Development and Application of Chemical Isotope Labeling Methods for Metabolomics Data Processing and Liquid Chromatography-Mass Spectrometry-Based Metabolomics

by

Tao Huan

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 the study of chemical processes involving small-molecule metabolites in a given biological system. These small molecules are downstream outputs of cellular activities. An understanding of the metabolome change can help us to visualize perturbations in the genome or proteome from the environmental impacts.

To study metabolomics, our lab has developed a differential chemical isotope labeling (CIL) platform. In this platform, dansylation reaction is used to label the amine/phenol submetabolome with a dansyl tag. The benefit of using this platform includes better LC separation efficiency, MS detection sensitivity, and metabolite quantification accuracy. We can typically detect over 1000 metabolites in human urine samples and over 600 metabolites in human serum samples.

In this thesis, Chapter 2 and Chapter 3 described the development of new data processing programs to better handle the large LC-MS metabolomics dataset from CIL LC MS platform. The two programs, zero-fill and IsoMS-Quant, were aimed to reduce the number of missing values in the LC-MS dataset and to improve the accuracy of the quantitative metabolomics result. Chapters, 4 and 5 focus on the development of metabolite identification methods. As described in Chapter 4, a retention time correction algorithm was combined with a dansyl labeled metabolite standard library, providing a possibility for quick metabolite identification through RT and m/z matching with 278 dns-standards. In Chapter 5, a library of predicted fragment-ion-spectra containing 383,830 possible human metabolites was developed, which allowed the search of experimental MS/MS spectra for metabolite identification. An application of these analytical techniques to biomarker discovery work is included in Chapter 6, specifically. The CIL LC-MS

platform was used in the study of potential biomarkers and diagnostic models for early-stage diagnosis of Alzheimer's disease and mild cognitive impairment.

Overall, these research activities have provided technique improvements and demonstrated the enhanced analytical performance as well as the capability of CIL LC-MS-based metabolomics methods. All these enabled analytical technique further our understanding of metabolomics and its role in system biology.

Preface

A version of Chapter 2 was published as: Tao Huan and Liang Li, "Counting Missing Values in a Metabolite-Intensity Dataset for Measuring the Analytical Performance of a Metabolomics Platform", Anal. Chem, 2015, 87, 1306-1313. I was responsible for the design of experiments, computer programing, data collection and processing, as well as manuscript preparation. L. Li supervised the project and edited the manuscript.

A version of Chapter 3 was published as: Tao Huan and Liang Li, "Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry", Anal. Chem, 2015, 87, 7011-7016. I was responsible for the design of experiments, programing, data collection and processing, as well as manuscript preparation. L. Li supervised the project and edited the manuscript.

Acknowledgements

First and foremost, I would like to thank my Ph.D. supervisor, Professor Liang Li, for his invaluable advice, supervision, guidance and encouragement throughout my graduate studies and the opportunity to contribute to his laboratory. It would be impossible for me to complete this thesis work without his incredible patience and support.

I would also like to express thanks to the members of my supervisory committee, Professor Michael J. Serpe, Professor Todd L. Lowary, Professor Guohui Lin and the external examiner, Professor Daniel Raftery from University of Washington for the time attending my oral examination and reviewing my thesis, and for their comments and suggestions on my research work. I would also like to thank Professor Mark T. McDermott for attending my candidacy examination and providing advice and suggestions on my research projects.

I would also like to thank the people with whom I collaborated, specifically, Professor Guohui Lin at the Department of Computer Science, University of Alberta, Professor Roger Dixon at the Department of Psychology, University of Alberta, Professor Michael Schultz at the Department of Biochemistry, University of Alberta, and Professor Francis CW Chan at the Faculty of Medicine, the Chinese University of Hong Kong.

The members of the Li group have contributed greatly during my five years Ph.D. study at the University of Alberta. I would like to appreciate all the senior group members who helped me during my early time in the group. In particular, I want to thank Dr. Nan Wang, Dr. Andy Luo, Dr. Kevin Guo, for their valuable advice and guidance. I would also like to express my appreciation to all the group members that I had the pleasure to work with: Dr. Avalyn Stanislaus, Dr. Azeret Zuniga, Dr. Jun Peng, Dr. Ruokun Zhou, Dr. Yiman Wu, Dr. Chiao-Li Tseng,

Zhendong Li, Tran Tran, Wei Han, Yunong Li, Xian Luo, Kevin Hooton, Dorothea Mung, Chemqu Tang, Jaspaul Tatlay, Shuang Zhao

In addition, I would like to thank Dr. Randy Whittal, Bela Reiz, and Jing Zheng in the mass spectrometry lab for their maintenance and assistance in running FT-ICR MS and other MS instruments. I would like to thank Gareth Lambkin in the biological service for his training on various biological techniques as well as providing biological fume hood. I must also acknowledge the peoples in the Electronics shop: Allan Chilton and Kim Do. With their expertise in electronic maintenance, we can focus better on the research.

Finally, I wish to express my special thanks to my parents, Mr. Wenlin Huan and Mrs. Guilan Yang. They have spent countless efforts to support me in the past 5 years. Words cannot express my deep appreciation to them for their love, support and confidence in me.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC	Alternating Current
ACN	Acetonitrile
ANOVA	Analysis of Variance
BPC	Base Peak Chromatogram
BSA	Bovine Serum Albumin
CE	Capillary Electrophoresis
CID	Collision-Induced Dissociation
CSF	Cerebrospinal Fluid
CV	Coefficient of Variation
Da	Dalton
DC	Direct Current
DIL	Differential Isotopic Labeling
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
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
PLS-DA	Partial Least Square Discriminant Analysis
ppm	part(s) per million
QC	Quality Control
Q-TOF-MS	Quadrupole Time-Of-Flight Mass Spectrometry
ROC	Receiver Operating Characteristic
RT	Retention Time
RPLC	Reversed Phase Liquid Chromatography
RSD	Relative standard derivation
S/N	Signal to noise ratio
SPE	Solid Phase Extraction
SVM	Support Vector Machine
TOF	Time-Of-Flight
UPLC	Ultra Performance Liquid Chromatography
UV	Ultra-violet
VIP	Variable Importance on the Projection
μM	Micro molarity

Chapter 1

Introduction

1.1 Introduction to metabolomics

Metabolomics is the systematic study of the entire small molecules in a biological system. These small molecules are the chemical products of cellular metabolism processes and include a range of endogenous and exogenous chemical entities, such as peptides, amino acids, nucleic acids, carbohydrates, organic acids, vitamins, polyphenols, alkaloids, minerals.[1] Metabolomics provides a unique scope of studying small biological molecules by allowing the simultaneous assessment of a large amount of chemicals in a biological system. This important feature leads to a growing interest in applying metabolomics to different areas of research. For example, disease states can be reflected by changes in metabolite concentration.[2] In metabolomics research based disease diagnosis, metabolites of significant concentration changes can be detected using the high throughput analytical techniques developed in this field.[3] Also, in clinical and pharmaceutical applications, quantifying and monitoring a large amount of small molecules can help evaluate the effect of drug metabolism and the toxicology. There are many in-depth reviews on the application of metabolomics in various research fields, including disease diagnosis,[4-9] drug discovery,[10-13] environmental assessment,[14, 15] as well as other biological related researches.[2, 16, 17]

From the system biology aspect, metabolomics together with genomics, transcriptomics, and proteomics, built up the “omics” research field.[18] The integrated study of “omics” can offer a better understanding of the entire system biology. Figure 1 shows the central dogma of biology and the omics cascade. Even through extensive biological information can be observed

on the genomics and proteomics side, the information has little correlation with the phenotype. On the other side, metabolites are the end product of the omics cascade and their concentration levels reflect the downstream biochemical response to genetic or environmental changes. Therefore, metabolomics is commonly considered as the linkage between genotypes and phenotypes and a direct signature of biochemical activities. More significantly, apart from genomics and proteomics, metabolomics research has become a new powerful approach to study the system biology.

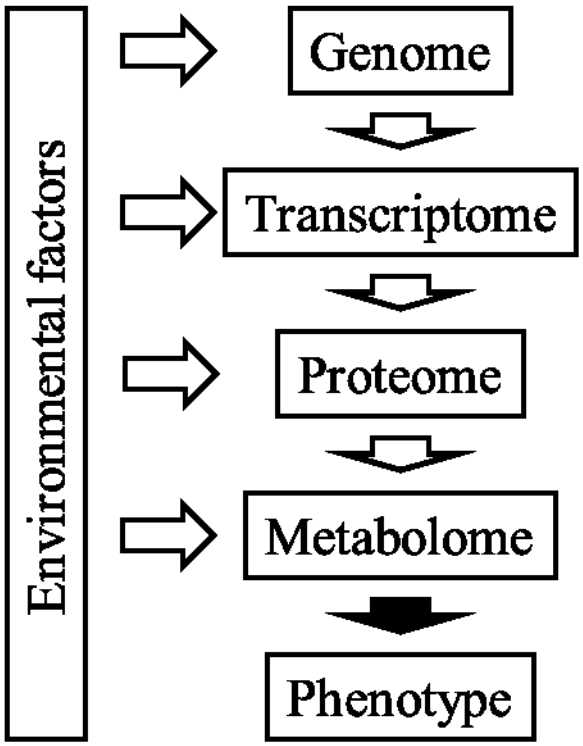


Figure 1.1The central dogma of biology and the omic cascade

Even through the study of metabolites can be traced back to ancient when people evaluated the glucose concentration in urine for diabetics diagnosis,[19] the term "metabolomics" was first brought up by S.G. Oliver and his colleagues [20] in the year of 1998. With the rapid development of recent analytical technologies in small molecule separation,

detection, and identification, the study of a whole metabolome became possible. Numerous research efforts have been made in the past decade to develop new tools and methods for metabolomics research. These efforts include the development of high resolution mass spectrometers[21, 22] and high performance separation system,[23] innovation of automated data processing programs and robust data analysis software,[24-27] as well as the construction of metabolomics database[1, 28, 29].

Metabolomics is still a relatively new research field and numerous technique developments have been published in the past decades. However, there are still some challenges in metabolomics.[30-32] The following context in the introduction chapter is going to specifically describe the achievements from the aspects of the analytical techniques, data processing methods and metabolite identifications in metabolomics. Also, the remaining challenges in each part will be addressed as well.

1.2 Analytical technologies in metabolomics analysis

Metabolomics is a rapidly evolving research field contributing to the development of highly sensitive analytical tools. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are two of the most commonly used technologies, relying on their high throughput and high sensitivity.

NMR is one of the first techniques used for metabolomics.[33] The advantage of NMR based metabolomics includes non-destructive, fast and highly robust as well as informative structural information.[34] However, the detection sensitivity of NMR is not as good as MS and large amounts of sample are required for the analysis.[35]

The fundamental mechanism behind MS instrument is that the *m/z* of charged molecules is measured in the MS analyzer. Even through the detection sensitivity of MS is several orders of magnitude higher than NMR, using MS alone for metabolite detection does not fulfill the high-throughput requirement in metabolomics research. The coupling of a separation technique with MS detection is more commonly applied. There are three most commonly used separation methods that are coupled with MS to achieve high throughput requirement: liquid chromatography (LC),[23, 36] gas chromatography (GC),[37] and capillary electrophoresis (CE).[38]

LC-MS is the most widely used method in metabolomics. The development of LC-MS significantly impacted metabolomics as it can detect metabolites of wide chemical variation yet only requires minimal amounts of sample. In the approach of using LC as the separation tool, the LC column needs to be carefully selected to achieve the best separation performance. RPLC and HILIC columns are two of the most widely used LC column types. In RPLC, the stationary phase is composed of non-polar chemicals covalently bonded onto the silica base microporous particles. Good separation performance can be achieved on moderately polar to non-polar metabolites using C18 based RPLC. In HILIC column, the stationary phase can be either unmodified bare silica or polar chemical bonded phase, such as amide (TSK gel Amide-80), aspartamide (PolyHYDROXYETHYLA), diol (YMC-pack Diol), cross-linked diol (Luna HILIC), cyano (Alltima Cyano), and cyclodextrin (Nucleodex β–OH) groups. Highly polar or ionic metabolites are usually not separated by RPLC due to their weak interaction with the stationary phase. When this problem is encountered, HILIC columns can be used.[39]

A smaller LC column particle size can increase the mass transfer rate and thus benefit the LC separation efficiency. An improved version of LC column, named UPLC column, was

invented back in 2004.[40] UPLC column uses sub-2 μm particles and the UPLC system operates with mobile phases at high linear velocities, causing higher pressures than those used in HPLC. A dramatic increase in resolution, speed and sensitivity can thus be achieved. The application of UPLC-MS for metabolome profiling has shown not only a faster separation speed compare with conventional HPLC-MS methods, but also a more superior metabolome coverage.[41]

The improvement of LC separation can also be reached by using multidimensional LC separation[42]. In a typical 2D LC separation, two chromatographic separations occur in sequence. Usually, selection of first dimensional and second dimensional columns is based on the principle that these two separations need to have different mechanism. In the case when the two separation mechanisms are completely unrelated, an orthogonal 2D separation can be achieved. For example, an orthogonal 2D separation based on strong cation exchange (SCX) and RPLC is widely used in proteomics analysis.[43]

Apart from LC-MS, GC-MS is a good choice for studying thermally stable and volatile metabolites or metabolites that can be volatized through chemical derivatization. GC-MS technique has been widely used to analyze organic acids, amino acids, and other volatile metabolites. There are two major advantages in GC-MS based metabolomics. First is the high resolution benefit from the separation mechanism of a GC column. Second is the consistent MS pattern benefit from the highly reproducible electron ionization (EI) fragmentation mechanism. Also, retention time in the GC-MS can easily be calibrated using the indexed retention time. However, for metabolites that are nonvolatile, labile or cannot be derivatized, it is difficult to detect those with GC-MS.

