisher Scientific.

4.2.2 Human urine sample preparation and dansyl labeling

A total of 36 human urine samples were collected from two individuals, one male and one female. A pooled sample was prepared by mixing equal volumes of all individual samples. The urine samples were filtered using a 0.22 um filter after collection and were centrifuged at 20,800 g for 10 min. Then, 25 μL of the supernatant of each individual sample were transferred into an Eppendorf for a labeling reaction. Next, 25 μL of 250 mM sodium carbonate/sodium bicarbonate buffer and 25 μL of ACN were added into the sample. The solution was vortexed, spun down, and mixed with 50 μL of freshly prepared ^{12}C-dansyl chloride solution (18 mg/mL, for light labeling) or ^{13}C-dansyl chloride solution (18 mg/mL, for heavy labeling). After 45 min incubation at 40 °C, 10 μL of 250 mM NaOH were added to the reaction mixture to quench the excess dansyl chloride, and the solution was incubated at 40 °C for another 10 min. Finally, 50 μL of formic acid (425 mM) in 50/50 ACN/H_2O were added to consume excess NaOH and to make the solution acidic. The ^{12}C- or ^{13}C-labeled sample was centrifuged at 14,000 rpm for 10 min before injecting onto LC-UV for quantification.[91] For LC-MS analysis, the ^{12}C and ^{13}C-labeled samples were mixed in equal amounts based on the quantification results.

4.2.3 LC-UV quantification and sample normalization

Inter-sample variations in the total metabolite amount must be minimized in order to assess the concentration differences caused by the factors being studied accurately. An LC-UV based method was applied to determine the total concentration of dansylated amine/phenol-containing metabolites based on the UV absorption of the dansyl group.[91] The experiment was performed with a Waters ACQUITY UPLC system UPLC (Waters, Milford, MA, USA) and a Phenomenex

Kinetex C18 column (2.1 mm × 5 cm, 1.7 μm particle size) (Phenomenex, Torrance, CA, USA). Two microliters of each dansyl-labeled individual or pooled sample were injected for a fast step-gradient run. Solvent A was 0.1% (v/v) formic acid in 5% (v/v) ACN/H_2O, and solvent B was 0.1% (v/v) formic acid in ACN. Starting at 0% B for 1 min, the gradient was then increased to 95% B within 0.01 min and held at 95% B for 1 min to ensure complete elution of all labeled metabolites. The flow rate was 0.45 mL/min, and the total UV absorption of dansyl-labeled metabolites in the sample was measured at 338 nm. The peak area, which can represent the total metabolite concentration in the sample, was integrated by the Empower software. According to the quantification results, the ^{12}C- and ^{13}C-labeled samples were mixed in equal amounts for the following LC-MS analysis.

4.2.4 "2:1" sample preparation

To test the accuracy of the ratio calculation, we selected one human urine sample and labeled it with both $^{12}C_2$- and $^{13}C_2$-dansyl chloride. The two labeling solutions were then mixed in a 2:1 ratio by volume followed by injections to LC-MS. A total of 20 injections were conducted at an increasing injection volume from 0.1 μL to 1 μL with 0.1 μL increments, and duplicate injections were conducted for each injection volume. In this dataset, the intensity of the light peak is expected to be twice the intensity of the heavy peak in each detected peak pair.

4.2.5 LC-MS analysis.

The ^{12}C- and ^{13}C-labeled human urine samples were mixed according to the LC-UV quantification results and centrifuged at 20,800 g for 10 min before being injected into a Bruker

maXis impact QTOF mass spectrometer (Billerica, MA, USA) linked to an Agilent 1100 series binary HPLC system (Agilent Palo Alto, CA). A Zorbax Eclipse Plus C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1% (v/v) LC-MS grade formic acid in 5% (v/v) LC-MS grade CAN, and solvent B was 0.1% (v/v) LC-MS grade formic acid in LC-MS grade ACN. The gradient elution profile was as follows: t=0.0 min 20% B, t=3.5 min, 35% B, t=18.0 min, 65% B, t=24 min, 99% B, t=28 min, 99% B. The flow rate was 180 μL/min.

4.3 Processing Algorithms

4.3.1 Peak pair detection with IsoMS

After data collection from a LC-MS instrument, the raw data are converted to a centroid peak list using a data analysis software provided by the manufacturer. All mass peaks above a user-defined signal to noise ratio (SNR) threshold are exported into a mass list csv file. Different from a label-free experiment in which feature picking is based on singular mass peaks, CIL data processing focuses on peak pair patterns in each spectrum, which requires the presence of at least four peaks (a light peak, a heavy peak, and their first natural isotope peaks, see in Figure 4.2). An in-house program, IsoMS,[68] was developed previously to extract the peak pairs from raw data. From the raw mass list, the IsoMS program performs peak pairing, peak pair filtering, peak pair grouping, and intensity ratio calculations. A set of rules is used during the peak pair feature selection, including light and heavy peak distance, isotopic pattern, charge state of the light and heavy peak, etc. After the first round of peak pairing, an exhausted peak pair list is created with all the qualified peak pairs found in each mass spectrum. Next, adduct ions derived from Na^+, K^+,

NH^{4+}, dimers, mutimers, and other common in-source fragment ion peaks are filtered. The user also can provide a list of background peak pairs for the program to filter from the peak pair list.

Most metabolites will be detected repeatedly in multiple spectra within the corresponding chromatographic peak. Based on the first-round peak pair list, IsoMS performs peak pair grouping by retaining the highest intensity peak pair for the same metabolite. For isomers with a similar retention time, IsoMS will examine the increase and decrease of peak intensity within one mass trail to determine if the signals belong to one metabolite or to multiple isomers. If two or more peak shapes are found in the adjacent mass spectra, multiple peak pairs will be saved, depending on the number of isomers found. As a result, a peak pair list is generated for each LC-MS data file with the mass and intensity information for each unique labeled metabolite. Table 4.1 shows the data format of a peak pair list that contains the retention time (RT), light and heavy peak intensities (sn_light and sn_heavy), number of tag (nTag), number of charge (nCharge), and peak pair ratio for each peak pair.

Table 4.1 Data format of a peak pair list generated by IsoMS processing. Each row contains the peak pair information of a unique metabolite.

Scan No	RT	mz_light	mz_heavy	sn_light	sn_heavy	nCharge	nTag	Ratio
1	RT_1	$mz_{1,\ light}$	$mz_{1,\ heavy}$	$sn_{light,1}$	$sn_{heavy,1}$	1	1	$ratio_1$
2	RT_2	$mz_{2,\ light}$	$mz_{2,\ heavy}$	$sn_{light,2}$	$sn_{heavy,2}$	2	2	$ratio_2$
.....	.....	.....	.....	------	------	------		------

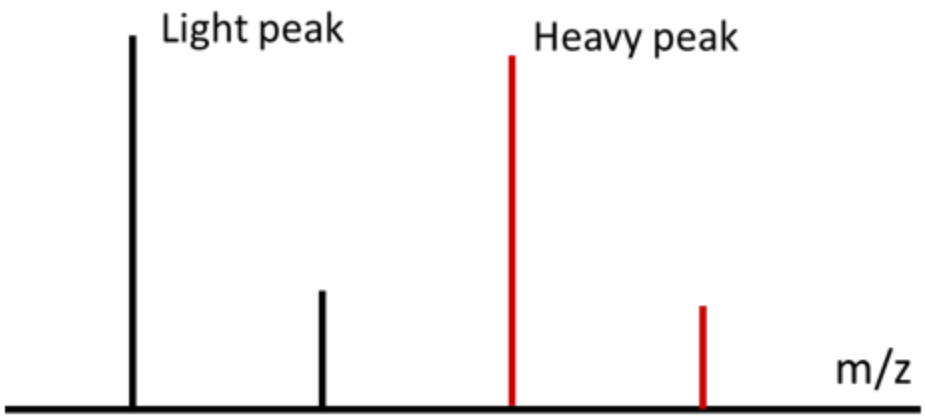


Figure 4.2 A peak pair pattern in the mass spectrum consisting of a light and a heavy peak, along with their first natural isotope peaks.

4.3.2 Peak pair alignment

An important task in metabolomics data processing is the data alignment for matching peak pairs representing the same metabolite in different samples. To do this, our group developed the IsoMS-align program that performs data alignment, combining the IsoMS peak pair lists from multiple samples into one metabolite table.[76] The peak pair matching is done by classifying peak pairs using their retention time, accurate mass, and heavy peak intensity. Since one peak pair potentially can match multiple peak pairs in another sample, depending on the matching tolerances used, a scoring method was used for finding the best match across all samples. The difference between two peak pairs is described by an alignment score,

$$alignment.score = \left(1 - \frac{rt.diff}{rt.tol}\right) + \left(1 - \frac{mz.diff}{mz.tol}\right) + (1 - int.diff) \qquad (4.1)$$

where

$$rt.diff = |rt.^{13}C.peak(a) - rt.^{13}C.peak(b)|$$

$$mz.diff = 1 \times 10^6 \times \frac{|mz.^{13}C.peak(a) - |mz.^{13}C.peak(b)|}{mz.^{13}C.peak(a)}$$

$$int.diff = |log(\frac{int.^{13}C.peak(a)}{int.^{13}C.peak(b)})|$$

The alignment score consists of three comparisons: retention time difference (rt.diff), mass difference (*mz.diff*), and intensity difference (*int.diff*). The differences are calculated based on the heavy peak in the two peak pairs since it usually gives a more consistent signal than the light peak. The retention time tolerance (*rt.tol*) and mass tolerance (*mz.tol*) are provided by the user, depending on the mass accuracy and retention time variation of the instruments.

The alignment processing starts with one IsoMS file as a template. For each additional file, it compares the peak pairs in the new file with those in the template and calculates the peak pair difference by the alignment score. If the score is larger than the score threshold (default value at 1.5), the new peak pair is aligned to the existing peak pair in the same row; otherwise, a new peak pair is created in the alignment table. Table 4.2 shows the data structure of an alignment table. The left side of the table contains the peak pair information, including retention time, accurate mass, and peak intensity (intensity column not shown); all values are averages from individual peak pairs. The right side of the table contains the peak pair ratios in each sample. Within each row, one can compare the relative concentrations of the metabolite in all samples.

Compared to other alignment methods[90,100,101] in which retention time and mass of each feature are compared, CIL data alignment additionally used the intensity of the heavy peak as the third comparison. This is because in the labeling experiment, the heavy labeled pooled sample is added to each individual sample with the same total sample amount.[91] For each labeled metabolite, although the light peak intensity can vary greatly, depending on the metabolite concentration in one sample, its heavy peak will be more consistent in all samples. Therefore, the heavy peak intensity difference was used in the score to increase the matching accuracy.

Table 4.2 Data format of an alignment table of *n* number of individual samples and *m* number of peak pairs. Peak pair ratios of each peak pair from different samples are aligned in the sample columns.

m x n	Retention time	mz_light	mz_heavy	Sample 1	……	Sample i	……	Sample n
Peak pair$_1$	RT$_1$	mz$_{1,\ light}$	mz$_{1,\ heavy}$	ratio$_{1,1}$	……	ratio$_{1,i}$	……	ratio$_{1,n}$
Peak pair$_2$	RT$_2$	mz$_{2,\ light}$	mz$_{2,\ heavy}$	ratio$_{2,1}$	……	ratio$_{2,i}$	……	ratio$_{2,n}$
……	……	……	……	……	……	……	……	……
Peak pair$_m$	RT$_m$	mz$_{m,\ light}$	mz$_{m,\ heavy}$	ratio$_{m,1}$	……	ratio$_{m,i}$	……	ratio$_{m,n}$

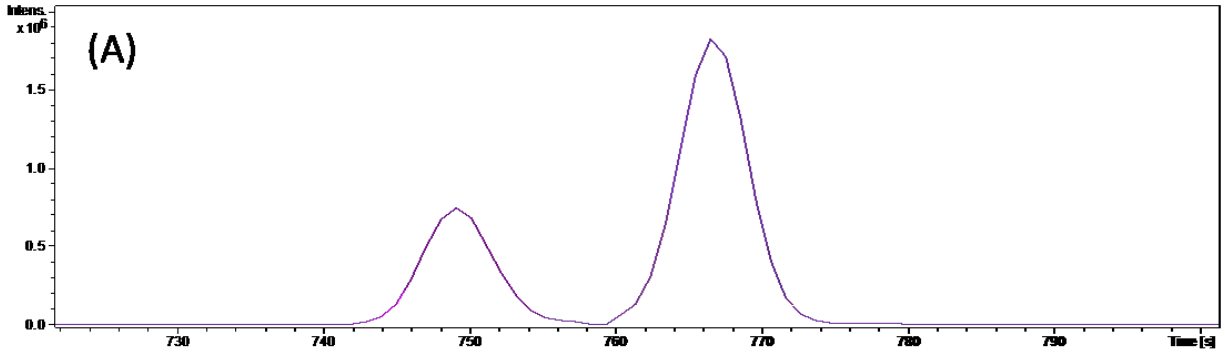
4.3.3 Identifying isomers

Ideally, the same metabolite should have the same measured m/z and retention time in different samples. However, due to instrumental limitations, retention time shift and mass error will exist from one sample to another. Therefore, mass tolerance and retention time tolerance were used in the alignment score to allow peak pair matching in different sample data files. On the other hand, if the retention time or the mass of one metabolite in one sample shifted enough to have its

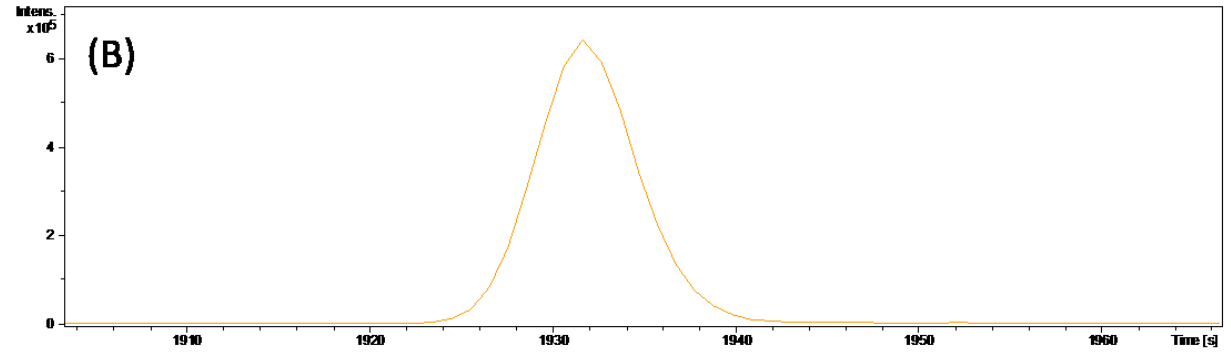
matching score lower than the score threshold, its peak pair will be recorded in the alignment table as a new peak pair. A peak pair in such a case is considered as a redundant peak pair. Redundant peak pairs can be similar to isomers that have similar retention times within a sample file. To distinguish between the two cases, we added one more step at the end of the alignment to identify all potential isomers.

True isomers are different peak pairs in one sample peak pair list after IsoMS processing. During the alignment, each of the isomers will be aligned into the alignment table with ratio values for all of them. For redundant peak pairs, only one of them can be present in a sample peak pair list. Thus, in the alignment table, only one ratio can be present for redundant peak pairs within each sample column. Figure 4.3 shows an example of the two cases. Leucine and isoleucine are two known isomers in the urine sample and were separated by 15 sec in the LC chromatogram. In the alignment table, both peak pairs have all ratios present in all sample columns. Another group of similar peak pairs is shown in Figure 4.3 (B) with only one ratio value available within a sample column. The extracted ion chromatogram showed only one distinctive peak present, indicating that the two peak pairs are likely to be redundant due to the retention time shift in some of the samples.

Based on this difference, we designed an algorithm to detect isomer peak pairs within an alignment table. To do this, the program first groups all peak pairs that are close in retention time and m/z using the same alignment RT and m/z window. Next, peak pairs are compared pair-wise within each group. If more than a certain percentage (20% by default) of the sample columns have both ratios showing up, the program labels the two peak pairs as isomers. The identified isomers will be reserved in the data table throughout the processing, and the true redundant peak pairs will be evaluated again later in the redundant peak pair merging step.



RT (s)	mz_light	mz_heavy	mz	int_heavy	nCharge	nTag	1	2	3	4	5	6	7	8	9	10
751	365.1534	367.1599	131.0951	893119	1	1	1.19	1.17	1.28	1.44	0.82	0.73	1.09	0.81	0.93	1.47
768	365.1535	367.16	131.0951	2207158	1	1	1.04	1.04	1.24	1.12	0.76	0.69	1.02	0.81	0.91	1.34



RT (s)	mz_light	mz_heavy	mz	int_heavy	nCharge	nTag	1	2	3	4	5	6	7	8	9	10
1936	491.3294	493.3355	257.271	201888	1	1	1.11	2.76	1.9	NA	2.53	3.47	0.55	1.65	NA	NA
1948	491.3273	493.3331	257.269	147124	1	1	NA	NA	NA	NA	NA	NA	NA	NA	4.24	0.92

Figure 4.3 Examples of isomers (A) and redundant peak pairs (B) in the alignment table. Leucine and isoleucine are shown in chromatogram (A) and another unknown peak is shown in chromatogram (B). The data tables shows the corresponding peak pair data from the alignment table.

4.3.4 Ratio zero-filling

In an alignment table, missing values may show up randomly in different peak pairs. These missing values can be caused by the absence of a certain metabolite in some of the samples or due to technical and data processing limitations. For example, the initial peak pair selection is

based on a set of criteria to balance the sensitivity and specificity in peak pair picking. With each paired light and heavy peaks, the IsoMS program searches further for the natural isotope peaks to ensure that the light and heavy peaks have the same charge state and number of tags. Only the peak pair meeting all selection rules can be saved in the sample peak pair list. For some of the low abundance peaks or other peaks not meeting all the IsoMS criteria, their ratios could be missing in the alignment table.

The initial peak pair searching in LC-MS raw data is based on the peak pair pattern in each mass spectrum. After alignment, a list of commonly detected metabolites are available in the data table. A peak pair with a missing ratio value will have a relatively high probability of presence in the raw data due to the same nature of the sample. To retrieve the missing ratios, a ratio zero-filling program was developed by re-analyzing the original LC-MS data in a targeted manner based on the peak pair information in the alignment table.[76] The zero-fill program searches the heavy peak of the peak pair in the original mass list based on its retention time, m/z, and peak intensity. The matching score is calculated by,

$$ratio.zerofilling.score = \frac{1 - \frac{rt.diff}{rt.tol}}{4} + \frac{1 - \frac{mz.diff}{mz.tol}}{2} + \frac{1 - int.diff}{4} \quad (4.2)$$

where

$$rt.diff = |rt.^{13}C.peak - rt.raw.data.peak|$$

$$mz.diff = 1 \times 10^6 \times \frac{|mz.^{13}C.peak - mz.rawdata.peak|}{mz.^{13}C.peak}$$

$$int.diff = |log(\frac{int.^{13}C.peak}{int.rawdata.peak})|$$

The weights allocated to the three comparisons, retention time, mass, and heavy peak intensity, are 25%, 50%, and 25%, respectively. We use a score threshold of 0.6 to determine a qualified match. With a matched heavy peak, the program continues to look for the light peak in the highest score heavy peak scan. A new peak pair ratio is calculated to fill the missing value if both light and heavy peak intensity are found.

4.3.5 Peak pair ratio calculation using chromatographic peak area

One limitation of the first-generation zero-filling algorithm is that it requires the presence of both the light and heavy peaks in one specific scan from the best matched heavy peak. It works in most cases, but in some low intensity peak pairs, the light peak can be missing in the highest score heavy peak scan.

On the other hand, a peak pair ratio calculated from one single data point may introduce a random measurement error. To estimate this error, we extract the peak pairs from an identified compound, dansyl proline, in a sample data. Peak pair ratios of dansyl proline were calculated in each mass spectrum where the signals of the peak pair can be detected. After a round of calculations, the list of ratios was plotted against the retention time in Figure 4.4; the red data points are the absolute peak intensities in the whole LC peak. We can see that the peak pair ratio is relatively stable regardless of the peak intensity changes; the average of the ratio is 1.57 ± 0.04 with a relative standard deviation of 2.49%.

In addition to the peak pair ratio, we also calculated the intensity ratio of the first natural isotope peak against the light peak. The relative peak intensity of the first isotope peak is dictated by the element composition of the compound; in theory, the intensity ratio of the first isotope peak

against the main peak should be a constant. Figure 4.5 shows the results of the intensity ratio as a function of retention time; the blue data points are intensity ratios calculated in each mass scan for the dansyl-proline light peak; the standard deviation and relative standard deviation are 0.016 and 7.9%, respectively.

These two experiments showed the precision of intensity ratio measurements. The error in the intensity ratio is due mainly to the uncertainty in the measurement of individual peak intensities. In conclusion, a random measurement error is associated with the peak pair ratio calculation, and a potentially large error can occur in the ratio calculated when using one spectrum.

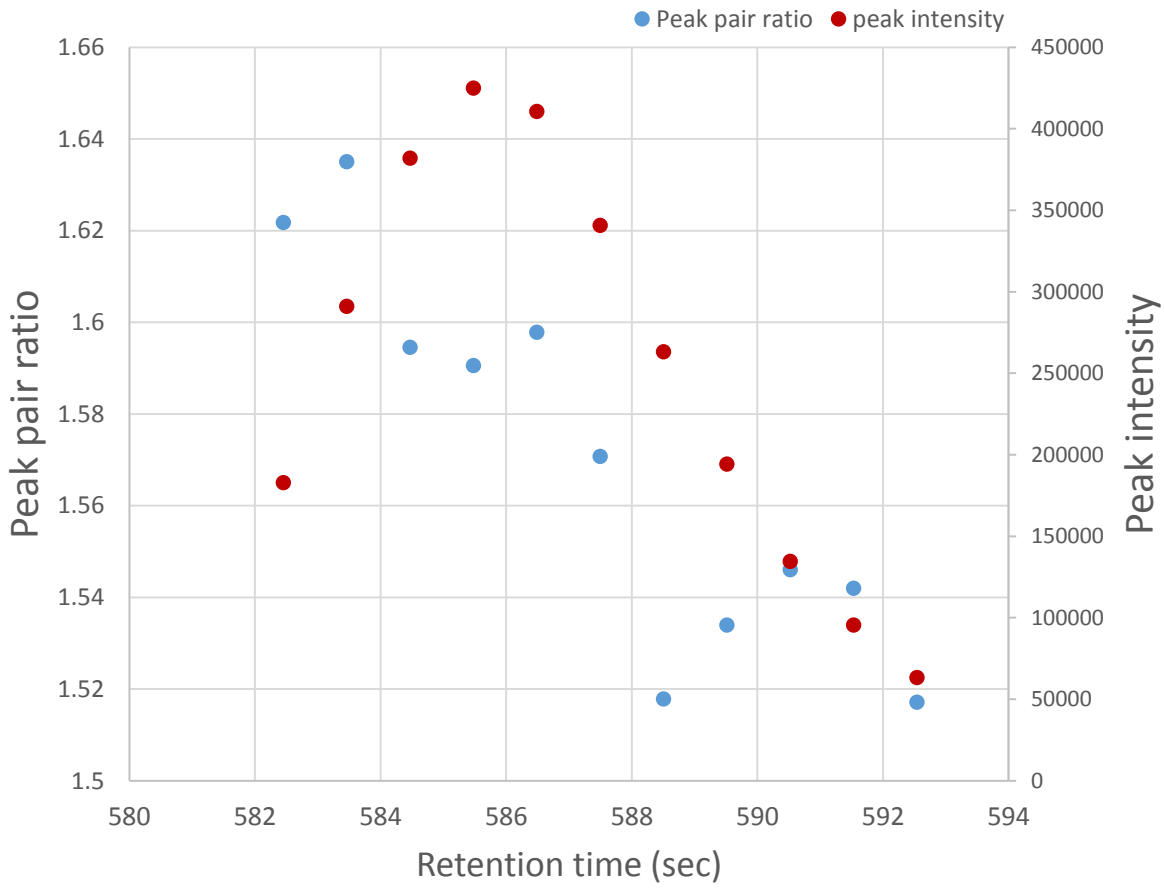


Figure 4.4 Peak pair ratios and light peak intensities of dansyl proline extracted from a sample data file. Blue data points are the peak pair ratios calculated in each mass spectrum and red data points are the light peak intensity in each spectrum.

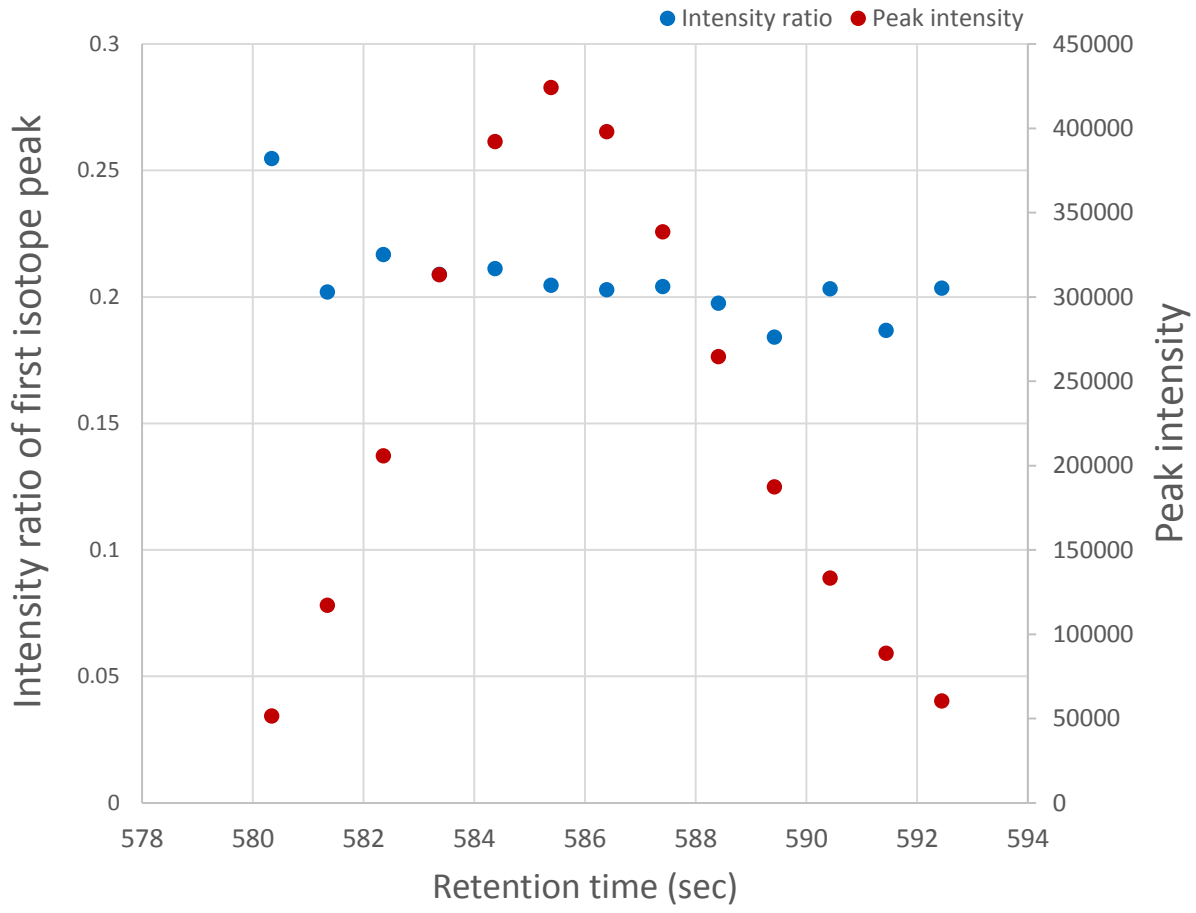


Figure 4.5 Peak intensity ratio of the first natural isotope peak versus the main peak for dansyl proline. Red data points are the proline main peak intensity, and blue data points are the intensity ratios in each mass spectrum.

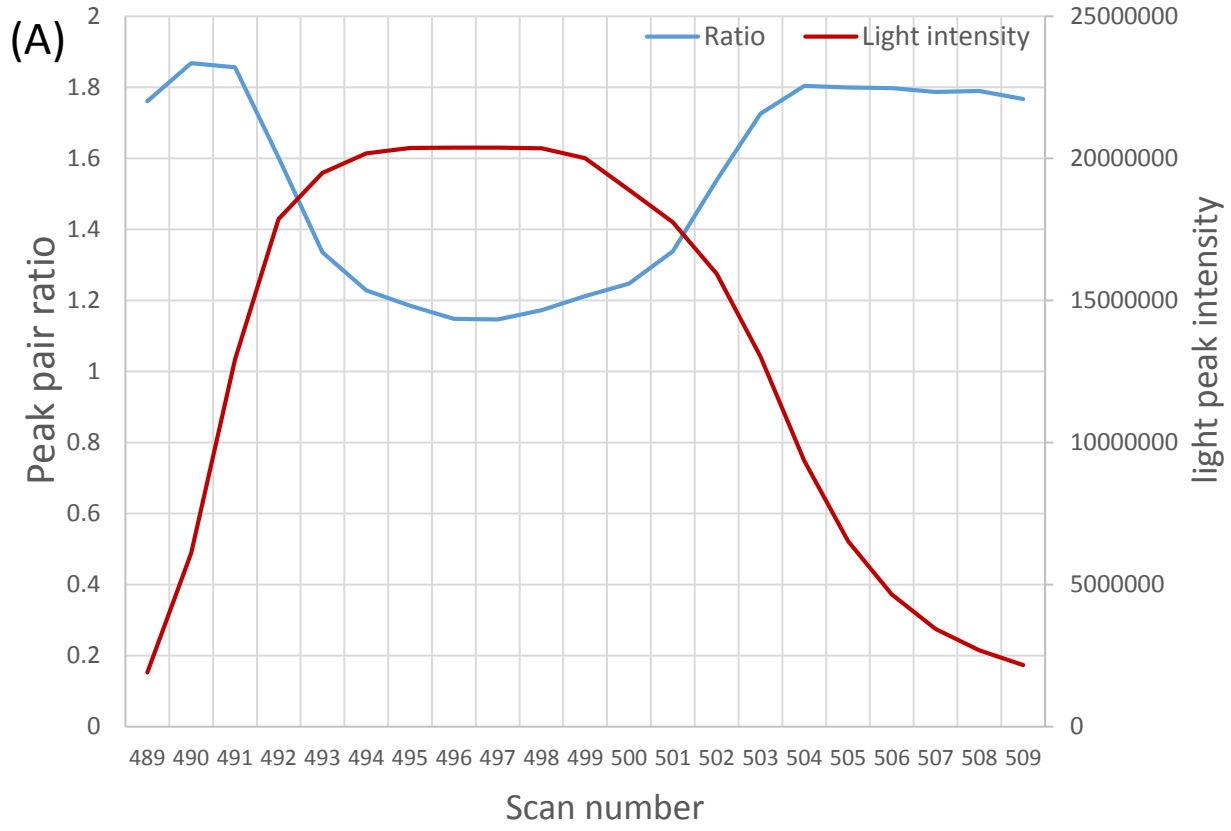
Fortunately in a CIL LC-MS experiment, the peak pair ratio is independent of the absolute peak intensity. To improve the peak pair ratio accuracy, we updated the zero-filling function by calculating the peak pair ratio in all mass scans with signals of the peak pair and used

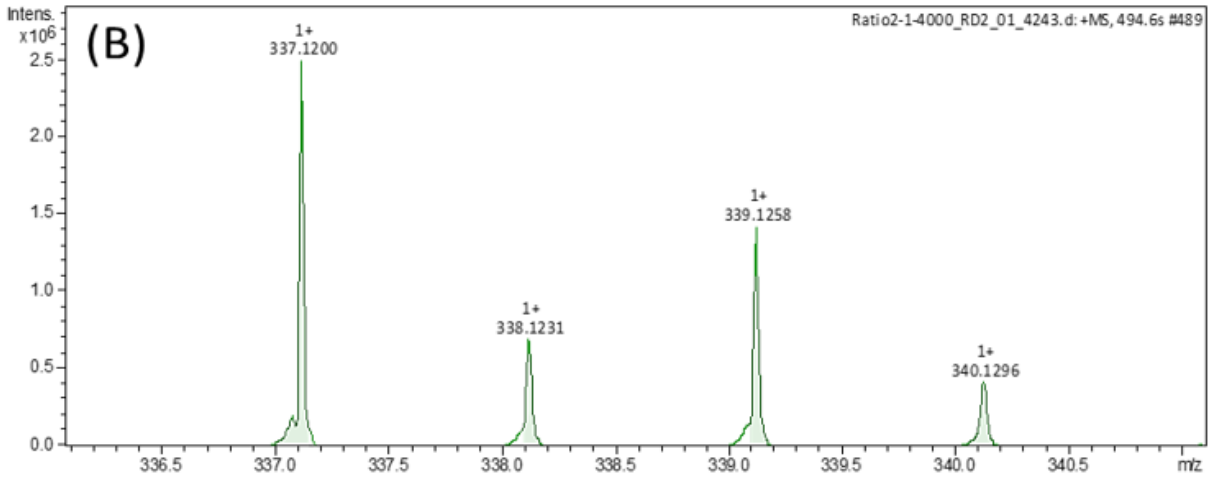
the average of the ratios to fill the alignment table. The program now searches all heavy peaks within the matching tolerances of retention time, accurate mass, and peak intensity. For each matched heavy peak, the program looks further for the light peak in the same mass spectrum. In this way, the data from the whole peak area of one metabolite is extracted and used for calculating the peak pair ratio average. If one of the light peaks is missing in one of the spectra, the program simply skips that scan. With the updated method, the accuracy of the peak pair ratio calculation was improved, and an increased number of missing values were retrieved from the raw LC-MS data.

4.3.6 Saturation signal determination

The detector of any mass spectrometer will have an upper limit for the ion intensity. The upper limit is the maximum intensity reading of one instrument and should not be confused with the upper limit of the linear range, which is always lower than the maximum detector intensity. To determine the upper limit of the linear range in one mass spectrometer, the data from a saturated peak pair (337/339) was extracted for an investigation. Figure 4.6 (A) shows the intensity ratio of the light peak over the heavy peak in different mass scans for the saturated peak pair; the red colored data are the light peak intensity for the peak pair. A drop in the intensity ratio can be observed when the light peak intensity increased to a certain value. Figure 4.6 (B) and (C) show the mass spectra of a saturated scan and an unsaturated scan. Since the light peak is more intense than the heavy peak, it became saturated earlier than the heavy peak as both intensities increased. Therefore, the intensity ratio would decrease after the light peak reached the saturation point. From Figure 4.6 (A), we can estimate the saturation intensity for the QTOF instrument to be at

around ~1.2×10^7, which is lower than the maximum display intensity of 2×10^7. In the peak pair ratio calculation, we will exclude peak pair data that have either peak intensity saturated.





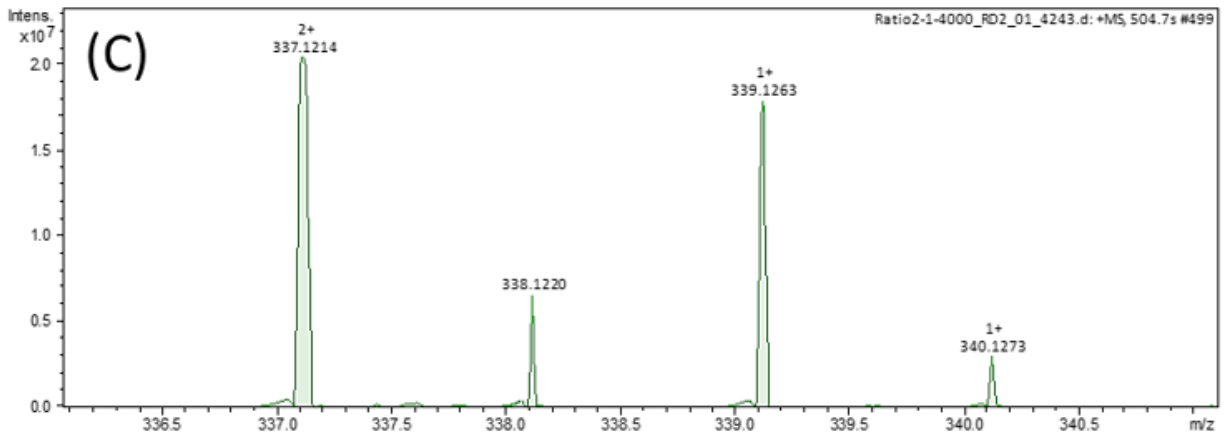


Figure 4.6 (A) Peak pair ratios for a peak pair 337/339 at different scans. Blue colored data are the peak pair ratios, and red colored data are the light peak intensities. (B) and (C) are mass spectra of an unsaturated scan and a saturated scan.

4.3.7 Peak pair validation

A biological sample may contain thousands of metabolites that can be detected using one labeling method. One should expect to see the peak pair signal for any labeled metabolite with a concentration above the detection limit. All the initial data processing steps followed the peak pair patterns to extract metabolite information. However, the complexity of the sample may introduce signals in which a "light" and a "heavy" peak from different molecules are paired falsely. These peak pairs are considered as false positive peak pair features.

Figure 4.7 shows a case of a suspicious false peak pair in a sample data. The three mass spectra were picked from three consecutive mass scans in one sample. The mass distances of peak pair A–B and peak pair B–C are in the mass window by 2.0067 Da. In principle, peaks A, B, and C can combine to have three different peak pairs as peak pairs A–B, B–C, and A–C (peak pair A–C can be a 2-tag 1-charge peak pair). To test the three potential peak pairs, we calculated the peak pair ratio of each possible combination in the three mass spectra (see Table 4.3). From

the result, we can observe a clear difference among the ratios for peak pairs A–B and B–C. Such a ratio difference from scan to scan indicated a falsely paired light and heavy peak. By examining more spectra, we found that peak A is an interfering peak that coincidently showed up near peak B by a 2.0067 Da distance. In this way, we concluded that only peak pair B–C is a true peak pair.