1.2.1 Current challenges in metabolomics analysis

There are several major analytical challenges in metabolomics studies that needs to be resolved for better understanding of the whole metabolome as well as the application. Most notably, there are three major analytical challenges; the unknown size of the overall metabolome; the wide concentration range, and diverse chemical properties of the metabolites.

The metabolome can be defined as a complete complement of all small molecules found in cell, organ, organism, and biofluids. The documentation of all the metabolites is fundamental for metabolomics research. However, it is still not clear how big the overall size of the metabolome is. The Human Metabolome Project (HMP) was launched in 2004 as part of an effort to identify and quantify all the detectable metabolites in the human body. A freely available electronic database, HMDB,[1] was released, containing records of 2180 endogenous metabolites. So far, HMDB is on its version 3.6 and contains 41,993 metabolite entries including water-soluble, lipid soluble, as well as predicted metabolites. However, the overall size of the entire metabolome is still unknown, causing issues for metabolite identification as well as completely understanding of the metabolic pathways.

The concentrations of metabolites also vary greatly in the biological samples. Take the human urine for example, the most concentrated metabolite, urea, normally has a concentration of over 10 mM, while there are some low abundance metabolites, with concentrations of sub pM. The concentration change from high concentration to low concentration is over 7 orders of magnitude. The analytical challenge over this wide concentration range is that if the instrument is tuned to detect these high concentration metabolites, the low concentration metabolites might not be detected due to ion suppression effect. On the other hand, if the instrument is tuned to be sensitive enough to detect low concentration metabolites, the high concentration metabolites

might over saturate the MS detector. Overall, it is very important to develop an analytical technique that is capable of covering a wide range of metabolome concentrations.

The diversity of chemical properties is another issue associated with metabolomics. Unlike gene, such as DNA, consisting of four building blocks, A, T, C, G; or proteins consisting of 20 amino acids, the building blocks for metabolites include but not limited to: amines, amino acids, carboxylic acids, ketone, and other chemical functional groups. Due to various chemical properties, their detection sensitivity can vary greatly, causing the difficulties of using one instrument to get good signals for all of them.

Focusing on resolving these analytical challenges, researchers have been working on developing new instruments and technologies. More sensitive and high throughput analytical tools have been developed to improve the coverage. For example, our research group has developed a divide-and-conquer approach by chemically isotope labelling metabolites with certain functional groups.[44, 45] Metabolite library has also been developed to facilitate the identification of the chemical isotope labeled metabolites.[46] The details of these approaches will be discussed in the following section.

1.2.2 Chemical Isotope Labeling LC-MS based metabolomics

A CIL LC-MS is a concept that instead of loading the metabolite samples directly into the LC-MS system, a chemical isotope labeling reagent is used to label the metabolites before LC-MS analysis. The original idea of the chemical isotope labeling method comes from the isotope internal standard. In LC-MS based quantitative metabolite analysis, for the purpose of overcoming matrix and ion suppression effects, isotope internal standards are widely used. To quantify the concentration of a certain metabolite, an isotope internal standard of that metabolite with known concentration is spiked into the solution. The concentration of the analyte can then

be calculated based on the MS signal intensity ratio between the analyte and the internal standard. For metabolomics research, a sufficiently high detection sensitivity is always favored but never gets satisfied by any of the currently detection technologies. Regarding the idea of isotope internal standard in metabolomics study, the approach of generating isotope internal standards for all the individual metabolite is expensive and impractical. An alternative approach is to use a chemical reaction to introduce an isotope tag on to the metabolite.

Chemical isotope labeling strategy has been widely used in metabolomics research by targeting at a specific chemical functional group.[47-64] Commonly targeted functional groups are amines, carboxylic acids, ketone, and carbonyls. An ideal chemical isotope labeling reagent should meet four requirements. Firstly, high labeling reaction efficiency is very important in CIL labeling. For example, in dansylation reactions, labeling efficiency of most of the amine/phenol-containing metabolites are over 90%. A high labeling efficiency is the first critical step of capturing low abundance metabolites. Without a good labeling efficiency, these trace metabolites can hardly be labeled by the reagent and thus cannot be MS detected. More importantly, a high and consistent labeling efficiency makes it possible to quantitatively compare the metabolites in different samples after isotope labeling. Secondly, the reaction byproducts during the CIL reaction need to be simple and low in abundance. If a large amount of byproducts are formed during the reaction, these byproducts may over saturate the LC separation, causing poor LC separation peaks and ion suppression in MS detection, resulting in undetected real metabolites. Thirdly, a clear isotope pattern is very important to distinguish real metabolite vs. back ground noise. For example, in the dansylation reaction, the lightly labeled peak and heavily labeled peak have a mass difference of 2.0067 Da. Using high resolution MS for metabolomics detection, this isotope pattern of labeled metabolites can be very distinctive to exclude background MS noise

from real metabolites. Last but not least, a good CIL labeling experiment also needs to be able to improve ESI sensitivity and LC separation. For example, the dansylation reaction procedure I used in my thsis work is a well-known reaction for chemical isotope labeling and has been successfully applied to metabolomics studies of various biological samples, such as urine, cerebrospinal fluid, saliva, and cell lysis. In this case, the derivatization strategy provides 10-1000 fold increase in sensitivity as well as good quantification precision. Also, upon the addition of the chemical tag, the hydrophobicity of the labeled metabolites is significantly improved. The labeled metabolites then spent longer time in the LC column and thus a better LC separation is achieved.

1.3 Metabolomic data analysis

After acquiring the raw LC-MS data, metabolic features need to be extracted for further analysis. The same metabolic feature from different sample analysis results needs to be aligned together to generate a metabolite-intensity table, which contains metabolite intensity in each sample column that is associated with metabolite ID in a row. Statistical analysis can then be performed on the table to find metabolites with significant statistical meanings. These statistically significant metabolites are then validated through internal or external approaches. Validated metabolites could then be used as biomarkers for disease diagnosis or other biological applications.

1.3.1 Metabolic feature extraction

Before data analysis, metabolic features or MS peaks need to be extracted from the raw LC-MS data. Several robust software tools have been developed for processing LC-MS data for

metabolomics, such as IsoMS,[27] XCMS,[24] and mzMine.[25] The key features of a peak picking software involve data conversion, peak detection, noise filtering, and alignment. These software tools have been proven to be very efficient in processing LC-MS based metabolomics data. In a chemical isotope labeling LC-MS metabolomics platform, a true metabolite is detected as a peak pair with a mass difference determined by the mass difference of light- and heavy-chain tags. Because of this unique feature, conventional data processing programs cannot handle CIL LC-MS data format very well. Our group developed a software tool, named IsoMS, for processing the data set generated from CIL LC-MS experiment.[27] Figure 1.2 shows the workflow of IsoMS data processing. In IsoMS, the first step is peak pairing. For each MS peak in a MS spectrum, its charge state and isotope distribution is determined. Then, two ions of same charge state with a user-defined mass difference (e.g., 2.0067 for $^{12}C_2/^{13}C_2$-dansyl metabolites) are then paired to each other. A confident level is also assigned to each peak pair. The second step in the IsoMS involves the filtering of adduct and byproduct peak pairs. Since all the noise peaks show up as singlet peaks, they have already been filtered out in the peak pairing step. However, labeled metabolites with adduct formation from Na^+, K^+, or MH_4^+, dimer and trimer, and in-source fragment ions (e.g., loss of CO_2) can also exist at the same mass spectrum together with the plus H^+ form. Fortunately, these peak pairs can be easily filtered out since they have a fixed mass difference with the MH^+ form. Byproducts generated during the chemical isotope labeling reaction process can also be filtered out by attaching a method blank in the filter background peak files implemented in the IsoMS. After peak pair filtering, only MH^+ peak pair from a metabolite is retained. The next step is the grouping of all the peak pairs belonging to the same metabolite in adjacent spectra according to mass and a user-defined m/z tolerance window. After this step, all the peak pairs belong to isotopic labeled metabolites were extracted from the

raw LC-MS data. A list of peak pair is then generated containing all the peak pairs with retention time, accurate m/z, ^{12}C intensity, $^{12}C_2$-/$^{13}C_2$- intensity ratio for each LC-MS data (attach a figure). For a metabolomics study involving the analysis of multiple samples, IsoMS offers an alignment function, IsoMS-align, to align the same peak pairs found in multiple runs based on accurate m/z, retention time, and intensity. A metabolite-intensity table will be generated after the last step (Figure 1.2). This table is then ready for further analysis.

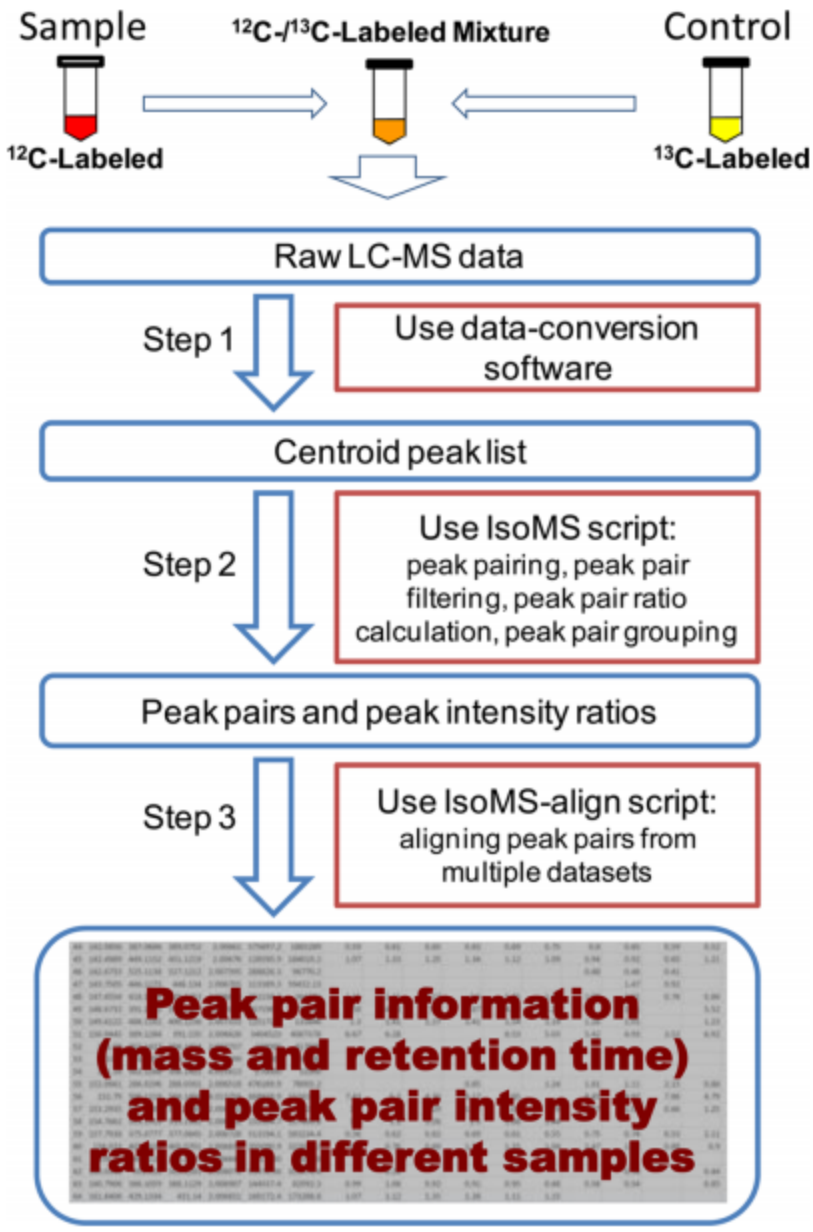


Figure 1.2 Workflow for IsoMS data processing (adapted from IsoMS[27])

To validate the peak pair picking efficiency, a manual inspection was performed using a file of 1388 peak pairs found in a differentially dansyl labeled human urine. In this validation test, only 24 false ones were detected. The false positive rate can thus be calculated as 1.7%,

indicating that a good specificity (<5%) could be obtained using IsoMS. This result indicates the good metabolic feature extraction ability of IsoMS.

1.3.2 Data preprocessing

After aligning all the LC-MS results together and generating the metabolite-intensity table, performing data preprocessing is usually recommended before stepping into the statistical analysis. In the data preprocessing stage, missing data refilling and data normalization are performed.

Missing intensity values is commonly observed in the LC-MS based metabolomics,[26, 65] where low intensity ions might not get detected due to detection sensitivity or ion suppression effect. A typical of 10-40% missing values are observed in a normal LC-MS dataset. The large amount of missing values can reduce the statistical performance. Conventionally, people would replace these missing values with a minimal value based on the hypothesis that these missing values are caused by low MS sensitivity. There are also other approaches available for treating these missing values, such as excluding the features with missing values, replacing the missing value with mean, or estimating the missing value using statistical calculations[66].

Data normalization is important in metabolomics data analysis in order to unbiasedly present metabolic features with equal importance and comparable scale for multivariate statistical analysis. Commonly used data normalization approaches are auto-scaling, pareto scaling, and range scaling[67] (see Table 1.1). Autoscaling is most widely used in preprocessing LC-MS based metabolomics data. In this method, a mean and standard deviation are calculated using all the quantitative values (absolute intensity or relative intensity ration) of a metabolic feature. Then, each quantitative value is subtracted by its mean and then divided by its standard deviation to get a scaled value. The benefit of using autoscaling is that all the quantitative values

unit variance and thus all the different metabolic features are of the same statistical weight. Pareto scaling method is similar to autoscaling. Instead of dividing by the standard deviation, square root of the standard deviation is used in the pareto scaling method. In a range scaling, each quantitative value is divided by the minimum and maximum range of that metabolic feature, after subtraing the mean.