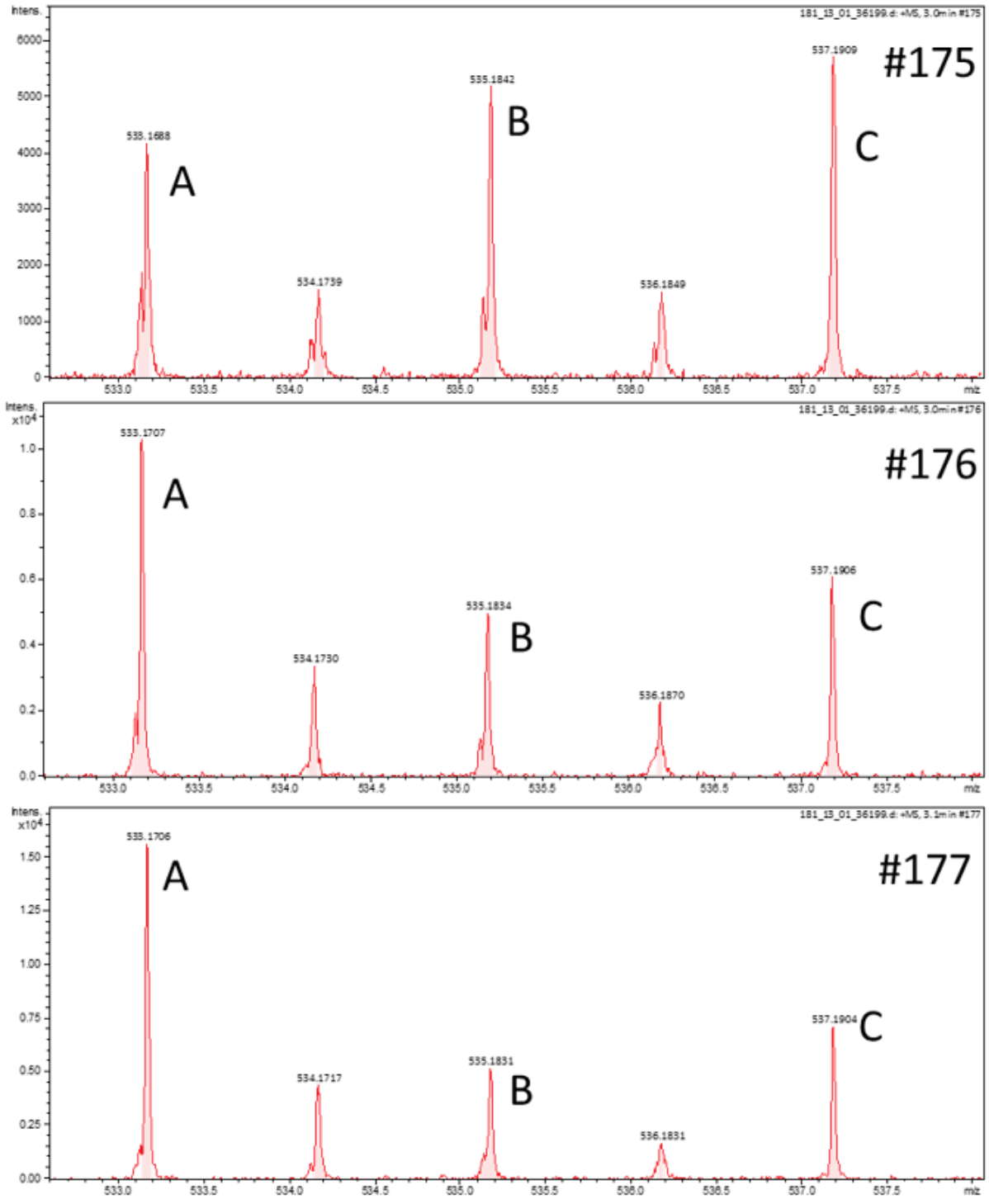


Figure 4.7 Three consecutive mass spectra in one sample data (scan number 175-177) zoomed in for the three mass peaks of interest. The mass distance of peak A to peak B and peak B to peak C are close to 2.0067 Da.

Table 4.3 Peak pair ratios calculated for the three peak pairs. The relative standard deviation (RSD) was calculated for the ratios in the three spectra.

Peak pair ratio	A–B	B–C	A–C
#175	0.95	0.98	0.86
#176	1.91	0.94	1.45
#177	2.23	0.89	2.05
RSD	0.39	0.05	0.41

To exclude all the falsely paired peak pairs in the alignment data table, we designed a peak pair validation step by testing the inter-dependence of the light and heavy peak for each collected peak pair. For each ratio in the alignment table, the program extracts the light and heavy peak in the whole peak area and calculates a list of ratios from all spectra. A validation score is calculated by,

$$validation.score = \left(1 - \frac{ratio.range}{range.tol}\right) \times \frac{1}{3} + \left(1 - \frac{ratio.RSD}{RSD.tol}\right) \times \frac{2}{3}$$
$$+ log_{10}(1 + \frac{sn.threshold}{average.sn.light}) + log_{10}(1 + \frac{sn.threshold}{average.sn.heavy})$$
$$+ (1 - \frac{scan.diff}{2}) \times \frac{1}{2} \quad (4.3)$$

The score consists of three parts. In the first part, it calculates the difference among all the ratios using the ratio range (*ratio.range*) and relative standard deviation (*ratio.RSD*). Next, a compensation score is added for a peak pair with a relatively low intensity. This is because the low intensity peaks tend to have a larger uncertainty for the measured intensity, giving the peak pair ratio a larger error. As the peak pair intensity (average of light or heavy peak intensity) gets close to the signal to noise ratio threshold (*sn.threshold*), an increasing score is added to compensate for the increasing uncertainty. The last part of the score looks at the differences in the retention time of light and heavy peaks. Since the isotope labeling (^{13}C-/^{12}C-labeling) will not affect the retention behavior of the labeled metabolites, a difference in the scan number (*scan.diff*) for light and heavy peaks can be another indicator of two independent peaks.

A validation score is calculated for each existing ratio after ratio zero-filling. We use a validation score of 0 as a threshold to label each ratio as TRUE or FALSE for one peak pair. If the ratios from all samples are labeled as FALSE (the peak pair is determined as a false peak pair in all samples), we then have a high confidence that the peak pair is a false peak pair. All the false peak pairs will be excluded from the data table at the end of the processing.

4.3.8 Redundant peak pair merging

The retention time and m/z of the same metabolite will have variations from sample to sample. In the alignment step, an alignment score was designed with a retention time and mass tolerance for matching the peak pair across samples. A threshold score value was optimized in the program to balance the sensitivity and specificity in peak pair matching and increase the accuracy of data alignment. However, due to experimental error, some of the peak pairs may have a relatively

large retention time or mass shift that gives them an alignment score larger than the score threshold. These peak pairs become redundant features in the alignment table. We have identified previously the isomer peak pairs that can be confused easily with redundant peak pairs. In this step, we will examine the rest of the peak pairs and remove the real redundant ones.

For the same metabolite, we have shown consistent peak pair ratio values within one sample. For two peak pairs with a similar retention time and m/z in an alignment table, we can compare further the ratio values column-wise to test the correlation of two peak pairs. If the ratio values are similar for two peak pairs in all sample columns, they are likely to be redundant peak pairs. Based on this strategy, the program first groups all peaks pairs with the retention time and m/z difference within the matching window. Within each group, a pairwise comparison is conducted between two peak pairs, peak pair (a) and peak pair (b). An overall peak pair ratio difference is calculated by,

$$average\ relative\ ratio\ error = (\sum_i^n \frac{|ration_{a,i} - ration_{b,i}|}{ration_{a,i}})/n \qquad (4.4)$$

where n is the total number of samples that have ratio values in both peak pair (a) and peak pair (b), and $ratio_{a,i}$ and $ratio_{b,i}$ are the peak pair ratios in the i^{th} sample column. The average of the relative ratio error is calculated as a measure of the ratio difference between two peak pairs. If the difference is less than the threshold (default at 20%), we determine the peak pairs to be redundant. Within each true redundant group, we keep the highest heavy peak intensity peak pair and merge the ratio values into the highest intensity peak pair. Table 4.4 shows an example of the redundant peak pair processing; Table 4.4 (A) is the alignment result and Table 4.4 (B) is the

table after the zero-filling processing. After ratio zero-filling, most of the missing value were recalculated from the peak pair's chromatographic data. If the two peak pairs are from the same LC peak, the ratio calculated from the peak area will similar. From Equation (4.4), we calculated an average relative ratio error between the two peak pairs at 1.58%. This result indicated a very similar ratio value for the two peak pairs, thus labeling them as true redundant peak pairs. Next, the ratios were merged into the highest intensity peak pair. In this example, the ratio value in sample column 8 from the removed peak pair was used to fill the missing value in the saved peak pair, as shown in Table 4.4 (C).

Table 4.4 (A) Example of two suspicious peak pairs after data alignment.

#	RT	mz_light	mz_heavy	intensity	1	2	3	4	5	6	7	8
914	396.1415	406.1416	408.1483	7824	NA	NA	NA	NA	NA	NA	NA	1.77
1036	414.1165	406.1425	408.149	252904	2.16	2.00	2.09	2.20	1.82	1.71	1.92	NA

Table 4.5 (B) The two suspicious peak pairs after ratio zero-filling.

#	RT	mz_light	mz_heavy	intensity	1	2	3	4	5	6	7	8
914	396.1415	406.1416	408.1483	7824	2.10	2.05	2.09	2.22	NA	1.69	1.88	1.77
1036	414.1165	406.1425	408.149	252904	2.16	2.00	2.09	2.20	1.82	1.71	1.92	NA

Table 4.6 (C) The merged peak pair after redundant peak pair merging. The ratio value of column 8 is filled using the data from the removed peak pair.

#	RT	mz_light	mz_heavy	intensity	1	2	3	4	5	6	7	8
1036	414.1165	406.1425	408.149	252904	2.16	2.00	2.09	2.20	1.82	1.71	1.92	**1.77**

Another source of redundant peak pairs can be from peak tailing. Peak tailing occurs for a relatively high intensity peak in which the signal of the peak can extend to a few minutes at the tail of the LC peak. The peak pair of one metabolite can be detected repeatedly within the tailing retention time, causing redundancy in the data table. To exclude a redundant peak pair caused by peak tailing, the program first searches all peak pairs with a heavy peak intensity larger than 10^6 and checks repeated peak pairs in a 3-min window after the main peak. If more than three repeated peak pairs are found in the 3-min tail with peak intensity lower than 5% of the main peak (see Figure 4.8), they are labeled as tailing peak pairs and deleted from the peak pair table.

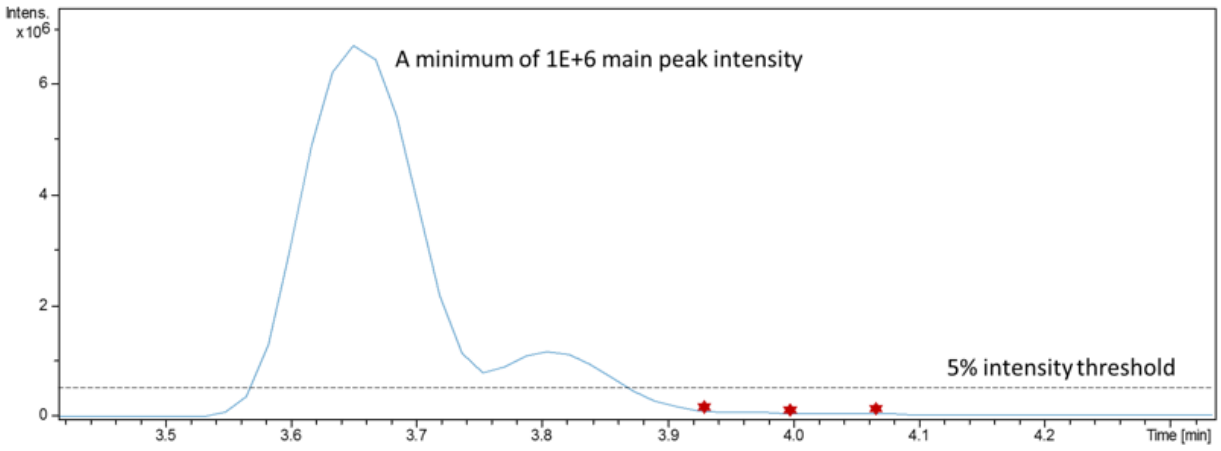


Figure 4.8 Peak tailing in the tailing peak pair removing process. The main peak must have an intensity larger than 10^6 (for Bruker QTOF mass spectrometer), and the tailing peak pair has to show up more than three (>=3) times within a 3-min tailing retention time window.

4.3.9 Missing value imputation

In studies of metabolomics, missing quantity values of metabolites are encountered commonly. They can originate from various sources, including analytical, computational, and biological.[102] These missing values can be problematic for statistical analysis as many analysis methods require a complete metabolite-intensity matrix. Methods for imputing missing data have been developed in recent years, but these were not designed usually for CIL LC-MS data in which peak pair ratio represents the relative abundance of each labeled metabolite in the metabolite-intensity table.

The zero-filling program can retrieve a significant portion of missing data based on their signals in the raw file. For the remaining missing ratios, we conducted a manual search in the raw data and found that most of the missing values either have their intensity totally missing or have an intensity much lower than those of existing peak pairs. A missing value in CIL LC-MS data can be divided into two categories: both peak intensities missing and one of the peak intensities missing.

The program first attempted to search for the possible peak intensities in each missing ratio. Two peak intensity tables were created first for the light and the heavy peak during the peak pair alignment step. For each ratio calculated in the metabolite-intensity table, its light and heavy peak intensities were stored in the light and heavy peak intensity tables through all processing steps. An intensity zero-filling program was developed to retrieve peak pair intensities from the raw data. Based on the two peak intensity tables, the light and heavy peak of a missing ratio are searched in the raw data by,

$$score.light = (1 - \frac{rt.diff}{rt.tol})/3 + (1 - \frac{mz.light.diff}{mz.tol}) \times 2/3$$

$$score.heavy = (1 - \frac{rt.diff}{rt.tol})/3 + (1 - \frac{mz.heavy.diff}{mz.tol}) \times 2/3 \qquad (4.5)$$

where $rt.diff = |rt.^{13}C.peak - rt.raw.data.peak|$

$$mz.heavy.diff = 1 \times 10^6 \times \frac{|mz.^{13}C.peak - mz.rawdata.peak|}{mz.^{13}C.peak}$$

$$mz.light.diff = 1 \times 10^6 \times \frac{|mz.^{12}C.peak - mz.rawdata.peak|}{mz.^{12}C.peak}$$

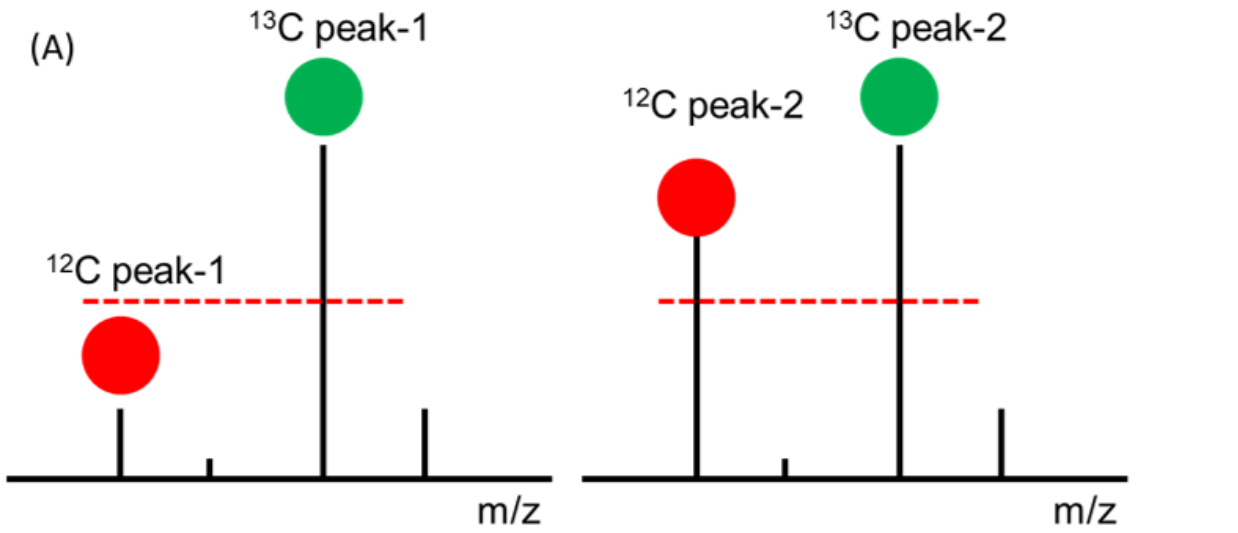
If the score is larger than the threshold value of 0.3, the intensity is filled in the intensity table. In this round of search, the score calculation is different from the ratio zero-fill in that the intensity difference restriction was removed from the peak matching. This is because the missing ratio is likely to have its light or heavy peak intensity much lower than other peak pairs. As a result, if both intensities are found in this round, a new ratio is filled in the ratio table.

The remaining missing ratio data is divided into three specific cases: light peak missing, heavy peak missing, and both peaks missing. Of the three cases, the light peak missing was found to be the most frequent one since the heavy peak from the pooled sample usually gives a consistent signal. The reason for not detecting the light peak could be that the metabolite concentration in this sample was too low or the ion intensity was below the detection limit due to a strong ion suppression effect.

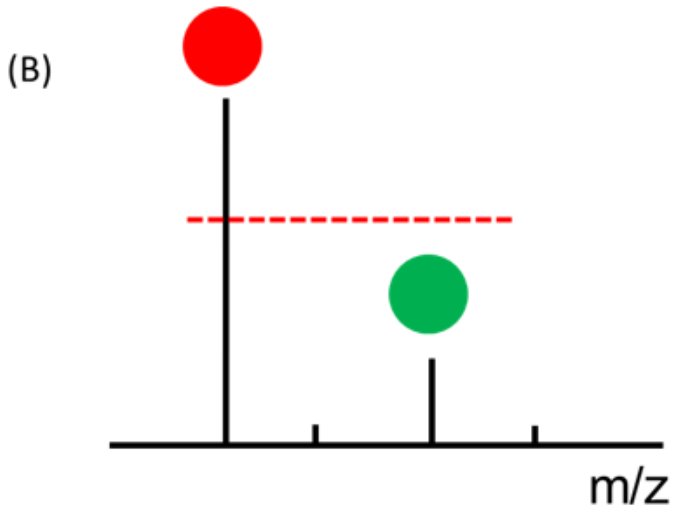
To estimate the missing ratio with better accuracy, we developed an algorithm to account for the ion suppression effect based on the comparison of the heavy peak intensity of the missing

value pair to that of another pair with the lowest light peak intensity. Figure 4.9 shows the details of the algorithms. In the light peak missing case, missing peak pair (1) has its light peak below the detection limit. Then, we find the peak pair (2) from another sample with the lowest light peak intensity. The ratio of the two heavy peaks is used as an estimate of the ion suppression difference in the two samples. The predicted light peak intensity is calculated based on the light peak intensity from peak pair (2) and another prediction constant. The prediction constant is a number between 0 and 1 to compensate for the fact that the light peak is not detected. After an estimated light peak intensity is calculated, the new ratio is calculated to fill the data table.

For the heavy peak missing case, the heavy peak intensity is likely to be below the detection limit. We use a value below the intensity threshold to replace the missing value. For both peak missing cases, we can assume only that both peaks are below the intensity threshold, so the peak pair ratio average in other samples is used to replace the missing value.



$$\text{missing } intensity = light\ peak\ 2 * \frac{heavy\ peak\ 1}{heavy\ peak\ 2} * prediction\ constant$$



$$\text{missing } intensity = intensity\ threshold\ * prediction\ constant$$

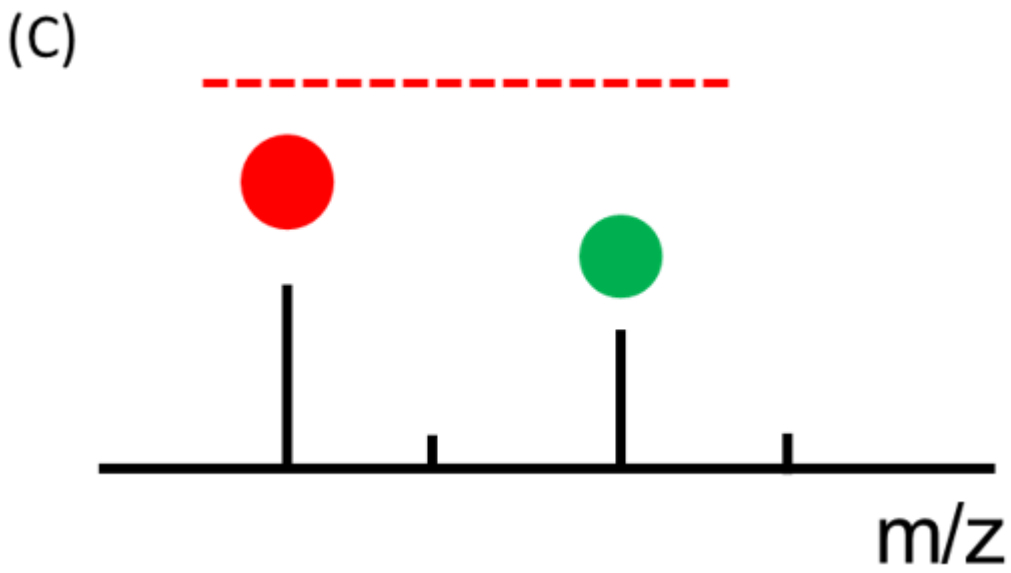


Figure 4.9 Missing value prediction algorithm: (A) a light peak missing, another existing ratio with the lowest light peak intensity is used for the intensity estimation, (B) a heavy peak missing, the heavy peak intensity is replaced by a small value below the intensity threshold, and (C) both peak intensities missing, the ratio is replaced by the average of ratios from that peak pair. The red dashed line indicates the level of the detection limit. The prediction constant is a number between 0 and 1 for the missing intensity prediction.

4.4 Results and Discussion

4.4.1 Saturation signal determination

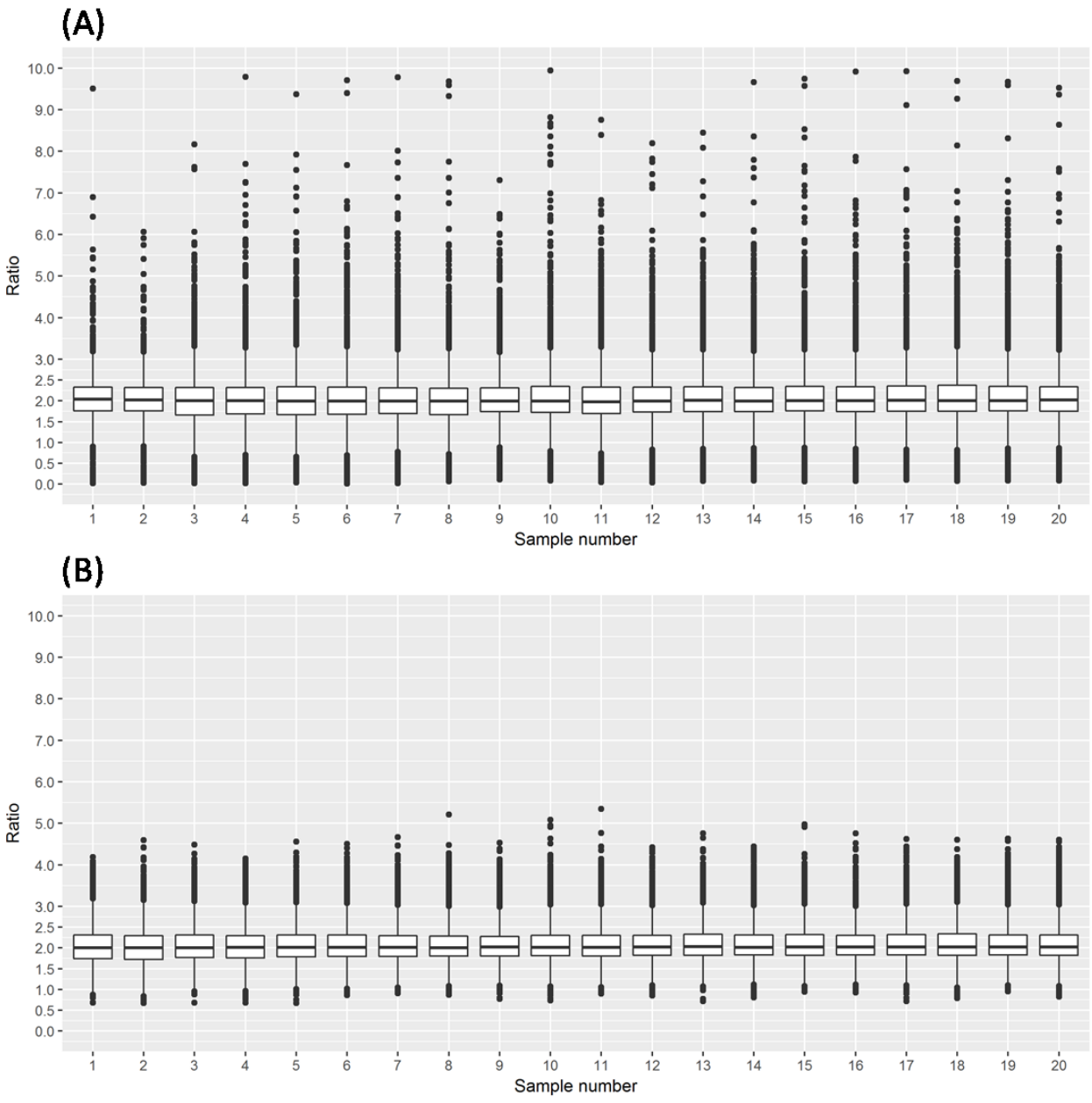
Before analyzing each individual sample in LC-MS, the injection amount was optimized first to allow the detection of low concentration metabolites and achieve maximum detectability. With the increasing injection amount, more peak pairs could be detected from one sample. However, a higher injection can cause the signal from the high concentration analytes to be saturated, making the ratio calculation inaccurate. To determine the saturation signal of one mass spectrometer, we collected data from a 2:1 mixture of a human urine sample (Section 4.2.4). After data collection, we selected a high intensity peak pair and calculated the intensity ratios of light peak over heavy peak in its chromatographic peak. Figure 4.6 shows the ratio change in the whole peak area. Since the light peak has a larger intensity than the heavy peak, saturation first occurred on the light peak and caused the ratio to decrease. The ratio curving point indicated a saturation signal at 1.2×10^7.

In general, one can use any saturated peak to determine the saturation intensity by calculating the intensity ratio of a saturated peak over its first natural isotope peak in the whole LC peak area. Since the relative intensity of the natural isotope peak for the same compound is a constant, the intensity ratio value can be used to estimate the saturation intensity. One should check the saturation signal when using a new mass method or after instrument maintenance. All saturation signals will be excluded from ratio calculations.

4.4.2 Performance of ratio zero-filling

We ran the alignment and ratio zero-filling method on the 2:1 mixture LC-MS data using both a previous zero-filling method and the updated one. After alignment of 20 sample files, we had 3,055 peak pairs in the metabolite-intensity table. The overall missing value percentage is 65.08% in the alignment table. After ratio zero-filling, the missing percentage is 16.74% with the first generation zero-filling method and 5.60% in the updated zero-fill method. The new ratio zero-fill showed a roughly 10% increase of the ratio values retrieved than the first generation zero-fill.

Figure 4.10 shows the ratio value distribution in each sample column for the 20 samples. Figure 4.10 (A) and Figure 4.10 (B) are the box plots of data generated by the first generation zero-fill and the updated ratio zero-fill, respectively. From the plots we can see that the new processing method removed most extreme outliers after peak pair validation. These outlier ratios were mostly from the low intensity peak pair. Their ratios tend to have a larger random error than high intensity peak pairs. Ratio medians are shown in Figure 4.10 (C), with all medians close to the theoretical value of 2.00. Figure 4.10 (D) shows the interquartile range of ratios in each sample column. An overall smaller interquartile range can be seen for data from the updated zero-fill method, indicating the improved accuracy after using ratios calculated from peak area data.


(A)
Ratio
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Sample number
(B)
Ratio
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Sample number

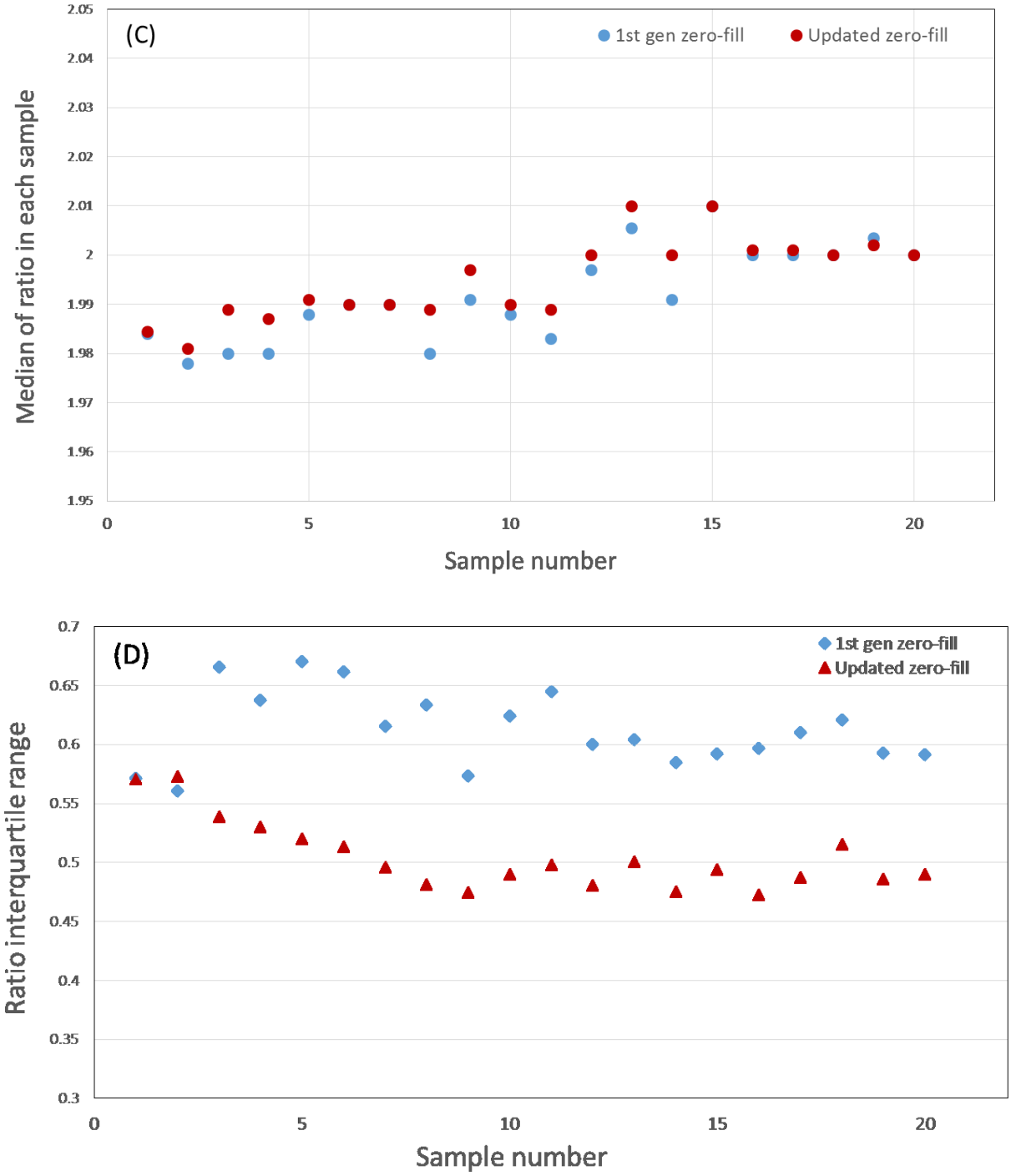


Figure 4.10 Peak pair ratio distribution in boxplot for data generated by (A) the first generation zero-fill method and (B) the updated zero-fill method. (C) The median value for the ratio in each sample column and (D) the interquartile range for each sample. The new method removed most of the extreme outliers and showed a smaller interquartile range.

Figure 4.11 shows the number of missing values in each sample. Duplication injections were conducted for each injection volume. The number of missing values decreases as the amount of sample injection increases. As injection volume increased, more peaks became saturated, causing a greater ion suppression to those low abundant peaks; this led to an increased missing value at over injection.

For the remaining 5.60% missing ratios, we applied the ratio value imputation method (Section 4.3.9) and generated a complete metabolite intensity table with no missing ratios. The 5.60% predicted values were extracted, and a box plot was created for a predicted value in each sample column in Figure 4.12. We can see that the medians for predicted values are close to the reference value of 2.00. Although the predicted ratio error can be relatively large for some of the data points, the overall prediction accuracy is good for each sample data.

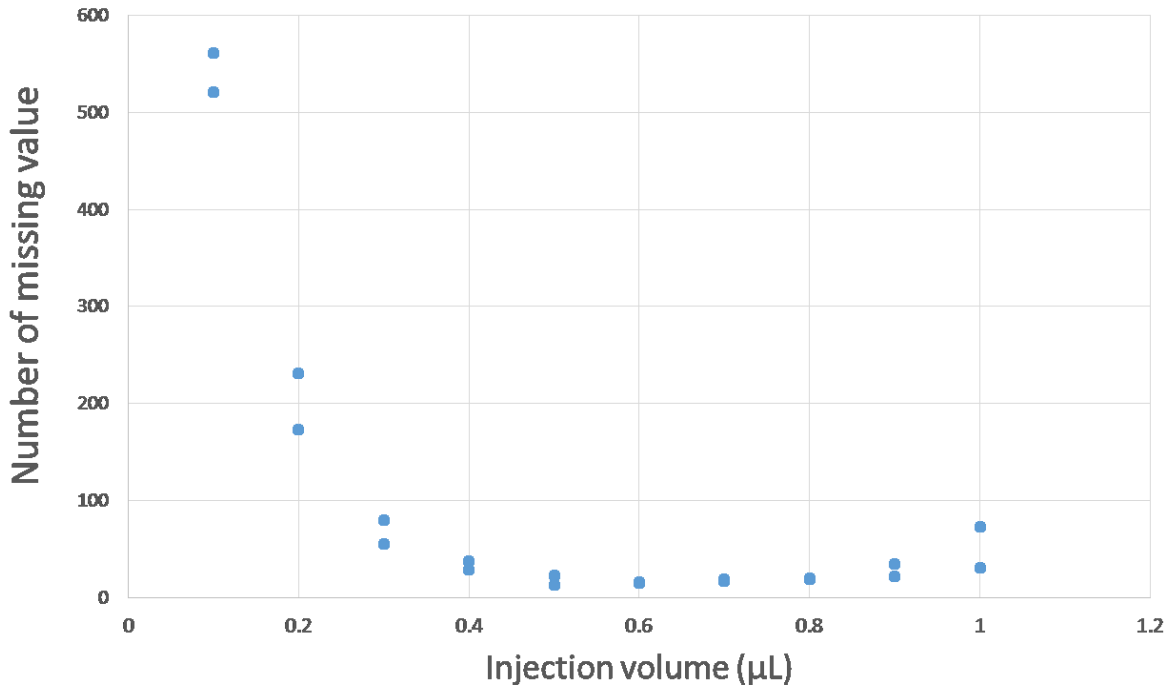


Figure 4.11 Number of missing values in each sample column after ratio zero-fill. Duplicate injections were conducted for each injection volume.

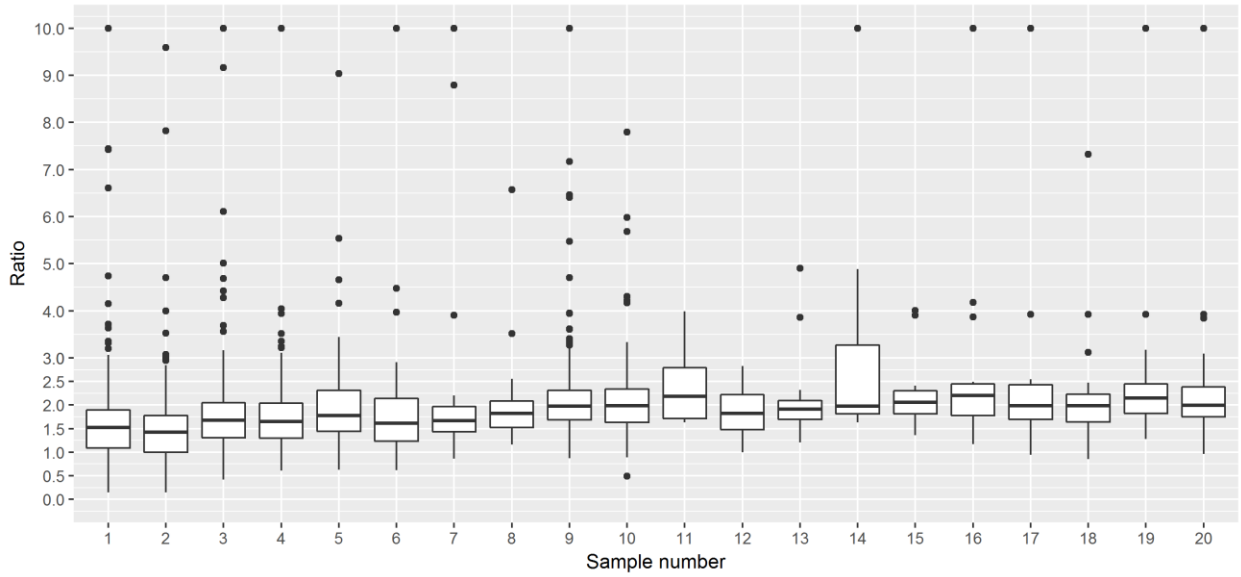


Figure 4.12 Box plots of predicted values in each sample column.

4.4.3 Human urine data

We applied the data processing workflow to the human urine data collected from two individuals. A total of 9,444 peak pairs were detected after aligning 18 sample data. On average, there are 77.76% ratio values missing in the ratio matrix after alignment. After ratio zero-filling, the missing value percentage dropped to 7.54%. In the peak pair validation and redundant peak pair merging steps, 395 and 1,122 peak pairs were deleted, respectively.

We filtered out the peak pairs that had the ratio missing in more than 50% of the samples after the ratio zero-fill and intensity zero-fill. A peak pair missing over 50% of its ratio values may have inadequate data for making an accurate prediction of the missing ratios. 101 peak pairs were excluded from the 50% filter, leaving a total number of 7,826 peak pairs. After missing data imputation, a complete data table was generated.

Figure 4.13 (A) and Figure 4.13 (B) are the principal component analysis (PCA) score plots for two missing data treatments: k nearest neighbor (KNN) method and zero-filling imputation processing method (plots generated by MetaboAnalyst[64]). Red and green data points are sample data from individual (1) and (2). Since the urine sample was collected from two genders, a separation is expected to be seen in a PCA score plot. In this result, the PCA plot (B) for data from the zero-filling imputation method shows a better separation between the two groups, with a smaller intra-group variation. Figure 4.13 (C) shows a comparison of the first two component percentages in the PCA plot using all available imputation methods in MetaboAnalyst and zero-filling imputation. A larger component of percentages can be seen in the result from the zero-filling ratio imputation method in comparison to all other imputation methods, showing the advantage of our method for predicting missing ratio data in a CIL LC-MS experiment.

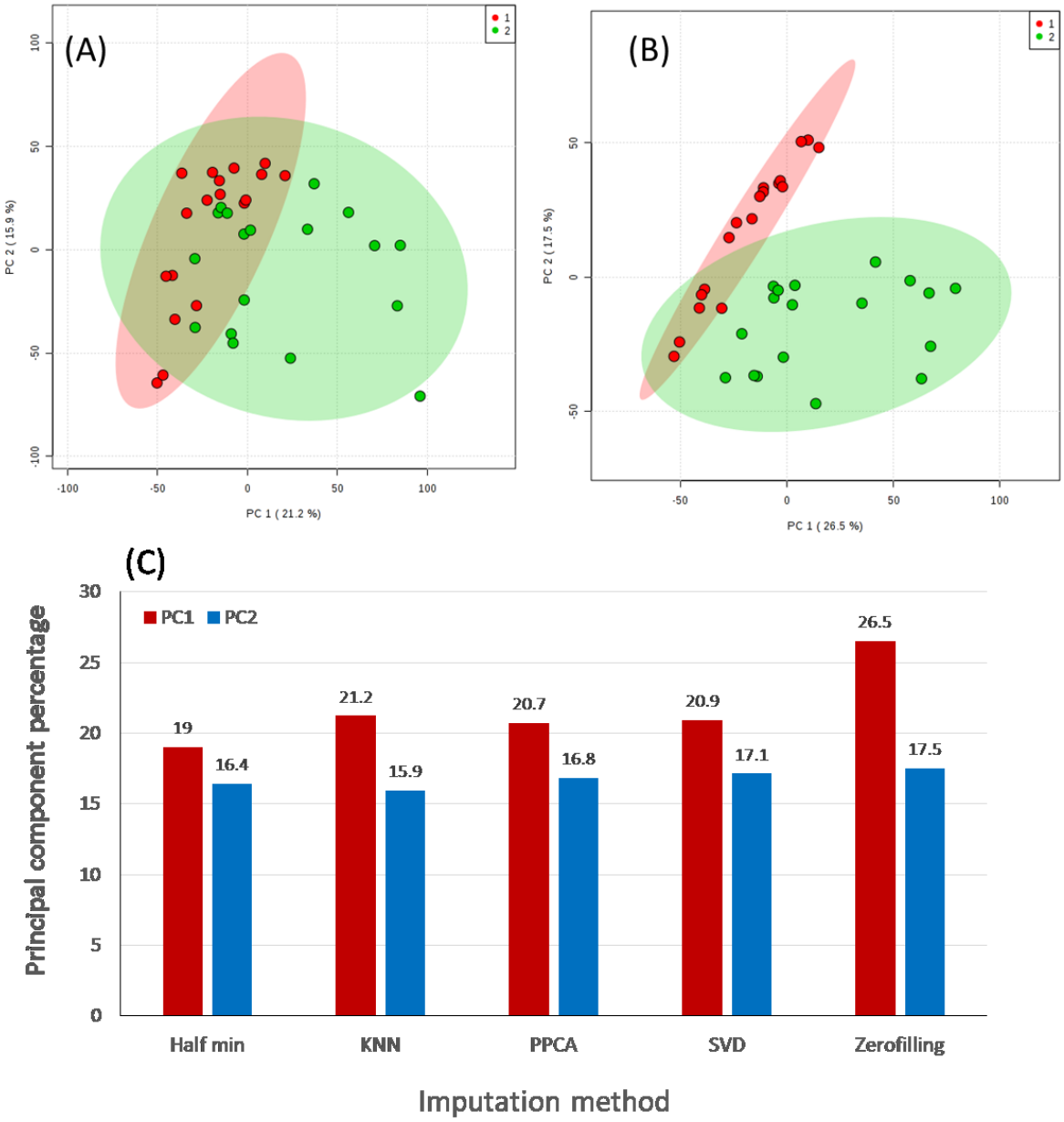


Figure 4.13 Principal component analysis (PCA) score plots using (A) the KNN imputation method and (B) the zero-filling processing method. (C) Comparison of PCA component 1 and 2 percentages using available imputation methods at metaboanalyst.ca.

4.5 Conclusions

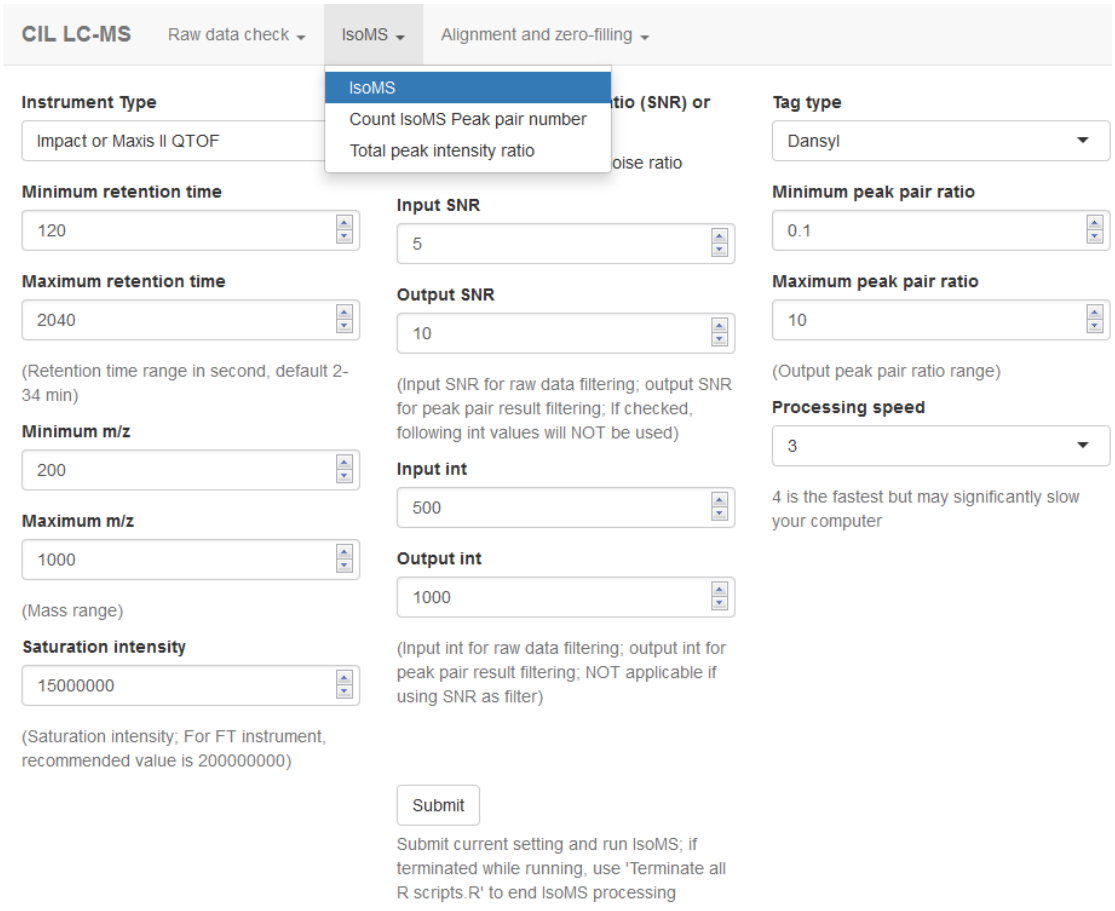
We reported an integrated data processing workflow for the chemical isotope labeling LC-MS experiment. In the new zero-fill algorithm, an extracted ion chromatogram was constructed for each ratio, and a new ratio was recalculated using the average of the peak pair ratios in the whole peak area. All saturated signals were excluded from the ratio calculation. A novel missing data imputation method was developed for the prediction of missing ratios based on the intensity data. The peak pair ratio imputation method showed a significant improvement in the separation of two biological groups in a PCA plot compared to other imputations methods.

In the data table, each peak pair should be representative of a labeled metabolite. However, due to random chances, there are a few cases where a light and a heavy peak were falsely paired together. A scoring method was designed to find all false positive peak pairs by evaluating the consistency of ratios from one peak pair and the overlap between light and heavy chromatographic peaks. The redundant peak pairs were examined also through the peak pair table. After ratio zero-filling using chromatographic peak area data, ratios for one metabolite should be the same in each sample. We tested the similarity among peak pairs based on differences in retention time, m/z, and within-sample ratio values to exclude all redundant peak pairs.

The new processing workflow was found to improve the peak pair ratio accuracy by using data from a whole peak area and minimizing noise data by excluding false positives and redundant peak pairs. These novel labeling data processing methods show the advantages of the chemical isotope labeling LC-MS platform. We also showed the performance of the workflow in a comparative metabolomics study using a human urine dataset collected from two individuals.

A better inter-group separation and a small intra-group variation were observed in the PCA score plots in comparison to data generated from previous workflow using imputation methods on MetaboAnalyst.

Moreover, we designed a graphical user interface based on the R shiny package and provided an easy access to all the processing functions; Figure 4.14 shows the current design of the program window. In comparison to the script format in the previous workflow, the graphical interface is more user friendly with all the adjustable parameters for each processing method. The program has been installed in our lab for processing data generated from different isotope labeling methods using different MS instruments. In future work, we will continue to integrate more data analysis methods in the workflow and provide a complete data analysis package for all types of metabolomics data.


CIL LC-MS
Raw data check
IsoMS
Alignment and zero-filling
IsoMS
Count IsoMS Peak pair number
Total peak intensity ratio
Instrument Type
Impact or Maxis II QTOF
Minimum retention time
120
Maximum retention time
2040
(Retention time range in second, default 2-34 min)
Minimum m/z
200
Maximum m/z
1000
(Mass range)
Saturation intensity
15000000
(Saturation intensity; For FT instrument, recommended value is 200000000)
Input SNR
5
Output SNR
10
(Input SNR for raw data filtering; output SNR for peak pair result filtering; If checked, following int values will NOT be used)
Input int
500
Output int
1000
(Input int for raw data filtering; output int for peak pair result filtering; NOT applicable if using SNR as filter)
Submit
Submit current setting and run IsoMS; if terminated while running, use 'Terminate all R scripts.R' to end IsoMS processing
Tag type
Dansyl
Minimum peak pair ratio
0.1
Maximum peak pair ratio
10
(Output peak pair ratio range)
Processing speed
3
4 is the fastest but may significantly slow your computer

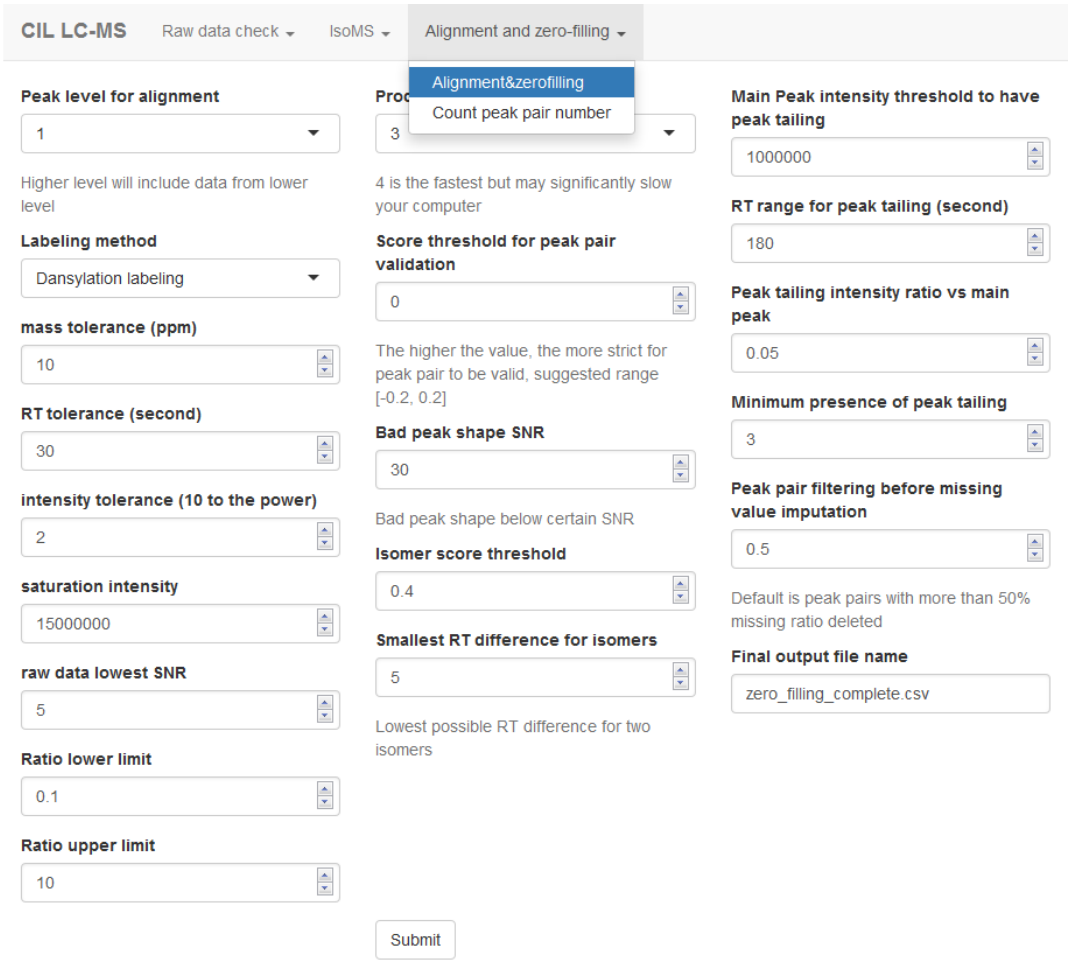


Figure 4.15 Graphical user interface for IsoMS, alignment and zero-fill.

Chapter 5 Improving Accuracy of Peak-Pair Intensity Ratio Measurement in Differential Chemical Isotope Labeling LC-MS for Quantitative Metabolomics

5.1 Introduction

Chemical isotope labeling (CIL) LC-MS has developed rapidly as a powerful tool for metabolomic profiling in metabolomics research. With differential isotope labeling of individual samples (e.g., $^{12}C_2$-labeling) and a control sample (e.g., $^{13}C_2$-labeling), the resulting mixtures ($^{12}C_2$-labeled sample and $^{13}C_2$-labeled control) can be analyzed using LC-MS to detect peak pairs of labeled metabolites from which the peak-pair intensity ratios can be measured to provide the basis of relative quantification of metabolites in different samples. A number of labeling reagents have been reported with varying degrees of enhancements in analytical performance.[30,103-105] With a rational design of the chemical structure of the labeling reagents used to derivatize a class of metabolites (i.e., a chemical-group-based submetabolome), both efficient LC separation and sensitive MS detection of labeled metabolites can be achieved, resulting in very high metabolomic coverage.[87,88] Using differential isotope labeling, where the light reagent is used to label the individual samples and the heavy reagent is used to label a control sample (e.g., a pooled sample for relative quantification or a standard with known concentration for absolute quantification), accurate and precise quantification of metabolites in comparative samples can be performed. To avoid the isotopic effect on chromatography separation of the light and heavy labeled metabolites, ^{13}C-encoded reagents are preferred over deuterium-containing reagents. However, ^{13}C-reagents may not be available at a reasonable cost compared to deuterium-based

isotope reagents. In some reagents, such as dansyl chloride (DnsCl) that has been widely used to label amine- and phenol-containing metabolites,[30] the relatively inexpensive form of ^{13}C-Dns contains only two ^{13}C atoms. As a result, the mass difference between the light ($^{12}C_2$-) and heavy ($^{13}C_2$-) labeled metabolite is 2 Da if one reagent tag is attached to a metabolite. Two other high-performance reagents, DmPA and dansylhydrazine, used to label carboxylic submetabolome[31,106] and carbonyl submetabolome,[88] respectively, also have a 2 Da difference between the light and heavy forms.

A mass difference of 2-Da apart between the light and heavy labeled metabolite may introduce an error in measuring the intensity ratio of the peak pair for relative metabolite quantification. This is because natural isotopologues of the light labeled metabolite may overlap with the heavy labeled peak and contribute their intensity to that of the heavy labeled peak, making the intensity ratio of $^{13}C_2$-peak vs. $^{12}C_2$-peak artificially higher. In this work, we have investigated the intensity contributions of the natural isotopes of various common elements present in human endogenous metabolites, including carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur. We report a data processing method that accounts for natural isotope contributions in the ratio calculation for $^{12}C_2$- and $^{13}C_2$-labeled peak pairs. It is shown that this method can improve the measurement accuracy for determining the peak intensity ratio of the light and heavy labeled metabolite in metabolomic profiling.

5.2 Materials and Methods

5.2.1 Chemicals and reagents

All the chemicals and reagents, unless otherwise stated, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). For the dansylation labeling reaction, the $^{12}C_2$-labeling reagent (dansyl chloride) was from Sigma-Aldrich, and the $^{13}C_2$-labeling reagent was synthesized in our lab using the procedure published previously.[30] LC-MS grade water, methanol, and acetonitrile (ACN) were purchased from ThermoFisher Scientific (Nepean, ON, Canada).

5.2.2 Human urine sample collection

Human urine samples were collected from two individuals to generate test data. Equal volumes of all the individual samples were mixed together to make a pooled sample. Each urine sample was centrifuged at 14,000 rpm for 10 min, and the supernatant was filtered twice through a 0.2 μm filter. The filtered urine was aliquoted and stored at -80°C until further use.

5.2.3 Dansylation labeling

A mixture solution of amino acid and peptide standards was prepared in 1:1 ACN/H_2O with a concentration of 0.1 mM for each. The frozen urine samples were thawed in an ice-bath. A 25 μL sample of urine or standard mixture solution was taken out for a labeling reaction in an Eppendorf tube. Then, 25 μL of 250 mM sodium carbonate/sodium bicarbonate buffer were added to the sample to introduce a basic environment for the labeling reaction. The solution was vortexed, spun down, and mixed with 50 μL of freshly prepared $^{12}C_2$-DnsCl solution (18 mg/mL) (for light labeling) or $^{13}C_2$-DnsCl solution (18 mg/mL) (for heavy labeling).[30] After the sample was incubated at 40 °C for 45 min, 10 μL of 250 mM NaOH were added to quench the excess dansyl chloride. The solution was incubated further at 40 °C for another 10 min to allow the

unreacted dansyl chloride to be hydrolyzed fully. Finally, 50 μL of formic acid (425 mM) in 1:1 ACN/H_2O were used to acidify the solution.

5.2.4 LC-MS

Each ^{12}C-labeled individual urine sample was mixed with the ^{13}C-labeled pooled urine sample in equal mole amounts based on the LC-UV measurement of labeled metabolites in individual and pooled samples.[107] The ^{12}C-labeled standard solution was mixed with an equal volume of the ^{13}C-labeled standard solution. The mixture was then ready to be analyzed by LC-MS using a Thermo Scientific Dionex Ultimate 3000 UHPLC System (Sunnyvale, CA) linked to a Bruker Impact quadrupole time-of-flight (Q-TOF) mass spectrometer (Bruker, Billerica, MA). The LC column was an Agilent reversed phase (RP) Eclipse Plus C18 column (2.1 mm × 10 cm, 1.8 μm particle size, 95 Å pore size). The LC gradient was: t = 0 min, 20% B; t = 3.5 min, 35% B; t = 18 min, 65% B; t = 24 min, 99% B; t = 34 min, 99% B. The flow rate was 0.18 mL/min.

5.2.5 Data analysis

All the spectra were first converted to .csv files by Bruker Daltonics Data Analysis 4.3 software. The peak pairs were extracted from .csv files by IsoMS.[108] Human urine data generated from multiple runs were aligned together based on the individual peak's accurate mass and retention time. The missing values in the aligned file were filled by Zerofill software.[76]

5.3 Results and Discussion

5.3.1 Peak ratio error

In differential CIL LC-MS using light and heavy reagents of 2-Da apart (e.g., $^{12}C_2$- and $^{13}C_2$-dansyl chloride), quantification is performed by analyzing the mixture of the $^{12}C_2$-labeled individual sample and $^{13}C_2$-labeled control. Since the same amount of ^{13}C-labeled control is spiked into the ^{12}C-labeled individual samples, the intensity ratio values of a labeled peak pair of a metabolite measured from different samples reflect the concentration differences of the metabolite in these comparative samples. If the $^{13}C_2$-labeled control is a standard of known concentration, the absolute concentration of this metabolite in a sample also can be determined. In both cases, measuring the peak intensity ratio of the ^{12}C-/^{13}C-labeled peak pair is needed. Table 5.1 shows the peak pair ratio values measured experimentally from LC-MS analysis of 1:1 (in mole) ^{12}C-/^{13}C-dansyl labeled standards. The expected ratio of the ^{13}C-labeled peak vs. the ^{12}C-labeled peak is 1.0. However, as Table 1 shows, the measured ratios calculated based on their intensities are greater than 1, with the highest ratio of 1.239 for dansyl labeled Gly-Gly-Phe-Leu. Thus, the error can be as high as 23.9%.

Table 5.1 Peak pair ratios of 1:1 (in mole) ^{12}C-/^{13}C-dansyl labeled standards and adjusted peak pair ratios after excluding the natural isotope intensity of the light labeled peak.

Name	mz_light	#C	#S	#O	Peak pair ratio	2nd isotope intensity by fitting function	Adjusted peak pair ratio	Theoretical 2nd isotope intensity	Adjusted peak pair ratio
L-Alanine	323.106	15	1	4	1.131	0.0525	1.078	0.053	1.078
Ammonium chloride	251.0849	12	1	2	1.096	0.0497	1.047	0.046	1.051
L-Arginine	408.17	18	1	4	1.135	0.0599	1.075	0.058	1.077
L-Aspartic acid	367.0958	16	1	6	1.061	0.0558	1.006	0.057	1.004
L-Glutamic acid	381.1115	17	1	6	1.084	0.057	1.027	0.059	1.025
Glycine	309.0903	14	1	4	1.066	0.0517	1.014	0.051	1.015

L-Histidine	389.1278	18	1	4	1.08	0.0578	1.022	0.058	1.022
L-Isoleucine	365.1529	18	1	4	1.11	0.0556	1.055	0.058	1.053
L-Methionine	383.1094	17	2	4	1.134	0.0572	1.077	0.092	1.043
L-Phenylalanine	399.1373	21	1	4	1.137	0.0589	1.078	0.063	1.074
L-Proline	349.1216	17	1	4	1.041	0.0543	0.987	0.056	0.985
L-Serine	339.1009	15	1	5	1.11	0.0536	1.057	0.054	1.056
L-Threonine	353.1166	16	1	5	1.124	0.0546	1.07	0.056	1.068
L-Valine	351.1373	17	1	4	1.116	0.0544	1.062	0.056	1.06
Trp-Gly-Gly	552.1911	27	1	6	1.022	0.0825	0.939	0.081	0.941
Gly-Gly-Phe-Leu	626.2643	31	1	7	1.239	0.0991	1.14	0.094	1.145

5.3.2 Theoretical isotopologue intensity

The ratio difference between the expected value and the measured value, as shown in Table 1, can be attributed to the contribution of natural abundance peaks from the light labeled metabolite. Human endogenous metabolites contain mainly hydrogen, carbon, oxygen, nitrogen, sulfur, and phosphorus.[109] The natural isotope abundance of these elements at a peak of +2-Da from the light labeled metabolite can be calculated readily. For example, for a compound containing n number of carbons (n=1, 2, 3 …), its mass spectrum, in theory, should have n number of ^{13}C isotopic peaks. The k^{th} isotopic peak (k=1, 2, 3…, n) occurs when the number of carbons, k, in this compound happens to be ^{13}C. Based on the binomial distribution model,[110] the probability of the k^{th} isotopic peak can be described as,

$$P(k) = \binom{n}{k} \times 1.109\%^{k} \times (1 - 1.109\%)^{(n-k)} \quad (5.1)$$

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \quad (5.2)$$

where $\binom{n}{k}$ calculates the number of possible combinations of k ^{13}C out of n carbons in a compound.

The relative peak intensity of the k^{th} ^{13}C isotopic peak vs. the main peak (all ^{12}C molecular peak) is equal to the probability ratio of these two isotope peaks. Therefore, the relative intensity of the k^{th} isotope peak can be calculated as,

$$k^{th}\ peak\ relative\ intensity = \frac{\binom{n}{k} \times 1.109\%^{k} \times (1 - 1.109\%)^{(n-k)}}{(1 - 1.109\%)^{n}} \qquad (5.3)$$

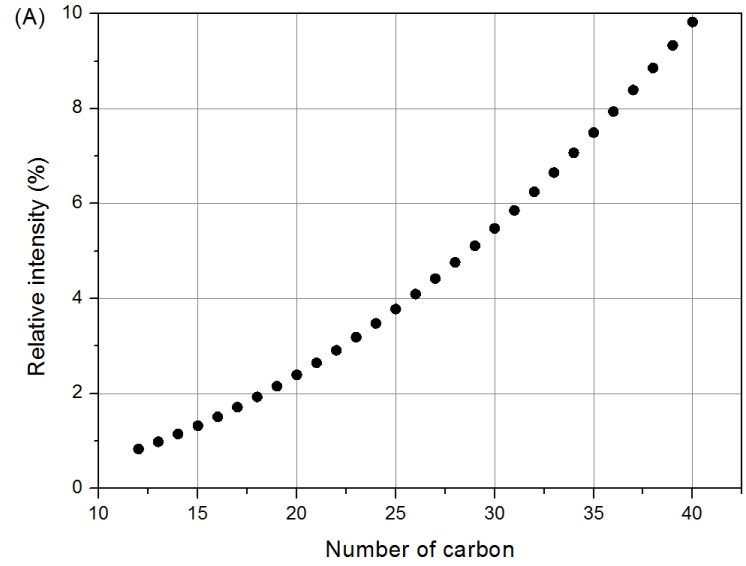
where $(1 - 1.109\%)^{n}$ is the probability of all carbons in the compound being ^{12}C. From Equation (5.3) we can calculate the theoretical relative intensity of the first and second ^{13}C isotope peak as 0.0112**n* and 0.000063**n**(*n*-1), respectively. For other elements, we can calculate their isotopologue relative intensity using equation (5.3) by substituting the isotope abundance. Using this method, we can calculate the theoretical relative peak intensity for any isotopologue. Table 5.2 lists the natural abundance of isotopes of common elements in human endogenous metabolites, including carbon, hydrogen, oxygen, nitrogen, and sulfur. Isotopes with an abundance less than 0.01% (e.g., phosphorus) are not listed. The Δ*m* column shows the monoisotopic mass difference of each isotope with respect to the most abundant isotope in that element. Due to mass defect, Δ*m* of adjacent isotopes in different elements can have a difference of a few mDa.

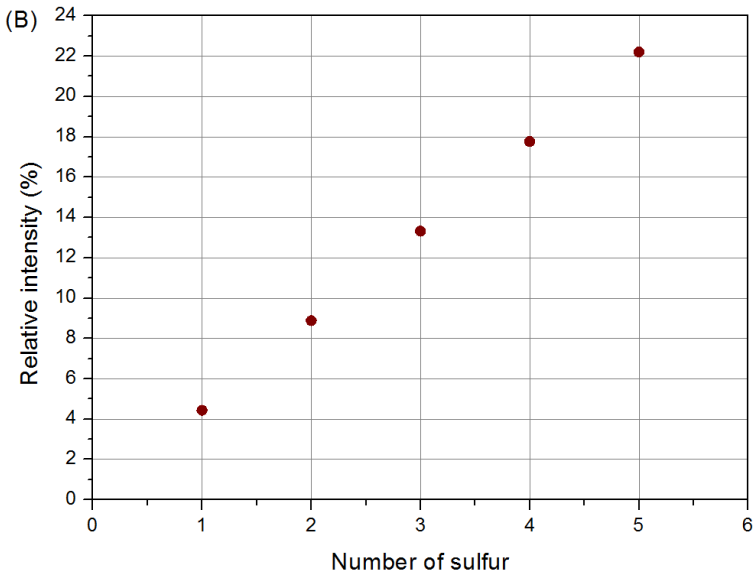
The $^{13}C_2$ and ^{34}S of the light ($^{12}C_2$-) labeled peak are found to be the major contributors to the heavy ($^{13}C_2$-) labeled peak in a quadrupole time-of-flight (QTOF) mass spectrometer, which generally has a resolving power of less than 50,000 for detecting low molecular mass ions (<1000 Da). The ^{34}S peak is separable from the $^{13}C_2$ peak when using ultrahigh resolution mass spectrometry, such as FT-ICR-MS or Orbitrap-MS. However, compared to these ion-trap-based high resolution devices, QTOF ion detection has a higher tolerance to ion saturation.[111] As a consequence, a larger amount of sample can be injected in order to detect the relatively lower abundance metabolites in a complex metabolome sample. For example, in our experience, the Bruker Impact QTOF instrument can detect about 20–25% more peak pairs, compared to the Bruker 9.4-T Fourier-transform mass spectrometer in $^{12}C_2$-/$^{13}C_2$-dansyl labeled samples such as human urine.

To evaluate the natural isotopologue interference with the +2-Da labeled peak, we plot the relative intensity of the $^{13}C_2$, ^{34}S and ^{18}O peak as a function of the number of atoms (for each element) in a dansyl labeled molecule. Other potential peaks at +2-Da, such as the $^{15}N^{13}C$, $^{2}H_2$, $^{2}H^{13}C$, and $^{33}S^{13}C$ peak, are not shown due to their low relative intensity for metabolites with molecular mass less than 1000 Da. The dansyl labeling tag will introduce 12 carbon, 1 sulfur and 2 oxygen atoms to the labeled molecule. Figure 5.1 clearly shows that the relative peak intensity of the natural abundance peak of the light labeled peak can be greater than 10% at the +2-Da heavy labeled peak. As the molecular mass of a metabolite increases, this intensity contribution or error would increase.

Table 5.2 Stable isotope abundance of common elements in endogenous human metabolites.

Isotope	Δm (Da)	Abundance
^{12}C	0	98.9%
^{13}C	1.00335	1.1%
^{1}H	0	99.98%
^{2}H	1.00628	0.02%
^{14}N	0	99.6%
^{15}N	0.99704	0.4%
^{16}O	0	99.76%
^{17}O	1.00422	0.04%
^{18}O	2.00425	0.20%
^{32}S	0	94.99%
^{33}S	0.99939	0.75%
^{34}S	1.9958	4.25%


(A)
Relative intensity (%)
Number of carbon
0
2
4
6
8
10
10
15
20
25
30
35
40


(B)
Relative intensity (%)
Number of sulfur
0
2
4
6
8
10
12
14
16
18
20
22
24
0
1
2
3
4
5
6

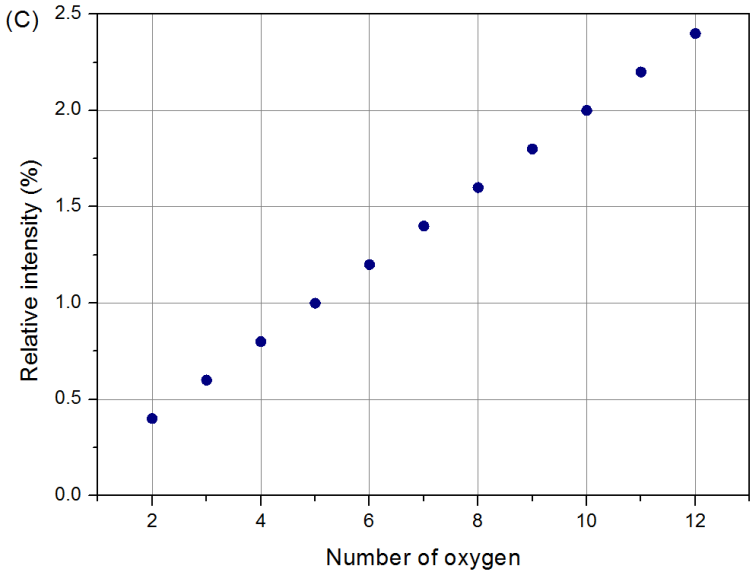


Figure 5.1 Relative intensity of +2-Da natural isotope peak as a function of (A) number of carbon, (B) number of sulfur, and (C) number of oxygen.

5.3.3 Excluding natural isotopologue contributions in ratio calculation

If one knows the chemical structure of the metabolite, the natural isotopologue contribution of the light labeled metabolite to the peak intensity of the +2-Da heavy labeled metabolite can be calculated (e.g., using Equation 5.3) and then subtracted from the measured intensity of the +2-Da peak to arrive at an accurate value of the $^{13}C_2$-labeled peak intensity. This approach is similar to those reported for correcting natural isotope contributions in lipid and metabolite analyses.[112-117] However, in untargeted metabolomic profiling, the chemical structures are not known, *a priori*, and many metabolites cannot be identified positively. Thus, we cannot rely on the elemental composition information to subtract out the natural isotope contribution.

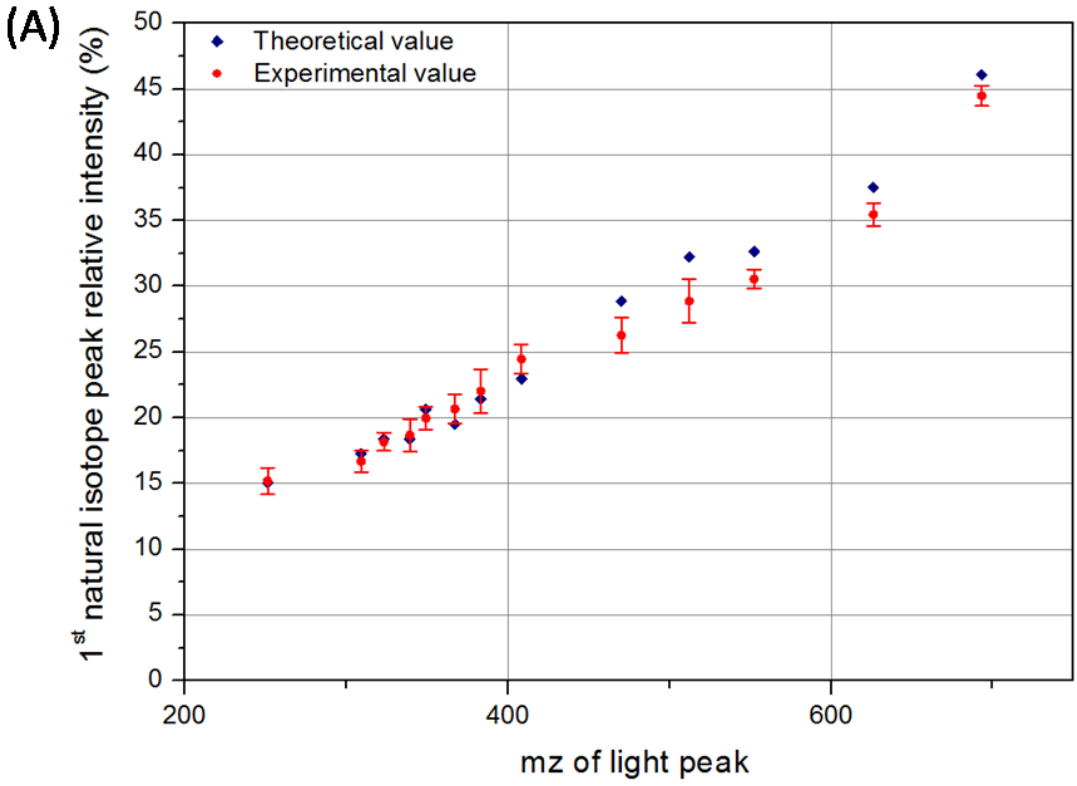
We examined the possibility of using the molecular mass (strictly m/z, but z is usually +1, for dansyl labeled ions) of the labeled metabolite to calculate the intensity contribution of the

natural isotopologue of the light labeled metabolite. Table 5.3 summarizes the results of the 1st and 2nd natural isotope peak relative intensities for 13 labeled standards. Figure 5.2(A) shows the intensity of the 1st natural isotope peak (+1-Da peak) of the light labeled standards as a function of their m/z values. The experimental relative intensity shows a linear relation with the m/z of the standards. Figure 5.2(B) shows the intensity of the 2nd natural isotope peak (+2-Da peak) of the light labeled standards as a function of their m/z values. The experimental value and the theoretical value have a quadratic relation with m/z. In both cases, as the molecular mass increases, the carbon number increases and the relative intensity of the 1st and 2nd isotope peaks increases. We note that there are some variations in replicate measurements of the relative intensity values shown in Figure 5.2. This is expected as the QTOF-MS instrumental setup was optimized to detect as many labeled metabolites as possible and not for targeted analysis of one or a few metabolites. Thus, the signal-to-noise ratio and signal integration time of a given metabolite ion may not be ideal to generate the perfect isotope abundance pattern. Overall, as the results of Figure 5.2 show, the QTOF-MS instrument measured the isotope abundances well. More importantly, the quadratic curve fit shown in Figure 5.2(B) can be useful in determining the relative intensity of the 2nd natural isotope peak of a light labeled metabolite based on its m/z value. Thus, without knowing the elemental composition of a metabolite, we propose to use the m/z value of the light labeled peak to estimate the relative intensity of its 2nd natural isotope peak.

We used a set of dansyl labeled dipeptides to test the accuracy of the quadratic equation for calculating the 2nd natural isotope intensity. Table 5.4 shows the calculated values, along with the theoretical values and the measured values. All three values are very close to each other, indicating the effectiveness of using the quadratic equation for determining the natural isotope peak intensity.

We applied the quadratic equation to calculate the 2nd isotope intensity of the dansyl labeled standards, and the results are shown in Table 5.1 (i.e., in the column entitled 2nd isotope intensity by fitting function). We subtracted this value from the heavy labeled peak intensity and then re-calculated the peak-pair intensity ratio ($^{13}C_2$-/$^{12}C_2$-labeled peak) to arrive at an adjusted peak pair ratio (see Table 5.1). This adjusted ratio is closer to 1.0, with an average ratio of 1.044 for these 15 labeled standards (methionine is not included; see discussion in the next Section for this special case) and a standard derivation of 0.046. Compared to an average ratio of 1.104 before subtracting the calculated contribution of the natural isotope peak of the lighted labeled metabolite, this adjusted ratio reduces the average error to 4.4% from 10.4% without adjustment.

Table 5.1 also shows the theoretical 2nd isotope intensity as the elemental compositions of these standards are known. The adjusted peak ratio using the theoretical intensity for each labeled metabolite, instead of the calculated intensity with the quadratic equation, is listed at the last column in Table 5.1. The average ratio found using the theoretical intensity values is 1.044, which is the same as the average ratio found using the calculated intensity. These results indicate that the proposed method of using the quadratic equation to calculate the natural isotope contribution and then subtract it from the measured intensity of the heavy labeled peak is an effective method to improve peak-ratio measurement accuracy.