Table 1.1 Different types of data normalization

Type	Algorithm
Autoscaling	$x_i' = \frac{x_i - \bar{x}}{SD}$
Pareto scaling	$x_i' = \frac{x_i - \bar{x}}{\sqrt{SD}}$
Range scaling	$x_i' = \frac{x_i - \bar{x}}{x_{max} - x_{min}}$

1.3.3 Statistical analysis

One of the important goals of metabolomics is to identify the meaningful metabolites that correlate with different biological stages in the dataset. Common statistical tools can be categorized into univariate analysis and multivariate analysis. In an univariate statistical analysis, each individual metabolic feature is statistically analyzed separately. Commonly used univariate statistical analysis tools are T test or ANOVA. These two univariate analyses can find metabolites of significant difference in pair-wise comparison or multiple group comparison. To quickly visualize and identify the metabolic features with significant concentration changes on a figure, volcano plot is more commonly used. In a volcano plot, the x axis is the fold change and

y axis is the p value calculated from ANOVA. Metabolites showing significantly different concentration in a particular group would be located on the top right or top left corner of the volcano plot. Univariate statistical tools are very convenient to identify a single or a list of metabolites that contribute to the biological difference of two metabolomic sample groups.

Benefiting from the high throughput and high sensitivity of LC-MS technique, metabolomics data normally contains thousands of metabolic features. Multivariate analysis allows the simultaneous analysis of multiple metabolic features in one statistical model. Multivariate analysis can be classified into non-supervised and supervised approaches. In the non-supervised approach, all the metabolic features were unbiasedly treated and the whole data set were projected into a two dimensional or three dimension al space. Each principle component (PC) is a linear combination of all these metabolic features. The most commonly used non-supervised method is principle component analysis (PCA)[68]. Non-supervised multivariate analysis is very useful in visualizing the overall separation of different groups of samples.

In supervised multivariate statistical analysis, the information of sample classification is provided for the model construction. Based on the classification information, the model can specifically look for metabolites that contribute the most to the classification. This important feature allows us to find significant metabolites (e.g., biomarkers) that can distinguish the two different biological states (disease vs. normal). Supervised multivariate analysis encompasses many methods, including partial least square (PLS),[69] orthogonal partial least square (OPLS) based discriminant analysis (DA)[70], soft-independent modeling of class analogy (SIMCA),[71] supportive vector machine, and binary logistic regression.

A potential important problem of using supervised multivariate statistical analysis is overfitting.[72] Since we provide the classification information to the model, it is likely that

sometimes the model use random noise to describe this classification if no other meaningful metabolic features are found. To reduce the risk of overfitting, internal validation and external validation are commonly used.[73] Commonly used internal validation methods are cross validation. Cross validation is performed by splitting the entire dataset into two parts, one for model development (training dataset) and another for model validation (validation dataset). The validation perform is used to indicate if the model is overfitting or not. In the case when overfitting happens, the model may separate the groups of samples in the training dataset very well but the separation of sample groups in the validation dataset is not as good. In an external validation experiment, another set of samples are collected independently and LC-MS analysis is performed following the same experimental protocol. Important metabolic features identified from the multivariate analysis were applied to the new set of LC-MS data, externally validating these features if the same kind of separation can be repeated in the external validation dataset.

1.4 Metabolite identification in MS-base metabolomics

The identification of unknown compounds is a main bottleneck in metabolomics. There are several commonly used compound database: KEGG Compound,[74] PubChem,[75] and Chemspider.[76] The use of accurate mass along with searching against metabolite library can results in many possible matches and it is difficult to tell the correct structure from other possible candidate merely based on the mass. However, if high resolution and high accuracy of mass and isotope ratio measurements are available, a correct elemental composition may be obtained and thus it is still possible to narrow down the candidate list. Seven Golden Rules developed by Tobias *et. al.*[77] summarized the rules that are required to generate a chemical formula based on the molecular mass obtained using high resolution MS. Most commonly, metabolite

identification is performed through the interpretation of structural information of a metabolite, which is generated using a tandem mass spectrometer.

Tandem MS/MS experiment is performed by isolating molecular ions with its particular precursor mass and then breaking down that ion into pieces with the help of collision gas. Mass signals of the fragments from that molecule provide unique structural information. This structural information is widely used for metabolite identification. The MS/MS spectrum generated from a tandem mass spectrometer experiment can be either interpreted manually by a MS expert to elucidate its structure or search against a standard MS/MS spectra library. The approach of manual interpretation is time-consuming and heavily relies on the experience of the MS/MS analyst. In the approach of searching against standard MS/MS spectra library, the spectra library can be either experimental or theoretical.

Currently, there are many experimental MS/MS library available for metabolite identification, such as NIST MS/MS library,[78] Metlin,[28], MassBank,[79] and HMDB.[1] In an experimental MS/MS standard spectra library, all the MS/MS spectra are acquired through the MS/MS experiments of metabolic standards. Instrument types and vendor information are also associated with these MS/MS data and user can choose the sub-database that has similar MS/MS experimental conditions for metabolite identification. The matching of experimental MS/MS spectrum with a standard MS/MS spectrum in a standard MS/MS spectra library is highly reliable for metabolite identification. Since the number of metabolite standards that can be acquired is limited, a major challenging in developing the experimental MS/MS standard spectra library is to expand the number of metabolites.

Since the coverage of experimental MS/MS standard libraries is far from covering the entire metabolome space, theoretical MS/MS library has been developed in the past decades. In

the theoretical MS/MS library, all the MS/MS spectra are generated based on prediction of fragmentation pattern according to a set of fragmentation rules. There are several approaches for generating predicted MS/MS spectra (more precisely a list of fragment ions with unit intensity), depending on the chemical bond breakage rules used and the number or level of fragment ions included in a predicted fragment ion spectrum.[80-91] There are also some commercial products (e.g., Mass Frontier from Thermo Scientific, Waltham, US and ACD/MS Fragmenter from Advanced Chemistry Labs, Toronto, Canada) as well as published tools (e.g., Metfrag[82], Fragment Identificator or FiD[81] and MIDAS[88]) for generating predicted MS/MS spectra with varying degrees of success.

1.5 Overview of thesis

Based on my research objective, this thesis can be divided into three parts. The first part (Chapter 2 and 3) describes the development of metabolomics data processing programs to better extract useful metabolomic information. The second part (Chapter 4 and 5) focuses on the development of metabolic libraries for metabolite identification. The third part (Chapter 6) discusses the applications of metabolic biomarker discovery work using our CIL LC-MS approach.

Specifically, Chapter 2 describes an R language based program, allowing the users to retrieve the missing values from raw LC-MS data. Our study shows that through filling in the low intensity metabolic information, this program can significantly improve the statistical performance. Chapter 3 describes another R language based program, aiming at better quantification accuracy in CIL LC-MS experiments. By re-constructing the 12C/13C peak patio

using the chromatographic area information, a more precise intensity ratio can be obtained. Specifically, in Chapter 4, a Dns-metabolite standards library, DnsID is developed. This library has the accurate mass, retention time and MS/MS spectra information for 275 common human endogenous metabolites after the dansylation labeling. A retention time correction function was embedded to correct systematic RT shifts of LC-MS setups in different labs. This library allows the automatic metabolite identification for dansylation based CIL LC-MS dataset. In Chapter 5, an in-silico fragmentation algorithm is developed to generate predicted MS/MS spectra for 8000 known human endogenous metabolites and 375,000 predicted human metabolites. This predicted MS/MS library allows for the metabolite identification with no experimental MS/MS spectra available. In Chapter 6 we applied the dansylation based CIL LC-MS to study saliva metabolic changes associated with Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). Potential biomarkers that distinguish these two disease stages from normal aging (NA) were discovered and identified. Metabolic diagnostic models were also developed for the purpose of clinical diagnosis. Overall, the thesis shows the development and application of analytical techniques to study metabolomics. Finally, Chapter 7 provides a conclusion of the thesis as well as a brief discussion on the future research directions.

Chapter 2

Counting Missing Values in a Metabolite-Intensity Dataset for Measuring the Analytical Performance of a Metabolomics Platform

2.1 Introduction

Missing intensity-values is common in a multiple-sample dataset generated by an "omics" analytical tool for genomics, proteomics and metabolomics applications.[92-98] One of the major roles of an omics study is to find genes, proteins or metabolites that have significant differences in different biological groups such as healthy vs. diseased samples. Analytical tools are used to generate a rectangular matrix or table containing an intensity (or quantity) value in each sample column that is associated with an individual gene, protein or metabolite in a row. Missing values in the table can cause problem in performing statistical calculation.[99] Genomics and proteomics researchers have devoted a considerable amount of efforts to understand and develop appropriate methods to handle the missing data.[92-95, 100-105] There is an increasing awareness of this problem in the field of metabolomics and several papers have been published on this topic,[96-98, 106-114] including the development of statistical tools to fill the missing values or simply disregard all the features with missing data. However, filling the missing values non-experimentally needs to be carefully performed.[97, 98, 114] There are debates on whether missing values should be filled and, if so, how best the missing values are filled (e.g., should we use the lowest intensity or a mean of all the measured values in a dataset to fill the missing values?).[110-112]

We echo the view of a growing number of researchers on the importance of dealing with missing values properly in metabolomics. In our view, the best approach to tackle the problem is

from the experimental side, i.e., developing and applying robust analytical tools to profile the metabolomes of many samples with the least number of missing values. In an ideal situation, there should be very few missing values if a metabolomic technique is capable of detecting and quantifying all the metabolites; missing values would indicate the true absence of the metabolites for biological reasons. However, due to technical limitations of current analytical methods, the extent of missing values can be large, even in replicate dataset of the same sample where metabolite concentrations should be the same. In the case of LC-MS based metabolomics research, low-concentration or not-easily-ionizable metabolites may not be detected due to detection sensitivity issue or ion suppression effect.[115] In addition, data processing including peak picking may cause the loss of peak intensity information.[110, 114, 116] There are several metrics including detection sensitivity, technical precision, quantification accuracy and the number of detectable metabolites that have been routinely used to measure the analytical performance of a metabolome profiling technique.[115, 117-119] We feel that the extent of missing values should be considered as another important parameter to gauge the performance of a method. In other words, the number of missing values should be reported, like the number of metabolites profiled, as a criterion to gauge the quality of a dataset.

In this work, we report an investigation of the issue of missing values in a chemical isotope labeling (CIL) LC-MS metabolomics platform. In high-performance CIL LC-MS, the isotope labeling reagents are rationally designed to improve both LC separation efficiency and MS detection sensitivity significantly.[45, 57, 58, 60-62, 120-122] For example, dansylation, targeting the amine/phenol submetabolome, allows the detection of labeled metabolites with a sensitivity improvement of 10 to 1000-fold over the un-labeled counterparts.[120] With the ability of detecting thousands of putative metabolites from an individual sample (e.g., human

urine) by using this platform, an important question rises on how well we can profile them consistently in multiple samples, as metabolomics requires analyzing many samples of usually the same type, not just one sample. To this end, we have developed a data processing workflow that explores a unique feature of peak-pair picking from mass spectra generated by differential CIL LC-MS in order to fill the missing values in a multiple-sample dataset. This method allows a significant reduction of missing values, enabling determination of a greater number of significant metabolites that separates different groups of samples, a common goal of many metabolomics studies in disease biomarker discovery and systems biology. To facilitate method comparison in terms of missing values, we propose a standardized approach of counting missing values in replicate dataset as a way of gauging the extent of missing values for a given analytical method.

2.2 Experimental Section

2.2.1 Dansylation Labeling

^{12}C-dansyl chloride for metabolite labeling was purchased from Sigma-Aldrich Canada (Markham, ON, Canada). ^{13}C-dansyl chloride was synthesized in our lab.[120] The labeling reaction was performed according to a protocol reported previously.[121]

2.2.2 LC-MS

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 series binary HPLC system (Palo Alto, CA, USA). A reversed-phase 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 ACN,

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 sample injection volume was 2 µL.

2.2.3 Zero-fill Program

The LC-MS data generated were first processed using a peak-pair picking software, IsoMS.[123] The level 1 peak pairs[123] were aligned from multiple runs by retention time match within 30 s and accurate mass match within 5 ppm to produce a CSV file or table. The zero-fill program was then used to fill the missing values in the CSV file. This program was written in R and is freely available from www.mycompoundid.org.

In zero-fill, finding the missing value of a peak pair in the raw data of a sample uses information of retention time (rt), m/z value (mz) and absolute intensity (int) of the ^{13}C-peak of the pair. The ^{13}C-peak is from a controlled sample (e.g., a ^{13}C-labeled pooled sample or ^{13}C-labeled universal-metabolome-standard) that is spiked into all the ^{12}C-labeled individual samples. Thus, the absolute intensity of this peak for a given labeled metabolite should be theoretically the same in all the samples. A matching score is used to find the peak pair based on similarities of these three parameters. It is defined as

$$\text{Score} = \left(1 - \frac{\text{rt. diff}}{\text{rt. tol}}\right)/4 + \left(1 - \frac{\text{mz. diff}}{\text{mz. tol}}\right)/2 + (1 - 2 \times \text{int. diff})/4$$

where

$$\text{rt. diff} = \text{abs}(\text{rt.}\,{}^{13}C\text{. peak} - \text{rt. rawdata. peak})$$

$$\text{mz. diff. light} = 1E6 \times \frac{\text{abs}\left(\text{mz.}\ ^{12}C\text{. peak} - \text{mz. rawdata. peak} + 2.0067\right)}{\text{mz.}\ ^{13}C\text{. peak}}$$

$$\text{mz. diff. heavy} = 1E6 \times \frac{\text{abs}\left(\text{mz.}\ ^{13}C\text{. peak} - \text{mz. rawdata. peak}\right)}{\text{mz.}\ ^{13}C\text{. peak}}$$

$$\text{int. diff} = \text{abs}\left(\log\left(\frac{\text{int.}\ ^{13}C\text{. peak}}{\text{int. rawdata. peak}}\right)\right)$$

The default rt tolerance (tol) is 30 s and the default mz tolerance is 5 ppm. A different weight (divided by 2 or 4) is assigned to each of the similarity equations in the score function; mz is deemed to be more important than rt and int and therefore given more weight. If the matching score is larger than 0.6, it will be considered as a match. This scoring algorithm was developed using several metabolomic datasets where missing values in metabolite-intensity tables had been manually picked from the raw data.