(A)
Theoretical value
Experimental value
1st natural isotope peak relative intensity (%)
mz of light peak
0
5
10
15
20
25
30
35
40
45
50
200
400
600

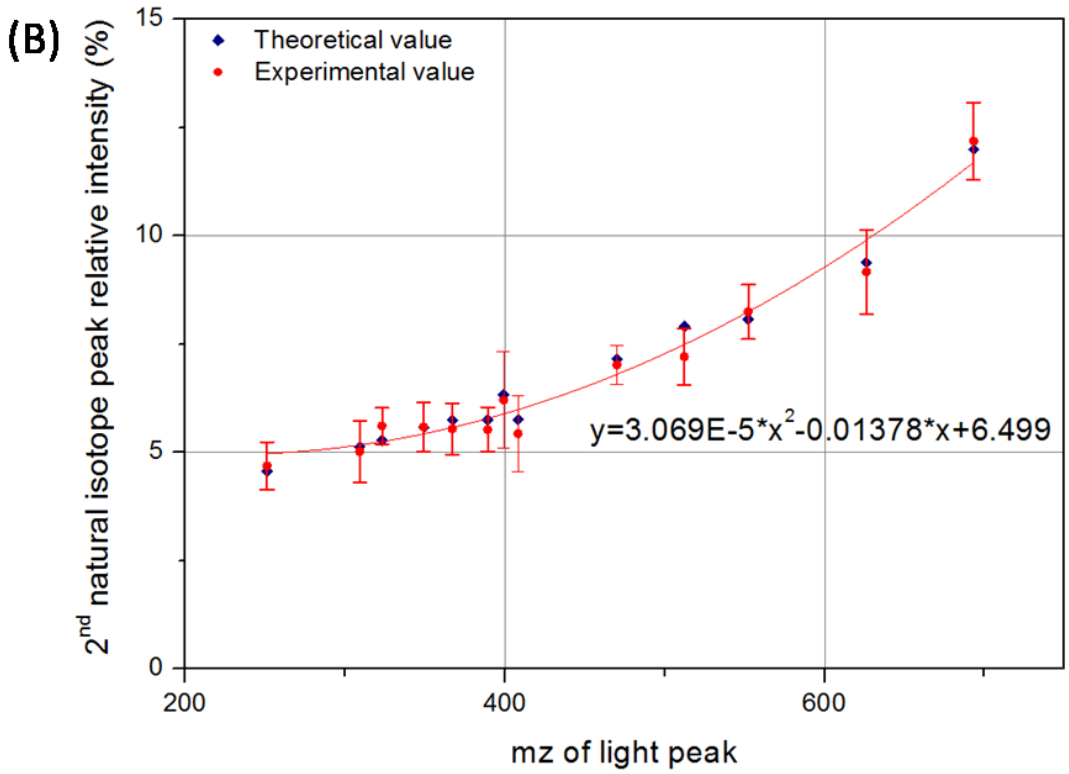


Figure 5.2 The theoretical and experimental (A) +1-Da and (B) +2-Da natural isotope peak relative intensity for standards with m/z of 250 to 700.

Table 5.3 Amino acid and dipeptide standards for investigating the relationship between the isotope peak intensity and the mass of the light labeled peak. Each experimental value was calculated from the peak intensities in multiple mass scans (*n*=5).

Name	mz_light	#C	# S	# O	#N	1st isotope peak theoretical intensity (%)	1st isotope peak exp't intensity (%)	1st isotope peak standard deviation (%)	2nd isotope peak theoretical intensity (%)	2nd isotope peak exp't intensity (%)	2nd isotope peak standard deviation (%)
L-Alanine	323.1060	15	1	4	2	18.394	18.17	0.69	5.275	5.61	0.42
Ammonium chloride	251.0849	12	1	2	2	15.034	15.19	1.00	4.563	4.69	0.55
L-Arginine	408.17	18	1	4	5	22.96	24.48	1.12	5.758	5.43	0.87
L-Aspartic acid	367.0958	16	1	6	2	19.514	20.67	1.08	5.746	5.54	0.59
Glycine	309.0903	14	1	4	2	17.274	16.69	0.82	5.134	5.02	0.72
L-Histidine	389.1278	18	1	4	4	22.558	23.22	0.61	5.758	5.53	0.51
Phenylalanine	399.1373	21	1	4	2	25.114	25.62	1.80	6.332	6.21	1.12
L-Proline	349.1216	17	1	4	2	20.634	19.97	0.87	5.587	5.59	0.57
Ala-Phe	470.1744	24	1	5	3	28.876	26.28	1.34	7.156	7.02	0.44
Leu-Phe	512.2214	27	1	5	3	32.236	28.86	1.66	7.91	7.21	0.65

Trp-Gly-Gly	552.1911	27	1	6	4	32.638	30.55	0.73	8.07	8.25	0.63
Gly-Gly-Phe-Leu	626.2643	31	1	7	5	37.52	35.47	0.87	9.378	9.17	0.97
Phe-phe-phe	693.2741	39	1	6	4	46.078	44.48	0.78	11.994	12.19	0.89

Table 5.4 Relative intensity of the +2-Da natural isotope peak of dipeptide standards for evaluating the quadratic fitting curve generated from Figure 5.2 (B).

Name	mz_light	Theoretical value (%)	Prediction (%)	Experimental value (%)	Error to prediction (%)
His-Ile	502.2128	7.129	7.319	6.704	0.616
Glu-Gly	438.1329	6.04	6.353	6.278	0.075
Gly-Asp	424.1173	6.211	6.175	6.127	0.048
Asp-Thr	468.1435	6.744	6.774	6.427	0.347
His-Thr	490.1755	6.836	7.118	6.835	0.283
His-Val	488.1962	6.897	7.086	6.464	0.623
Trp-Trp	624.2275	9.996	9.856	9.865	0.01
Thr-Ala	424.1537	6.232	6.175	5.793	0.382

5.3.4 Sulfur natural isotopologue intensity contribution

The above discussion only considered the natural isotope peak contributions of carbon. One of the standards in Table 5.1 is methionine, which contains one sulfur atom. The use of molecular mass to calculate the contribution of the $^{13}C_2$ natural isotope peak to the heavy labeled peak, as discussed in Section 5.3.3, works well if there is no additional S present in a metabolite. For an untargeted analysis of a metabolome, we do not know whether a metabolite contains S or not. The presence of one sulfur atom in a metabolite will add 4.47% to the heavy labeled peak. The question is whether this small contribution will introduce a bias in the relative quantification of metabolites among different comparative samples. In metabolomics, the first task of metabolomic profiling is to determine the significant metabolites that can differentiate two or more groups of samples; relative quantification is required for this task. If one S is present in a

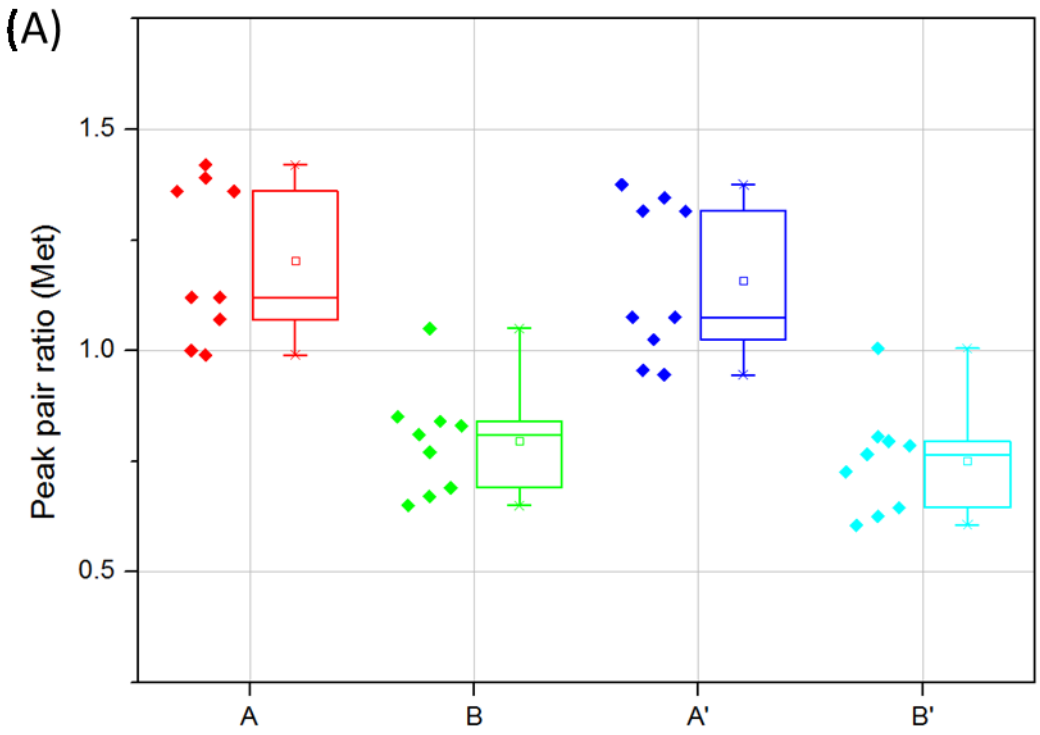
metabolite, does it cause an error in relative quantification? We used an example of analyzing ^{13}C-/^{12}C-dansyl labeled human urine to answer this question.

Figure 5.3 (A) shows the box plots of peak pair ratios of methionine (Met) in two groups of samples (group A, n=9; group B, n=9), and their corresponding ratio values are shown in Table 5.5. In Table 5.5, columns A and B show the peak pair ratios that were calculated from light and heavy peak intensities that were adjusted using the isotope intensity fitting line. Each peak pair was adjusted further by removing the extra sulfur contribution, and their values are shown in columns A' and B'. The fold change of group B over group A is 0.661 and 0.648 before and after subtracting the sulfur contribution, respectively. Although different in the exact fold change value, both results show a significant change of methionine in the two biological groups. This example illustrates that the presence of one sulfur atom in a metabolite does not cause a significant difference in calculating the peak ratio change or in the relative metabolite quantification.

The presence of two or more S atoms in a metabolite would increase the ^{34}S isotope peak contribution to the heavy labeled metabolite (n×4.47% where n = number of sulfur atoms); the value 4.47% is calculated by the abundance ratio of S34/S32=4.25%/94.99%. Table 5.6 shows ratio data of a putatively identified compound methionyl-methionine (Met-Met) that contains two sulfur atoms. Columns A' and B' show the ratios after excluding sulfur contribution. Figure 5.3 (B) shows the box plot of the ratio data in all groups. The folder change of group B over group A is 1.07 and 1.08 before and after removing the S contribution, respectively.

The above discussion indicates that the presence of S (1 or 2) does not cause a major problem in relative quantification; however, it can affect the absolute quantification. For absolute quantification, we determine the absolute concentration of a metabolite in a pooled sample using

a labeled standard of known concentration. We then use the peak ratio of an individual sample vs. the pooled sample of the metabolite multiplied by the absolute concentration of the pool to determine the absolute concentration of this metabolite in the individual sample. Fortunately, at the research stage where we want to determine the absolute concentration of a metabolite (e.g., this metabolite is considered to be a biomarker of a certain disease), the metabolite structure would be known and the number of S atoms present in a molecule is known. We can calculate and subtract out the contribution of the S atoms to the heavy labeled peak by n×4.47%, where *n* = number of sulfur atoms, to determine the absolute concentration of the metabolite in individual samples.



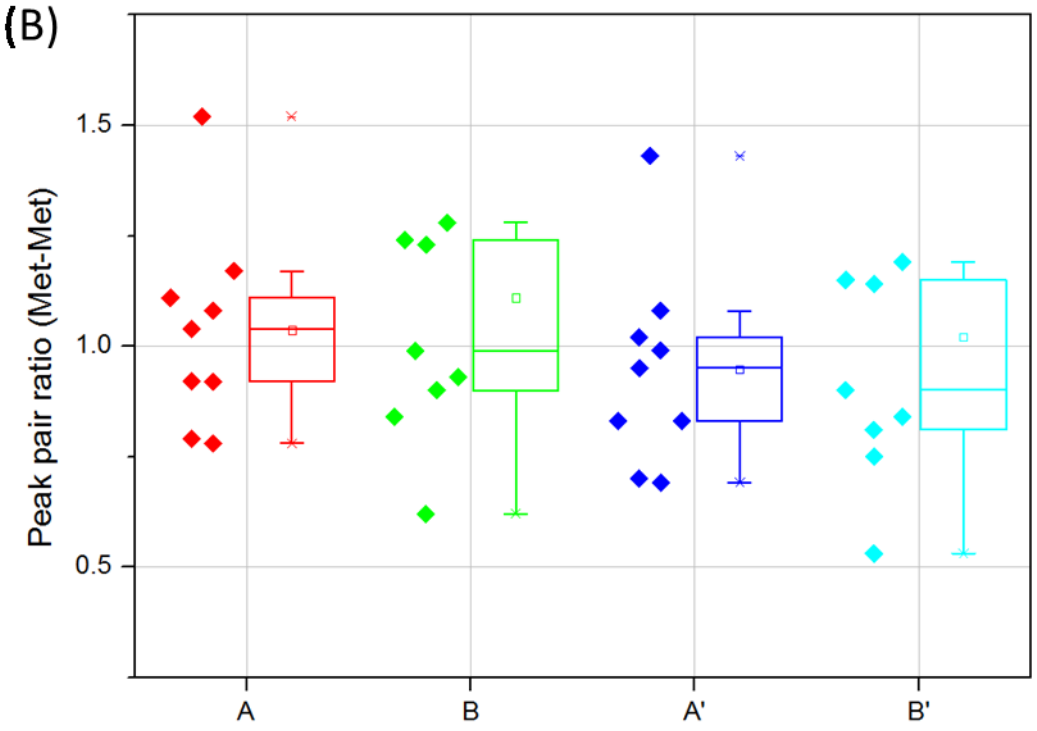


Figure 5.3 Boxplot of peak intensity ratio data for (A) Met and (B) Met-Met for two groups of human urine, A and B. A' and B' are the data after excluding the sulfur contribution to the heavy peak intensity.

Table 5.5 Peak pair ratio of Met in group A and group B. A' and B' show the ratio values after removing the sulfur isotope contribution.

	A	B	A'	B'
Peak pair ratio	1.07	0.67	1.03	0.63
	1.39	0.84	1.35	0.80
	1.36	0.81	1.32	0.77
	1.12	1.05	1.08	1.01
	0.99	0.83	0.95	0.79
	1.00	0.65	0.96	0.61
	1.36	0.77	1.32	0.73
	1.12	0.69	1.08	0.65
	1.42	0.85	1.38	0.81

Average	1.20	0.80	1.16	0.75
Standard deviation	0.18	0.12	0.18	0.12

Table 5.6 Peak pair ratio of Met-Met in group A and group B. A' and B' show the ratio values after removing the isotope contribution of two sulfur atoms.

	A	B	A'	B'
Peak pair ratio	1.08	0.62	0.99	0.53
	1.04	1.23	0.95	1.14
	1.17	0.90	1.08	0.81
	1.11	0.99	1.02	0.90
	0.78	1.24	0.69	1.15
	0.92	0.93	0.83	0.84
	1.52	0.84	1.43	0.75
	0.79	1.28	0.70	1.19
	0.92	1.96	0.83	1.87
Average	1.08	1.11	0.95	1.02
Standard deviation	0.21	0.36	0.13	0.27

5.4 Conclusions

We report a detailed study of the natural isotopologue intensity of a light ($^{13}C_2$-) labeled metabolite and its potential interference in the peak pair ratio calculation in chemical isotope labeling LC-MS using QTOF-MS. For a carbon natural isotope contribution, a new algorithm based on the m/z value to estimate the 2nd natural isotope peak intensity was developed to reduce its interference with the heavy ($^{13}C_2$-) labeled peak intensity. For a sulfur natural isotope contribution, with one or two sulfur atoms in a metabolite, a relative quantification of peak ratios from two groups of samples still can be performed without a correction. However, for determining the absolute concentration of a metabolite containing sulfur, a sulfur natural

isotopologue contribution needs to be subtracted from the +2-Da heavy labeled peak in order to reduce the peak ratio error.

Chapter 6 Intensity-dependent Mass Search in Liquid Chromatography Mass Spectrometry Based Metabolomics

6.1 Introduction

Mass spectrometry (MS)-based metabolomics has advanced rapidly in recent years. One of the central challenges in metabolomics is metabolite identification. Regardless of whether one uses targeted or untargeted metabolomics, eventually all paths lead to the requirement of identifying and quantifying certain key metabolites. Without metabolite identification, the results of any metabolomic analysis are biologically and chemically uninterpretable. Given the chemical diversity of most metabolomes and the characteristics of most metabolomic data, metabolite identification is intrinsically difficult. Over the past decade, a great deal of effort in metabolomics has been focused on making metabolite identification better, faster, and cheaper. The fast growing metabolite databases have facilitated the metabolite identification in metabolomics studies greatly.[2,70,118-122] However, a mass-based database search is still a challenging step in metabolomics as many potential structures can match to a single query mass within the search window. Currently, a mass-based library search relies on accurate mass information from the user along with a mass tolerance window. The mass window varies according to the type of instrument and MS setups used in the metabolome analysis. It requires the highest possible mass accuracy and a carefully chosen mass tolerance to obtain accurate match results.

The choice of search tolerance is often derived by one's experience of the mass error from a specific MS instrument and MS settings. Too large a tolerance may lead to many possible

matches to one query mass, while a narrow tolerance would lower the number of successful matches. To increase the number of metabolites identified, one can increase the mass tolerance in the search, however, this will increase the manual effort in interpreting the results.

The precision of a mass measurement is dependent on the number of ions sampled in the measurement and is likely to be different for every measurement. Due to the complexity of the metabolome, a wide range of signal intensities can be detected from different metabolites in a given sample. The difference in the peak intensity can influence the mass peak shape and affect the mass precision further.[123] In the experimental data, metabolites with a different signal intensity were observed to give a different mass error. The current use of a fixed tolerance in the search assumed the same mass error for all query masses. This is true when analyzing standards with an optimal injection amount in which the detected intensities are relatively high. In real sample analysis, the intensity of the metabolites can vary by a few magnitudes. Compared to the data from chemical standards, the biological sample data may contain more low intensity peaks; in this case, the accuracy of the mass measurement can vary for different peaks.

In this work, an intensity-dependent mass accuracy was studied and a correlation of mass error with mass peak intensity was observed in TOF-MS based data. Mass error distribution was investigated using data from chemical standards at different peak intensities. An intensity-dependent mass tolerance method was developed and applied in a mass-based data search. All programs were developed using R language. We will implement the function at the public website (www.mycompoundid.org). The new search function allows users to generate an intensity-dependent mass tolerance for each query mass and ultimately improve the accuracy in the library search results.

6.2 Materials and Methods

6.2.1 Preparation of sodium formate solution for mass calibration

A sodium formate calibration solution was prepared by mixing formic acid and sodium hydroxide in 50% isopropanol in water. Different orders of sodium formate adducts will be produced in the solution in the form of Na(NaCOOH)*n* (*n*=1, 2, 3, ...). Table 6.1 shows the full list of sodium formate adducts used in the calibration for an m/z range below 1000. The concentration and injection volume were adjusted so that all sodium formate adduct peaks are detected with a relatively high intensity to ensure the quality of mass calibration (see in Figure 6.1). The sodium formate calibration solution was injected into the mass spectrometer during the first 2 min of each LC-MS analysis and served as an external mass calibration segment.

Table 6.1 The sodium formate adducts list used in positive mode mass calibration.

Formula	m/z
$Na(NaCOOH)_3$	226.951493
$Na(NaCOOH)_4$	294.938917
$Na(NaCOOH)_5$	362.926341
$Na(NaCOOH)_6$	430.913765
$Na(NaCOOH)_7$	498.901189
$Na(NaCOOH)_8$	566.888613
$Na(NaCOOH)_9$	634.876037
$Na(NaCOOH)_{10}$	702.863461
$Na(NaCOOH)_{11}$	770.850884
$Na(NaCOOH)_{12}$	838.838308
$Na(NaCOOH)_{13}$	906.825732
$Na(NaCOOH)_{14}$	974.813156

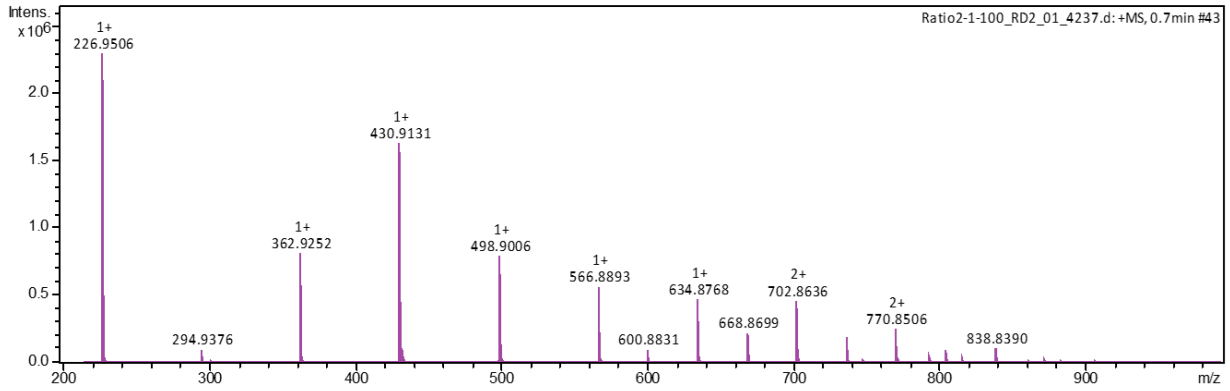


Figure 6.1 An example of the mass spectrum of sodium formate adducts for mass calibration.

6.2.2 Sodium formate series injection

To study the mass accuracy at different peak intensities, a sodium formate solution was diluted into a series of concentrations to generate data with various intensities covering the linear range of the MS detector. The analysis was done using two Bruker mass spectrometers of maXis impact and maXis II quadrupole time-of-flight (Q-TOF) (Bruker, Billerica, MA).

6.2.3 Standard mixture analysis

A standard mixture solution was prepared by mixing 22 selected standards (see Table 6.2) with the same concentration of each. The resulting sample was labeled with ^{12}C-/^{13}C-isotope dansyl chloride using the labeling protocol.[30] The light and heavy labeled samples were mixed in a 1:1 ratio, followed by the analysis of LC-QTOF-MS. The mass of the dansylated standards cover a mass range from 250 to 700, consistent with most small metabolites. The mixture was diluted into different concentrations to obtain standards with mass peaks of different intensities.

Table 6.2 Twenty-two standards used for the preparation of the standard mixture solution.

Name	Accurate mass	m/z light peak	m/z heavy peak	RT (sec)
L-Alanine	89.0477	323.1060	325.1127	377
Ammonium chloride	17.0266	251.0849	253.0916	317
L-Arginine	174.1117	408.1700	410.1767	154
L-Aspartic acid	133.0375	367.0958	369.1025	282
L-Cystine	240.0238	354.0702	356.0769	510
L-Glutamic acid	147.0532	381.1115	383.1182	279
Glycine	75.0320	309.0903	311.0970	341
L-Histidine	155.0695	389.1278	391.1345	114
L-Isoleucine	131.0946	365.1529	367.1596	500
L-Lysine	146.1055	307.1111	309.1178	556
L-Methionine	149.0510	383.1094	385.1161	457
L-Phenylalanine	165.0790	399.1373	401.1440	493
L-Proline	115.0633	349.1216	351.1283	445
L-Serine	105.0426	339.1009	341.1076	258
L-Threonine	119.0582	353.1166	355.1233	308
L-Tyrosine	181.0739	324.5953	326.6020	630
L-Valine	117.0790	351.1373	353.1440	459
Ala-Phe	236.1160	470.1744	472.1811	476
Leu-Phe	278.1630	512.2214	514.2281	543
Try-Gly-Gly	318.1328	552.1911	554.1978	371
Gly-Gly-Phe-Leu	392.2060	626.2643	628.2710	476
Phe-phe-phe	459.2158	693.2741	695.2808	579

6.2.4 HPLC-QTOF-MS analysis of human urine

Human urine samples were collected from six healthy individuals and filtered using a 0.22 μm pore size filter (Millipore Corp., MA). A pooled sample was prepared by mixing equal volumes of all individual urine samples. The individual and pooled samples were labeled with ^{12}C-/^{13}C-isotope dansyl chloride using the labeling protocol.[30] After centrifugation at 20,800 g for 10 min, 25 μL of the supernatant of each individual sample were transferred into an Eppendorf for a labeling reaction. Next, 25 μL of 250 mM sodium carbonate/sodium bicarbonate buffer and 25 μL of ACN were added to the sample. The solution was vortexed, spun down, and mixed with 50 μL of freshly prepared ^{12}C-dansyl chloride solution (18 mg/mL, for light labeling) or ^{13}C-dansyl chloride solution (18 mg/mL, for heavy labeling). After 45 min incubation at 40 °C, 10 μL of 250 mM NaOH were added to the reaction mixture to quench the excess dansyl chloride. The solution was then incubated at 40 °C for another 10 min. Finally, 50 μL of formic acid (425 mM) in 50/50 ACN/H_2O were added to consume excess NaOH and to make the solution acidic. Equal volumes of the light and heavy labeled pooled urine were mixed to generate a QC sample and were injected every 10 sample runs to monitor the instrument stability. During the data analysis, QC data was used for metabolite identification as it contains all the metabolites from the individuals. LC-MS analysis was performed on the Bruker QTOF-MS equipped with an Agilent 1100 HPLC system (Palo Alto, CA). A reversed-phase Zorbax Eclipse C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1% (v/v) formic acid in water with 5% (v/v) ACN, and solvent B was 0.1% (v/v) formic acid in ACN. The gradient elution profile was as follows: t=0.0min, 20%B; t=3.5min, 35% B; t=18.0 min, 65% B; t=24 min, 99% B; t=28 min, 99% B. The flow rate was 180 μL/min. The sample injection

volume was optimized for the highest detectability. All the spectra were collected using the positive ion mode.

6.3 Results and Discussion

6.3.1 Overall workflow

Figure 6.2 shows the overall workflow of an intensity dependent mass search. During instrumental analysis, the sodium formate calibration solution was injected into the mass spectrometer during the LC dead time for a post-acquisition mass calibration on each LC-MS run. The mass data from the calibration segment was extracted first from all sample data in order to investigate the relationship between mass error and peak intensity. The mass error was calculated for each sodium formate peak at different peak intensities. A mathematical relationship was derived between the mass tolerance and the peak intensity. Finally, an intensity dependent mass tolerance was assigned for each mass in the sample data table to be used in the standards library.

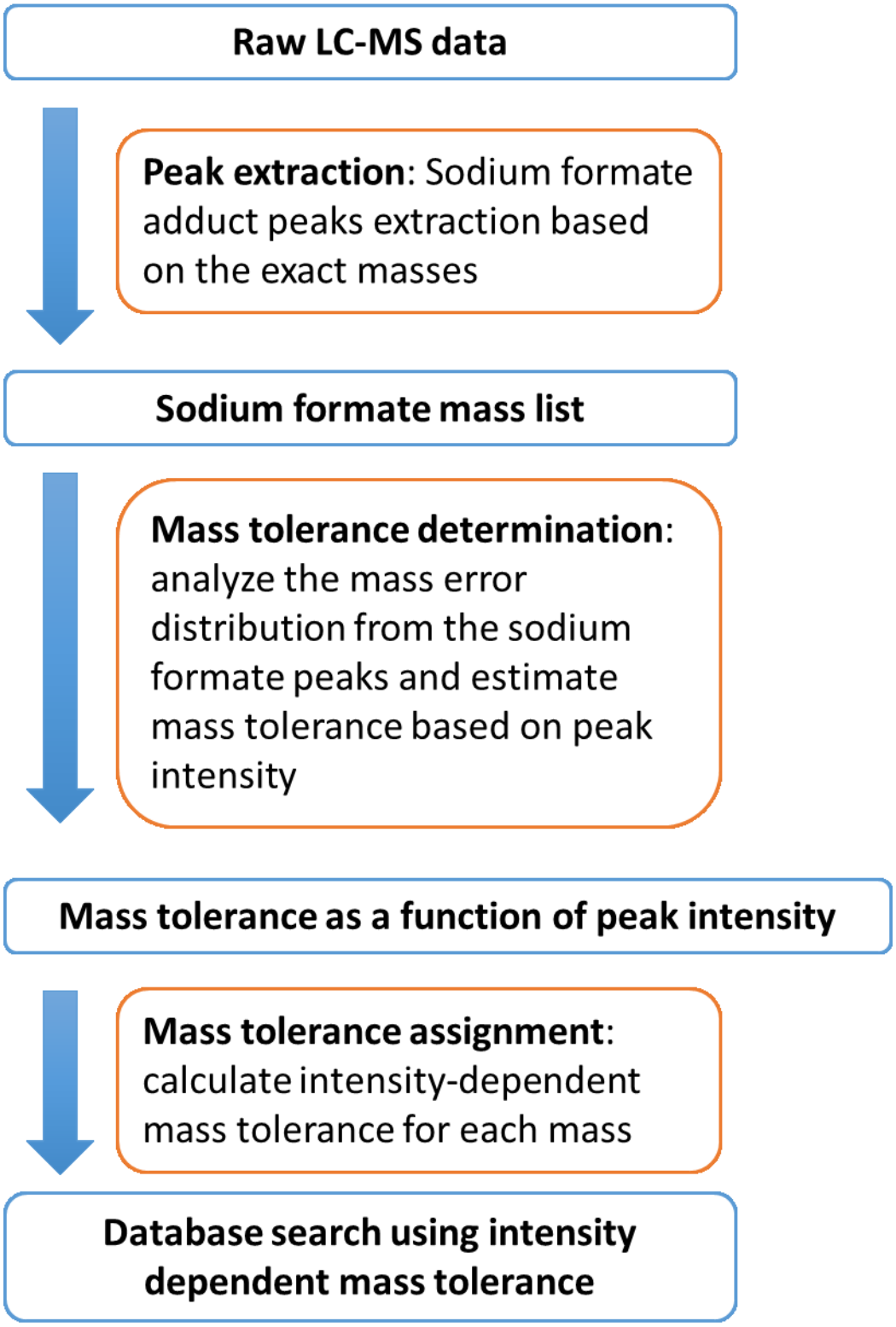


Figure 6.2 Overall workflow for an intensity dependent mass search.

6.3.2 Mass calibration and mass error distribution

The error for a measured mass is composed of two components, the systematic error (the accuracy of the measurement) and the statistical or random error (the precision of the measurement). A quadrupole time-of-flight mass spectrometer (QTOF) in general has a resolving power of up to 50,000 for detecting low molecular mass ions (m/z<1000). With proper mass calibration, the mass accuracy can reach low ppm values for small metabolites.

In a chemical isotope labeling LC-MS workflow, an external mass calibration with sodium formate solution was conducted frequently to maintain the mass accuracy of one QTOF instrument. An injection of sodium formate also was implemented at the beginning of each LC-MS run to obtain mass calibration spectra. Each LC-MS data file was calibrated after data acquisition using its calibration segment before exporting peaks to a mass list csv file. Table 6.3 shows a typical calibration status result using sodium formate in the positive mode. The mass error was calculated at each sodium formate standard. At the end, a standard deviation of the mass errors was calculated with a less than 1 ppm value to show a good calibration status.

The calibration status sheet provides only the mass error for calibration standards with a relatively high peak intensity. To monitor the mass fluctuations at a different peak intensity, we picked a background mass peak (average measured m/z at 335.144) and plotted the peak intensity and measured mass in each mass scan in Figure 6.3. From the extracted chromatogram, we can see that the mass intensity ranged from the low end of the detection limit to over 10^4 over the whole chromatogram. The mass error plot shows the mass error distribution at different intensities. At a relatively high peak intensity, the mass shows a better precision with a much narrower error distribution, and the random mass error increased for low intensity peaks; this indicated that the peak intensity can affect the mass precision.

Table 6.3 An example of a mass calibration status sheet for a LC-MS file.

Date:	11/17/2016 10:57		
Calibration spectrum:	+MS, 0.0-2.0min #1-119: Scan		
Reference mass list:	ESI: Na Formate (pos)		
Calibration mode:	HPC Calibration		
Reference m/z	Resulting m/z	Intensity	Error [ppm]
226.9515	226.9515	3357893	-0.067
294.9389	294.9390	402100	0.213
362.9263	362.9263	2798971	-0.172
430.9138	430.9137	3184750	-0.124
498.9012	498.9013	993770	0.229
566.8886	566.8886	997247	-0.037
634.876	634.8760	697186	-0.006
702.8635	702.8634	908381	-0.103
770.8509	770.8509	692145	0.045
838.8383	838.8384	660032	0.071
906.8257	906.8257	585708	-0.063
974.8132	974.8132	450964	0.015
Standard deviation: 0.189			

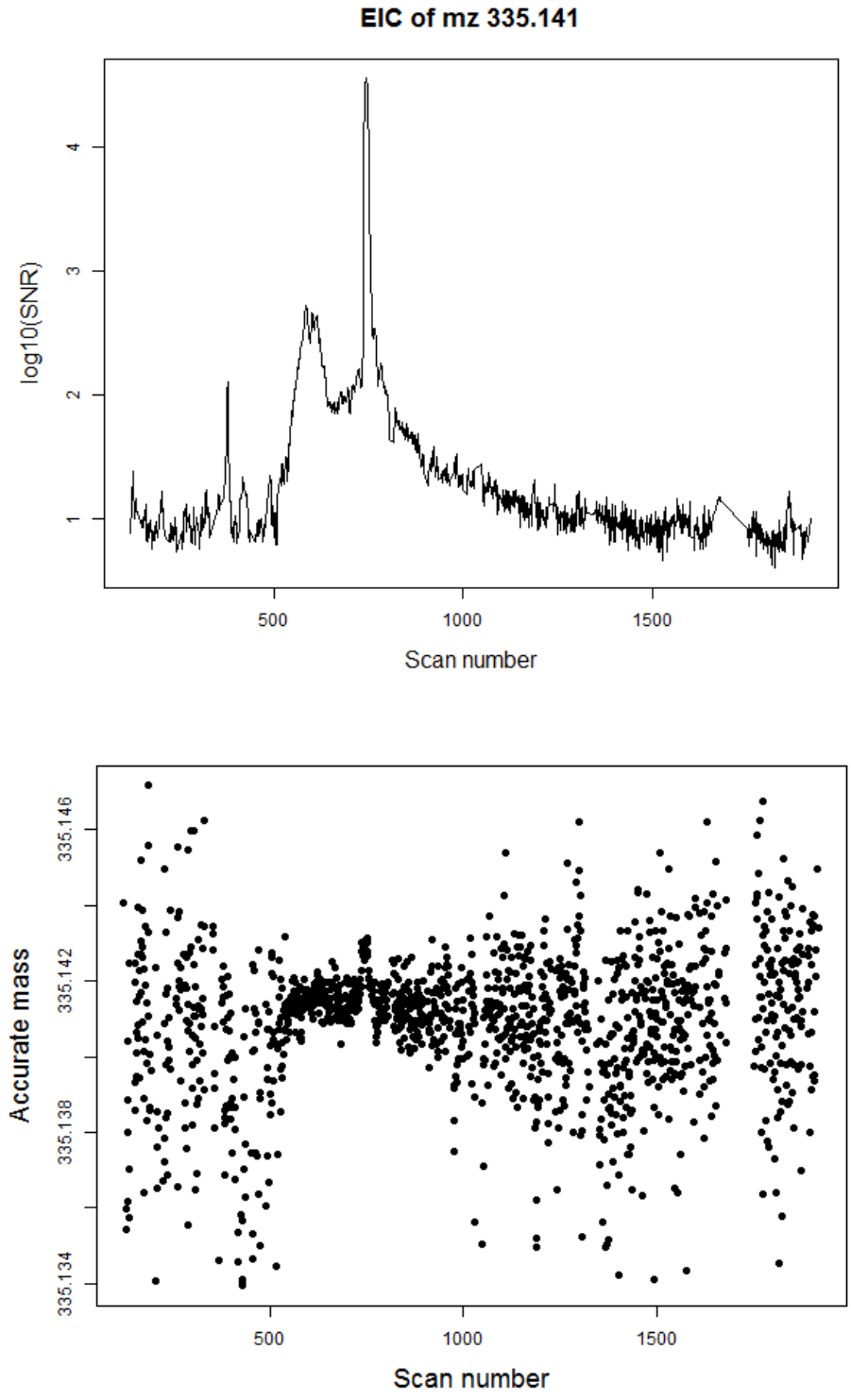


Figure 6.3 An extracted ion chromatogram of a background mass peak and its measured mass at each mass scan after mass calibration.

6.3.3 The influence of signal intensity on mass accuracy

To investigate the influence of peak intensity on mass error, different concentrations of sodium formate were prepared and analyzed using two models of QTOF mass spectrometers from Bruker. The intensity and accurate mass from sodium formate were extracted from each LC-MS run. A broad range of signal intensities was collected for each calibration standard. In total, 11,306 mass peaks were collected from impact QTOF, and 3,771 mass peaks from maXis II QTOF.

For each sodium formate adduct, the mass error was calculated as the measured mass minus the exact mass. The mass errors were plotted against the logarithm of the signal to noise ratio (SNR). Data from impact QTOF and Maxis II QTOF are shown in Figure 6.4 and Figure 6.5, respectively. In each Figure, 12 plots were created for each sodium formate adduct. Each data point represents a mass error from a sodium formate adduct peak. Fewer data points were observed for higher molecule weight sodium formate adducts due to relatively lower concentrations and the mass setting used in this experiment. From Figure 6.4 and Figure 6.5, we can observe an increasing trend of random mass error as the peak intensity decreases.

Figure 6.6 shows the mass error standard deviation for peaks in different intensities for data generated with impact QTOF. The standard deviation was calculated using mass peaks within each small intensity window. For different masses, we can observe a uniform trend of increasing mass error variation with decreasing peak intensity; this shows that the relation between mass error and peak intensity is relatively independent of the m/z value.