2.2.4 Statistical Analysis

Multivariate statistical analysis was carried out using SIMCA-P+ 12 (Umetrics AB, Umea, Sweden). Volcano plot was plotted using Origin 8.5.

2.3 Results and Discussion

2.3.1 IsoMS and Missing Values

Figure 2.1 shows the workflow for processing isotope labeled LC-MS data. IsoMS is used to perform peak picking, peak pairing, peak-pair filtering and peak intensity ratio calculation.[123] Using IsoMS-align script, information on the peak pair IDs and their peak ratio values from multiple LC-MS runs is extracted to produce a CSV file. This metabolite-intensity

data file can be opened as an Excel table for data inspection or further statistical analysis. In picking the peak pairs, IsoMS classifies the peak pairs into three groups, namely level 1, 2 or 3.[123] Level 1 peak pairs are the most confident pairs where the ^{13}C-natural-isotope peaks are accompanied with the light- and heavy-chain labeled metabolite within a pair. Level 2 peak pairs miss one of the ^{13}C-natural-isotope peaks. Level 3 peak pairs are the least confident pairs with both ^{13}C-natural-isotope peaks missing. To reduce the extent of false positive peak pairs found by IsoMS, only level 1 peak pairs are retained in the metabolite-intensity table. In doing so, the false positive rate (FPR) is usually less than 5%.

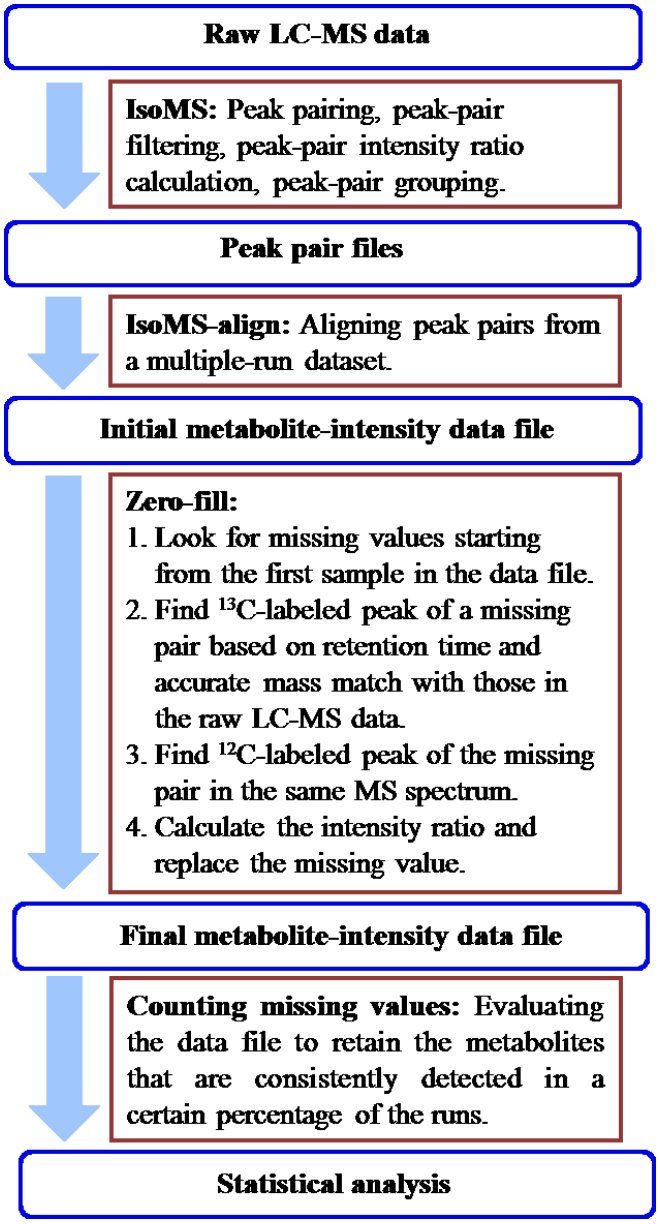
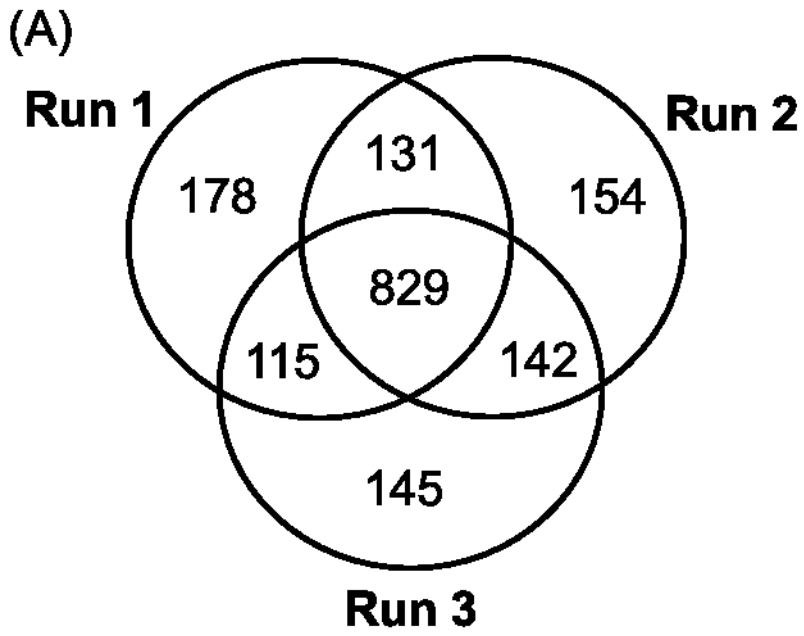


Figure 2.1 Workflow for processing CIL LC-MS data that incorporates the zero-fill program

Inspecting the metabolite-intensity table generated by IsoMS, it is apparent that there are many missing values in the table from a multiple sample dataset, even in replicate runs of the same sample. As an example, Figure 2.2A shows a distribution of the number of peak pairs found in ^{12}C-/^{13}C-dansyl labeled human urine samples (i.e., experimental triplicate runs of the same urine). Among the 1549 peak pairs found in run 1 and run 2, 960 pairs or 62% are in

common. Comparing run 1 and run 3, 944 out of 1540 pairs (61%) are in common. There are 971 common pairs out of 1516 pairs (64%) found in run 2 and run 3. As the sample number increases, the number of commonly detected metabolites decreases (see below). In metabolomics work, it is common to use a criterion such as 50%-rule to retain the metabolites with missing intensity values in no more than 50% of the samples for statistical analysis. Currently there is no consensus on what this percentage limit should be.[97, 98, 114]



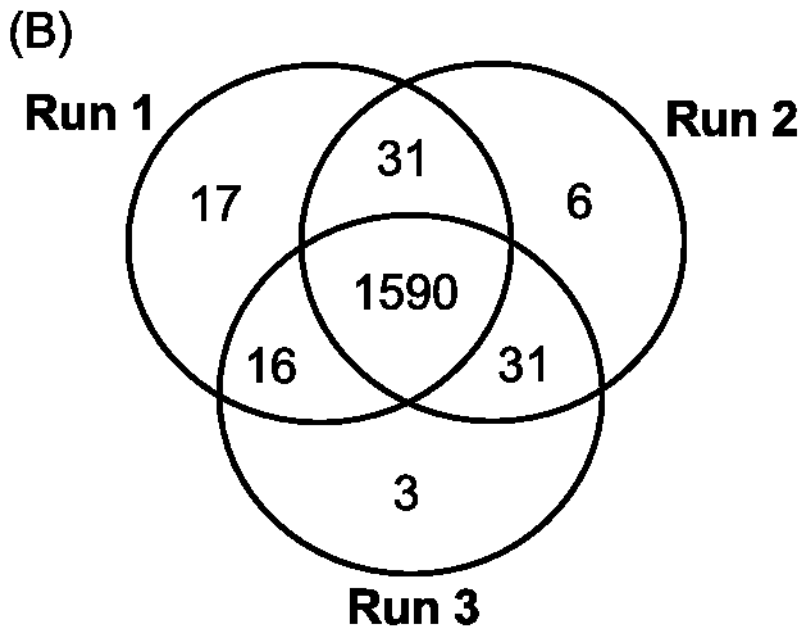


Figure 2.2 Venn diagrams of the number of peak pairs detected from experimental triplicate analysis of 13C-/12C-dansyl labeled human urine samples: (A) without zero-fill and (B) with zero-fill.

Missing values in replicate runs are mainly caused by technical and data processing limitations. To reduce the number of missing values, measurement should be done using a technique that gives very high reproducibility. However, even for a very reproducible technique, data processing can still be the limiting factor. In processing LC-MS data (with or without CIL), because of the need to balance the sensitivity and specificity in peak picking and intensity measurement, some low-abundance peaks or other peaks not meeting a set of criteria in the peak picking algorithm are missing in the metabolite-intensity table. Re-analyzing the original LC-MS data may help filling the missing values in the table. This can be done manually by inspecting the original spectrum or chromatogram. Because this is a time-consuming process, manual filling of missing values is best done for selected metabolites that have already been found to be

significant in statistical analysis of the initial metabolite-intensity data file. However, this approach will not alter the initial metabolite-intensity table used to perform statistical analysis for finding the significant metabolites in the first place. Alternatively, an algorithm may be developed to automate the re-analysis process to detect and fill the missing values (i.e., zero-fill). However, this is not easy to implement due to the fact that it is often difficult to differentiate the metabolite peaks from the background peaks when the signal intensity is very low, even with a high resolution instrument. Solvents, impurities, salts, etc., and their multi-mers and clusters can give arise many peaks at the low mass region (m/z<300) where metabolite ions are detected.

2.3.2 Zero-fill Program

CIL LC-MS offers an opportunity to overcome the difficulty of implementing an automated zero-fill process. In CIL LC-MS, the metabolite ion mass is shifted to a higher mass (m/z>300) by adding the labeling group to a metabolite (e.g., the mass of dansyl group is 234.0583 Da). This reduces the extent of background interference. More importantly, all the metabolite peaks in differential CIL LC-MS are detected in pairs and thus can be distinguished from the singlet background peaks. In addition, a ^{13}C-labeled control sample is spiked to all ^{12}C-labeled individual samples. As a consequence, the absolute intensity of the ^{13}C-peak of a metabolite peak-pair should be similar for all the samples, providing another differentiation parameter. We have developed a zero-fill program to re-analyze the CIL MS data after the initial generation of the metabolite-intensity data file by IsoMS.

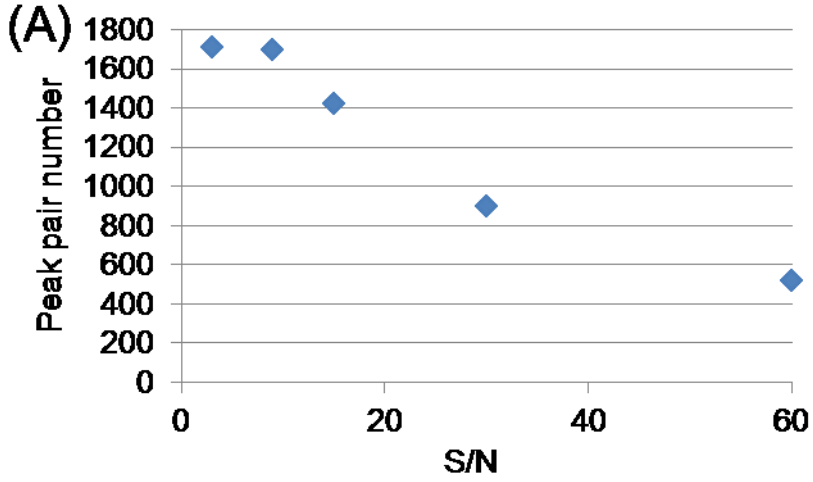
As Figure 2.1 shows, the zero-fill program first reads the metabolite-intensity data file, and then looks for missing values starting from the first sample run. Once a missing value is found, it goes back to the raw peak list file of the LC-MS run. Based on the matching of retention time, m/z and ^{13}C-labeled-peak intensity (see Experimental Section) of the missing-

value peak pair with those in the raw data file, the program finds the correct ^{13}C-peak. In case that the ^{13}C-peak is not available in the raw data, the program stops the search for the ^{12}C-labeled peak of the peak pair to avoid generating any false positive result. If the ^{13}C-peak exists in the data, the program would go ahead to search for the ^{12}C-peak also based on retention time, m/z and intensity, as well as that the ^{12}C-peak must exist in the same mass spectrum as the ^{13}C-peak. Once both peaks are picked, the zero-fill program calculates the peak intensity ratio. This ratio is entered into the metabolite-intensity data file to replace the missing value. To distinguish the ratios determined by IsoMS and the zero-filled ratios, 2 decimal places are kept for the ratios from IsoMS, while 8 decimal places are kept for the ratios from the zero-fill program. This helps glance at the table to obtain a visual impression of the extent of zero-fill.