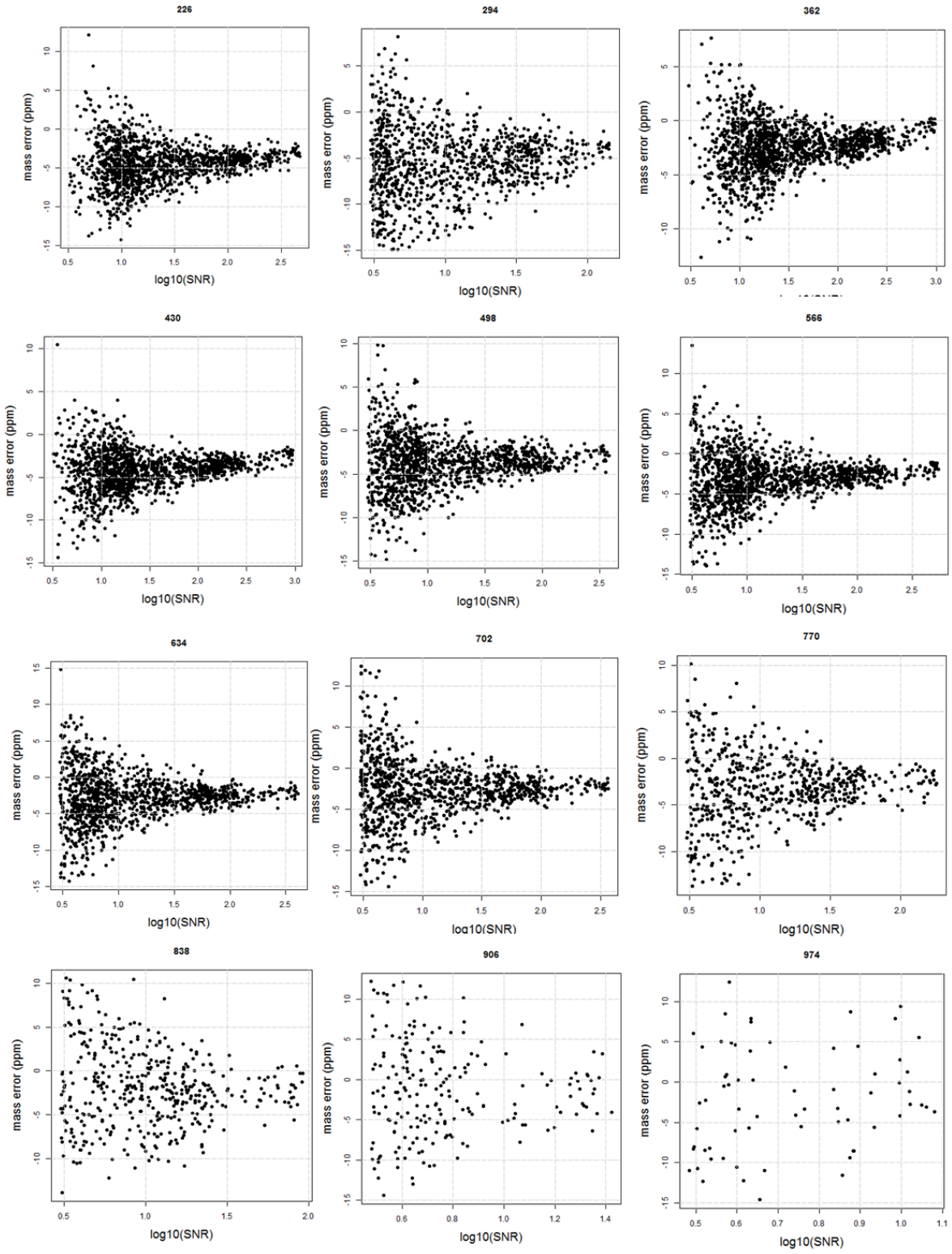


Figure 6.4 Mass error against peak signal to noise (SNR) for each sodium formate adduct peak in Bruker impact QTOF.

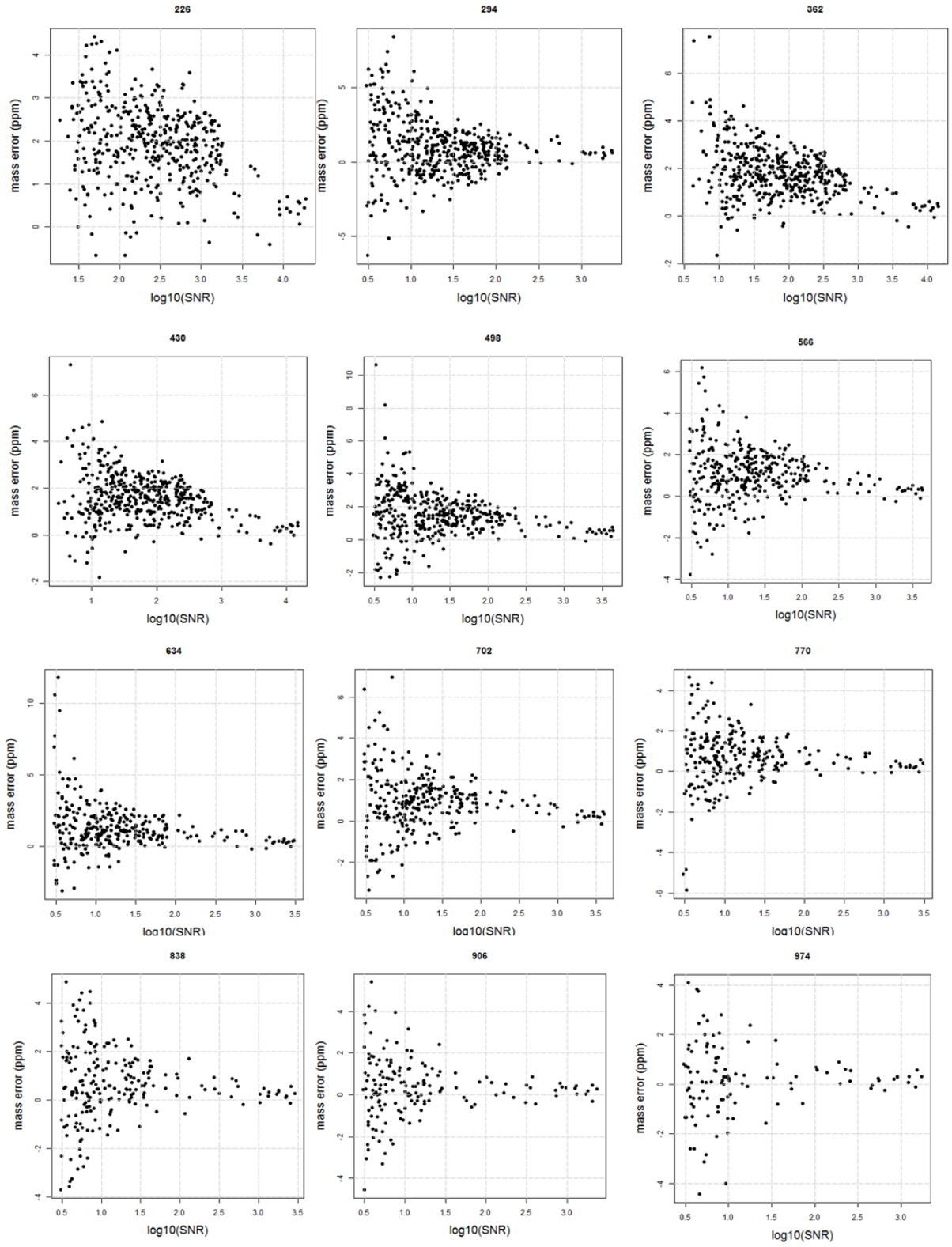


Figure 6.5 Mass error against peak signal to noise (SNR) for each sodium formate adduct peak in Bruker maXis II QTOF.

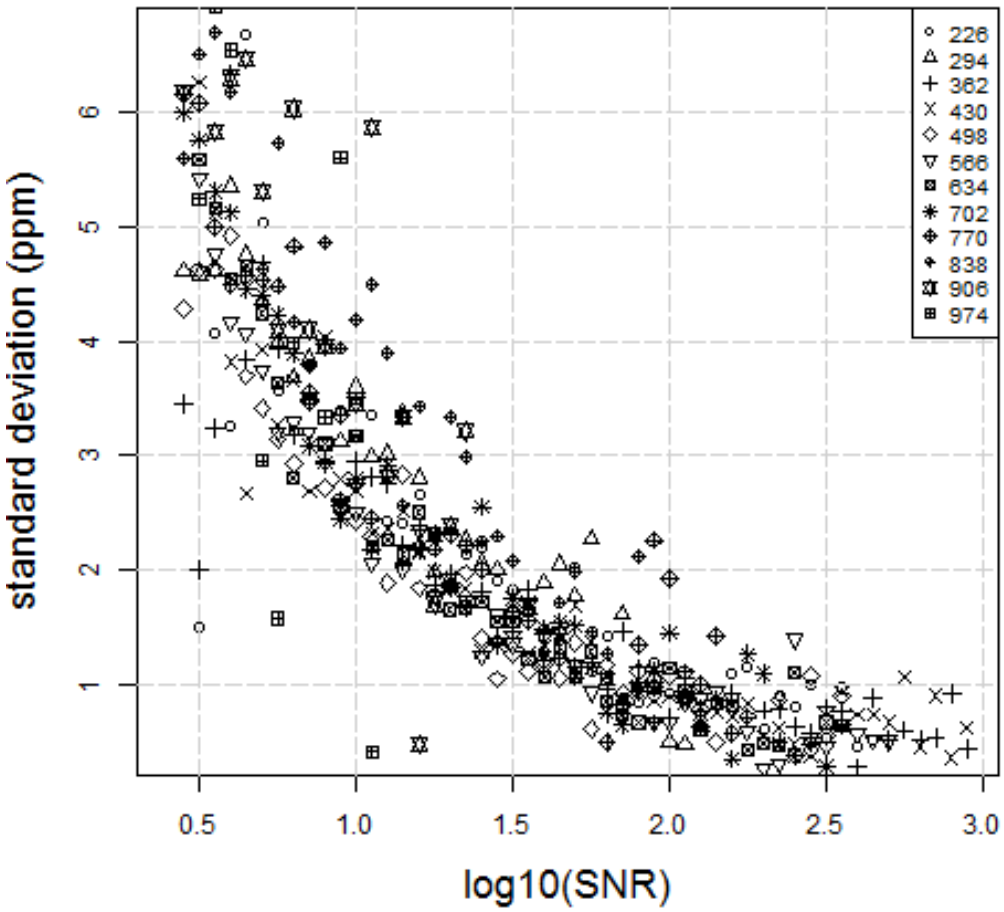


Figure 6.6 The mass error standard deviation at different peak intensities for all 12 sodium formate peaks. Each data point is the standard deviation of the mass error within a window of 0.05 length on the $\log_{10}$ (SNR) axis.

To better understand the error distribution and the increased random mass error at lower peak intensity, we checked some of the mass peak signals in the LC-MS data manually. We selected a compound, dansyl-Phe-Phe-Phe, that showed a large mass error in some of its spectra. Three of its mass peaks were picked in Figure 6.7. In the three continuous mass scans, the measured mass of dansyl-Phe-Phe-Phe gave mass errors of -0.28, 7.89, and -8.00 ppm, showing an abnormal mass shift in a small time window. By carefully inspecting the mass peaks, we can see that the distortion of the peak shape caused the error in the peak centroid calculation.

For a real sample analysis, the sample matrix can be a lot more complicated, and mass measurements of metabolites often are affected by coeluting interfering peaks. Thus, low concentration metabolites can be affected more easily by the noise signal. As a result, random mass error increased for those peaks with decreased peak intensity, as shown in Figure 6.4 and Figure 6.5; this indicated an intensity dependent mass error in a real dataset.

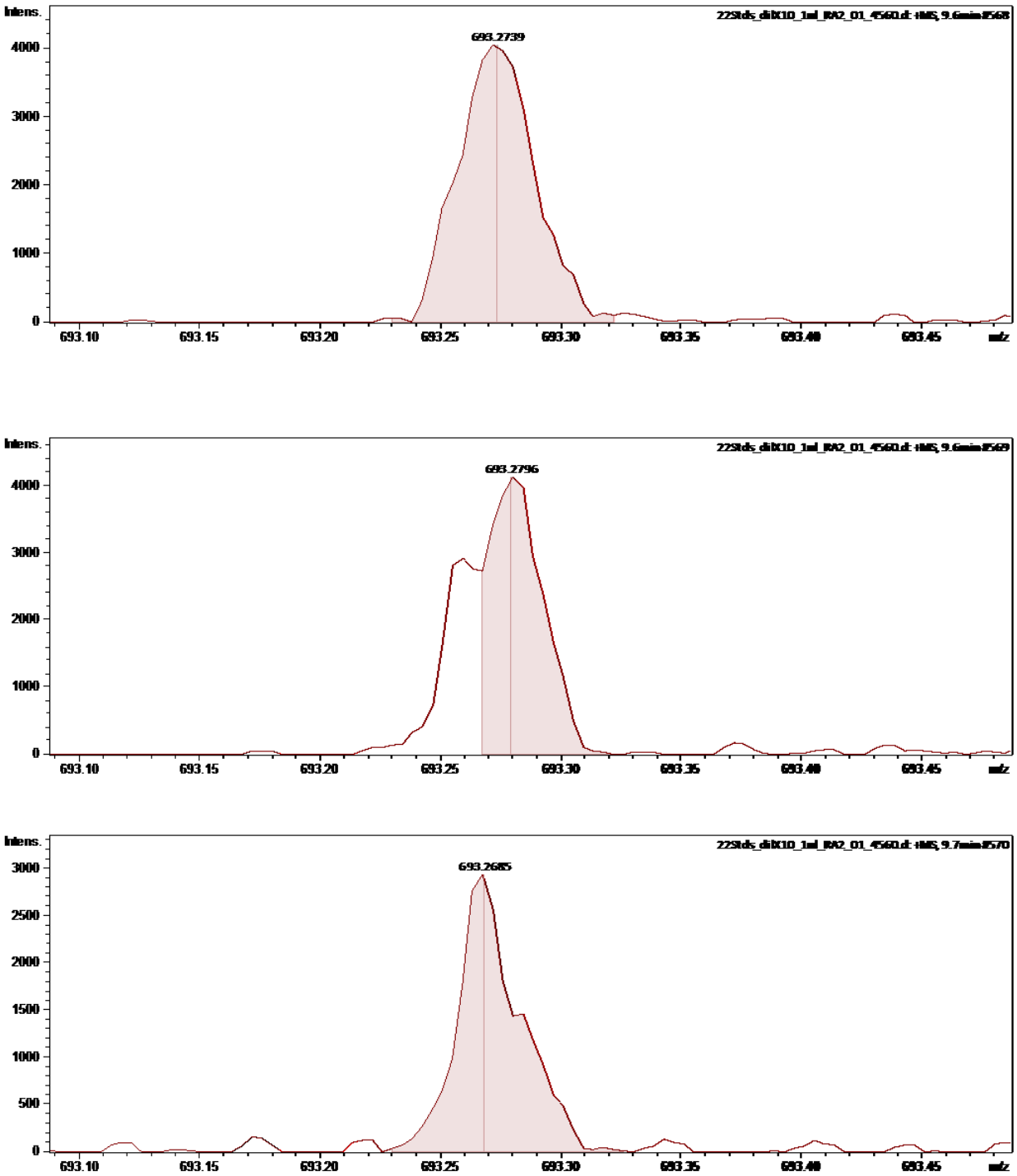


Figure 6.7 Mass peaks of dansyl-Phe-Phe-Phe in three consecutive scans. Mass errors are -0.28, 7.89, and -8.00 ppm for the three peaks.

6.3.4 Mass tolerance estimation for database search

Since mass error is related to peak intensity, the mass tolerance used in the mass database search should also be different for different masses. To be able to estimate the mass error at different peak intensities in the actual sample dataset, we extracted the data from the sodium formate calibration segment and used it as training data to establish a mathematical relationship between mass error and peak intensity.

In this experiment, different concentrations of the 22-standard mixture (see Table 6.2) solutions were prepared and analyzed using LC-QTOF-MS. In total, 94 sample data files were generated. Figure 6.8 (a) shows the sodium formate data by combining the sodium formate peaks in all 94 runs. The mass errors were plotted against $\log_{10}$(SNR). To determine the mass tolerance at different intensity levels, we divided the dataset by the peak intensity using a window size of 0.1 in the $\log_{10}$(SNR) axis. For data in each intensity range, we calculated the 95^{th} percentile mass error point in Figure 6.8 (b). We then applied a quadratic equation to construct the tolerance curve using these 95^{th} percentile points in each peak intensity range. The results are shown as the curve in Figure 6.8 (b). From the tolerance curve fitting function, the mass tolerance can be estimated for any peak intensity. A minimum of 2 ppm is set for the tolerance upper limit.

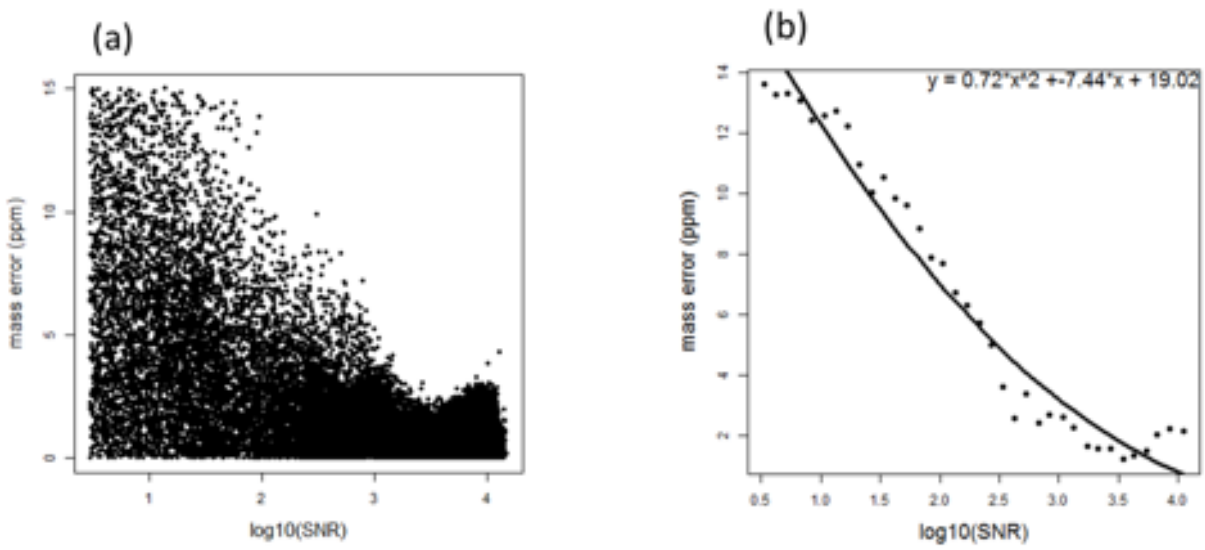


Figure 6.8 (a) The mass error of sodium formate at different peak intensities, and (b) the 95th percentile point of mass error in each 0.1 length window in the $\log_{10}$(SNR) axis.

To validate the tolerance curve derived from the training dataset, we examined the mass errors from the 22 standards and compared them with the estimated mass tolerance. From the 94 LC-MS runs, we extracted peak pairs from all standards. The test data from 22 standards were divided into three groups according to their dilution factors so that we could have data with three intensity levels. In each group, the data were aligned into one metabolite intensity table. The mass tolerance was calculated based on the tolerance curve generated by the training dataset. The actual mass error for each standard also was calculated. The result is plotted in Figure 6.9; red data points are the actual mass error, and blue data points are the calculated mass tolerance. The result showed an accurate tolerance estimation as most of the actual errors are below the corresponding mass tolerance for three data tables. Using the intensity-dependent tolerance, we could achieve a 100%, 95.2%, and 95.2% correct matching rate in the three groups of data in searching the 22 standards in the library.

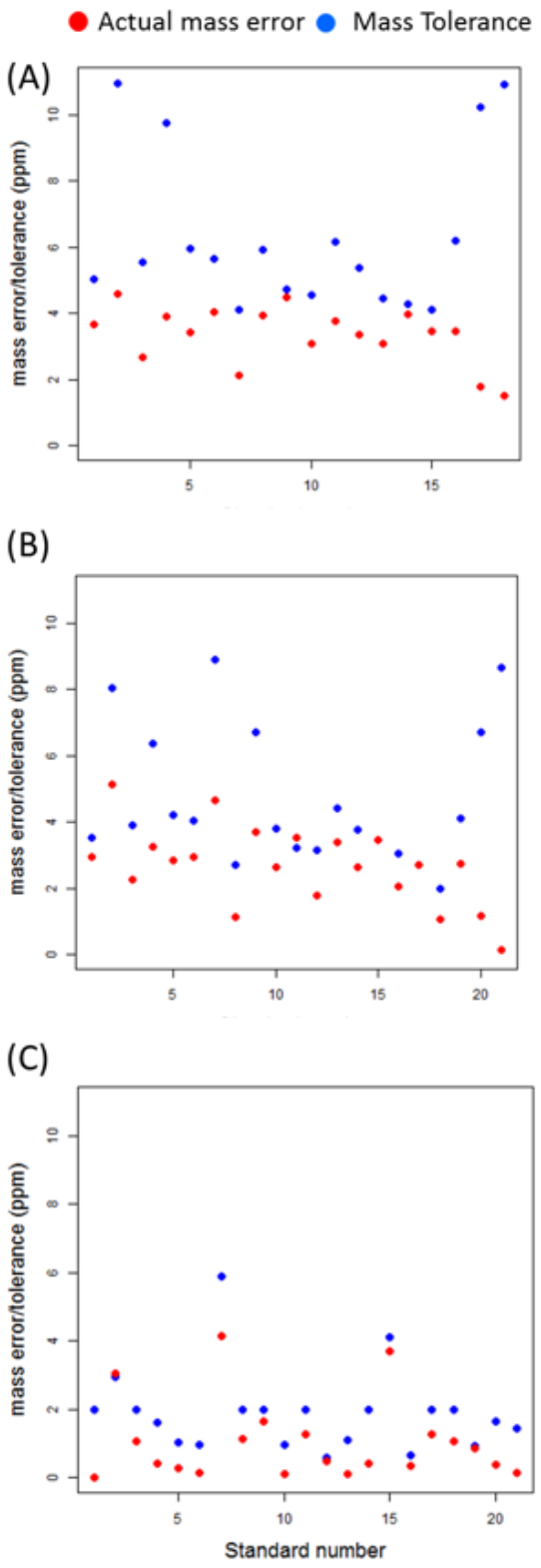


Figure 6.9 Mass tolerance and actual mass error calculated for 22 standards. (A), (B) and (C) are data generated from different concentrations of the standard mixtures.

We applied this algorithm to a human urine dataset. We aligned 9 QC data files to generate an alignment data table to be used in a library search. In this dataset, the signal to noise ratio of the detected signal ranged from 10 to 10^4. A mass tolerance curve was derived first from the sodium formate data from the same batch of sample data. Based on the peak intensity of the heavy peak in each peak pair, mass tolerances were calculated for each mass.

Next, a database search was performed against the MyCompoundID[70] database using the intensity-dependent mass tolerance along with other fixed tolerance searches. Table 6.4 summarizes the number of matches using different parameters. "Number of query mass with match from library" shows the number of query mass with at least one matched compound in the library, and "total number of matched compounds from library" is the sum of all the matches from each query mass and includes the multiple match cases. Compared to the fixed tolerance, the intensity-dependent tolerance search gave a larger number of matched query masses and a relative small number of total matched compounds, showing an improvement over other arbitrary fixed tolerance searches.

From the matching results, we took 133 compounds that were identified definitely using the dansyl library[86] and evaluated the accuracy of the search further. Table 6.5 shows the number of correct matches out of these 133 compounds using different search tolerances in MyCompoundID database. The intensity-dependent method gave a 94.7% (126 out of 133) correct matching rate.

Table 6.6 shows the false positive and false negative rate of the search result. A false positive case is defined as a wrong match resulting from a query mass, and a false negative case is defined as a missed match from a query mass. The intensity-dependent search gave a zero false positive result and showed a relative low false negative rate.

We tested the search method in MyCompoundID website (www.mycompoundid.org). Figure 6.10 shows the user interface on the webpage. R programs will be posted on the website for free downloading. Three different searching modes are provided on this page, including fixed tolerance search, user-defined intensity-dependent tolerance search, and data derived intensity-dependent tolerance search. In this way, the user can input one's own intensity-dependent search tolerance if sodium formate data was not available.

Table 6.4 MyCompoundID Library search results using different search tolerances.

Search tolerance (ppm)	2	5	10	15	30	Intensity-dependent
Number of query mass with match from library	202	476	711	855	1072	708
Total number of matched compounds from library	537	1199	1716	2030	2480	1608

Table 6.5 Number of correct matches out of 133 identified compounds.

Search tolerance (ppm)	5	10	15	20	30	Intensity-dependent
Number of correct match	91	116	121	124	129	126

Table 6.6 False negative/positive rate of the intensity-dependent search and other fixed tolerance searches.

	Intensity-based	5 ppm	10 ppm	15 ppm	20 ppm	30 ppm
False negative rate	5.22%	31.34%	12.69%	8.96%	6.72%	2.99%
False positive rate	0.00%	0.00%	0.00%	2.24%	4.48%	6.72%

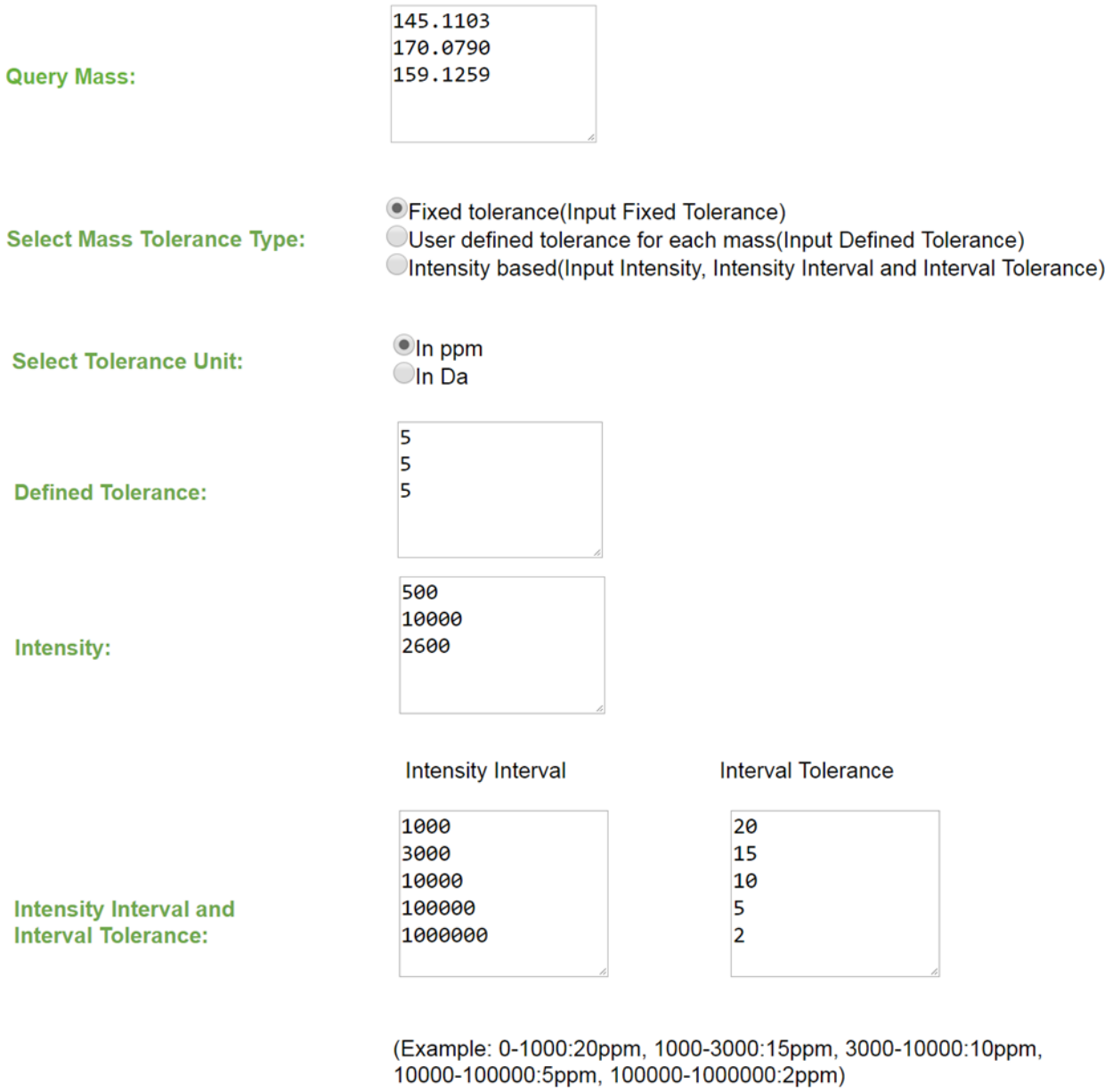


Figure 6.10 MyCompoundID webpage for an intensity-dependent mass-based library search. The database provides three searching options: 1) fixed tolerance, 2) user-defined tolerance by intensity intervals, and 3) pre-defined tolerance for each query mass.

6.4 Conclusions

In this study, we investigated the influence of peak intensity on mass accuracy. The random mass measurement error was observed to increase with the decrease of peak intensity. By investigating the relationship between peak intensity and mass error, we designed an algorithm to predict a mass tolerance for each measured mass accurately, and use them in the database searching. The method does not require additional experimental steps and used the sodium formate calibration data from the sample data to determine the mass tolerances.

Compared to an arbitrary tolerance, the new method provided an accurate estimate of the mass search tolerance and improved the metabolite identification efficiency during the database search. The searching function will be added to the MyCompoundID database search webpage in future work.

Chapter 7 Development of Chemical Isotope Labeling LC-MS for Wine Metabolomics

7.1 Introduction

In recent years, metabolomics has been used increasingly for nutritional studies in the production of nutritional products and the relation between diet and health.[124-126] Nutritional metabolomics involves the characterization of the food metabolomes and the investigation of their effects on the metabolic profile of a specific organism.[127] Wine is a widely consumed alcoholic beverage throughout the world and represents an important food commodity of a relatively high commercial value. Therefore, it requires a fast and sensitive analytical method for the detection of a wide range of molecules in a wine sample for the wine quality and authenticity control.

Wine is a complex matrix of its major components, water, alcohol, and abundant organic and inorganic contents. The wine making process involves the harvest of grapes from different grape varieties and the fermentation process that transforms grape juice into an alcoholic beverage.[128] During the fermentation, sugars in the grapes juice are turned into ethanol, and a number of compounds are produced through the metabolism of the yeast. The product after the metabolic fermentation will go through aging to allow further chemical reactions within the wine solution, giving the wine a more complex flavor. The wine metabolites are the combination of grape metabolites, yeast metabolites, and the interaction of both through a series of steps in the wine making process. The concentration levels of the compounds in wine are influenced significantly by many factors, including grape variety, climate, grape-growing area, and the winemaking process. Numerous studies have been published on the wine topic describing the

applications of the instrumental analysis for detecting various groups of compounds to enable the wine classification and quality control.[129-134]

The chemical isotope labeling (CIL) LC-MS method has been used widely in metabolic profiling of different biological samples. CIL LC-MS is a general strategy of using chemical labeling to improve separation, detection, and quantification of metabolites.[30] CIL targets a particular sub-metabolome based on a shared chemical group; for example, dansylation labeling has been shown to be effective in analyzing the amine/phenol sub-metabolome. In this Chapter, we develop a workflow using the CIL LC-MS method for the profiling of wine samples and demonstrate the overall analytical performance of the method in differentiating wines of different brands.

7.2 Materials and Methods

7.2.1 Chemicals and reagents

All the chemicals and reagents, unless otherwise stated, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). In the dansylation labeling reaction, the ^{12}C-labeling reagent (dansyl chloride) was from Sigma-Aldrich, and the ^{13}C-labeling reagent was synthesized and purified in our lab using the procedure published previously.[30] LC-MS grade water, methanol, and acetonitrile (ACN) were purchased from ThermoFisher Scientific.

7.2.2 Red wine sample collection and preparation

All wines were purchased from a local certified liquor store. Each wine was aliquoted in a 1.5 mL vial and stored at -80 °C until use. Different brands of wine were selected from various grape varieties and geographical origins (see Table 7.1); all are red wines, except the Pinot Grigio, which is a white wine. A 1 mL volume of wine sample was centrifuged at 14,000 rpm for 15 min, and a 50 μL of supernatant was taken into a 1.5 mL vial for dansylation labeling. A pooled wine sample was prepared by mixing all the individual samples in equal volumes.

A dealcoholized red wine sample was prepared using a sample from BBR red wine to test the interference of alcohol contents on the sample analysis. After centrifugation, a 50 μL of supernatant solution was transferred to a 1.5 mL vial in the SpeedVac for drying. The drying process will remove water and other volatile alcohol components. The dried sample was reconstituted with 50 μL of water as the dealcoholized sample.

Table 7.1 Wine sample list.

Grape variety	Sample ID	Vintage	Region
Cabernet Sauvignon	CS1	2012	South Australia
	CS2	2014	ON, Canada
	CS3	2015	ON, Canada
Cabernet Merlot	CM	2012	BC, Canada
Shiraz	SZ	2014	BC, Canada
Big Bold Red	BBR	2013	CA, USA
Pinot Grigio	PG	2013	South Australia

7.2.3 Dansylation labeling

Dansyl chloride (DnsCl) was used as the labeling reagent to react mainly with amine- and phenol-containing metabolites to form dansyl-amine or dansyl-phenol derivatives. For a labeling reaction, 50 μL of sample were mixed with 25 μL of ACN and 25 μL of 250 mM sodium carbonate/sodium bicarbonate buffer, which introduced a basic environment for the labeling reaction. The solution was vortexed, spun down, and mixed with 50 μL of freshly prepared ^{12}C-DnsCl solution (18 mg/mL, for light labeling) or ^{13}C-DnsCl solution (18 mg/mL, for heavy labeling). After the sample was incubated at 40 °C for 45 min, 10 μL of 250 mM NaOH were added to quench the excess dansyl chloride. The solution was incubated further at 40 °C for another 10 min to allow the unreacted dansyl chloride to be hydrolyzed fully. Finally, 50 μL of formic acid (425 mM) in 1:1 ACN/H_2O were used to acidify the solution.

7.2.4 Sample normalization by LC-UV

Sample normalization is a necessary step to minimize the inter-sample variations. An LC-UV based method was applied to determine the total concentration of dansylated amine/phenol-containing metabolites based on the UV absorption of the dansyl group.[91] The experiment was performed with a Waters ACQUITY UPLC system UPLC (Waters, Milford, MA, USA) and a Phenomenex Kinetex C18 column (2.1 mm × 5 cm, 1.7 μm particle size, Phenomenex, Torrance, CA, USA). Two microliters of each dansyl-labeled individual or pooled sample were injected for a fast step-gradient run. Solvent A was 0.1% (v/v) formic acid in 5% (v/v) ACN/H_2O, and solvent B was 0.1% (v/v) formic acid in ACN. Starting at 0% B for 1 min, the gradient was then increased to 95% B within 0.01 min, and held at 95% B for one min to ensure complete elution

of all labeled metabolites. The flow rate was 0.45 mL/min, and the total UV absorption of dansyl-labeled metabolites in the sample was measured at 338 nm. The peak area, which can represent the total metabolite concentration in the sample, was integrated by the Empower software. According to the quantification results, the ^{12}C- and ^{13}C-labeled samples were mixed in equal amounts for the following LC-MS analysis.

7.2.5 LC-MS analysis

Figure 7.1 shows the workflow of dansylation isotope labeling LC-MS analysis for wine samples. Each individual sample was labeled by the ^{12}C-DnsCl, and the pooled sample was labeled by the ^{13}C-DnsCl. The mixture of the two labeling solutions was analyzed by LC-MS with an Agilent 1100 HPLC system (Palo Alto, CA) connected to Bruker Impact HD quadrupole time-of-flight (QTOF) mass spectrometer (Billercia, MA) with an ESI source. The MS settings were: end plate: 500 V, capillary voltage 4500 V, nebulizer: 1.0 bar, dry gas: 8.0 L/min, dry temperature: 230 °C, scan range: 220-1000. MS spectra were acquired in the positive ion mode. Chromatographic separations were performed on an Agilent C18 column (2.1 mm × 100 mm, 1.7 μm). Mobile phase A consisted of 5% (v/v) acetonitrile and 0.1% (v/v) formic acid in water. Mobile phase B was 0.1% (v/v) formic acid in acetonitrile. The 32-min gradient conditions were: 0 min (20% B), 0-3.5 min (20-35% B), 3.5-18 min (35-65% B), 18-24 min (65-99% B), and 24-32 min (99% B). The column was re-equilibrated at 20% B for 10 min, and the flow rate was 180 μL/min.

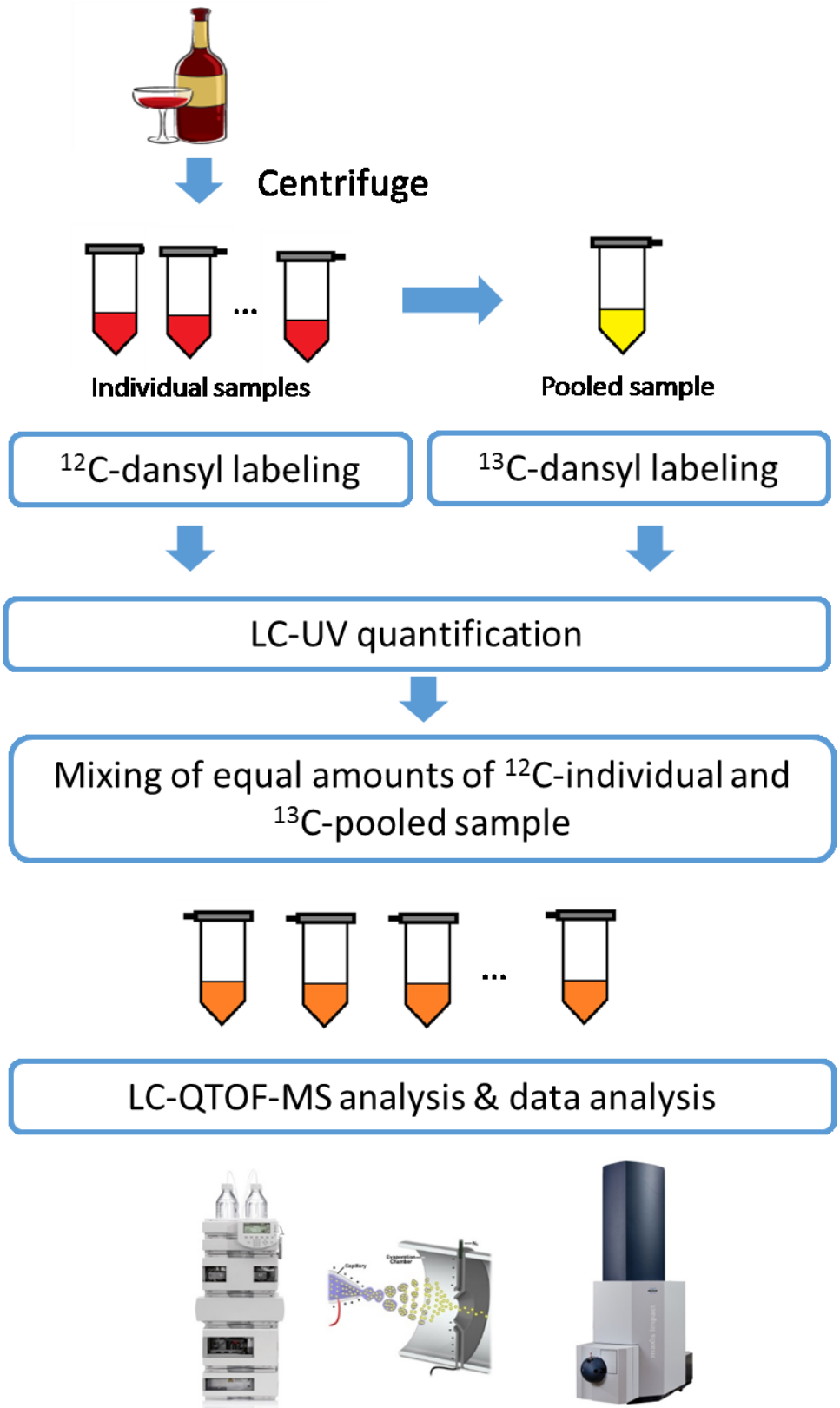


Figure 7.1 Workflow of dansylation isotope labeling LC-MS for a wine sample.