2.3.3 Performance of Zero-fill

We have systematically evaluated the performance of the zero-fill program with an objective of extracting a maximal number of peak pairs within an acceptable level of FPR (i.e., <5%) from a multiple-run dataset. In the workflow shown in Figure 2.1, IsoMS is first used to process the dataset using a chosen intensity or S/N threshold for extracting the peak pairs. The value of this threshold has a large effect on the number of peak pairs picked by this program. Figure 3 shows the total number of level 1 peak pairs, FPR, and the number of missing values as a function of threshold value for peak-pair picking. These results were obtained from an experimental triplicate dataset of dansyl labeled human urine. Figure 2.3A shows an overall decrease in the peak-pair number as the S/N threshold increases. The FPR level (see Figure 2.3B, without zero-fill) does not change significantly except that it is lower at the threshold of S/N 60 from which only the very high abundance peaks are picked. These results indicate that IsoMS is able to pick the level 1 peak pairs with FPR of <4% even at a very low threshold (S/N 3).

However, the numbers of peak pairs detected using S/N 3 and 9 thresholds are similar, suggesting that lowering the threshold from 9 to 3 cannot increase the peak pair number anymore. Manual inspection of the results indicates that many of the peak pairs with S/N<9 are not belonging to the level 1 group. The plot in Figure 2.3C (without zero-fill) shows that the average number of missing values in each run decreases as the threshold increases. This is consistent with the notion that the high abundance peaks are more reproducible. Considering that the performance of using S/N 9 is similar to that of S/N 3 and IsoMS data processing is faster with S/N 9 (i.e., 5 min per run using S/N 9 vs. 20 min per run using S/N 3), we choose a threshold of S/N 9 to carry out the IsoMS data processing to generate the initial metabolite-intensity data file.



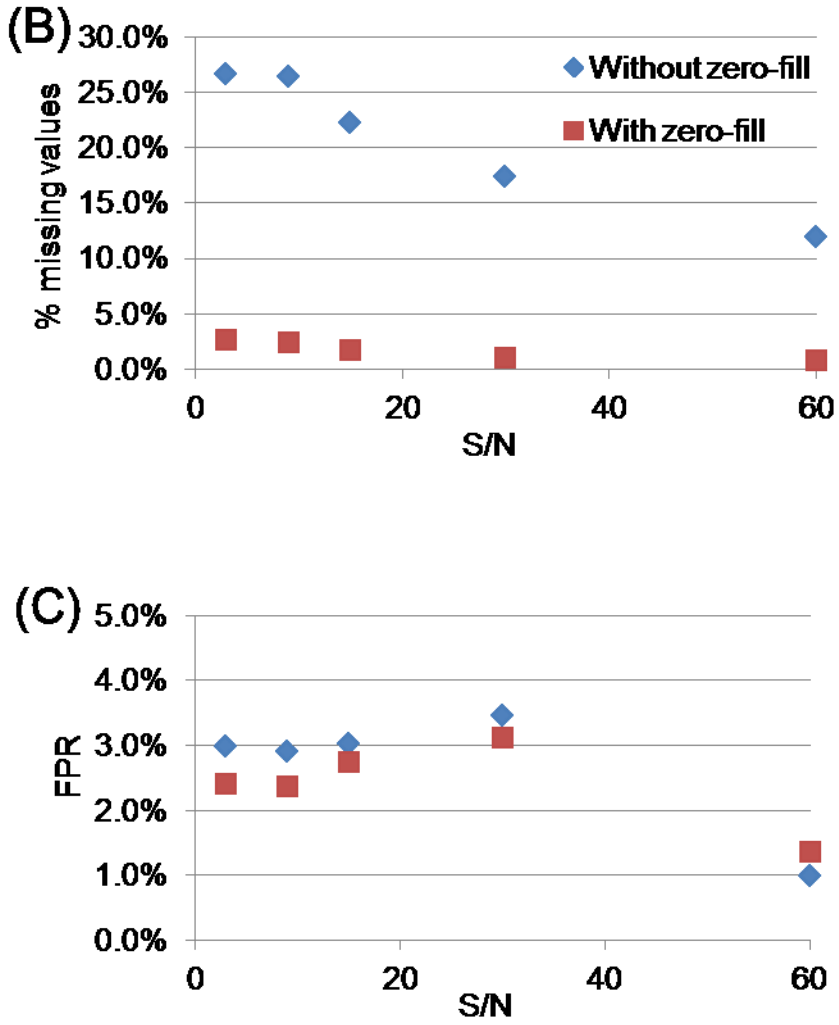
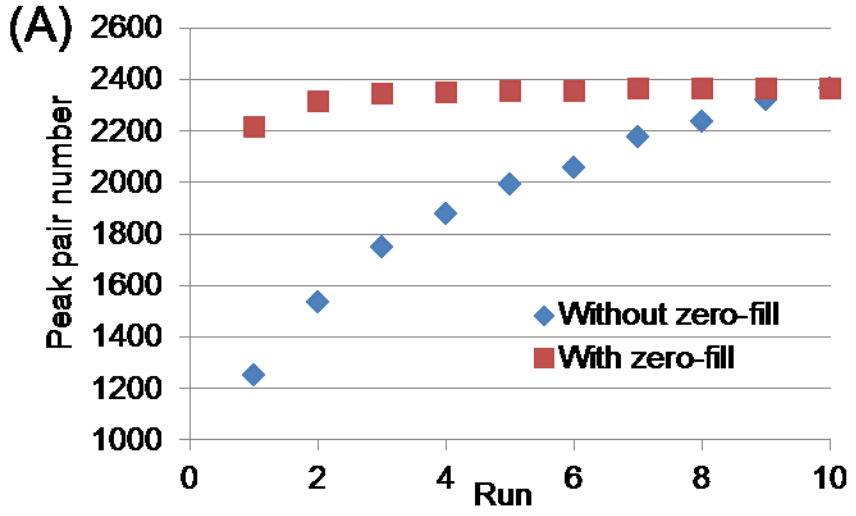


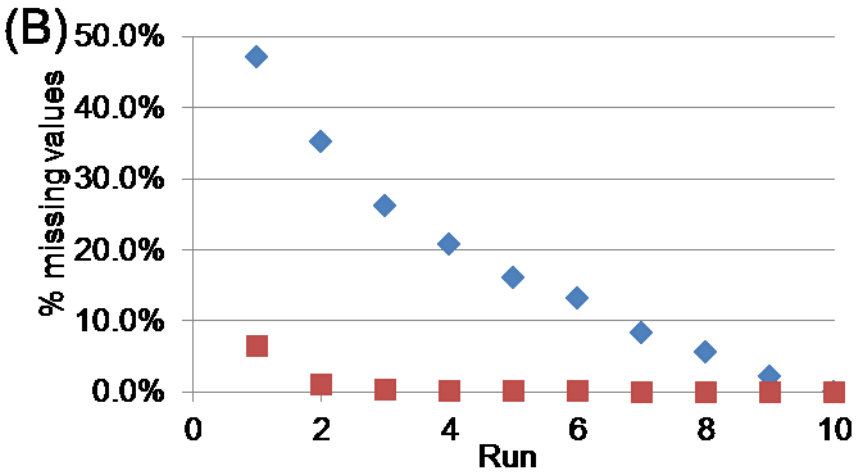
Figure 2.3 (A) Number of peak pairs detected, (B) percentage of missing values and (C) false positive rate (FPR) as a function of S/N used for IsoMS data processing of the experimental triplicate dataset of labeled urine.

Applying the zero-fill program to re-analyze the triplicate dataset, the average percentage of missing values drops dramatically from 26.4% to 2.5%. This can be more clearly seen in Figure 2.2B where the distribution of the number of peak pairs found in the three runs is shown. The common peak pairs found in the three runs increases from 829 (48.9%) to 1590 (93.9%). The average run-to-run reproducibility is 98%, compared to 67% ± 1% without using zero-fill

(see Figure 2.2A). Many of the retrieved values can be manually confirmed by inspecting the peak pairs in the raw mass spectra. In fact, with zero-fill, the FDR drops from 2.9% to 2.4% (see Figure 2.3B). Thus, the zero-fill program can retrieve missing values from the raw data very effectively.

We have studied the performance of zero-fill in a dataset containing 10 replicate injections of the same dansyl urine sample. Figure 2.4A shows the number of peak pairs detected with and without zero-fill as a function of cumulative injection number. Without zero-fill, the cumulative number of peak pairs increases gradually and then reaches a near-plateau after 9 injections. The number of missing values also gradually reduces as more replicate data are included in the combined runs (Figure 2.4B). However, with zero-fill, both the total number of peak pairs detected and the number of missing values reach the plateau much faster. In fact, the results of duplicate injections with zero-fill are similar to those of 9 or 10 injections without zero-fill. Even using one injection, 2217 peak pairs can be detected, compared to 2368 peak pairs from the combined results of duplicate injections. As Figure 2.4C shows, with zero-fill, the FPR decreases as more replicate data are included, while without zero-fill, the FPR increases.


(A)
Peak pair number
2600
2400
2200
2000
1800
1600
1400
1200
1000
0
2
4
6
8
10
Run
Without zero-fill
With zero-fill


(B)
% missing values
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
2
4
6
8
10
Run

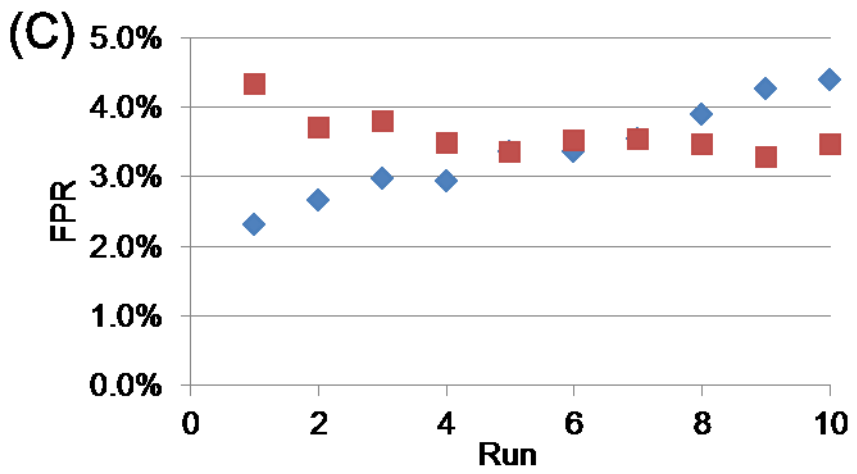
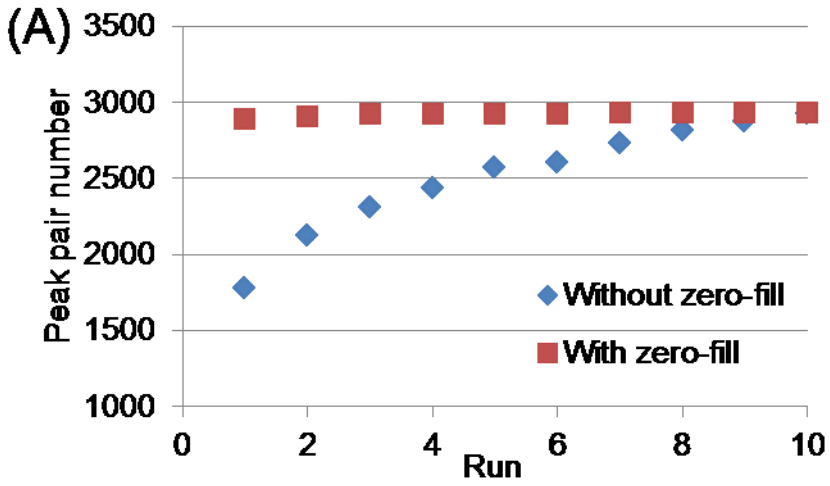


Figure 2.4 (A) Number of peak pairs detected, (B) percentage of missing values and (C) false positive rate (FPR) as a function of S/N used for IsoMS data processing of the 10-run replicate injection dataset of labeled urine.

We have also analyzed the performance of zero-fill on a dataset of 30 LC-MS runs from experimental triplicate of dansyl labeled samples with 10 injections for each sample. The results of experimental replicates measure the overall experimental variations, not just instrumental variation which is gauged by repeat injections of the same sample. Figure 2.5 shows the plots where the y-axis represents the injection number. The combined results of three replicate samples from each injection are used. For example, for injection 1, the total number of peak pairs detected in the three samples with the first injection is used (three LC-MS runs). For injection 2, the combined total number of peak pairs detected in the three samples with the first and second injections is plotted (6 LC-MS runs). As Figure 2.5 shows, the trends of changes in the number of peak pairs, missing values and FPR are similar to the injection replicate dataset shown in Figure 2.4. However, in the experimental triplicate results, even after 10 replicate runs for each sample, there are still about 15% missing values (~450 peak pairs) if zero-fill is not performed

(see Figure 5B). These are the peak pairs with variations caused by the sample handing process. For example, some low abundance metabolites might be labeled with slightly different efficiencies in the triplicate samples, which can result in signal intensity reduction in one of the ^{13}C-natural-isotope peaks to a level that the peak pair is no longer belonging to the level 1 group. In contrast, with zero-fill, the percentage of missing values drops much faster and reaches almost zero after two injections of each sample. Even with one injection, most of the peak pairs from the combined results are detected (See Figure 2.5B).