7.2.6 Data analysis

Bruker DataAnalysis software 4.3 was used to extract MS spectral peaks. We ran a raw data quality check on the raw mass list for the mass accuracy and retention time shift. An in-house software IsoMS was used to extract unique peak pairs from each individual sample file. The peak pair lists were aligned and zero-filled using the processing methods discussed in Chapter 4. The same peak pairs detected from multiple samples were aligned to produce a CSV file containing the metabolites information and peak ratios relative to a control (i.e., a pooled sample). False peak pairs and redundant peak pairs were evaluated and excluded from the alignment data table. The missing ratio data was imputed using the ratio imputation method discussed in Chapter 4. The final complete metabolite-intensity data table was uploaded to MetaboAnalyst[135] for multivariate analyses.

Positive metabolite identification was performed based on mass and retention time matching to the dansyl standards library consisting of 665 entries.[86] The database includes the dansyl library in MyCompoundID and the data of 400 dansyl peptides from all combinations of 20 common amino acids. Putative identification was done based on the accurate mass match to the metabolites in the human metabolome database (HMDB) (8,021 known human endogenous metabolites) and the Evidence-based Metabolome Library (EML) (375,809 predicted human metabolites with one reaction) using MyCompoundID.[70] The mass accuracy tolerance window was set at 10 ppm, and the retention time tolerance window was set to 20 sec for the data alignment and library searches. For multifunction compounds (e.g., containing two amines), labeling by one or more reagent molecules to generate multiple products may happen to some metabolites, although in most cases the complete labeling was found. If multiple products were

found from one metabolite, they could be spotted readily in the final list of significant metabolites for differentiating different groups as they would be matched to the same metabolite.

7.3 Results and Discussion

7.3.1 Sample normalization for a red wine sample

Different varieties of red wine can have different concentrations of metabolites due to the different grape varieties and manufacturing processes. For a labeling reaction with a red wine sample, we had to determine the metabolite concentration first in order to choose a proper amount of starting material. Also, for a comparative analysis, it is necessary to normalize the total amount of metabolites in order to minimize the variation caused by the difference in the total concentration of metabolites. Our group previously developed a LC-UV based sample quantification method, which measures the UV peak area of dansyl labeled metabolites for the quantification of total amount of analytes.[91] It assumed that the dansyl group contributed most to the UV absorption and that the native chemical structures would not affect the accuracy of the UV quantification. The red wine, however, contains many polyphenols that can have UV absorption over a broad range of wavelengths; this may affect the quantification result of the LC-UV experiment.

To study the UV absorption of native red wine metabolites, we analyzed the unlabeled and labeled red wine sample using the established LC-UV method. Figure 7.2 (A) shows the integrated UV peak areas of the unlabeled and labeled red wine samples from the BBR red wine. Each data point is an average of peak areas from duplicate injections. Samples of 5 µL to 50 µL of red wine were used in this experiment. Each sample was diluted to 50 µL, followed by the

dansyl labeling protocol. Table 7.2 summarizes the UV peak areas from different sample volumes and the area ratios of the unlabeled red wine to the labeled one. The labeled red wine sample has a much higher UV absorption since the wavelength was optimized for the dansyl group. However, the UV absorption of native red wine contributed also to the total UV area in each sample volume. We subtracted the UV area of unlabeled red wine from labeled red wine in each sample volume and plotted it in Figure 7.2 (B).

From Table 7.2, we can see that the UV peak area ratio of the unlabeled red wine increased with an increase in sample volume. This indicates that the labeling efficiency decreased with higher sample amounts. As a result, 25 μL was selected as the volume for the labeling reaction. Figure 7.3 shows the total concentration for different types of wine calculated from the amino acid calibration curve.[91] Red wines from different grape varieties and brands show a similar total concentration for the labeled metabolites, and less amine and phenol containing metabolites were found in the white wine sample PG.

Table 7.2 The UV integrated area for the dansyl labeled red wine sample at different sample volumes and the ratios of the unlabeled red wine peak area.

Sample volume (μL)	UV integrated area (labeled red wine)	Ratio of UV area (unlabeled/labeled)
5	1,146,204	0.07102
10	2,154,313	0.07788
20	3,977,676	0.08645
30	5,281,741	0.1027
40	6,491,637	0.1149
50	7,598,724	0.1209

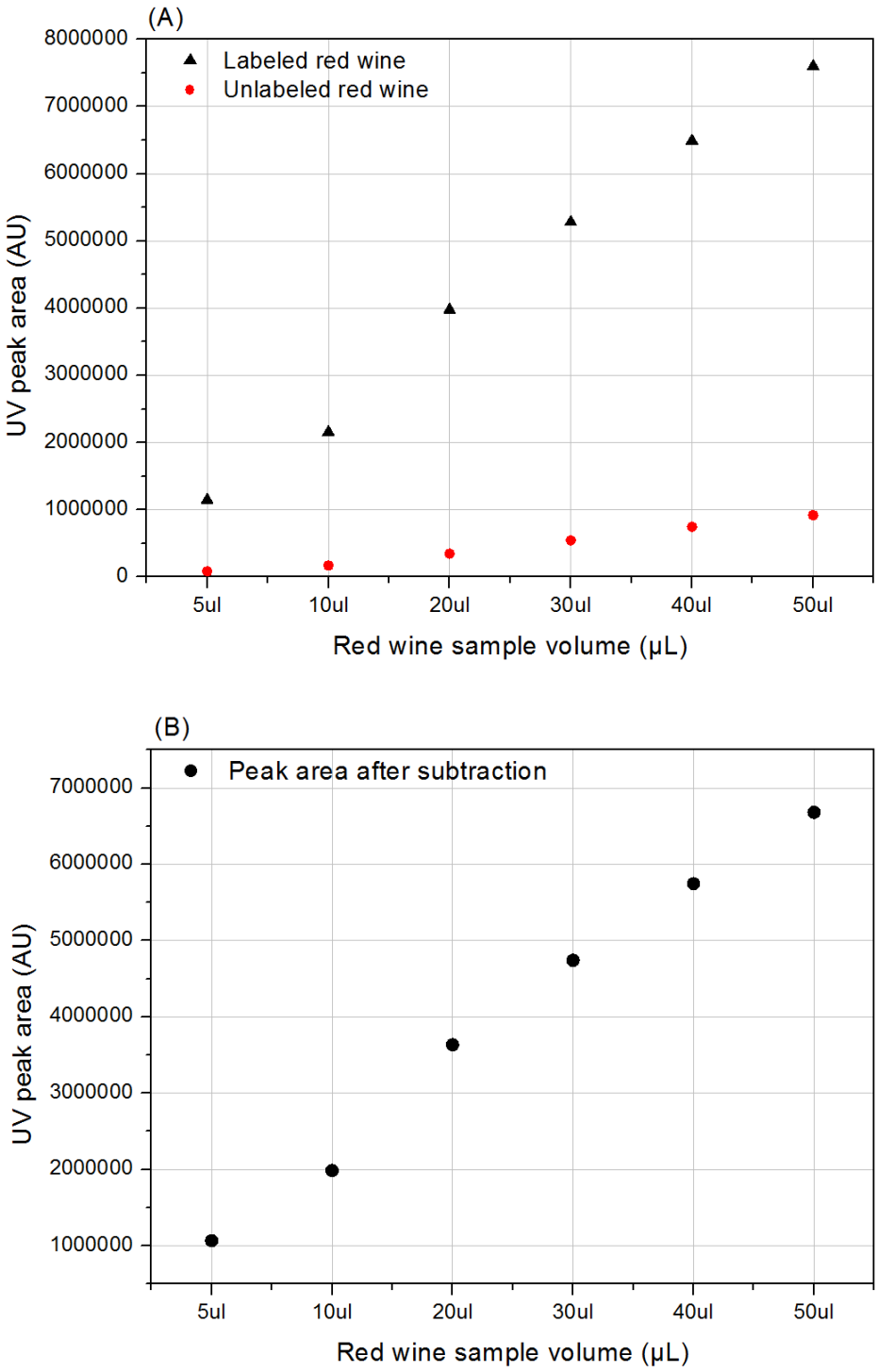


Figure 7.2 (A) The UV integrated peak area of the labeled and the unlabeled red wine, and (B) The UV integrated peak area of the labeled red wine after subtraction of the peak area from the unlabeled red wine. Different sample volumes were used in each injection. Each data point is an average of duplicate injections for one sample volume.

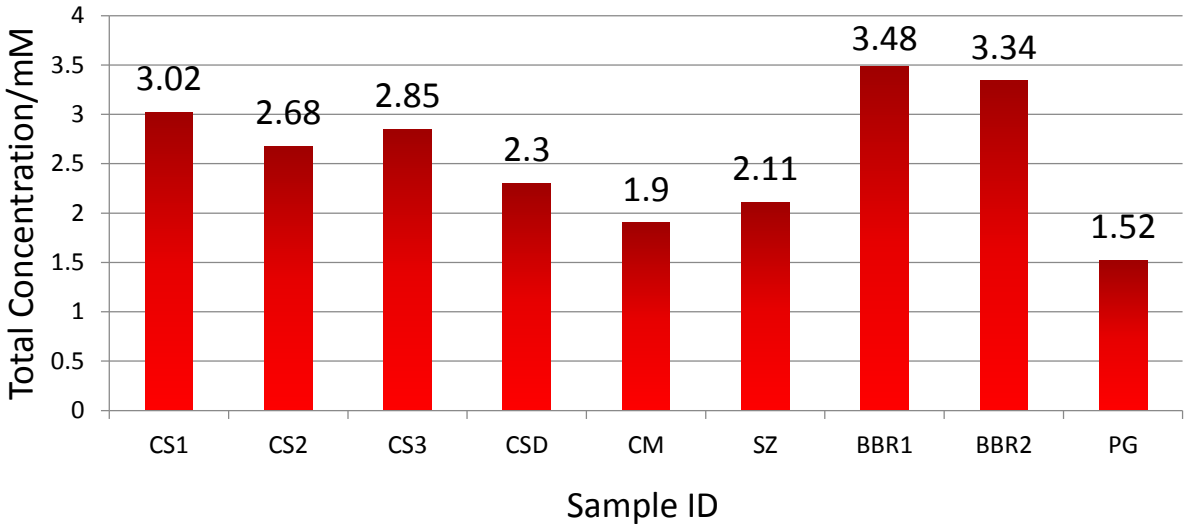


Figure 7.3 The total concentration of labeled metabolites in different types of wine. See Table 7.1 for sample ID information. CSD is the dealcoholized sample prepared from CS1. BBR1 and BBR2 are two different batches of the BBR red wine.

7.3.2 Alcohol interference

Alcohol makes up 12.5% to 14.5 % (by volume) of red wine content. Although the dansyl labeling reaction is not targeted at alcohol compounds, at a relatively high concentration, ethanol still can be labeled. Figure 7.4 shows a LC chromatogram of a labeled red wine sample. The red colored extracted ion chromatogram shows the signal from the dansyl ethanol. The presence of ethanol and other alcohol compounds potentially can suppress the signals of other labeled metabolites.

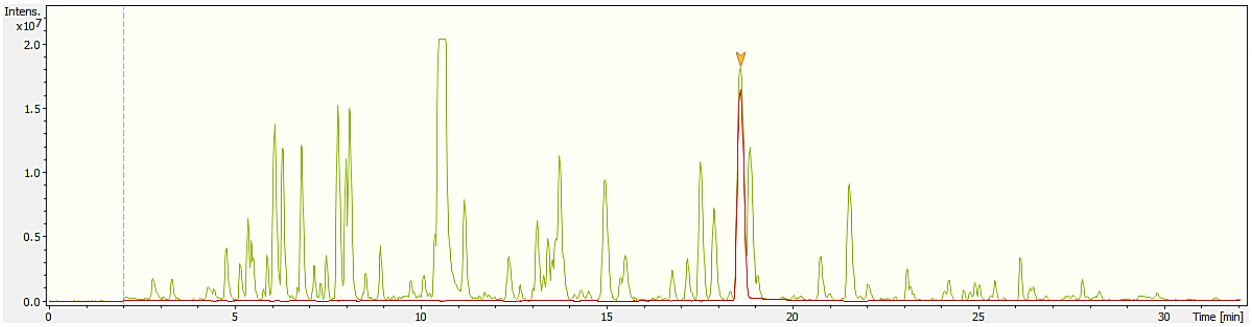


Figure 7.4 A LC chromatogram of a dansyl labeled red wine sample. A dansyl ethanol peak (in red color) shows up at 18.65 min.

To test the interference of ethanol in the data collection, we designed a study to compare the red wine sample with and without the alcohol components. Since alcohol compounds are not the targets of the dansylation reaction, we can remove them at the beginning of the sample preparation. After a complete drying down of the sample in the SpeedVac, ethanol, along with other volatile alcohols, were removed from the sample. Then, water was added to the sample vial for a reconstitution of the sample solution. The dealcoholized red wine sample was labeled with ^{13}C-dansyl chloride, and the raw red wine sample was labeled with ^{12}C-dansyl chloride. We mixed the two labeling solutions in equal volumes; therefore, the peak pair ratio of each metabolite can reflect the concentration changes in the drying process. The raw red wine sample was labeled also with light and heavy dansyl chloride and mixed in equal volumes as a control sample. The mixtures were analyzed by the LC-MS in triplicate injections.

After data collection, we extracted all the peak pairs detected in each injection and calculated the peak pair ratio (see Table 7.3). Sample 0 shows the data from the control experiment with the raw red wine. The peak pair ratio of each metabolite in the control sample was expected to be close to 1. Samples 1–3 are the results from the dried red wine sample in experimental triplicates. We can observe a lower total peak pair number and a higher peak pair

ratio average in the dried sample compared to the raw sample. The increased peak pair ratio average indicates an overall smaller relative peak intensity of the heavy peak. Since each of the light peaks are from the raw sample and the heavy peaks are from the dried sample, the ratio increase with respect to 1 indicates the sample loss during the drying step. Also, the Venn diagram in Figure 7.5 shows a larger number of metabolites in the raw red wine that were not detected in the dried wine sample.

In conclusion, the sample drying process can eliminate the alcohols in the wine sample. However, the drying step can cause the loss of certain metabolites and increase the error of the quantification results. In considering the speed and convenience in sample handling, we decided to use the raw wine sample for the labeling reaction.

Table 7.3 Peak pair number and average peak pair ratio in data collected from red wine and dried red wine sample.

Sample number	# peak pair	Average ratio
0	1649±58 (n=3)	1.031
1	1463	1.478
2	1516	1.168
3	1523	1.211

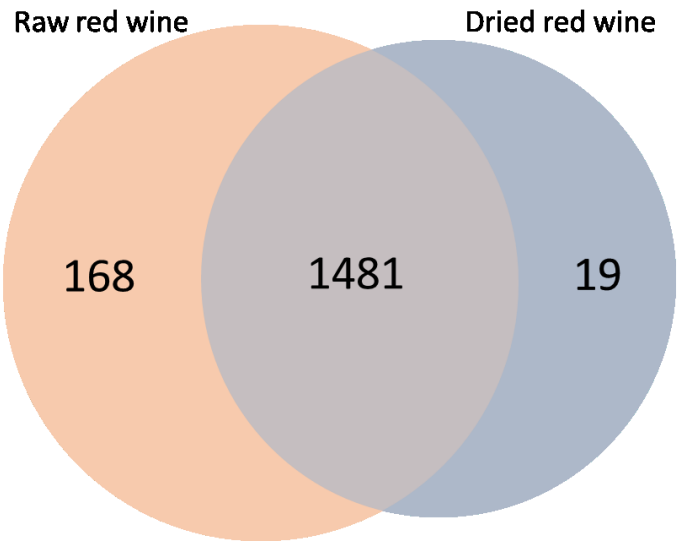


Figure 7.5 Venn diagram of peak pair distribution in raw red wine and dried red wine.

7.3.3 Injection optimization

In LC-MS operation, using an optimal sample injection amount is critical in order to detect the maximum number of labeled metabolites. With dansyl labeling, the total concentration of labeled metabolites in each sample can be measured by LC-UV, as described in the experimental section. One benefit of knowing the total concentration of labeled metabolites is that the exact amount of sample injected into LC-MS can be controlled well. To determine the optimal injection amount, a 1:1 ^{12}C-/^{13}C-labeled red wine sample with a known concentration measured by LC-UV was injected from 1 to 30 µL (1.032 to 30.96 nmol) to the LC-MS.

Experimental triplicate runs were performed for gauging the technical reproducibility. Figure 7.6 shows the plot of the peak pair number detected as a function of sample amount injected. Peak pair number saturation occurred when 5.16 nmol of sample in 5 µL was injected. Thus, in subsequent experiments, we injected 5 nmol of labeled sample for each of the LC-MS for analyses.

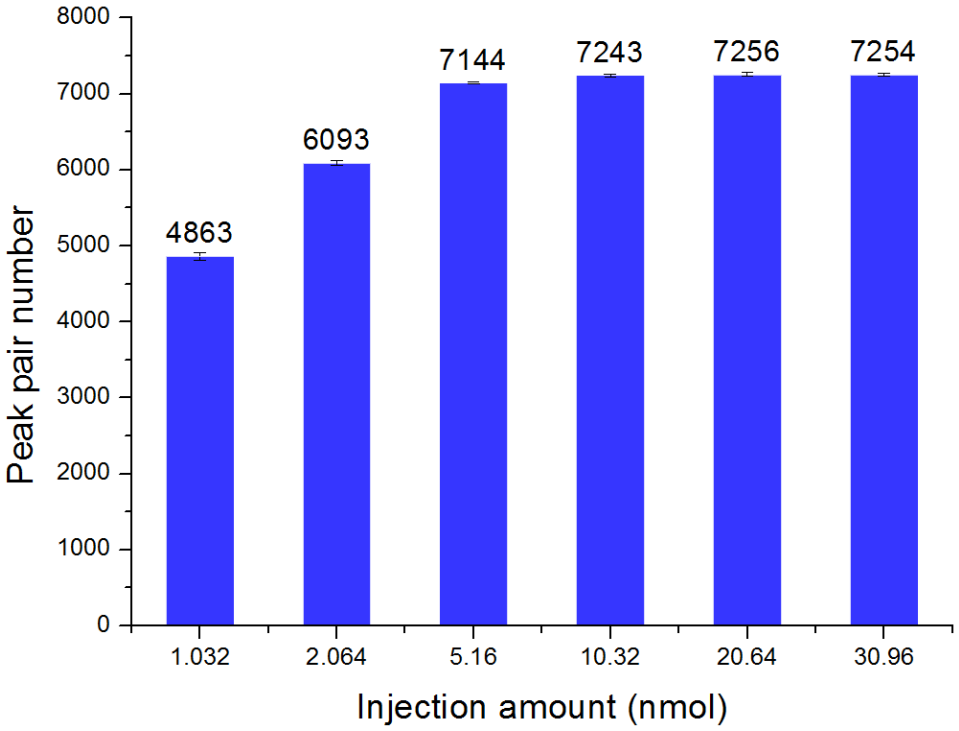


Figure 7.6 Number of peak pair commonly found in triplicate injections at different injection amounts.

7.3.4 Metabolic profiling of different red wines

To demonstrate the applicability of the workflow for red wine metabolomics, we examined the metabolome differences among different groups of red wine samples. The profiling work involved the analysis of seven wine samples collected from six red wine brands and a white wine brand (see Table 7.1), among which, CS2 and CS3 are the products from the same brand but of a different production year. An unknown sample was selected randomly from the seven wine samples to test the classification results. Experimental triplicates were performed for each sample, except for the BBR red wine, in which four labeled samples were prepared. In total, 25 samples were prepared and analyzed by LC-MS.

Figure 7.7 shows a base peak chromatogram of a wine sample. After exporting the instrument data to a CSV file, we checked the mass accuracy and retention time shift using the internal compounds from the samples. Two internal compounds were used for the mass accuracy check: dansyl ammonia and dansyl proline; Figure 7.8 shows the mass accuracy results. Each data point is the average of the measured mass in a sample file. The exact mass of dansyl ammonia and dansyl proline are 251.0849 and 349.1216, respectively. We can see that most of the mass average values are within the 5 ppm error window compared to the theoretical value, except for the data in File 1, which shows a larger mass error in both compounds. Figure 7.9 (A) and (B) show the measured masses of dansyl ammonia and dansyl proline in each mass spectrum in File 1. We can observe a mass shift towards a lower value over the course of the analysis, which caused the large mass error in Figure 7.8 (A). Fortunately, there was enough sample volume for a re-analysis, and the problematic file was replaced by the newly collected data. Figure 7.8 (C) shows the results after the re-analysis of the sample. The new file shows a similar mass error as the rest of the files.

In retention time analysis, common amino acids were selected to evaluate the retention shift in all the sample data files. Figure 7.10 shows the retention times of three standards: threonine, proline, and tyrosine. The range of the retention time for each standard is well within a 20-sec window, which is expected from the LC system. From the results of the mass accuracy check and retention time shift analyses, we determine the mass tolerance and retention time tolerance to be 10 ppm and 20 sec, respectively, for data alignment and library searches.

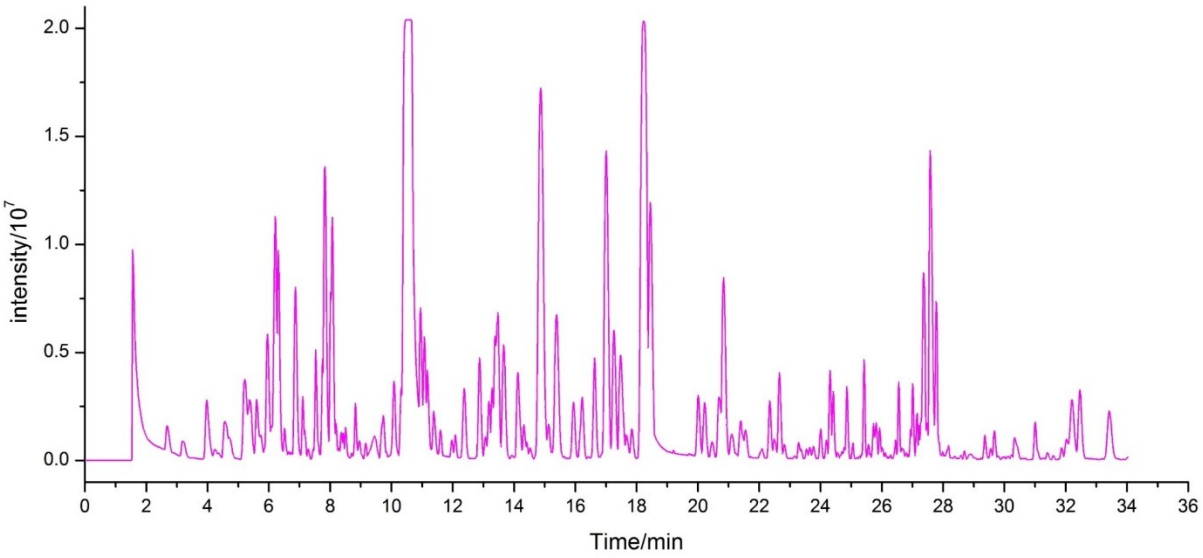
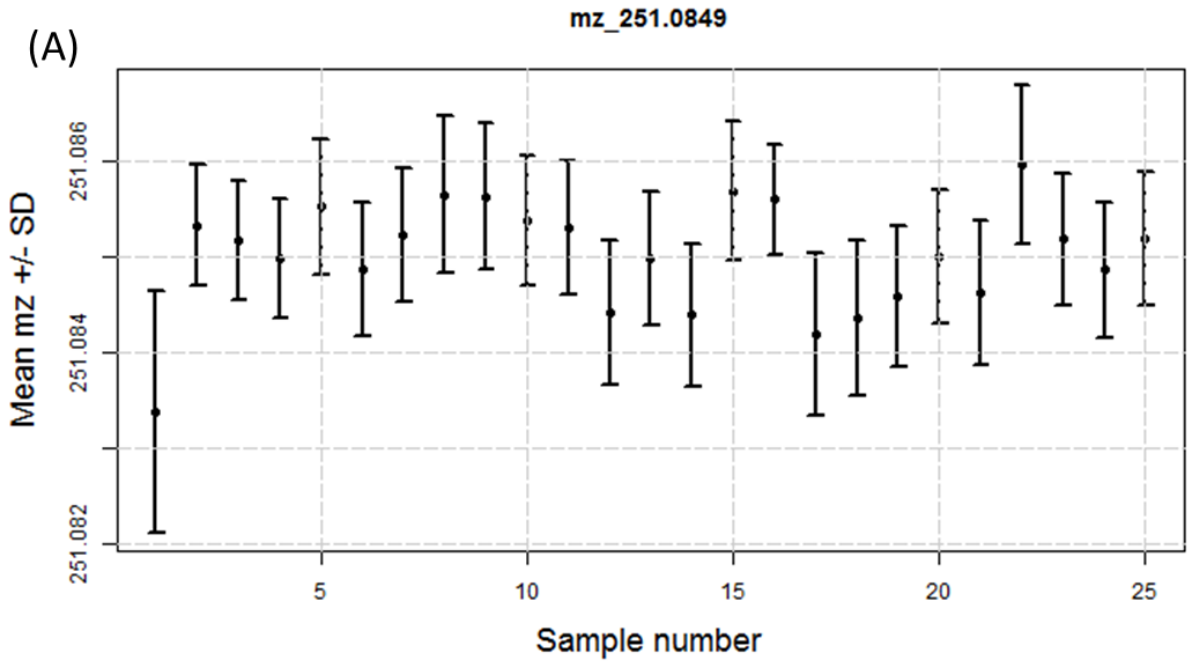


Figure 7.7 An example of the base peak chromatogram in a LC-MS analysis of a red wine sample.



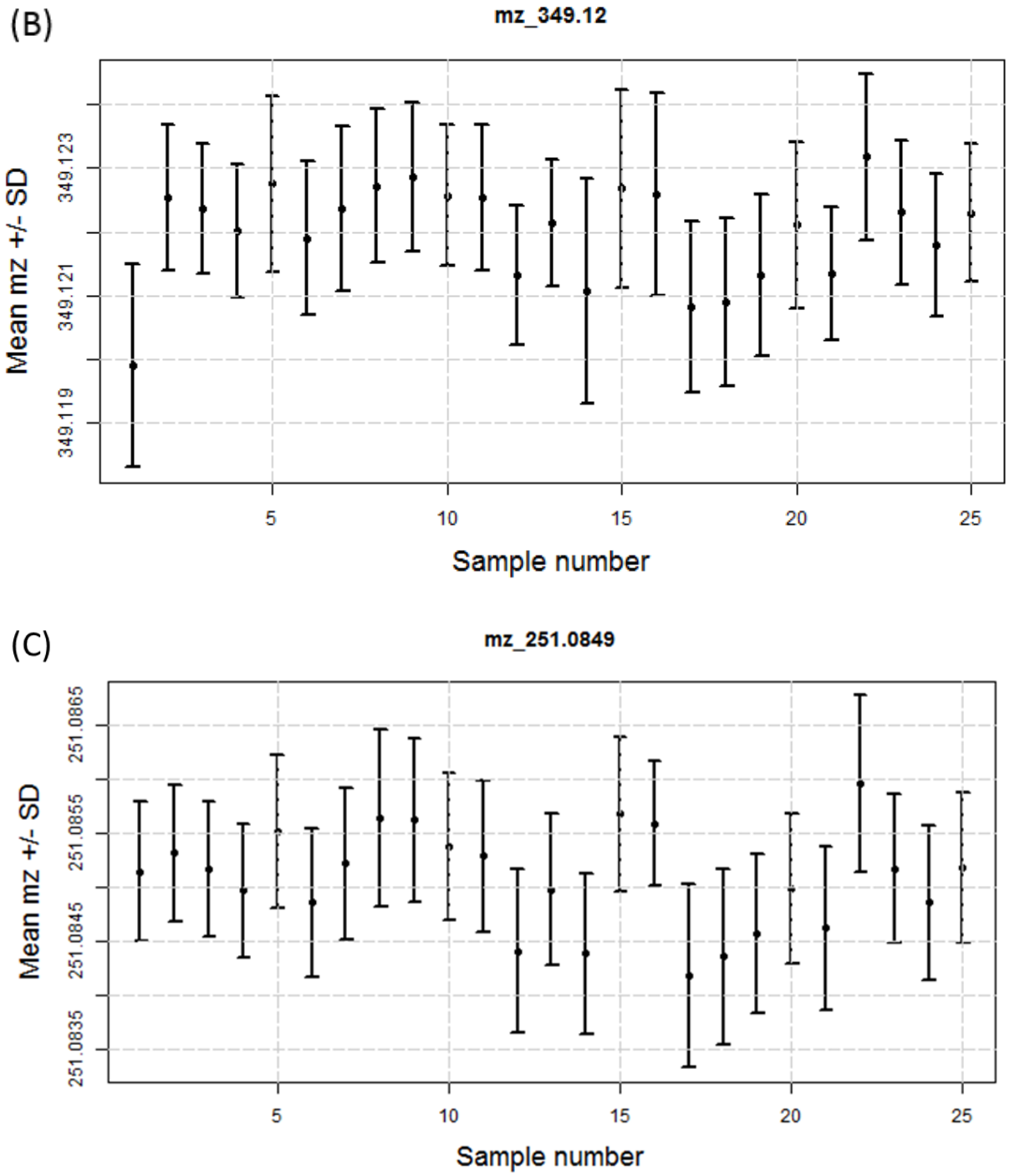


Figure 7.8 Mass accuracy results using two different internal compounds from the background mass peaks in the red wine data: (A) dansyl ammonia, (B) dansyl proline, and (C) mass check results after replacing the file with a mass accuracy issue with the newly collected data.

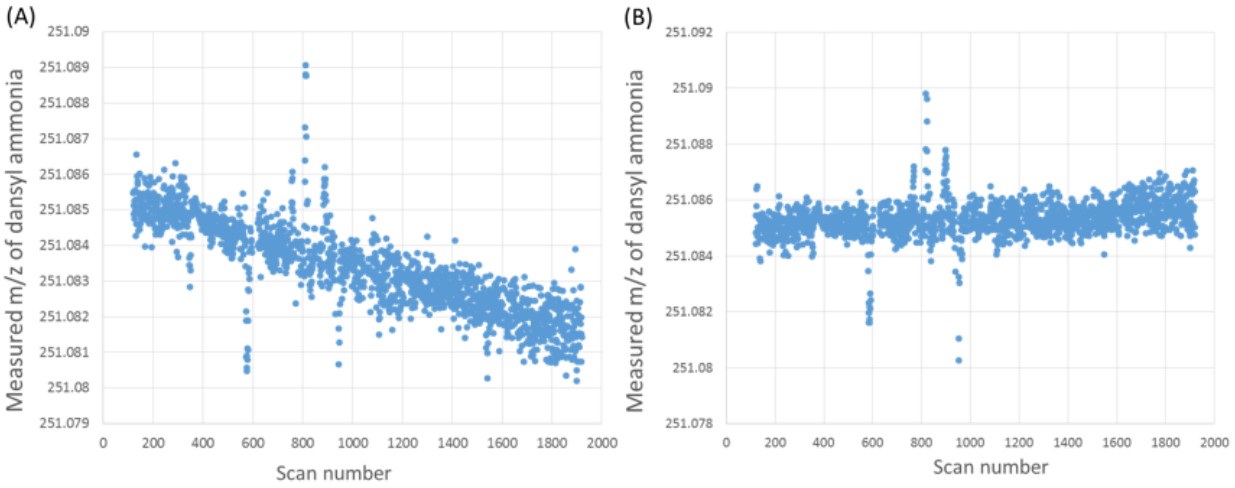
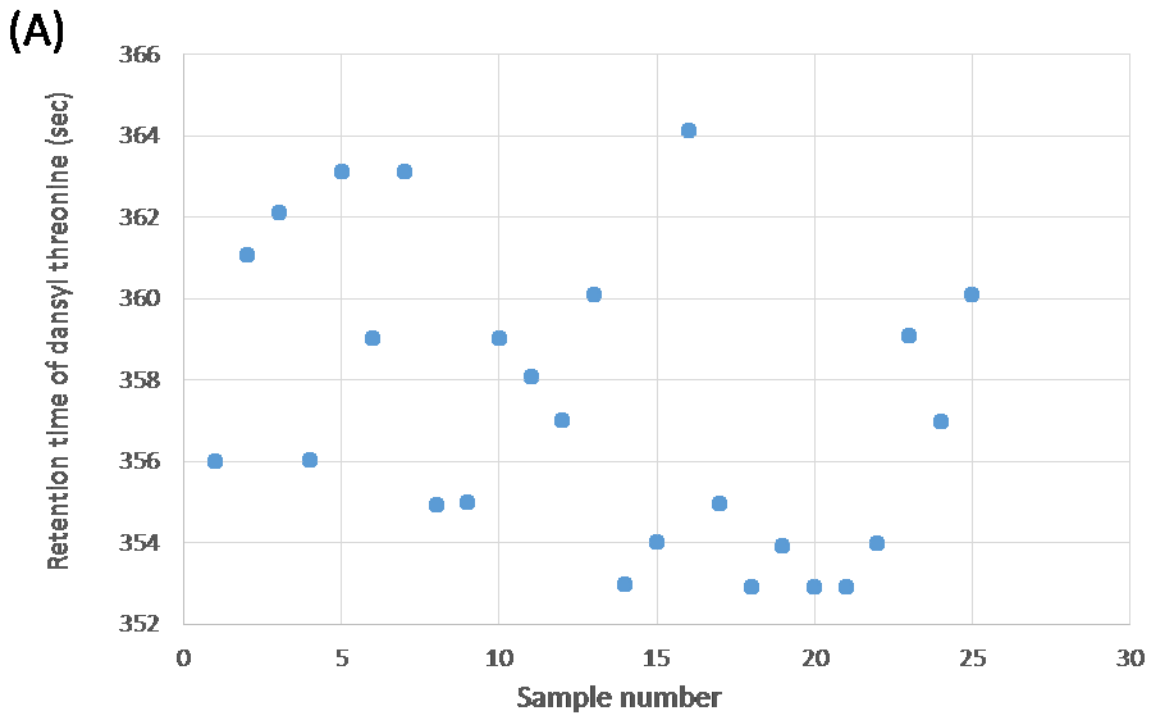


Figure 7.9 Measured masses of dansyl ammonia in (A) File 1, and (B) data re-collected using the same sample.



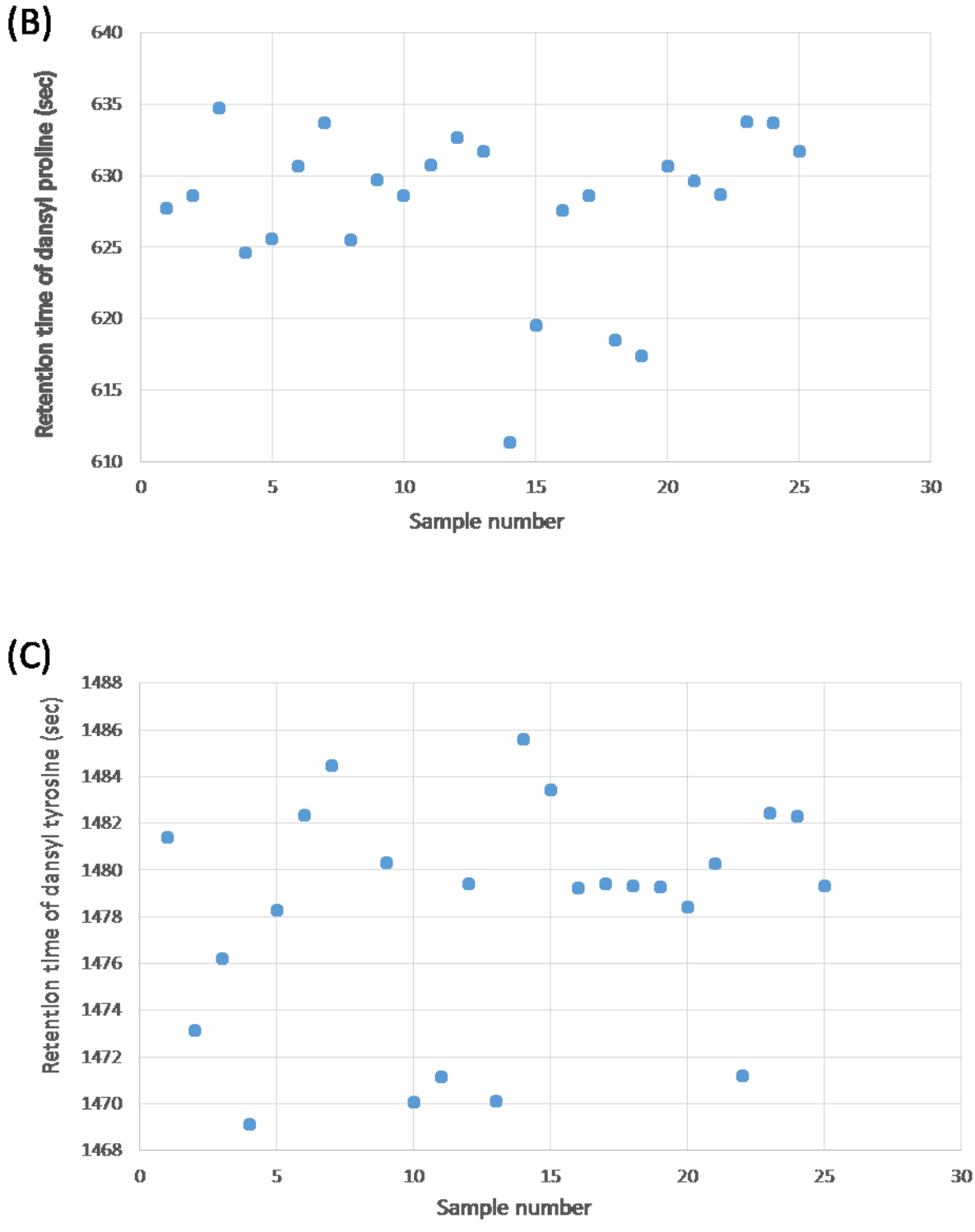


Figure 7.10 Retention time analysis using (A) threonine, (B) proline, and (C) tyrosine. Each data point is the retention time extracted from one sample data file.

After running the IsoMS program to extract the peak pairs in each individual sample, a peak pair list was created for each data file, and Figure 7.11 shows the number of peak pairs found in each sample from the experimental triplicates. On average, the number of peak pairs found in the red wine sample and the white wine sample were 3,178 ± 180 (n=19) and 1,983 ± 17 (n=3), respectively. The main difference in the white wine making procedure is that the grape skin is removed before the fermentation. The lower number of peak pairs found in the white wine sample, PG, indicated that a number of metabolites were missing in the wine after removing the grape skin.