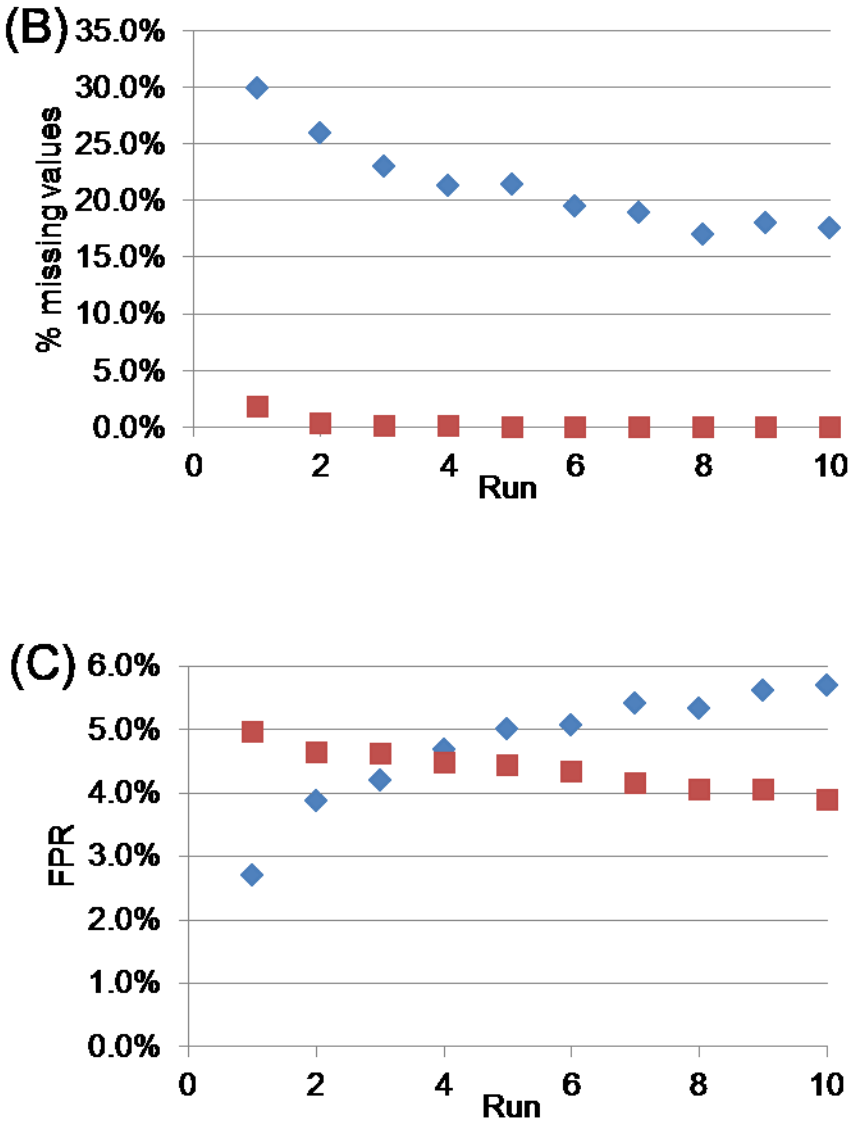


Figure 2.5 (A) Number of peak pairs detected, (B) percentage of missing values and (C) false positive rate (FPR) as a function of S/N used for IsoMS data processing of the 30-run dataset of labeled urine.

The above results indicate that with zero-fill the number of peak pairs detected in each run can reach a near-maximal number even without performing replicate runs for each sample. However, the maximal number of peak pairs detectable within a dataset is dependent on the

number of runs present in the dataset. Comparing the total maximal number of peak pairs detected in the 3-run dataset (Figure 2.3A) to the 10-run dataset (Figure 2.4A) and the 30-run dataset (Figure 2.5A), it is clear that the maximal number increases as the number of LC-MS runs increases. This is understandable considering the fact that each run adds some unique peak pairs to the total. However, there appears to be a diminished return as the number of runs increases beyond a certain value. For example, using 10 runs, instead of 3 runs, the peak pair number increases from 1700 to 2350 (i.e., 38% with a net gain of 650 pairs). However, using 30 runs, instead of 10 runs, the pair number increases from 2350 to 2900 (i.e., 23% with a net gain of only 550 pairs). Thus, performing replicates merely for the purpose of increasing the peak pair number in a dataset needs to be considered within the context of instrumental time available. In a clinical metabolomics study involving the profiling of hundreds of samples, one may choose not to perform replicate runs in order to save instrument time. On the other hand, for a cellular metabolomics work where only a few samples are profiled, it may be well justified to perform replicate runs. In any case, with zero-fill, we can recover the missing values in a dataset very effectively and efficiently.

2.3.4 Characterization of Missing Values

As indicated earlier, the source of missing values in replicate run dataset is mainly from the measurement and data processing processes which can be influenced much more by the low abundance peaks than the high abundance ones. We have characterised the missing values in terms of signal intensity in the 10-run dataset. While peak ratio is used to measure the relative concentration in CIL LC-MS, the absolute intensity of a peak is related to abundance and detection sensitivity of the metabolite. It should be noted that detection sensitivity of different metabolites becomes more uniform after dansylation labeling. For example, the difference in MS

signal intensity for 17 dansyl amino acid standards is within one order of magnitude, compared to more than three orders of magnitude for unlabeled amino acids.[120] Thus, the absolute intensity of labeled metabolites is a good indication of analyte abundance in a sample. Figure 2.6 shows a histogram of the peak pair distribution as a function of the absolute intensity measured by S/N. The S/N values are binned in log9 to distribute the number of peak pairs found in each bin evenly across the y-axis. In the low S/N bins, there are significantly more pairs detected with zero-fill. For example, at around S/N 9 (i.e., 1.0 in log9), about 200 pairs are detected with zero-fill, compared to 100 pairs detected without zero-fill. In the high S/N bins, the number of peak pairs found with and without zero-fill is similar. Thus, the zero-fill process recovers mainly the low intensity or low abundance metabolites that fail to detect in the 1st path of data analysis by IsoMS.

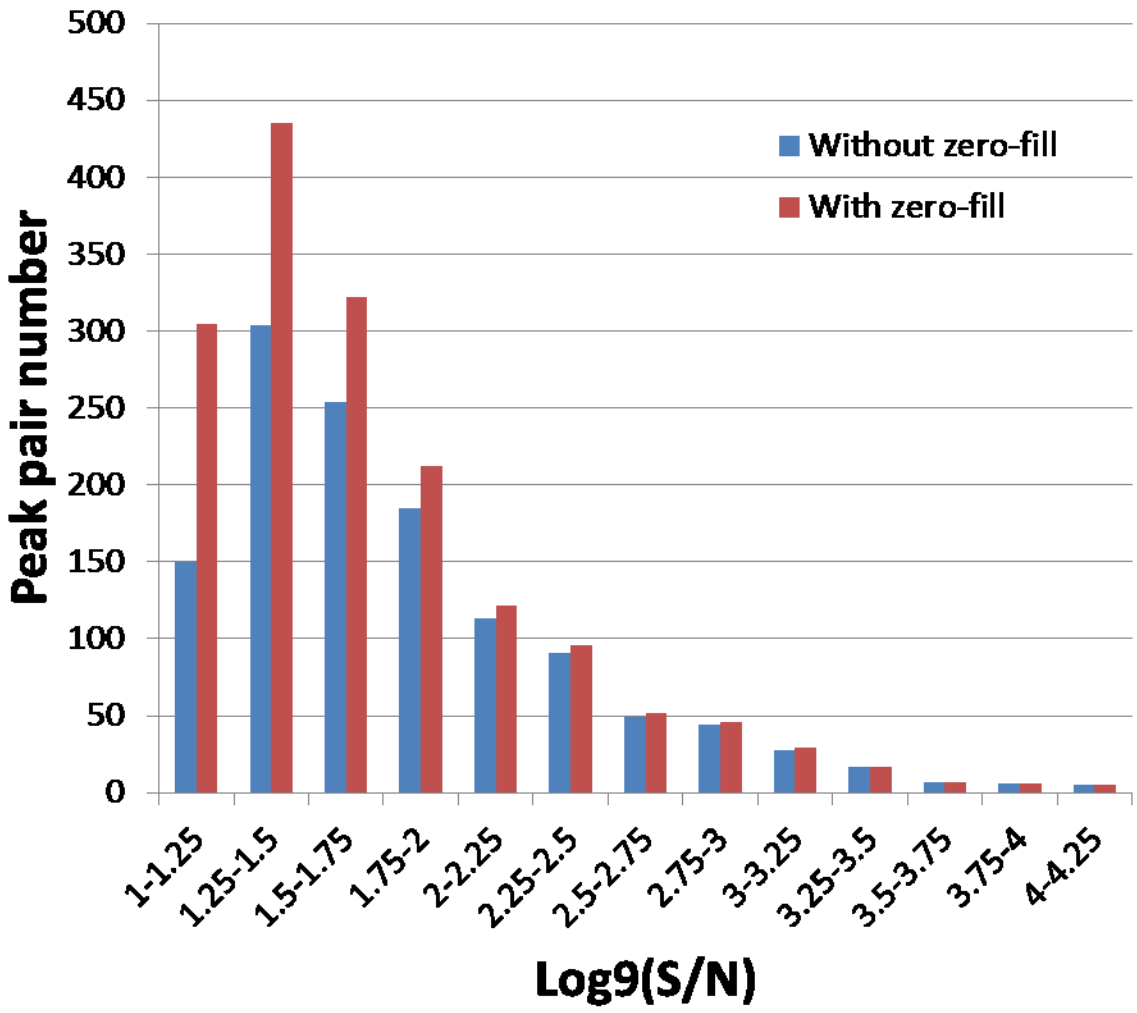


Figure 2.6 Number of peak pairs as a function of log9(S/N)

2.3.5 Standardization of Counting Missing Values

Because missing values are mainly from the low abundance peaks which are more difficult to reproducibly detect, the extent of missing values in a dataset should be a good indicator to judge the overall analytical performance of a metabolome profiling method. We propose to use an experimental triplicate dataset (e.g., the data shown in Figure 2.2) and a 10-run injection replicate dataset (e.g., the data shown in Figure 2.4A) of the same sample to measure the performance of a method regarding the missing values. Although using data of different

samples has the benefit of evaluating how well a method quantifies the same metabolites of different concentrations in different samples, it requires a set of standard samples available for method evaluation. Replicate data of the same sample is readily generated in a lab. Using the same type of sample (e.g., human urine), the performance of different methods in terms of missing values can still be compared, at least within the context of performing metabolomics study using this type of sample. Recent development of standard samples such as NIST serum standard should facilitate future work of comparing different methods, if such a standard is used across different platforms and methods.[124]

Using the replicate dataset, we propose that the performance indicators be 1) number of peak pairs detected per run and the total number of peak pairs detected within a dataset (triplicate or 10-run replicate), 2) intensity dynamic range from the lowest absolute signal intensity giving a quantity result to the highest intensity giving an intensity value, and 3) number of missing values and percentage of missing values in triplicate and 10-run replicate datasets. Table 2.1 and 2.2 show the summary of the results for the triplicate and 10-run datasets obtained by the dansylation CIL LC-MS method, respectively.

Table 2.1 The summary of the results for the triplicate dataset obtained by dansylation CIL LC-MS method

	number of pairs	min intensity	max intensity	no. of missing value	percent of missing value
Run 1	1678	1510	1872500	16	0.94%
Run 2	1683	1510	1872500	11	0.65%
Run 3	1675	1510	1872500	19	1.12%
Average	1679	1510	1872500	15	0.91%
std	4	0	0	4	0.24%
Overall performance	1694	1510	1872500	46	

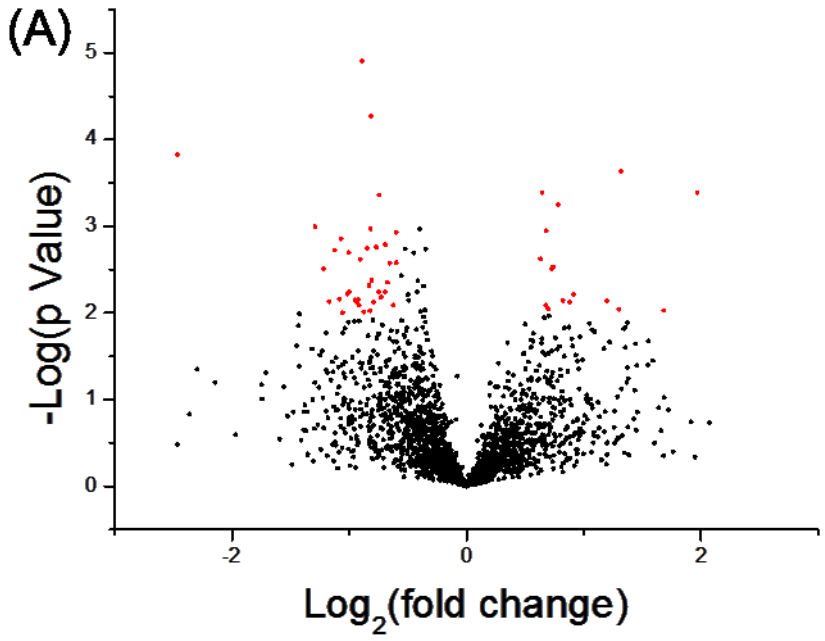
Table 2.2 The summary of the results for the 10-run dataset obtained by dansylation CIL LC-MS method