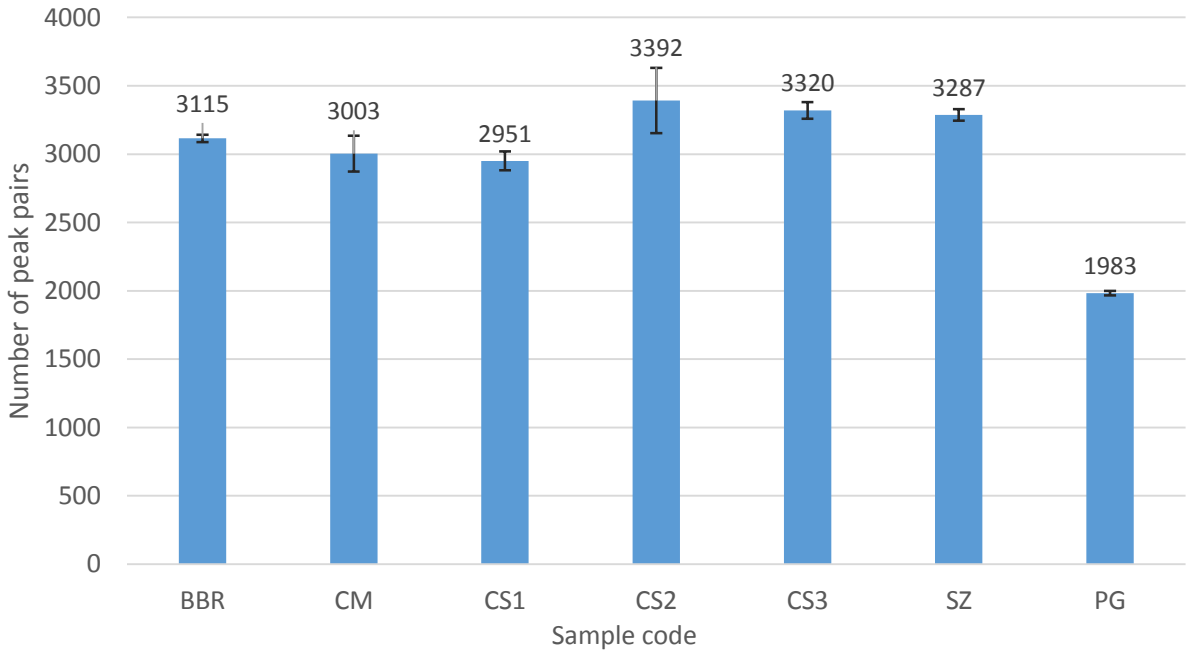


Figure 7.11 Number of peak pairs detected in each type of wine sample.

The peak pair lists were aligned and zero-filled using the processing method discussed in Chapter 4. After alignment of the peak pair data from 25 samples, a total of 13,467 peak pairs

were recorded in the metabolite-intensity table. After peak pair validation, 495 peak pairs were excluded as false peak pairs. In the redundant peak pair merging, 3,111 peak pairs were deleted as repeated peak pairs, and 58 peak pairs were deleted as tailing peak pairs. After these peak pair checking steps, 9,803 peak pairs were reserved in the data table. Figure 7.12 shows the missing value distribution in the ratio matrix. A white area represents the missing data, and a black area indicates the true ratios. We can see that a significant portion of ratio data can be retrieved by the zero-filling method. Three columns from the white wine samples can be observed to have a larger area of missing value in the NA maps, since some of the metabolites could not be detected in the white wine samples.

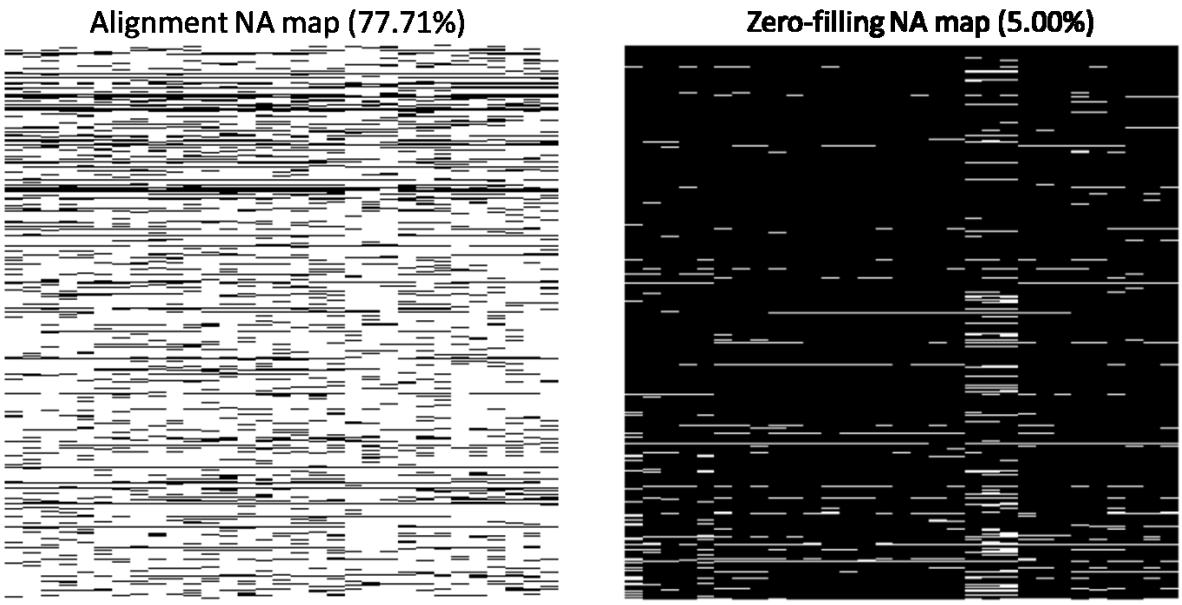


Figure 7.12 Missing value maps for ratio matrix after alignment and zerofilling.

The 9,803 peak pairs were searched against the dansyl standards library using a search window of 20 sec and 10 ppm, by which 305 metabolites were positively identified (see Table 7.4). Some of the peak pairs were matched to two dipeptide isomers in the library since the

retention time of some of the dipeptide isomers are within the retention time search window. In such cases, MS/MS data can be collected in the future to confirm the identity of the compounds. By using the accurate mass against the HMDB and EML libraries, 1,536 and 6,303 peak pairs were matched to one or a few chemical structures, respectively. It should be noted that the human metabolome database was used in the identification of wine metabolites since the database for wine metabolites is not available.

After applying the ratio imputation method to the ratio matrix of the aligned data table, we generated the complete data table and conducted the multivariate analysis with MetaboAnalyst. Figure 7.13 shows the principal component analysis (PCA) score plot. Each data point represents the data collected from one sample. The triplicate data points within a sample group cluster together tightly, indicating a good reproducibility of the method. We can observe a much greater separation of the white wine type PG, reflecting a much different metabolome between red wine and white wine. The unknown samples, which were originally collected from the CS1 sample, show a good clustering with the CS1 sample points. The CS2 and CS3 samples from the same brand with different production years were separated in the plot, showing a good sensitivity of the method to detect difference among different batches of red wines. Figure 7.14 shows the peak pair ratios of two identified compounds from the dansyl library. With the ratio results, we can monitor the concentration difference of each identified compound in different wine samples easily.

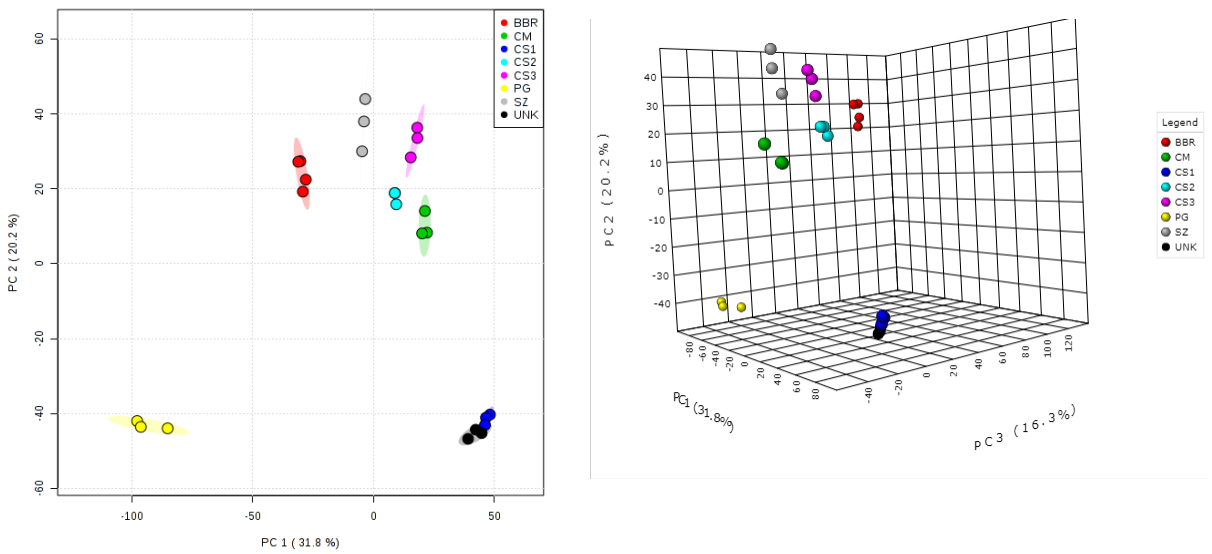


Figure 7.13 PCA score plots (2-D and 3-D plots) of wine samples grouped based on the brand of the wine.

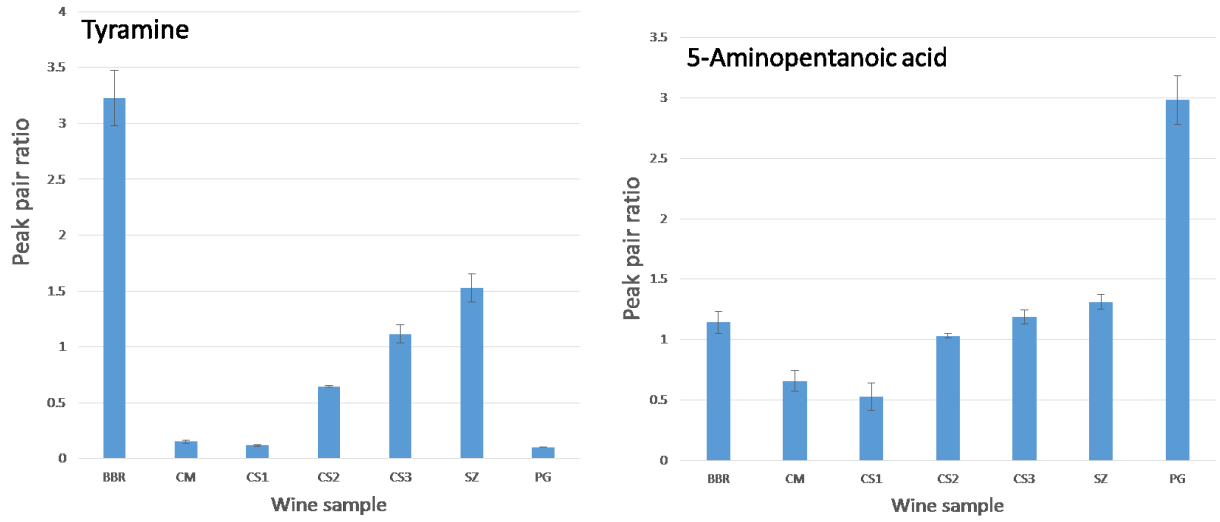


Figure 7.14 Peak pair ratio changes in different wine samples using two identified compounds: tyramine and 5-aminopentanoic acid.

7.4 Conclusions

We have developed an analytical workflow to address the current challenges of performing comprehensive and quantitative profiling of the wine metabolome. The sample preparation was optimized, and the dansylation isotope labeling LC-MS method was applied successfully in the wine metabolomics study.

The data processing methods described in Chapters 2–4 were used in the wine data processing. With the mass accuracy check and retention time shift analysis, we were able to pick out the file with a mass accuracy issue quickly after raw data were exported. The sample was re-analyzed, and the new data were used to replace the issue file. Mass and retention time tolerances were determined from the raw data checking results.

After data alignment and zero-filling, a total of 9,308 peak pairs were aligned in the metabolite-intensity table, with each peak pair representing a unique labeled metabolite commonly detected in the wine samples. Among all the detected peak pairs, 305 of them were positively identified using the dansyl library. The high-coverage and quantitative metabolome data were used to reveal small differences in wine metabolome profiles. The application of the workflow potentially can be applied to a comprehensive and quantitative wine metabolomics studies in the future for the quality control and product authentication in the wine industry.

Table 7.4 Search results using the dansyl library.

No	Input mass	Calibrated RT (sec)	Name	mz of light peak	Monoisotopic mass	Library RT (sec)

1	408.1709	146.74	L-Arginine	408.1700	174.1117	2.44
2	318.0780	1441.10	3,4-Dihydroxybenzeneacetic acid	318.0794	168.0423	23.90
3	364.1810	278.12	Agmatine	364.1802	130.1218	4.52
4	319.1115	800.31	Gamma-Aminobutyric acid - H2O	319.1144	103.0633	13.57
5	317.6082	1535.28	3-Methoxytyramine	317.6056	167.0946	25.49
6	317.6082	1535.28	Phenylephrine	317.6056	167.0946	25.39
7	302.6001	1550.74	Tyramine	302.6004	137.0841	25.83
8	329.1059	735.99	Diaminopimelic acid	329.1060	190.0954	12.30
9	329.1057	768.95	Diaminopimelic acid - Isomer	329.1060	190.0954	12.96
10	364.6259	785.10	Lysyl-Aspartate	364.6246	261.1325	13.00
11	356.1321	1544.41	4-Ethylphenol	356.1315	122.0732	25.63
12	337.1223	462.38	Gamma-Aminobutyric acid	337.1216	103.0633	7.79
13	363.1023	561.09	L-Glutamic Acid - H2O	363.1009	147.0532	9.46
14	366.1132	178.48	L-Asparagine	366.1118	132.0535	3.00
15	278.1082	1266.97	1,4-diaminobutane	278.1083	88.1000	21.27
16	400.1222	1087.97	Desaminotyrosine	400.1213	166.0630	18.04
17	371.6324	785.37	Lysyl-Glutamate	371.6324	275.1481	13.05
18	289.0775	1599.28	pyrocatechol	289.0767	110.0368	26.70
19	438.1484	681.70	L-Tryptophan	438.1482	204.0899	11.44
20	351.1378	523.83	5-Aminopentanoic acid	351.1373	117.0790	8.68
21	505.2227	253.10	Arginyl-Proline	505.2228	271.1644	4.23
22	546.2048	988.32	Phenylalanylphenylalanine	546.2057	312.1474	16.49
23	368.0875	948.79	L-Homocystine	368.0859	268.0551	15.82
24	521.2538	388.18	Arginyl-Leucine	521.2541	287.1957	6.54
25	510.1905	130.51	Saccharopine	510.1905	276.1321	2.26
26	460.1178	511.11	Uridine - H2O	460.1173	244.0695	8.67
27	456.1593	696.49	Glycyl-Phenylalanine	456.1588	222.1004	11.65
28	361.1336	646.44	4-Guanidinobutanoic acid - H2O	361.1329	145.0851	11.00
29	446.1750	546.68	L-prolyl-L-proline	446.1744	212.1161	9.08
30	335.6218	911.48	Glycyl-Lysine	335.6218	203.1270	15.20
31	342.6295	929.28	Lysyl-Alanine	342.6296	217.1426	15.53
32	383.1106	649.09	L-Methionine	383.1094	149.0510	10.89
33	436.1902	653.86	Alanyl-isoleucine	436.1900	202.1317	10.86
34	465.1793	374.13	Asparaginyl-Valine	465.1802	231.1219	6.21
35	307.1110	1047.88	L-Lysine	307.1111	146.1055	17.47
36	450.2058	690.61	Valyl-Valine	450.2057	216.1474	11.45

37	353.1063	1278.59	Glycyl-Tyrosine	353.1060	238.0954	21.63
38	436.1903	676.62	Alanyl-Leucine	436.1901	202.1317	11.36
39	422.1869	165.31	Homo-L-arginine	422.1856	188.1273	3.00
40	374.1301	1373.29	Tyrosyl-Valine	374.1295	280.1423	22.83
41	365.1543	783.43	L-Alloisoleucine	365.1529	131.0946	13.20
42	365.1543	783.43	L-Isoleucine	365.1529	131.0946	13.06
43	452.1850	449.42	Isoleucyl-Serine	452.1850	218.1267	7.46
44	452.1850	449.42	Threoninyl-Valine	452.1850	218.1267	7.61
45	464.2214	819.31	Leucyl-Valine	464.2214	230.1630	13.70
46	464.2214	819.31	Valyl-Isoleucine	464.2214	230.1630	13.60
47	464.2214	819.31	Valyl-Leucine	464.2214	230.1630	13.95
48	464.2212	825.42	Valyl-Leucine	464.2214	230.1630	13.95
49	410.1381	324.22	Glycyl-Threonine	410.1380	176.0797	5.10
50	399.1379	772.70	L-Phenylalanine	399.1373	165.0790	12.74
51	339.1019	262.21	L-Serine	339.1009	105.0426	4.40
52	381.1362	1422.01	Leucyl-Tyrosine	381.1373	294.1580	23.98
53	353.1063	1192.71	Tyrosyl-Glycine	353.1060	238.0954	20.19
54	462.2056	782.28	Leucyl-Proline	462.2057	228.1474	12.99
55	480.1620	706.10	Prolyl-Methionine	480.1621	246.1038	11.64
56	452.1850	526.40	Serinyl-Leucine	452.1850	218.1267	8.90
57	452.1850	526.40	Serylisoleucine	452.1850	218.1267	8.66
58	551.2325	948.80	Leucyl-Tryptophan	551.2323	317.1739	15.77
59	346.0867	680.03	Uracil	346.0856	112.0273	11.34
60	480.1607	678.49	Methionyl-Proline	480.1621	246.1038	11.16
61	464.2198	792.09	Isoleucyl-Valine	464.2214	230.1630	13.14
62	365.1528	801.89	L-leucine	365.1529	131.0946	13.36
63	470.1382	400.98	Methionyl-Serine	470.1414	236.0831	6.86
64	360.1140	1301.05	Alanyl-Tyrosine	360.1138	252.1110	21.85
65	328.1002	1401.91	Phenol	328.1002	94.0419	23.16
66	345.0921	813.06	Allocystathionine	345.0920	222.0674	13.33
67	345.0921	813.06	Allocystathionine - Isomer	345.0920	222.0674	13.61
68	345.0921	813.06	L-Cystathionine	345.0920	222.0674	13.34
69	345.0921	813.06	L-Cystathionine - Isomer	345.0920	222.0674	13.69
70	498.2055	879.60	Valyl-Phenylalanine	498.2057	264.1474	14.63
71	460.1634	1055.31	Alanyl-Histidine	460.1649	226.1066	17.62
72	500.1850	642.75	Threoninyl-Phenylalanine	500.1850	266.1267	10.70
73	381.1363	1429.45	Tyrosyl-Isoleucine	381.1373	294.1580	23.76
74	381.1363	1429.45	Tyrosyl-Leucine	381.1373	294.1580	23.77
75	373.0864	909.66	6-Hydroxynicotinic acid - Isomer	373.0853	139.0269	15.32

76	371.6330	813.62	Glutamyl-Lysine	371.6324	275.1481	13.60
77	394.1440	384.05	Alanyl-Alanine	394.1431	160.0848	6.10
78	337.1199	542.90	2-Aminoisobutyric acid	337.1216	103.0633	8.91
79	337.1199	542.90	D-Alpha-aminobutyric acid	337.1216	103.0633	9.23
80	337.1199	542.90	L-Alpha-aminobutyric acid	337.1216	103.0633	9.13
81	478.2368	909.35	Isoleucyl-Isoleucine	478.2370	244.1787	15.14
82	422.1749	658.15	Glycyl-Isoleucine	422.1744	188.1161	10.78
83	422.1749	658.15	Glycyl-L-leucine	422.1744	188.1161	10.97
84	373.1220	1354.20	Tyrosyl-Proline	373.1216	278.1267	22.60
85	318.0794	1498.64	Homogentisic acid	318.0794	168.0423	24.84
86	512.2214	995.00	Leucyl-Phenylalanine	512.2214	278.1630	16.59
87	494.1953	562.54	Glutamylisoleucine	494.1955	260.1372	9.44
88	375.1206	1135.91	Tyrosyl-Threonine	375.1191	282.1216	19.10
89	479.2071	158.73	Alanyl-Arginine	479.2071	245.1488	2.43
90	479.2071	158.73	Arginyl-Alanine	479.2071	245.1488	2.78
91	337.1221	517.95	3-Aminoisobutanoic acid	337.1216	103.0633	8.67
92	454.1465	537.24	Alanyl-Methionine	454.1465	220.0882	9.00
93	454.1465	537.24	Methionyl-Alanine	454.1465	220.0882	8.84
94	355.6380	1069.01	Prolyl-Lysine	355.6375	243.1583	17.92
95	553.1760	478.59	Tryptophyl-Aspartate	553.1751	319.1168	7.74
96	342.6286	954.38	Alanyl-Lysine	342.6296	217.1426	15.83
97	364.1329	705.45	Lysyl-Asparagine	364.1325	260.1485	11.77
98	514.1646	591.23	Aspartyl-Phenylalanine	514.1642	280.1059	9.97
99	514.1646	591.23	L-Aspartyl-L-phenylalanine	514.1642	280.1059	10.07
100	511.1745	434.36	Arginyl-Cysteine	511.1792	277.1209	7.08
101	512.2208	928.40	Phenylalanyl-Isoleucine	512.2214	278.1630	15.44
102	537.2123	121.78	Arginyl-Glutamic acid	537.2126	303.1543	1.97
103	537.2123	121.78	Glutamylarginine	537.2126	303.1543	2.03
104	335.6217	874.97	Lysyl-Glycine	335.6218	203.1270	14.61
105	393.1246	1565.97	Histidinyl-Tyrosine	393.1247	318.1328	26.37
106	351.1373	645.59	L-Valine	351.1373	117.0790	10.81
107	357.6352	853.13	Threoninyl-Lysine	357.6349	247.1532	14.18
108	356.6454	1097.96	Lysyl-Valine	356.6453	245.1739	18.44
109	494.1957	586.05	Glutamylleucine	494.1955	260.1372	9.84
110	381.1352	1434.60	Isoleucyl-Tyrosine	381.1373	294.1580	23.98
111	410.1387	261.48	Alanyl-Serine	410.1380	176.0797	4.54
112	410.1387	261.48	Serylalanine	410.1380	176.0797	4.36
113	466.2007	599.41	Threoninyl-Leucine	466.2006	232.1423	10.18
114	355.6378	1052.52	Lysyl-Proline	355.6375	243.1583	17.65

115	452.1850	473.49	Leucyl-Serine	452.1850	218.1267	7.93
116	436.1905	577.36	Isoleucyl-Alanine	436.1901	202.1317	9.64
117	356.6457	1081.14	Valyl-Lysine	356.6453	245.1739	18.08
118	478.2365	928.59	Isoleucyl-Leucine	478.2370	244.1787	15.51
119	478.2365	928.59	Leucyl-Isoleucine	478.2370	244.1787	15.57
120	360.1024	657.90	Imidazoleacetic acid	360.1012	126.0429	11.12
121	382.1099	1151.71	Aspartyl-Tyrosine	382.1087	296.1008	19.35
122	410.1387	245.78	Threoninyl-Glycine	410.1380	176.0797	4.20
123	466.2008	534.70	Leucyl-Threonine	466.2006	232.1423	8.84
124	358.1113	1354.60	Guaiacol	358.1107	124.0524	22.54
125	399.1051	241.12	Methionine Sulfoxide - Isomer	399.1043	165.0460	4.20
126	478.2369	944.17	Leucyl-Leucine	478.2370	244.1787	15.80
127	399.1055	210.77	Methionine Sulfoxide	399.1043	165.0460	3.72
128	479.1953	486.18	Asparaginyl-Leucine	479.1959	245.1376	7.93
129	293.0949	943.66	2,4-Diaminobutyric acid	293.0954	118.0742	15.80
130	389.1177	1145.57	Glutamyltyrosine	389.1166	310.1165	19.30
131	493.2112	553.55	Glutaminylisoleucine	493.2115	259.1532	9.32
132	408.1593	586.85	Gly-Norvaline	408.1588	174.1005	9.51
133	315.1081	827.23	5-Hydroxylysine	315.1085	162.1004	13.88
134	416.1169	867.04	Hydroxyphenyllactici acid	416.1162	182.0579	14.39
135	402.1013	1047.77	Vanillic acid	402.1006	168.0423	17.34
136	360.1142	1242.82	Tyrosyl-Alanine	360.1138	252.1110	20.86
137	521.2541	375.61	Arginyl-Isoleucine	521.2541	287.1957	6.21
138	466.1988	573.46	Threoninyl-Isoleucine	466.2006	232.1423	9.44
139	388.6233	993.02	Tyrosyl-Glutamine	388.6246	309.1325	16.71
140	494.1957	481.93	Isoleucyl-Glutamate	494.1955	260.1372	8.01
141	422.1751	568.27	Alanyl-Valine	422.1744	188.1161	9.44
142	422.1751	568.27	Leucyl-Glycine	422.1744	188.1161	9.37
143	370.0986	512.86	Hypoxanthine - multi-tags	370.0968	136.0385	8.73
144	512.2210	948.21	Phenyl-Leucine	512.2214	278.1631	15.90
145	408.1593	531.52	Glycyl-Valine	408.1588	174.1004	9.19
146	382.1080	1089.36	Tyrosyl-Aspartate	382.1087	296.1008	18.44
147	363.6536	1170.98	Isoleucyl-Lysine	363.6531	259.1896	19.41
148	480.1800	404.88	Valyl-Glutamate	480.1799	246.1216	6.57
149	368.1117	1087.82	Tyrosyl-Serine	368.1113	268.1059	18.26
150	371.1402	696.33	Lysyl-Glutamine	371.1404	274.1641	11.58
151	375.6415	1458.70	Histidinyl-Lysine	375.6405	283.1644	24.12
152	375.6415	1458.70	Lysyl-Histidine	375.6405	283.1644	24.15
153	486.1809	1191.97	Prolyl-Histidine	486.1806	252.1222	20.09

154	537.2137	838.09	Valyl-Tryptophan	537.2166	303.1583	13.98
155	478.1280	463.18	Uridine	478.1279	244.0695	7.84
156	409.1546	205.80	Citrulline	409.1540	175.0957	3.74
157	368.1115	1169.12	Seryltyrosine	368.1113	268.1059	19.59
158	465.1913	152.67	Glycyl-Arginine	465.1915	231.1331	2.39
159	326.0777	1307.64	3,4-Dihydroxymandelic acid	326.0769	184.0372	21.73
160	512.2213	970.42	Isoleucyl-Phenylalanine	512.2214	278.1630	16.17
161	495.1715	673.62	Glycyl-Tryptophan	495.1697	261.1113	11.19
162	496.1874	779.45	L-phenylalanyl-L-proline	496.1901	262.1317	12.94
163	356.0960	1304.20	3-Hydroxymandelic acid - COOH	356.0951	168.0423	21.64
164	422.1742	523.02	Isoleucyl-Glycine	422.1744	188.1161	8.65
165	420.1600	453.93	Alanyl-Proline	420.1588	186.1004	7.89
166	500.1850	526.01	Phenylalanyl-Threonine	500.1850	266.1267	8.73
167	480.1801	503.93	Leucyl-Aspartate	480.1799	246.1216	8.37
168	466.2004	506.40	Isoleucyl-Threonine	466.2006	232.1423	8.47
169	464.1491	406.37	Aspartyl-Proline	464.1486	230.0903	6.50
170	464.1491	406.37	Prolyl-Aspartate	464.1486	230.0903	6.58
171	438.1695	381.89	Valyl-Serine	438.1693	204.1110	6.06
172	488.1958	1183.25	Valyl-Histidine	488.1962	254.1379	19.88
173	504.1547	861.42	Histidinyl-Aspartate	504.1547	270.0964	14.13
174	462.2047	738.39	Isoleucylproline	462.2057	228.1474	12.61
175	363.6541	1181.70	Leucyl-Lysine	363.6531	259.1896	19.73
176	381.6158	1058.88	Asparaginyl-Tyrosine	381.6167	295.1168	17.95
177	462.2057	833.67	Prolyl-Isoleucine	462.2057	228.1474	13.91
178	459.1331	436.75	Cytidine - H2O	459.1333	243.0855	7.38
179	454.1642	241.35	Threoninyl-Threonine	454.1642	220.1059	4.05
180	452.1606	858.10	N-Acetylserotonin	452.1638	218.1055	14.32
181	454.1282	164.41	Serylaspartic acid	454.1279	220.0695	2.81
182	480.1800	469.74	Glutamylvaline	480.1799	246.1216	7.82
183	480.1800	469.74	Isoleucyl-Aspartate	480.1799	246.1216	7.65
184	466.1644	392.24	Valyl-Aspartate	466.1642	232.1059	6.33
185	490.1750	979.87	Threoninyl-Histidine	490.1755	256.1172	16.36
186	311.0716	1482.16	Gentisic acid - multi-tags	311.0716	154.0266	24.69
187	311.0716	1482.16	Protocatechuic acid	311.0716	154.0266	24.51
188	466.1640	457.29	Aspartyl-Valine	466.1642	232.1059	7.61
189	493.2109	410.23	Isoleucyl-Glutamine	493.2115	259.1532	6.86
190	374.1301	1401.42	Valyl-Tyrosine	374.1295	280.1423	23.30
191	373.0860	1417.84	4-Nitrophenol	373.0853	139.0269	23.45
192	424.1179	259.85	Aspartyl-Glycine	424.1173	190.0590	4.20
193	424.1179	259.85	Glycyl-Aspartate	424.1173	190.0590	4.40

194	386.0931	529.26	Xanthine	386.0917	152.0334	8.95
195	468.1423	180.47	Glutamylserine	468.1435	234.0852	3.27
196	502.2114	1282.13	Histidinyl-Isoleucine	502.2119	268.1535	21.38
197	502.2114	1282.13	Histidinyl-Leucine	502.2119	268.1535	21.62
198	502.2114	1282.13	Isoleucyl-Histidine	502.2119	268.1535	21.20
199	502.2114	1282.13	Leucyl-Histidine	502.2119	268.1535	21.57
200	388.1200	1431.89	3,5-Dimethoxyphenol	388.1213	154.0630	23.75
201	438.1324	284.94	Alanyl-Aspartic Acid	438.1329	204.0746	4.83
202	438.1324	284.94	Glycyl-Glutamate	438.1329	204.0746	4.87
203	363.6532	1206.95	Lysyl-Leucine	363.6531	259.1896	20.28
204	498.1403	440.91	Aspartyl-Methionine	498.1363	264.0780	7.63
205	498.1403	440.91	Methionyl-Aspartate	498.1363	264.0780	7.16
206	476.1597	908.81	Serylhistidine	476.1598	242.1015	15.14
207	365.1179	455.27	5-Aminolevulinic acid	365.1166	131.0582	7.59
208	408.1590	432.80	Valyl-Glycine	408.1588	174.1004	7.07
209	363.6533	1192.08	Lysyl-Isoleucine	363.6531	259.1896	19.98
210	493.2115	445.63	Leucyl-Glutamine	493.2115	259.1532	7.19
211	398.1283	1449.18	Phenylalanyl-Tyrosine	398.1295	328.1423	24.22
212	398.1283	1449.18	Tyrosyl-Phenylalanine	398.1295	328.1423	24.08
213	528.1800	520.21	Phenylalanyl-Glutamate	528.1799	294.1216	8.54
214	318.1019	1164.85	Pyridoxamine	318.1033	168.0899	19.47
215	411.1060	1159.51	Chlorogenic acid	411.1059	354.0951	19.45
216	479.1926	433.25	Leucyl-Asparagine	479.1959	245.1376	7.16
217	363.1384	796.72	D-Pipecolic acid	363.1373	129.0790	13.23
218	363.1384	796.72	L-Pipecolic acid	363.1373	129.0790	13.45
219	486.1695	582.18	Serinyl-Phenylalanine	486.1693	252.1110	9.38
220	360.1019	797.32	Thymine	360.1012	126.0429	13.21
221	480.1801	544.14	Aspartyl-Isoleucine	480.1799	246.1216	9.17
222	480.1803	569.11	Aspartyl-Leucine	480.1799	246.1216	9.53
223	496.1927	843.00	Methionyl-Isoleucine	496.1934	262.1351	13.94
224	496.1927	843.00	Methionyl-Leucine	496.1934	262.1351	14.26
225	536.1961	1285.65	Phenylalanyl-Histidine	536.1962	302.1379	21.64
226	450.1693	404.15	Prolyl-Threonine	450.1693	216.1110	6.91
227	365.1512	836.02	L-Norleucine	365.1529	131.0946	14.11
228	335.1065	556.95	L-Homoserine - H2O	335.1060	119.0582	9.26
229	323.1063	452.26	L-Alanine	323.1060	89.0477	7.57
230	462.2059	844.85	Prolyl-Leucine	462.2057	228.1474	14.09
231	416.1152	990.56	Homovanillic acid	416.1162	182.0579	16.51
232	436.1544	387.77	Prolyl-Serine	436.1537	202.0954	6.19
233	535.2012	834.77	Prolyl-Tryptophan	535.2010	301.1426	13.89
234	496.1900	911.14	L-prolyl-L-phenylalanine	496.1901	262.1317	15.11

235	406.1440	461.24	Prolylglycine	406.1431	172.0848	7.57
236	450.1695	390.76	Threoninyl-Proline	450.1693	216.1110	6.44
237	364.1703	418.48	N-Acetylputrescine	364.1689	130.1106	7.25
238	451.1659	206.22	Alanyl-Glutamine	451.1646	217.1063	3.72
239	509.1708	150.63	Glutaminylglutamic acid	509.1701	275.1117	2.67
240	509.1708	150.63	Glutamylglutamine	509.1701	275.1117	2.69
241	285.1161	1345.87	Cadaverine	285.1162	102.1157	22.39
242	514.1651	506.60	Phenylalanyl-Aspartate	514.1642	280.1059	8.54
243	375.1182	1235.16	Threoninyl-Tyrosine	375.1191	282.1216	20.72
244	525.1817	457.75	Tryptophyl-Serine	525.1802	291.1219	7.39
245	380.6449	1218.65	Lysyl-Phenylalanine	380.6453	293.1739	20.42
246	481.1752	157.39	Threoninyl-Glutamine	481.1751	247.1168	2.67
247	452.1516	304.60	Alanyl-Glutamic acid	452.1486	218.0903	4.97
248	476.1602	834.76	Histidinyl-Serine	476.1598	242.1015	14.01
249	478.1649	417.89	Prolyl-Glutamate	478.1642	244.1059	6.77
250	373.1225	1415.62	Prolyl-Tyrosine	373.1216	278.1267	23.50
251	309.0908	393.65	Glycine	309.0903	75.0320	6.59
252	388.0853	1036.69	Gentisic acid	388.0849	154.0266	17.11
253	470.1745	725.92	Alanyl-Phenylalanine	470.1744	236.1161	12.11
254	353.1173	362.07	L-Threonine	353.1166	119.0582	5.79
255	411.1060	1259.28	Chlorogenic acid - Isomer	411.1059	354.0951	21.24
256	440.1495	189.36	Serylthreonine	440.1486	206.0903	3.40
257	440.1495	189.36	Threoninyl-Serine	440.1486	206.0903	3.25
258	420.1599	509.99	Prolyl-Alanine	420.1588	186.1004	8.47
259	446.1488	916.53	Histidinyl-Glycine	446.1493	212.0909	15.32
260	324.5954	1362.12	L-Tyrosine	324.5953	181.0739	22.65
261	324.5954	1362.12	o-Tyrosine	324.5953	181.0739	22.38
262	300.1027	1004.14	Ornithine	300.1033	132.0899	16.58
263	380.6459	1185.47	Phenylalanyl-Lysine	380.6453	293.1739	19.92
264	390.1192	1370.31	Tyrosyl-Methionine	390.1155	312.1144	22.79
265	482.1596	208.29	Threoninyl-Glutamate	482.1592	248.1008	3.76
266	295.1116	377.11	Ethanolamine	295.1111	61.0528	6.00
267	432.1113	763.50	Vanillylmandelic acid	432.1111	198.0528	12.81
268	477.1447	369.79	Cytidine	477.1438	243.0855	5.87
269	486.1709	483.31	Phenylalanyl-Serine	486.1693	252.1110	8.02
270	390.1147	1409.12	Methionyl-Tyrosine	390.1155	312.1144	23.37
271	426.1289	155.10	Serylserine	426.1329	192.0746	2.69
272	319.1115	987.25	3-Aminoisobutanoic acid - H2O	319.1110	103.0633	16.29
273	380.1280	196.87	L-Glutamine	380.1275	146.0691	3.32
274	372.0912	1061.64	4-Hydroxybenzoic acid	372.0900	138.0317	17.57
275	394.1574	1088.00	Tryptamine	394.1584	160.1000	18.03

276	494.1957	516.47	Leucyl-Glutamate	494.1955	260.1372	8.55
277	518.1702	861.80	Histidinyl-Glutamate	518.1704	284.1121	14.23
278	428.1162	1111.99	trans-Ferulic acid	428.1162	194.0579	18.47
279	359.0751	137.00	Taurine	359.0730	125.0147	2.24
280	365.1493	862.56	L-Norleucine	365.1529	131.0946	14.11
281	467.1598	139.19	Glutaminylserine	467.1595	233.1012	2.32
282	467.1598	139.19	Serylglutamine	467.1595	233.1012	2.33
283	495.1540	131.08	Glutaminylaspartic acid	495.1544	261.0961	2.49
284	388.0855	989.39	2-Pyrocatechuic acid	388.0849	154.0266	16.31
285	353.1177	242.81	L-Homoserine	353.1166	119.0582	4.05
286	323.1063	556.03	Sarcosine	323.1060	89.0477	9.34
287	342.1163	1485.72	p-Cresol	342.1158	108.0575	24.54
288	436.2015	184.05	Symmetric dimethylarginine	436.2013	202.1430	3.05
289	406.1426	427.12	Glycylproline	406.1431	172.0848	7.17
290	594.2538	417.04	Arginyl-Tryptophan	594.2493	360.1910	6.75
291	530.1779	822.75	Phenylalanyl-Methionine	530.1778	296.1195	13.87
292	470.1739	621.72	Phenylalanyl-Alanine	470.1744	236.1161	10.58
293	398.1064	1115.01	m-Coumaric acid	398.1057	164.0473	18.51
294	512.1531	459.46	Glutamylmethionine	512.1520	278.0936	7.77
295	422.1748	407.04	N-Alpha-acetyllysine	422.1744	188.1161	6.79
296	432.1117	1079.68	Syringic acid	432.1111	198.0528	18.10
297	400.1487	1174.49	Lysyl-Tryptophan	400.1507	332.1848	19.83
298	323.1067	433.28	Beta-Alanine	323.1060	89.0477	7.24
299	452.1813	428.84	Valyl-Threonine	452.1850	218.1267	6.96
300	339.1347	328.48	Diethanolamine	339.1373	105.0790	5.49
301	402.1007	986.14	5-Methoxysalicylic acid	402.1006	168.0423	16.38
302	389.1289	1086.68	L-Histidine	389.1278	155.0695	18.09
303	381.6165	1012.87	Tyrosyl-Asparagine	381.6167	295.1168	16.96
304	481.1384	136.92	Asparaginyl-Aspartic acid	481.1388	247.0804	2.34
305	402.0979	788.49	3-Hydroxymandelic acid	402.1006	168.0423	12.94

Chapter 8 Conclusions and Future Work

Metabolomics is an active research field concerned with developing methods for the analysis of low molecular weight compounds in biological systems. In metabolomics analysis, large amounts of data are produced routinely in order to characterize a sample consisting of hundreds to thousands of metabolites. The conclusions drawn from the resultant data rely on the accuracy of the metabolite concentration measurements, the coverage of the detection method, and the completeness of data to include all the metabolite signals. Therefore, a number of challenges are associated with the data processing methods specific to each experimental platform.