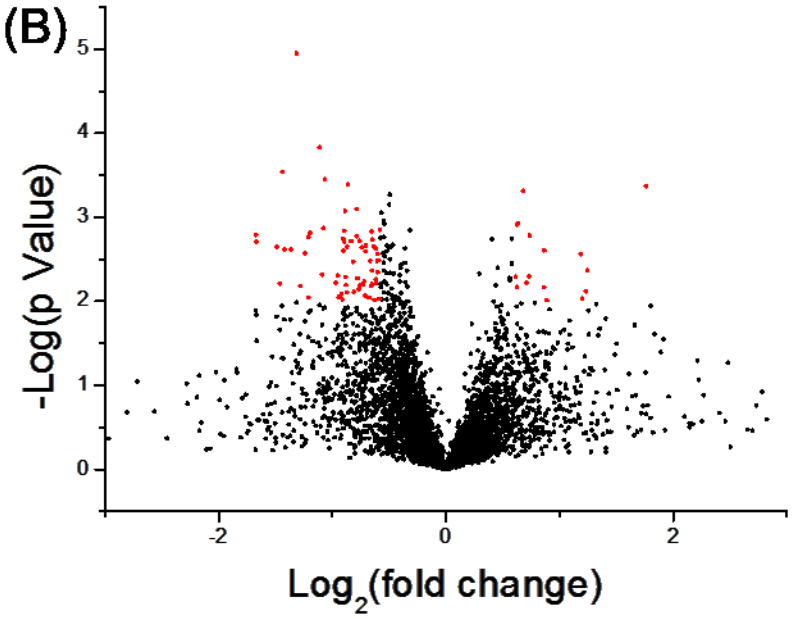
	number of pairs	min intensity	max intensity	no. of missing value	percent of missing value
Run 1	2217	1510	1850000	151	6.38%
Run 2	2235	1510	1850000	133	5.62%
Run 3	2138	1510	1850000	230	9.71%
Run 4	2233	1500	1850000	135	5.70%
Run 5	2257	1500	1850000	111	4.69%
Run 6	2168	1500	1850000	200	8.45%
Run 7	2088	1500	1850000	280	11.82%
Run 8	2226	1500	1850000	142	6%
Run 9	2168	1500	1850000	200	8.45%
Run 10	2198	1500	1850000	170	7.18%
Average	2193	1503	1850000	175	7.40%
std	52	5	0	52	2.20%
Overall performance	2368	1500	1850000	1752	

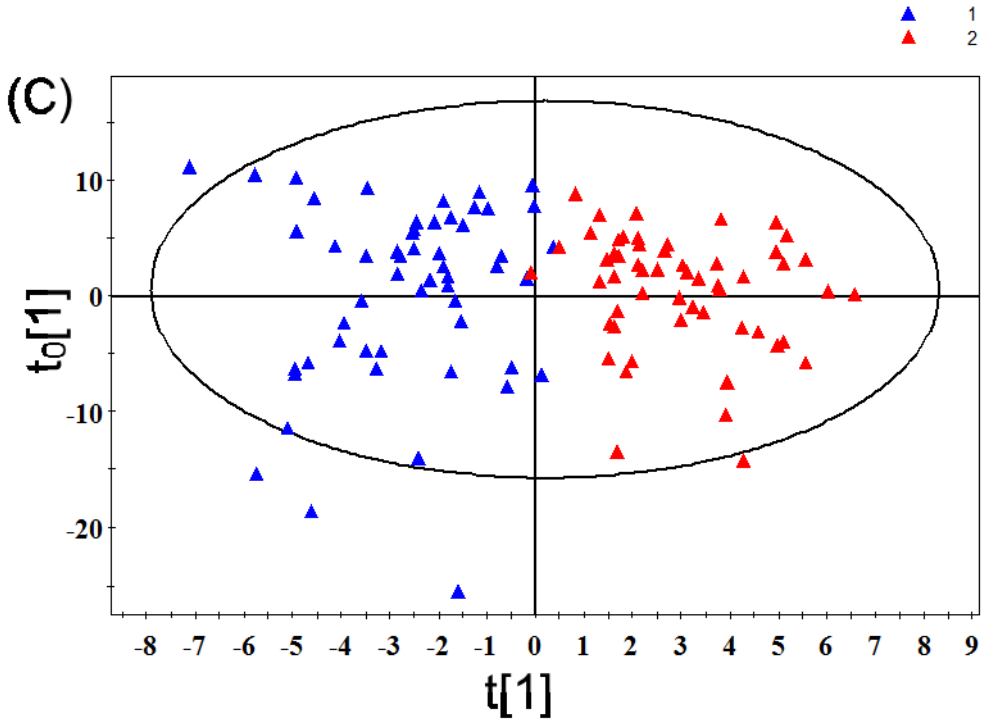
2.3.6 Metabolomics Application

Finally, we have applied the zero-fill program in a metabolomics study to demonstrate the benefits of using zero-fill for disease biomarker discovery. In this case, we applied zero-fill to a set of LC-MS data generated from a human bladder cancer metabolomics study.[125] It consists of 109 LC-MS runs of dansyl labeled urine samples collected from 55 bladder cancer patients and 54 controls. Individual samples were separately labeled with ^{12}C-danylation and then mixed with ^{13}C-dansylated universal metabolome-standard of human urine. The individual ^{13}C-/^{12}C-labeled mixtures were separated and analyzed using reversed phase LC and Bruker 9.4-Tesla Fourier transform ion cyclotron resonance mass spectrometer.[125] Supplemental Table

T2.1 (see Appendix folder) shows the original metabolite-intensity table generated using IsoMS from the 109 runs. Supplemental Table T2.2 (see Appendix folder) shows the table after applying the zero-fill program to the dataset. The volcano and OPLS-DA plots of the datasets with and without zero-fill are shown in Figure 2.7.




(B)
-Log(p Value)
Log2(fold change)


(C)
1
2
t0[1]
t[1]

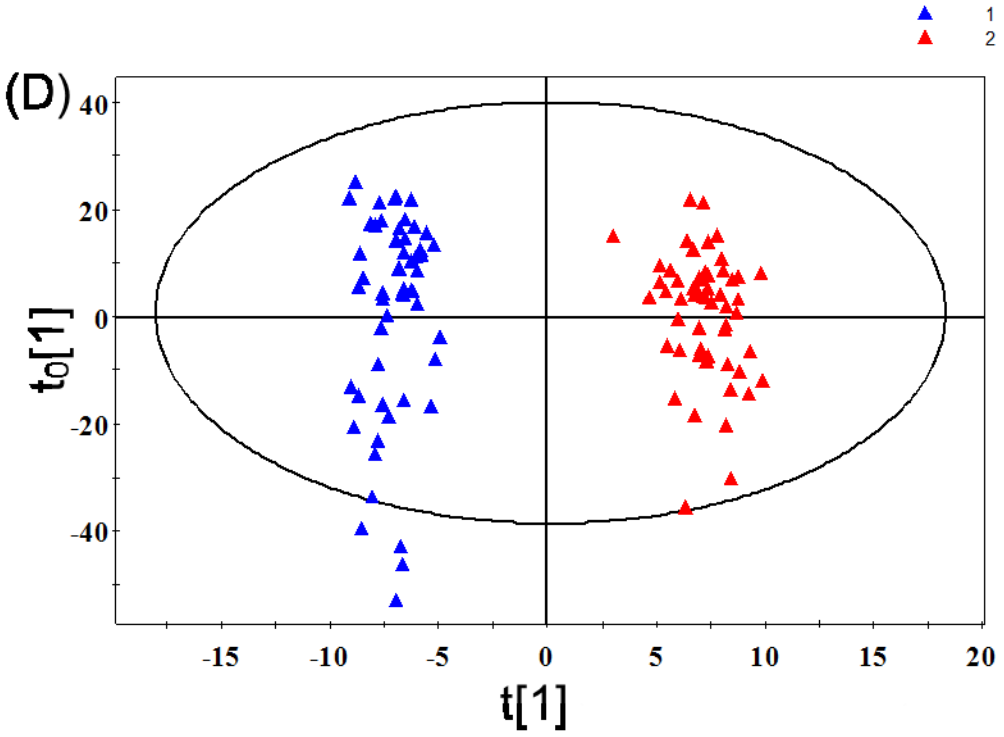


Figure 2.7 Volcano plots of the 109-sample data set from a bladder cancer biomarker discovery study: (A) without zero-fill and (B) with zero-fill. The red dots represent a metabolite with a fold change of > 1.5 and p-value < 0.01. OPLS-DA plots of the 109-sample dataset: (C) without zero-fill and (D) with zero-fill

As Figure 2.7A, B shows, more significant metabolites are detected in the volcano plot of the zero-filled data. There are 81 metabolites with fold change of greater than 1.5 and p-value of less than 0.01 in the dataset with zero-fill, compared to 65 metabolites without zero-fill. A similar observation is found in the OPLS-DA analysis. There are 385 significant metabolites (VIP score of ≥1.5) found from the zero-filled data, compared to 53 metabolites without zero-fill. Supplemental Tables T2.3-T2.6 (see Appendix folder) list the significant metabolites; some of them were putatively identified based on accurate mass match against the Human Metabolome Database (HMDB) and the Evidence-Based Metabolome Library (EML) by using the

MyCompoundID program. As Figure 2.7C,D shows, a much better separation of the cancer and control groups is obtained with the zero-filled data (without zero-fill: R^2X=0.389, R^2Y=0.745, Q^2=0.562; with zero-fill: R^2X=0.366, R^2Y=0.972, Q^2=0.621).

The above results clearly show a significant improvement of the quality of statistical analysis after applying zero-fill to the 109-sample dataset, enabling the detection of more and better-discriminating significant metabolites to differentiate two cohorts of samples. To measure the quality of the metabolite-intensity data in terms of missing values, we plot the percentage of common peak pairs detectable in cumulative samples as a function of sample runs in a dataset (see Figure 2.8). This plot is informative for determining the consistency of metabolite detection among all the runs. For example, 2858 peak pairs or about 60% of the total number of peak pairs found in the zero-filled dataset (4761) can be consistently quantified in half of the samples (109/2), while without zero-fill only 395 or 8.3% of the total (4761) are commonly detected. In our view, this type of plot should be presented, along with the metabolite-intensity table, when reporting the metabolome profiling data in a metabolomics study. This would assist in judging the overall coverage of the metabolomic profiles in a study.

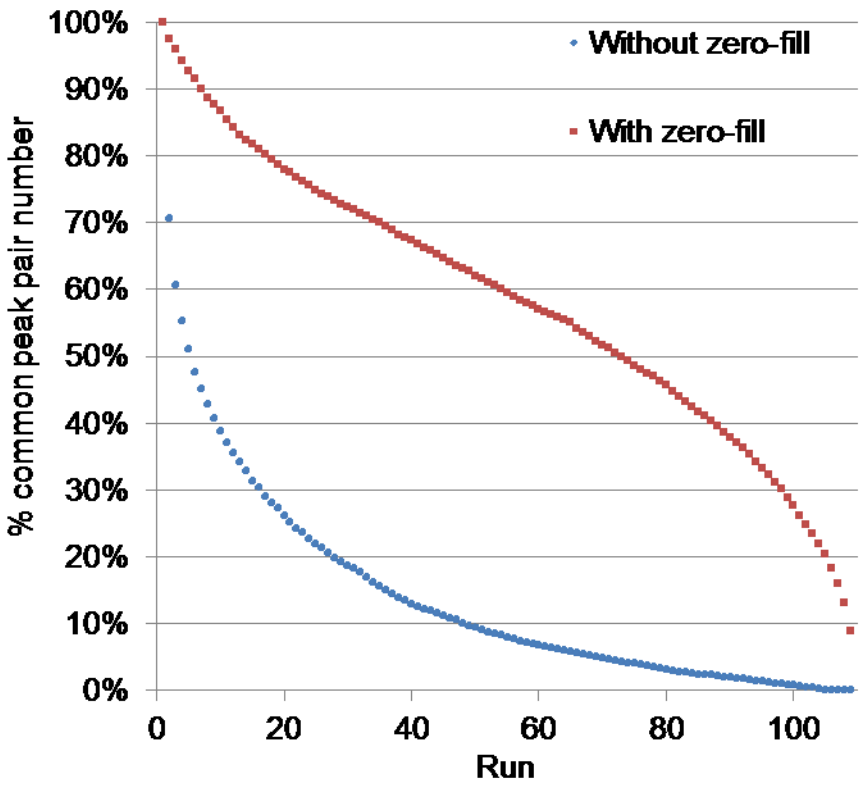


Figure 2.8 Percentage of common peak pairs detected in cumulative runs as a function of sample runs. The total number of pairs detected from 109 runs is 4761.

2.4 Conclusions

We report a workflow to reduce the extent of missing values in chemical isotope labeling LC-MS metabolomics platform. A zero-fill program, freely available at MyCompoundID.org, has been developed to retrieve missing values in the initial metabolite-intensity table generated by IsoMS. Missing values were found to be mainly from the metabolites with low signal intensity in mass spectra. The zero-fill program developed based on the unique features of peak pairing and consistency of absolute intensity of the ^{13}C-labeled peaks from a ^{13}C-labeled control sample spiked into all ^{13}C-labeled individual samples allows significant reduction in missing

values. This reduction affords the detection of more and better-discriminating significant metabolites in a metabolomics study involving the metabolomic profiling of 109 samples for bladder cancer biomarker discovery.

Because the extent of missing values can have a profound effect on metabolomics results, we feel that counting missing values should be considered as an important metrics for measuring the analytical performance of a metabolomics platform. To facilitate method comparison in terms of missing values, we proposed and illustrated the use of two datasets, one from experimental triplicate and another one from 10 replicate injections of the same sample, to measure the extent of missing values. Finally, in reporting metabolomics data, we feel that it is important to include the result of missing value analysis (e.g., a plot of number or percentage of common metabolites detected in cumulative samples as a function of sample runs). This analysis result, along with the metabolite-intensity table containing all the metabolites and their intensity values from the entire dataset, measures the level of commonly quantifiable metabolites in a metabolomics study. At a chosen % threshold (e.g., metabolites commonly quantifiable in more than 50% of all the samples), the number of metabolites retained for statistical analysis should be reported. In this regard, future work is still needed to examine the issue of selecting the most appropriate % threshold for data inclusion in statistical analysis.