The LC-MS technique has been used widely in the application of metabolomics due to its high sensitivity and high throughput. However, the traditional LC-MS platforms are limited by the coverage of the detection and the less reproducible quantification results. Thus, the chemical isotope labeling LC-MS method was developed in our group for an improvement of the metabolite separation and a higher detection sensitivity of a broad range of metabolites in a biological sample. In the labeling LC-MS method, each labeled metabolite will generate a peak pair signal in the mass spectrum, with the light peak from the individual sample and the heavy peak from the pooled sample. Accordingly, a customized data processing method is required in dealing with the data generated by different chemical isotope labeling LC-MS experiments. The design of the data processing algorithms have to consider the specific experimental methods and instrumental settings to serve the objectives of the metabolomics study. For this reason, my thesis work focuses primarily on the development of data processing methods to address the challenges from the growing data processing tasks in the chemical isotope labeling LC-MS. This

thesis work focuses on the novel aspects of the data processing area in the CIL LC-MS metabolomics workflow and provides a number of original algorithms design in response to the current challenges in metabolomics data processing.

Chapter 1 provides an overview of metabolomics, its relationship with other omics studies, and the different applications using metabolomics technologies. We compared the different analytical methods and discussed specifically the CIL LC-MS method to show the advantage in its detection sensitivity and quantification accuracy. We presented the current development of data processing in CIL LC-MS; possible solutions were proposed to the data processing challenges that arose with the development in the experimental methods.

Chapter 2 discuss the methods for the data quality check in the LC-MS raw data, which is not limited to the labeling method. Instead of using an internal standard in the sample preparation, the method exploited the internal compounds that are shared by different samples in their LC-MS raw data to monitor the instrument fluctuation in terms of the mass accuracy. The background mass peaks usually are the signals to be excluded from the LC-MS data in the initial data processing. In our method, we designed a program to find the background mass peaks that constantly show up in most of the mass spectra and used them as the internal references to calculate the mass shift over the course of data acquisition. Compared to the traditional approach that shows the mass errors of a few calibration standards after mass calibration, the background mass monitoring method examines the mass accuracy in every single mass spectrum and provides a more comprehensive evaluation of the mass accuracy. Although an internal mass calibration was not achieved with the background mass peaks due to the relatively low peak intensity, the mass accuracy monitoring method can help to pick out any LC-MS data file with a major mass shift and assist the user to correct or replace the issue data either by an additional

mass calibration or a re-analysis of the same sample. The program is generally applicable to any LC-MS data, and background mass peak was first used in the software for gauging mass accuracy.

In Chapter 3, in developing the retention time shift analysis, a similar strategy was applied by using the commonly detected internal compounds in the same type of sample to monitor the inter-sample retention time changes. Compared to other MS method without the labeling step, our method used the highly confident identified compound by their peak pair signals in the retention time analysis and correction. After the retention time of the internal compounds were extracted from each individual sample, the retention time shift at different retention time points can be viewed from the scattered plots generated by the program (see plot shown in Figure 8.1). A retention time correction program was designed to normalize the retention time of multiple LC-MS data based on the retention time of the selected internal compounds.

The LC-MS raw data check program provides a new way to determine the mass and retention time tolerances for data alignment and library search. One challenge in comparative metabolomics is to align the metabolites detected in different samples. In the LC-MS data, the retention time and mass are used as the parameters to assign the signals of peak pairs in different samples to the same compound. However, the retention time and mass measurement always will show some inter-sample variation due to the limitations of the analytical instruments. The selection of the alignment tolerances often is based on one's experience of the instruments and settings applied in the data acquisition. The measurement error can change from one experiment to another, and the same alignment tolerances may not be applicable to the data generated from different experiments. With the results from the LC-MS raw data check, the actual mass error

and retention time shift in all the sample data are presented in detail; this can be used to estimate the tolerances for data alignment and library search. Figure 8.1 shows an example of this strategy using the retention time analysis of dansyl threonine in a total of 25 sample data files. The range of the retention time data shows the overall retention time change at the dansyl threonine peak. Therefore, a retention time window of ±15 sec could be estimated roughly from the distribution of the dansyl threonine's retention times. Since the retention time shift can be different for different peaks, one can examine the results from all standards to determine the retention time window used in the subsequent data processing.

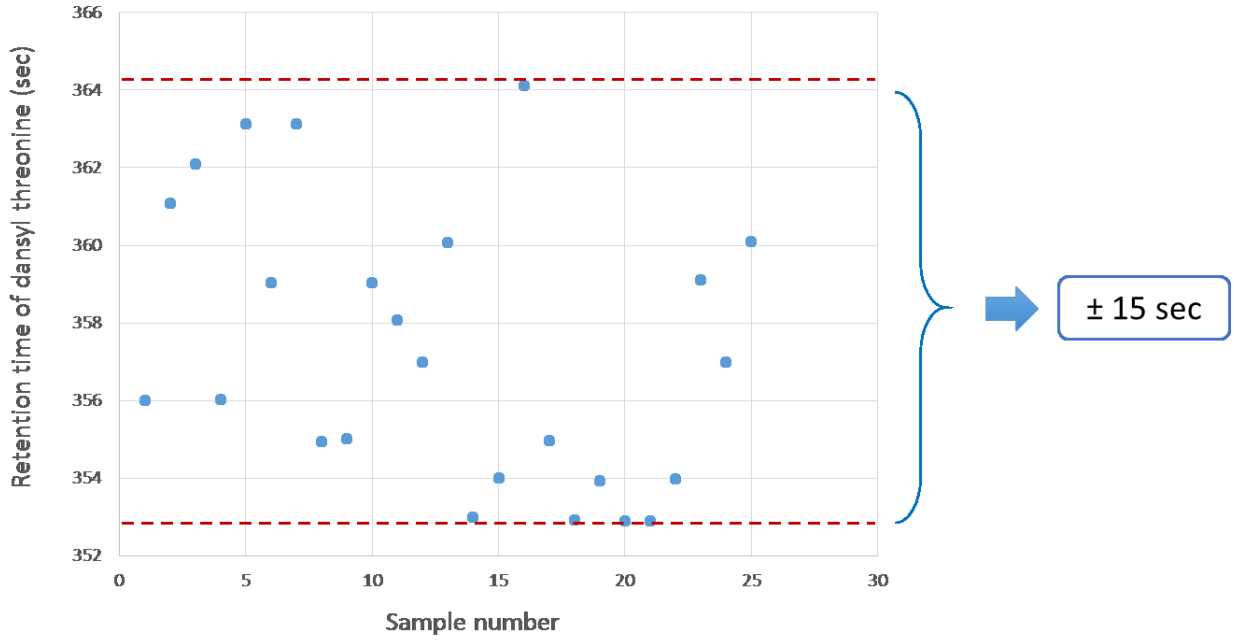


Figure 8.1 Determination of the retention time window using the results from the retention time shift analysis. The retention times of dansyl threonine were extracted from 25 sample data files and are shown in the scattered plot. From the distributions of the retention time at different retention time points, one can determine the retention time tolerance to be used in data alignment and database search.

Chapter 4 discusses the data processing algorithms in CIL LC-MS. A number of topics were involved in the design of the processing pipeline, including isomer compounds, false peak pair, redundant peak pair, and missing ratio data. False metabolite features are encountered often in the metabolomics data in which the noise signal can be selected as a metabolite feature. With the chemical isotope labeling method, the metabolites of interest are labeled selectively in the sample with an enhanced sensitivity in the detection. The peak pair pattern from each labeled metabolite facilitates greatly the feature selection and helps to exclude all the unlabeled signals. For each peak pair detected, we checked also the inter-dependency of the light and heavy peak to remove any falsely paired one. Moreover, the repeated peak pairs are removed by examining the similarities of mass, retention time, and within-sample ratios. The methods are unique to the chemical isotope labeling method and ensure that each peak pair data entry represents a true labeled metabolite.

The ratio matrix in the data table represents the relative concentration of each metabolite in different samples. Although most of the peak pair ratios can be calculated after a thorough inspection of LC-MS raw data, a missing value still can occur in the resultant data table. These missing data could have a great influence on the conclusions drawn from different data analysis methods. In CIL LC-MS, the abundance of one metabolite is presented as a peak pair ratio value calculated from two peak intensities. The mathematical meaning of the ratio value is different from the peak intensity value, making most of the current missing data imputation methods inapplicable in predicting the missing ratio data. To generate a complete metabolite intensity table, we investigated the origin of the missing ratio in CIL LC-MS data and developed a missing value prediction method based on the intensity of light- and heavy-labeled peaks in the LC-MS data. We demonstrated the prediction accuracy by comparing the predicted values

against the experimental values, and the ratio imputation method showed a much improved accuracy than other imputation methods.

Peak pair ratio calculation is a critical step in the CIL LC-MS data processing and determines directly the results of the statistical analyses. The ratio measurement has been shown to be reproducible regardless of the changes in the absolute peak intensity. To reduce the random measurement error, the peak pair ratio calculation was improved by using the peak pair signals in multiple scans, and the average of the ratios was used in the alignment table. Another ratio error can be from the natural isotopologues that are present for any mass peak. These natural occurring peaks often can overlap with the signal of the heavy- labeled peak and introduce a systematic bias in the peak pair ratio calculation. Chapter 5 discusses the intensity of the natural isotopologues from different elements existing in human endogenous metabolites, including carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorous. We reported a data processing method that accounts for natural isotope contributions in ratio calculations for $^{12}C_2$- and $^{13}C_2$-labeled peak pairs. It was shown that this method can improve the measurement accuracy for determining the peak intensity ratio of the light- and heavy- labeled metabolite in metabolomic profiling.

Metabolite identification is another important step in metabolomics. Without metabolite identification, the results of any metabolomic analysis are biologically and chemically uninterpretable. The identification of the metabolite is essential for the explanation of the biological meanings in a metabolomics study. Chapter 6 presents a mass-based database searching algorithm to improve the accuracy of the metabolite identification. Mass accuracy was observed to vary depending on the peak intensities. For mass peaks with relatively low peak intensity, the mass measurement is more likely to give a larger error than peaks with higher

intensity. An intensity-dependent mass tolerance was calculated to facilitate the mass search using a compound database.

The identification of a metabolite usually requires a multi-layer confirmation that involves the accurate mass, tandem mass data, and retention time. A continuous effort is needed to expand the current database by collecting data of different groups of standards. For example, the small peptide compounds have been shown to have an important role in various biological processes,[136-139] and different biological samples can have a specific collection of metabolites. Structure specific and sample specific metabolite databases could be created in the future to increase the efficiency and confidence of the metabolite annotation step further.[140]

Reference

(1) Xia, J.; Broadhurst, D. I.; Wilson, M.; Wishart, D. S., Translational biomarker discovery in clinical metabolomics: an introductory tutorial, *Metabolomics,* **2013**, *9*, 280-299.
(2) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S., HMDB: the human metabolome database, *Nucleic acids research,* **2007**, *35*, D521-D526.
(3) Fiehn, O. In *Functional genomics*; Springer, 2002, pp 155-171.
(4) Hartwell, L. H.; Hopfield, J. J.; Leibler, S.; Murray, A. W., From molecular to modular cell biology, *Nature,* **1999**, *402*, C47.
(5) Dettmer, K.; Aronov, P. A.; Hammock, B. D., Mass spectrometry‐based metabolomics, *Mass spectrometry reviews,* **2007**, *26*, 51-78.
(6) Crick, F., Central dogma of molecular biology, *Nature,* **1970**, *227*, 561.
(7) Schreiber, S. L., Small molecules: the missing link in the central dogma, *Nature chemical biology,* **2005**, *1*, 64.
(8) Naylor, S.; Taylor & Francis, 2003.
(9) Lewis, G. D.; Asnani, A.; Gerszten, R. E., Application of metabolomics to cardiovascular biomarker and pathway discovery, *Journal of the American College of Cardiology,* **2008**, *52*, 117-123.
(10) Lindon, J. C.; Holmes, E.; Bollard, M. E.; Stanley, E. G.; Nicholson, J. K., Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis, *Biomarkers,* **2004**, *9*, 1-31.
(11) Mishur, R. J.; Rea, S. L., Applications of mass spectrometry to metabolomics and metabonomics: Detection of biomarkers of aging and of age‐related diseases, *Mass spectrometry reviews,* **2012**, *31*, 70-95.
(12) Botstein, D.; Risch, N., Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nature genetics,* **2003**, *33*, 228.
(13) Maher, B., Personal genomes: The case of the missing heritability, *Nature News,* **2008**, *456*, 18-21.
(14) Mokdad, A. H.; Marks, J. S.; Stroup, D. F.; Gerberding, J. L., Actual causes of death in the United States, 2000, *Jama,* **2004**, *291*, 1238-1245.
(15) Rappaport, S. M.; Barupal, D. K.; Wishart, D.; Vineis, P.; Scalbert, A., The blood exposome and its role in discovering causes of disease, *Environmental health perspectives,* **2014**, *122*, 769.
(16) Levine, A. J.; Puzio-Kuter, A. M., The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes, *Science,* **2010**, *330*, 1340-1344.
(17) Collins, F. S.; Varmus, H., A new initiative on precision medicine, *New England Journal of Medicine,* **2015**, *372*, 793-795.
(18) Jameson, J. L.; Longo, D. L., Precision medicine—personalized, problematic, and promising, *Obstetrical & Gynecological Survey,* **2015**, *70*, 612-614.

(19) Beger, R. D.; Dunn, W.; Schmidt, M. A.; Gross, S. S.; Kirwan, J. A.; Cascante, M.; Brennan, L.; Wishart, D. S.; Oresic, M.; Hankemeier, T., Metabolomics enables precision medicine:"a white paper, community perspective", *Metabolomics,* **2016**, *12*, 149.
(20) Walther, Z.; Sklar, J., Molecular tumor profiling for prediction of response to anticancer therapies, *The Cancer Journal,* **2011**, *17*, 71-79.
(21) Ward, P. S.; Patel, J.; Wise, D. R.; Abdel-Wahab, O.; Bennett, B. D.; Coller, H. A.; Cross, J. R.; Fantin, V. R.; Hedvat, C. V.; Perl, A. E., The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting α-ketoglutarate to 2-hydroxyglutarate, *Cancer cell,* **2010**, *17*, 225-234.
(22) Dunn, W. B.; Bailey, N. J.; Johnson, H. E., Measuring the metabolome: current analytical technologies, *Analyst,* **2005**, *130*, 606-625.
(23) Robertson, J. G., Mechanistic basis of enzyme-targeted drugs, *Biochemistry,* **2005**, *44*, 5561-5571.
(24) Lu, W.; Bennett, B. D.; Rabinowitz, J. D., Analytical strategies for LC–MS-based targeted metabolomics, *Journal of Chromatography B,* **2008**, *871*, 236-242.
(25) Choi, H.-K.; Choi, Y. H.; Verberne, M.; Lefeber, A. W.; Erkelens, C.; Verpoorte, R., Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique, *Phytochemistry,* **2004**, *65*, 857-864.
(26) Ellis, D. I.; Goodacre, R., Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy, *Analyst,* **2006**, *131*, 875-885.
(27) Ellis, D. I.; Dunn, W. B.; Griffin, J. L.; Allwood, J. W.; Goodacre, R., Metabolic fingerprinting as a diagnostic tool, **2007**.
(28) Wenk, M. R., The emerging field of lipidomics, *Nature reviews Drug discovery,* **2005**, *4*, 594.
(29) Smith, M. T.; de la Rosa, R.; Daniels, S. I., Using exposomics to assess cumulative risks and promote health, *Environmental and molecular mutagenesis,* **2015**, *56*, 715-723.
(30) Guo, K.; Li, L., Differential 12C-/13C-Isotope Dansylation Labeling and Fast Liquid Chromatography/Mass Spectrometry for Absolute and Relative Quantification of the Metabolome, *Analytical chemistry,* **2009**, *81*, 3919-3932.
(31) Guo, K.; Li, L., High-performance isotope labeling for profiling carboxylic acid-containing metabolites in biofluids by mass spectrometry, *Analytical chemistry,* **2010**, *82*, 8789-8793.
(32) Zhou, B.; Xiao, J. F.; Tuli, L.; Ressom, H. W., LC-MS-based metabolomics, *Molecular BioSystems,* **2012**, *8*, 470-481.
(33) Dunn, W. B.; Ellis, D. I., Metabolomics: current analytical platforms and methodologies, *TrAC Trends in Analytical Chemistry,* **2005**, *24*, 285-294.
(34) Wilson, D.; Burlingame, A.; Cronholm, T.; Sjövall, J., Deuterium and carbon-13 tracer studies of ethanol metabolism in the rat by 2H, 1H-decoupled 13C nuclear magnetic resonance, *Biochemical and biophysical research communications,* **1974**, *56*, 828-835.
(35) Wu, J.; An, Y.; Yao, J.; Wang, Y.; Tang, H., An optimised sample preparation method for NMR-based faecal metabonomic analysis, *Analyst,* **2010**, *135*, 1023-1030.

(36) Reo, N. V., NMR-based metabolomics, *Drug and chemical toxicology,* **2002**, *25*, 375-382.
(37) Wishart, D. S., Quantitative metabolomics using NMR, *TrAC Trends in Analytical Chemistry,* **2008**, *27*, 228-237.
(38) Villas‐Bôas, S. G.; Mas, S.; Åkesson, M.; Smedsgaard, J.; Nielsen, J., Mass spectrometry in metabolome analysis, *Mass spectrometry reviews,* **2005**, *24*, 613-646.
(39) Yuan, M.; Breitkopf, S. B.; Yang, X.; Asara, J. M., A positive/negative ion–switching, targeted mass spectrometry–based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue, *Nature protocols,* **2012**, *7*, 872.
(40) Brown, S. C.; Kruppa, G.; Dasseux, J. L., Metabolomics applications of FT‐ICR mass spectrometry, *Mass spectrometry reviews,* **2005**, *24*, 223-231.
(41) Ma, B.; Zhang, Q.; Wang, G.-j.; Ji-Ye, A.; Wu, D.; Liu, Y.; Cao, B.; Liu, L.-s.; Hu, Y.-y.; Wang, Y.-l., GC-TOF/MS-based metabolomic profiling of estrogen deficiency-induced obesity in ovariectomized rats, *Acta Pharmacologica Sinica,* **2011**, *32*, 270.
(42) Adams, R. P.; Sparkman, O. D., Review of Identification of Essential Oil Components by Gas Chromatography/Mass Spectrometry, *Journal of the American Society for Mass Spectrometry,* **2007**, *18*, 803-806.
(43) Wells, R. J., Recent advances in non-silylation derivatization techniques for gas chromatography, *J Chromatogr A,* **1999**, *843*, 1-18.
(44) Finglas, P.; Faulks, R., Critical review of HPLC methods for the determination of thiamin, riboflavin and niacin in food, *Journal of micronutrient analysis,* **1987**.
(45) Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. M., HPLC-MS-based methods for the study of metabonomics, *Journal of Chromatography B,* **2005**, *817*, 67-76.
(46) Wu, Y.; Li, L., Development of Isotope Labeling Liquid Chromatography–Mass Spectrometry for Metabolic Profiling of Bacterial Cells and Its Application for Bacterial Differentiation, *Analytical chemistry,* **2013**, *85*, 5755-5763.
(47) Luo, X.; Zhao, S.; Huan, T.; Sun, D.; Friis, R. M. N.; Schultz, M. C.; Li, L., High-Performance Chemical Isotope Labeling Liquid Chromatography–Mass Spectrometry for Profiling the Metabolomic Reprogramming Elicited by Ammonium Limitation in Yeast, *Journal of Proteome Research,* **2016**, *15*, 1602-1612.
(48) Hooton, K.; Han, W.; Li, L., Comprehensive and Quantitative Profiling of the Human Sweat Submetabolome Using High-Performance Chemical Isotope Labeling LC–MS, *Analytical chemistry,* **2016**, *88*, 7378-7386.
(49) Hooton, K.; Li, L., Nonocclusive Sweat Collection Combined with Chemical Isotope Labeling LC–MS for Human Sweat Metabolomics and Mapping the Sweat Metabolomes at Different Skin Locations, *Analytical chemistry,* **2017**, *89*, 7847-7851.
(50) Mung, D.; Li, L., Development of chemical isotope labeling LC-MS for milk metabolomics: comprehensive and quantitative profiling of the amine/phenol submetabolome, *Analytical chemistry,* **2017**, *89*, 4435-4443.
(51) Mung, D.; Li, L., Applying quantitative metabolomics based on chemical isotope labeling LC-MS for detecting potential milk adulterant in human milk, *Analytica chimica acta,* **2018**, *1001*, 78-85.

(52) Wang, X.; Han, W.; Yang, J.; Westaway, D.; Li, L., Development of chemical isotope labeling LC-MS for tissue metabolomics and its application for brain and liver metabolome profiling in Alzheimer's disease mouse model, *Analytica Chimica Acta,* **2018**.
(53) Sapkota, S.; Huan, T.; Tran, T.; Zheng, J.; Camicioli, R.; Li, L.; Dixon, R. A., Alzheimer's Biomarkers From Multiple Modalities Selectively Discriminate Clinical Status: Relative Importance of Salivary Metabolomics Panels, Genetic, Lifestyle, Cognitive, Functional Health and Demographic Risk Markers, *Frontiers in aging neuroscience,* **2018**, *10*.
(54) Huan, T.; Tran, T.; Zheng, J.; Sapkota, S.; MacDonald, S. W.; Camicioli, R.; Dixon, R. A.; Li, L., Metabolomics Analyses of Saliva Detect Novel Biomarkers of Alzheimer's Disease, *Journal of Alzheimer's Disease,* **2018**, 1-16.
(55) Peng, J.; Guo, K.; Xia, J.; Zhou, J.; Yang, J.; Westaway, D.; Wishart, D. S.; Li, L., Development of isotope labeling liquid chromatography mass spectrometry for mouse urine metabolomics: quantitative metabolomic study of transgenic mice related to Alzheimer's disease, *Journal of proteome research,* **2014**, *13*, 4457-4469.
(56) Han, W.; Sapkota, S.; Camicioli, R.; Dixon, R. A.; Li, L., Profiling novel metabolic biomarkers for Parkinson's disease using in‐depth metabolomic analysis, *Movement Disorders,* **2017**, *32*, 1720-1728.
(57) Huan, T.; Troyer, D. A.; Li, L., Metabolite Analysis and Histology on the Exact Same Tissue: Comprehensive Metabolomic Profiling and Metabolic Classification of Prostate Cancer, *Scientific reports,* **2016**, *6*, 32272.
(58) Chen, D.; Su, X.; Wang, N.; Li, Y.; Yin, H.; Li, L.; Li, L., Chemical Isotope Labeling LC-MS for Monitoring Disease Progression and Treatment in Animal Models: Plasma Metabolomics Study of Osteoarthritis Rat Model, *Scientific reports,* **2017**, *7*, 40543.
(59) Lommen, A., MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, *Analytical chemistry,* **2009**, *81*, 3079-3086.
(60) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M., MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, *BMC bioinformatics,* **2010**, *11*, 395.
(61) Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D., LC‐MS data processing with MAVEN: a metabolomic analysis and visualization engine, *Current protocols in bioinformatics,* **2012**, *37*, 14.11. 11-14.11. 23.
(62) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S., MetaboAnalyst: a web server for metabolomic data analysis and interpretation, *Nucleic acids research,* **2009**, *37*, W652-W660.
(63) Xia, J.; Mandal, R.; Sinelnikov, I. V.; Broadhurst, D.; Wishart, D. S., MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis, *Nucleic acids research,* **2012**, *40*, W127-W133.
(64) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S., MetaboAnalyst 3.0—making metabolomics more meaningful, *Nucleic acids research,* **2015**, *43*, W251-W257.

(65) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G., XCMS Online: a web-based platform to process untargeted metabolomic data, *Analytical chemistry,* **2012**, *84*, 5035-5039.
(66) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, *Analytical chemistry,* **2006**, *78*, 779-787.
(67) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B., Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses, *Analytical chemistry,* **2014**, *86*, 6931-6939.
(68) Zhou, R.; Tseng, C.-L.; Huan, T.; Li, L., IsoMS: Automated Processing of LC-MS Data Generated by a Chemical Isotope Labeling Metabolomics Platform, *Analytical chemistry,* **2014**, *86*, 4675-4679.
(69) Huan, T.; Li, L., Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform, *Analytical chemistry,* **2014**, *87*, 1306-1313.
(70) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S., MyCompoundID: using an evidence-based metabolome library for metabolite identification, *Analytical chemistry,* **2013**, *85*, 3401-3408.
(71) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I., YMDB: the yeast metabolome database, *Nucleic acids research,* **2011**, *40*, D815-D820.
(72) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic acids research,* **2006**, *34*, D668-D672.
(73) Huan, T.; Li, L., Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform, *Analytical chemistry,* **2015**, *87*, 1306-1313.
(74) Huan, T.; Li, L., Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry, *Analytical chemistry,* **2015**, *87*, 7011-7016.
(75) Gepner, P.; Kowalik, M. F. In *Parallel Computing in Electrical Engineering, 2006. PAR ELEC 2006. International Symposium on*; IEEE, 2006, pp 9-13.
(76) Huan, T.; Li, L., Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform, *Analytical chemistry,* **2015**, *87*, 1306-1313.
(77) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap, *Mol. Cell. Proteomics,* **2005**, *4*, 2010-2021.
(78) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A., An introduction to quadrupole‐time‐of‐flight mass spectrometry, *Journal of mass spectrometry,* **2001**, *36*, 849-865.

(79) C., B. S.; Gary, K.; Jean‐Louis, D., Metabolomics applications of FT‐ICR mass spectrometry, *Mass Spectrometry Reviews,* **2005**, *24*, 223-231.
(80) Ledford, E. B.; Rempel, D. L.; Gross, M., Space charge effects in Fourier transform mass spectrometry. II. Mass calibration, *Analytical chemistry,* **1984**, *56*, 2744-2748.
(81) Domon, B.; Aebersold, R., Mass spectrometry and protein analysis, *science,* **2006**, *312*, 212-217.
(82) Holland, R.; Wilkes, J.; Rafii, F.; Sutherland, J.; Persons, C.; Voorhees, K.; Lay Jr, J., Rapid identification of intact whole bacteria based on spectral patterns using matrix‐assisted laser desorption/ionization with time‐of‐flight mass spectrometry, *Rapid Commun Mass Sp,* **1996**, *10*, 1227-1232.
(83) Benton, H. P.; Want, E. J.; Ebbels, T. M., Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data, *Bioinformatics,* **2010**, *26*, 2488-2489.
(84) Andreas, S.; Rudolf, V. E., Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry, *Journal of Mass Spectrometry,* **2003**, *38*, 523-525.
(85) Olsen, J. V.; de Godoy, L. M. F.; Li, G. Q.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap, *Mol. Cell. Proteomics,* **2005**, *4*, 2010-2021.
(86) Huan, T.; Wu, Y.; Tang, C.; Lin, G.; Li, L., DnsID in MyCompoundID for Rapid Identification of Dansylated Amine- and Phenol-Containing Metabolites in LC–MS-Based Metabolomics, *Analytical chemistry,* **2015**, *87*, 9838-9845.
(87) Zhao, S.; Luo, X.; Li, L., Chemical Isotope Labeling LC-MS for High Coverage and Quantitative Profiling of the Hydroxyl Submetabolome in Metabolomics, *Analytical chemistry,* **2016**, *88*, 10617-10623.
(88) Zhao, S.; Dawe, M.; Guo, K.; Li, L., Development of High-Performance Chemical Isotope Labeling LC–MS for Profiling the Carbonyl Submetabolome, *Analytical chemistry,* **2017**, *89*, 6758-6765.
(89) Frenzel, T.; Miller, A.; Engel, K.-H., A methodology for automated comparative analysis of metabolite profiling data, *European Food Research and Technology,* **2003**, *216*, 335-342.
(90) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, *Analytical chemistry,* **2006**, *78*, 779-787.
(91) Wu, Y.; Li, L., Determination of Total Concentration of Chemically Labeled Metabolites as a Means of Metabolome Sample Normalization and Sample Loading Optimization in Mass Spectrometry-Based Metabolomics, *Analytical chemistry,* **2012**, *84*, 10723-10731.
(92) Duran, A. L.; Yang, J.; Wang, L.; Sumner, L. W., Metabolomics spectral formatting, alignment and conversion tools (MSFACTs), *Bioinformatics,* **2003**, *19*, 2283-2293.

(93) Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data, *J Chromatogr A,* **2002**, *961*, 237-244.
(94) Eilers, P. H., Parametric time warping, *Analytical chemistry,* **2004**, *76*, 404-411.
(95) Tayyari, F.; Gowda, G. N.; Gu, H.; Raftery, D., 15N-Cholamine · A Smart Isotope Tag for Combining NMR-and MS-Based Metabolite Profiling, *Analytical chemistry,* **2013**, *85*, 8715-8721.
(96) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B., Missing value estimation methods for DNA microarrays, *Bioinformatics,* **2001**, *17*, 520-525.
(97) Hrydziuszko, O.; Viant, M. R., Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline, *Metabolomics,* **2012**, *8*, S161-S174.
(98) Di Guida, R.; Engel, J.; Allwood, J. W.; Weber, R. J.; Jones, M. R.; Sommer, U.; Viant, M. R.; Dunn, W. B., Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling, *Metabolomics,* **2016**, *12*, 93.
(99) Armitage, E. G.; Godzien, J.; Alonso‐Herranz, V.; López‐Gonzálvez, Á.; Barbas, C., Missing value imputation strategies for metabolomics data, *Electrophoresis,* **2015**, *36*, 3050-3060.
(100) Katajamaa, M.; Orešič, M., Processing methods for differential analysis of LC/MS profile data, *BMC Bioinformatics,* **2005**, *6*, 179.
(101) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L., Optimized Time Alignment Algorithm for LC−MS Data: Correlation Optimized Warping Using Component Detection Algorithm-Selected Mass Chromatograms, *Analytical chemistry,* **2008**, *80*, 7012-7021.
(102) Hrydziuszko, O.; Viant, M. R., Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline, *Metabolomics,* **2012**, *8*, 161-174.
(103) Tayyari, F.; Gowda, G. A. N.; Gu, H.; Raftery, D., 15N-Cholamine—A Smart Isotope Tag for Combining NMR- and MS-Based Metabolite Profiling, *Analytical chemistry,* **2013**, *85*, 8715-8721.
(104) Hao, L.; Zhong, X.; Greer, T.; Ye, H.; Li, L., Relative quantification of amine-containing metabolites using isobaric N,N-dimethyl leucine (DiLeu) reagents via LC-ESI-MS/MS and CE-ESI-MS/MS, *Analyst,* **2015**, *140*, 467-475.
(105) Huang, Y.-Q.; Liu, J.-Q.; Gong, H.; Yang, J.; Li, Y.; Feng, Y.-Q., Use of isotope mass probes for metabolic analysis of the jasmonate biosynthetic pathway, *Analyst,* **2011**, *136*, 1515-1522.
(106) Peng, J.; Li, L., Liquid–liquid extraction combined with differential isotope dimethylaminophenacyl labeling for improved metabolomic profiling of organic acids, *Analytica Chimica Acta,* **2013**, *803*, 97-105.

(107) Wu, Y. M.; Li, L., Determination of Total Concentration of Chemically Labeled Metabolites as a Means of Metabolome Sample Normalization and Sample Loading Optimization in Mass Spectrometry-Based Metabolomics, *Anal. Chem.,* **2012**, *84*, 10723-10731.
(108) Zhou, R.; Tseng, C. L.; Huan, T.; Li, L., IsoMS: Automated Processing of LC-MS Data Generated by a Chemical Isotope Labeling Metabolomics Platform, *Anal. Chem.,* **2014**, *86*, 4675-4679.
(109) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G., et al., HMDB: the human metabolome database, *Nucleic Acids Res.,* **2007**, *35*, D521-D526.
(110) Wadsworth, G. B.; Wadsworth, J. *Introduction to Probability and Random Variables*; McGraw Hill: New York, 1960 p292.
(111) Zhou, R. K.; Li, L., Effects of sample injection amount and time-of-flight mass spectrometric detection dynamic range on metabolome analysis by high-performance chemical isotope labeling LC-MS, *J. Proteomics,* **2015**, *118*, 130-139.
(112) Han, X. L.; Gross, R. W., Quantitative analysis and molecular species fingerprinting of triacylglyceride molecular species directly from lipid extracts of biological samples by electrospray ionization tandem mass spectrometry, *Anal. Biochem.,* **2001**, *295*, 88-100.
(113) Liebisch, G.; Lieser, B.; Rathenberg, J.; Drobnik, W.; Schmitz, G., High-throughput quantification of phosphatidylcholine and sphingomyelin by electrospray ionization tandem mass spectrometry coupled with isotope correction algorithm, *Biochim. Biophys. Acta Mol. Cell Biol. Lipids,* **2004**, *1686*, 108-117.
(114) Han, X. L.; Yang, K.; Gross, R. W., Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses, *Mass Spectrom. Rev.,* **2012**, *31*, 134-178.
(115) Eibl, G.; Bernardo, K.; Koal, T.; Ramsay, S. L.; Weinberger, K. M.; Graber, A., Isotope correction of mass spectrometry profiles, *Rapid Commun. Mass Spectrom.,* **2008**, *22*, 2248-2252.
(116) Moseley, H. N. B., Correcting for the effects of natural abundance in stable isotope resolved metabolomics experiments involving ultra-high resolution mass spectrometry, *BMC Bioinformatics,* **2010**, *11*, 6.
(117) Midani, F. S.; Wynn, M. L.; Schnell, S., The importance of accurately correcting for the natural abundance of stable isotopes, *Anal. Biochem.,* **2017**, *520*, 27-43.
(118) Zhang, P.; Foerster, H.; Tissier, C. P.; Mueller, L.; Paley, S.; Karp, P. D.; Rhee, S. Y., MetaCyc and AraCyc. Metabolic pathway databases for plant research, *Plant physiology,* **2005**, *138*, 27-37.
(119) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G., METLIN: a metabolite mass spectral database, *Therapeutic drug monitoring,* **2005**, *27*, 747-751.
(120) Pence, H. E.; Williams, A.; ACS Publications, 2010.

(121) Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S., LIPID MAPS online tools for lipid research, *Nucleic acids research,* **2007**, *35*, W606-W612.
(122) Huan, T.; Wu, Y.; Tang, C.; Lin, G.; Li, L., DnsID in MyCompoundID for rapid identification of dansylated amine-and phenol-containing metabolites in LC–MS-based metabolomics, *Analytical chemistry,* **2015**, *87*, 9838-9845.
(123) Blom, K. F., Estimating the precision of exact mass measurements on an orthogonal time-of-flight mass spectrometer, *Analytical chemistry,* **2001**, *73*, 715-719.
(124) Mezzano, D.; Leighton, F.; Martinez, C.; Marshall, G.; Cuevas, A.; Castillo, O.; Panes, O.; Munoz, B.; Perez, D.; Mizon, C., Complementary effects of Mediterranean diet and moderate red wine intake on haemostatic cardiovascular risk factors, *European journal of clinical nutrition,* **2001**, *55*, 444.
(125) Simonetti, P.; Gardana, C.; Pietta, P., Plasma levels of caffeic acid and antioxidant status after red wine intake, *Journal of agricultural and food chemistry,* **2001**, *49*, 5964-5968.
(126) Marfella, R.; Cacciapuoti, F.; Siniscalchi, M.; Sasso, F.; Marchese, F.; Cinone, F.; Musacchio, E.; Marfella, M.; Ruggiero, L.; Chiorazzo, G., Effect of moderate red wine intake on cardiac prognosis after recent acute myocardial infarction of subjects with Type 2 diabetes mellitus, *Diabetic Medicine,* **2006**, *23*, 974-981.
(127) Jones, D. P.; Park, Y.; Ziegler, T. R., Nutritional metabolomics: Progress in addressing complexity in diet and health, *Annual review of nutrition,* **2012**, *32*, 183-202.
(128) Amerine, M. A. *The technology of wine making*, 1980.
(129) Arvanitoyannis, I.; Katsota, M.; Psarra, E.; Soufleros, E.; Kallithraka, S., Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics), *Trends in Food Science & Technology,* **1999**, *10*, 321-336.
(130) Bevin, C. J.; Fergusson, A. J.; Perry, W. B.; Janik, L. J.; Cozzolino, D., Development of a rapid "fingerprinting" system for wine authenticity by mid-infrared spectroscopy, *Journal of agricultural and food chemistry,* **2006**, *54*, 9713-9718.
(131) Baxter, M. J.; Crews, H. M.; Dennis, M. J.; Goodall, I.; Anderson, D., The determination of the authenticity of wine from its trace element composition, *Food Chemistry,* **1997**, *60*, 443-450.
(132) Wine-omics, *Nature,* **2008**, *455*, 699.
(133) Roullier-Gall, C.; Witting, M.; Gougeon, R. D.; Schmitt-Kopplin, P., High precision mass measurements for wine metabolomics, *Frontiers in Chemistry,* **2014**, *2*.
(134) Roullier-Gall, C.; Witting, M.; Tziotis, D.; Ruf, A.; Gougeon, R. D.; Schmitt-Kopplin, P., Integrating analytical resolutions in non-targeted wine metabolomics, *Tetrahedron,* **2015**, *71*, 2983-2990.
(135) Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J., MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, *Nucleic acids research,* **2018**.
(136) Adibi, S. A., The oligopeptide transporter (Pept-1) in human intestine: biology and function, *Gastroenterology,* **1997**, *113*, 332-340.

(137) Cheung, H.-S.; Wang, F.-l.; Ondetti, M. A.; Sabo, E. F.; Cushman, D. W., Binding of peptide substrates and inhibitors of angiotensin-converting enzyme. Importance of the COOH-terminal dipeptide sequence, *Journal of Biological Chemistry,* **1980**, *255*, 401-407.
(138) Ding, C.; Yuan, L.-F.; Guo, S.-H.; Lin, H.; Chen, W., Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions, *J Proteomics,* **2012**, *77*, 321-328.
(139) Ohyama, K.; Ogawa, M.; Matsubayashi, Y., Identification of a biologically active, small, secreted peptide in Arabidopsis by in silico gene screening, followed by LC‐MS‐based structure analysis, *The Plant Journal,* **2008**, *55*, 152-160.
(140) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A., Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics, *Analyst,* **2009**, *134*, 1322-1332.