Chapter 3

Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry

3.1 Introduction

Chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS) uses differential isotope mass tags to label a metabolite in two comparative samples (e.g., ^{12}C-labeling of an individual sample and ^{13}C-labeling of a pooled sample), followed by mixing and LC-MS analysis. Individual metabolites are detected as peak pairs in mass spectra. The intensity ratio of a peak pair can be used to measure the relative concentration of the same metabolite in two samples. CIL LC-MS can significantly increase the detectability of metabolites by rationally designing the labeling reagents to target a group of metabolites (e.g., all amine-containing metabolites or amine submetabolome) to improve both LC separation and MS sensitivity.[45, 120] It can also overcome the technical problems such as matrix effects, ion suppression and instrument drifts to generate more precise and accurate quantitative results, compared to conventional LC-MS.[60, 61, 122, 126] There are a number of new advances reported[53-58, 61, 62, 122, 126-136] in the area of developing CIL LC-MS for targeted and untargeted metabolomics, particularly for improving labeling chemistries and extending the utility of CIL LC-MS to analyze a broad range of metabolites. However, proper processing of CIL LC-MS data is also critical to maintain high sensitivity (i.e., extracting as many peak pairs as possible from a dataset), high specificity (i.e., keeping low false-positive rate), and high performance quantification (i.e., achieving high precision and accuracy).[123] To this end, we have been

involved in developing data processing methods specifically for handling CIL LC-MS data. The software tools related to these methods including IsoMS[123] and Zero-fill[137] are freely available from the www.mycompoundid.org website.

In our data processing workflow, the raw mass spectral data, instead of the chromatographic peak data, are used for metabolite peak detection, peak pairing, peak-pair filtering and peak ratio calculation by IsoMS.[123] This MS-centric approach allows us to detect more peaks, as many regions of the baseline in a chromatogram still contain mass spectra with low abundance ion peaks. Using a chromatographic peak threshold for peak picking will not detect these peaks. Moreover, it is easier and more reliable to group or remove peaks from the same metabolite using a mass spectrum. This is because the salt/solvent adducts, mono- or hetero-dimers, multimers, common fragment ions (e.g., $-H_2O$ and $-CO_2$) of a molecular ion are present in the same mass spectrum and thus can be readily detected and filtered out. Finally, the Zero-fill program[137] can be used to detect a missing peak pair in a mass spectrum based on the similarity of retention time, accurate mass and ^{13}C-peak intensity (the same amount of ^{13}C-labeled pool is spiked to each sample) to those of the other samples where the peak pair is detected. This algorithm would be difficult to implement using chromatographic peak information.

While processing mass spectral data directly provides some advantages, it is not optimal for extracting quantitative information from the mass spectral peak intensities. Currently, in IsoMS, the peak ratio of a peak pair from a ^{13}C-/^{12}C-labeled metabolite is calculated from a mass spectrum.[123] If the same peak pair shows up in multiple neighbouring scans or spectra, only the highest intensity peak pair is kept. Its peak ratio is calculated and then entered in the metabolite-intensity table. In order to utilize all the peak pairs intensity information, we have

now developed a program, IsoMS-Quant, to reconstruct two chromatographic peaks, one for ^{12}C- or light-labeled metabolite and another one for ^{13}C- or heavy-labeled metabolite, for each peak pair shown in the metabolite-intensity table. The area ratio of the two chromatographic peaks measured by the sums of ^{13}C- or ^{12}C-labeled peak intensities is calculated as a measure of relative concentration of the metabolite in light-labeled sample vs. heavy-labeled sample. Using chromatographic peaks for quantification smoothes out signal fluctuations associated with mass spectral peak intensities in multiple scans, thereby providing better quantification. For targeted metabolite quantification, chromatographic peaks of an analyte are often used. In this report, we describe the IsoMS-Quant program and how it can be used to generate quantitative information in CIL LC-MS. Using examples of urine and serum metabolome analysis, we demonstrate that this program can improve untargeted quantitative metabolome profiling as well as targeted metabolite quantification significantly. The IsoMS-Quant program is freely available at www.mycompoundid.org and this program, along with IsoMS and Zero-fill, forms a complete data processing tool for the CIL LC-MS quantitative metabolomics platform.

3.2 Experimental Section

3.2.1 Dansylation Labeling and LC-MS

The labeling reaction (see Figure 3.1 for the reaction scheme) and LC-MS analysis on a Bruker Impact HD QTOF mass spectrometer (Billerica, MA, USA) linked to an Agilent 1100 HPLC system (Palo Alto, CA, USA) and an electrospray ionization source were performed according to a protocol reported previously.[130, 137]

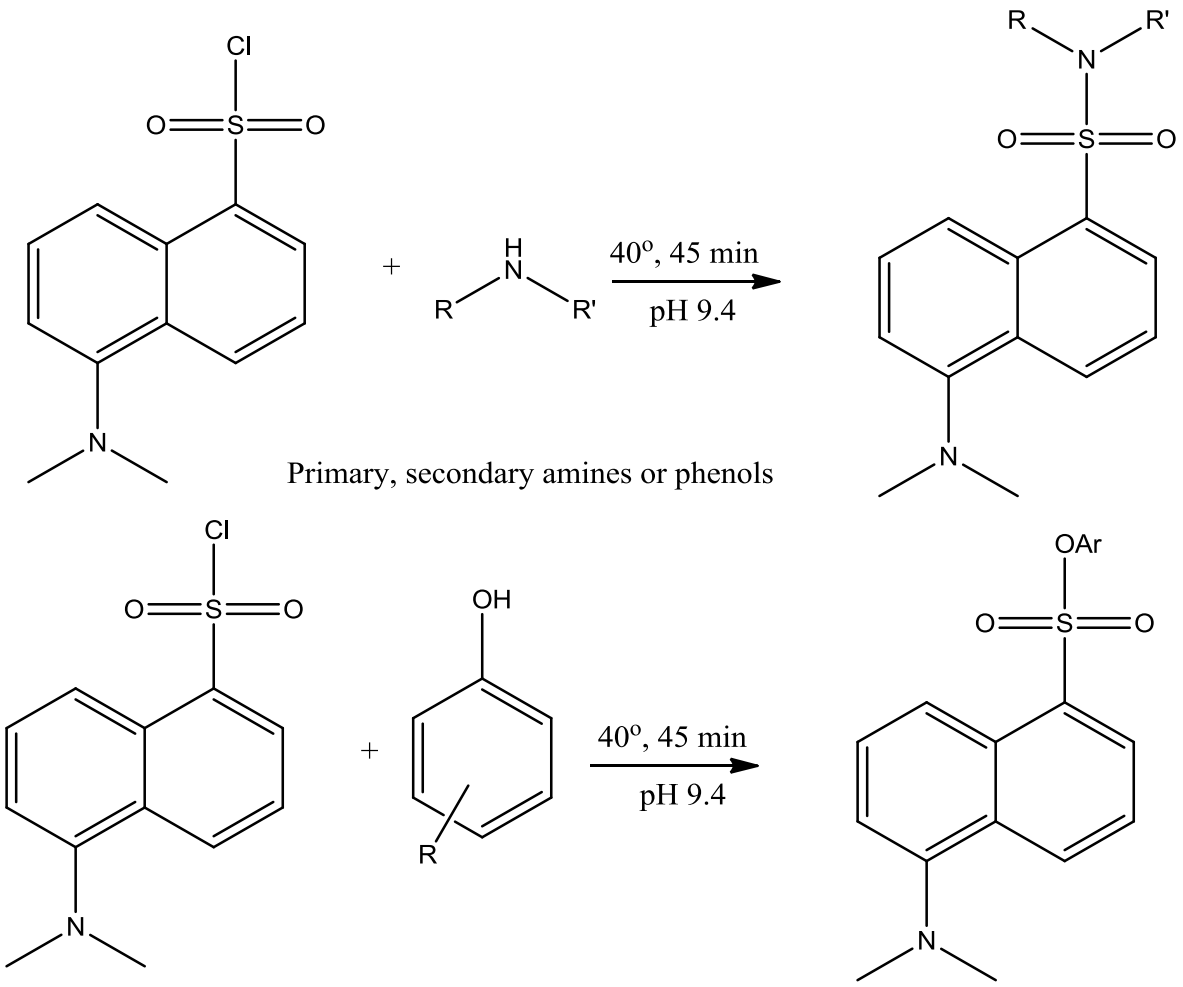


Figure 3.1 Dansylation reaction scheme

3.2.2 IsoMS-Quant

The IsoMS-Quant program was developed using R, an open source language and environment used in data processing and statistical programming. The user manual is provided in "IsoMS-Quant user manual" (see Appendix). Figure 3.2 shows the overall workflow for CIL LC-MS data processing. The raw LC-MS data are first processed using a peak-pair picking software, IsoMS. The high-confident level 1 peak pairs (i.e., the pair with two labeled peaks accompanied with their corresponding ^{13}C natural abundance peak) are aligned from multiple LC-MS runs to produce a metabolite-intensity CSV file or table. The Zero-fill program is then used to fill the

missing values in the CSV file. The IsoMS-Quant program is applied to the final metabolite-intensity table after the zero-fill process. Although we use the overall workflow shown in Figure 3.2 to illustrate how IsoMS-Quant is implemented in processing CIL LC-MS data, this program, in principle, should be applicable to other peak-picking software. While it is beyond the scope of this work, comparing different software packages for processing CIL LC-MS data should be valuable from a user's perspective.

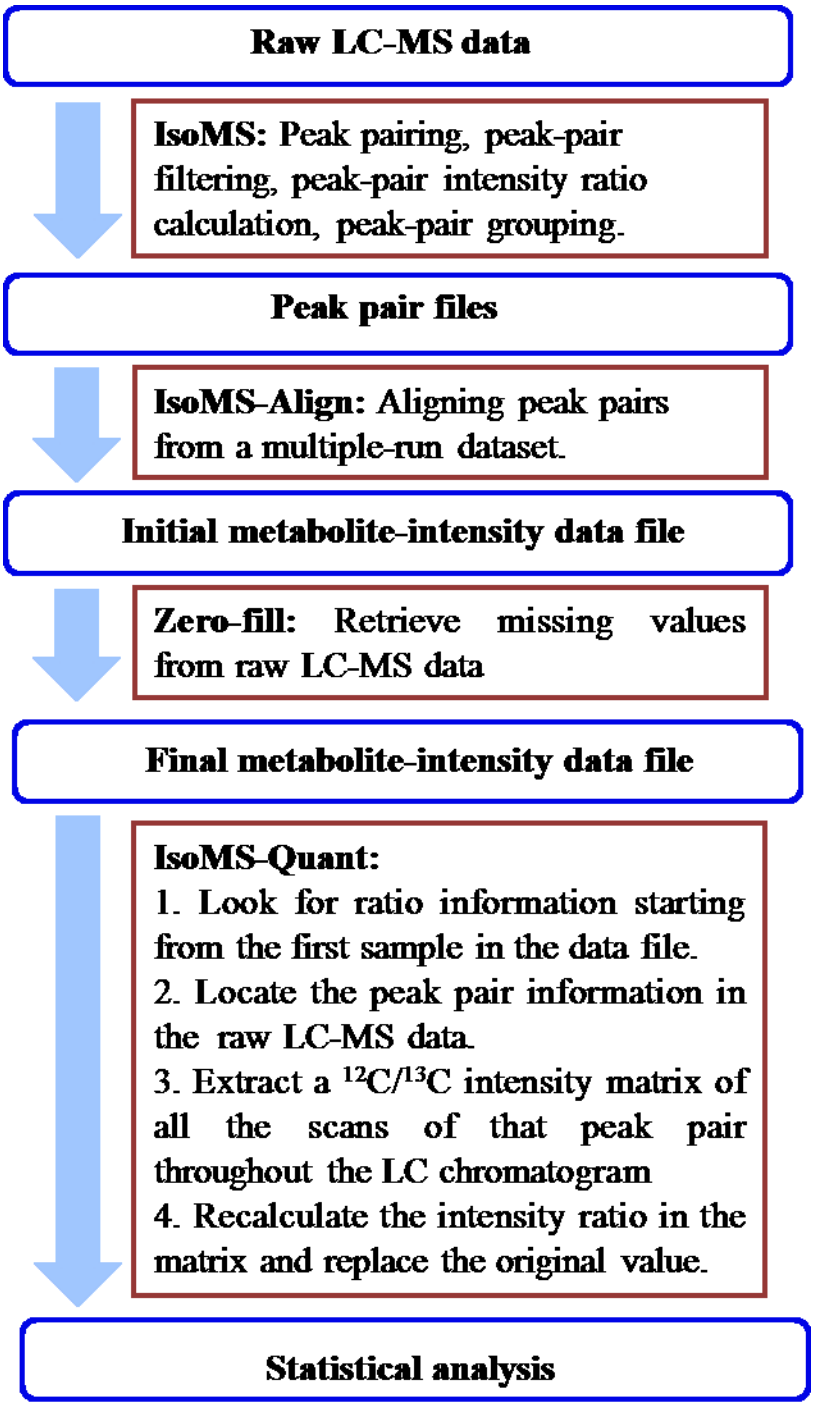


Figure 3.2 Workflow for processing chemical isotope labeling LC-MS data for quantitative metabolomic profiling.

During the IsoMS-Quant processing, the program loops through all the available MS-peak-intensity ratios starting from the first sample in the metabolite-intensity table. For each peak ratio, its associated retention time (rt), mz_light, mz_heavy, and ^{13}C-labeled MS-peak intensity (int) are used to locate this peak in the raw MS peak list from the original LC-MS dataset. A matching score is used to find the corresponding ^{13}C-labeled peak that was used to calculate the MS-peak-intensity ratio entered into the metabolite-intensity table. The matching score is defined as:

$$\text{Score} = \left(1 - \frac{\text{rt. diff}}{\text{rt. tol}}\right)/4 + \left(2 - \frac{\text{mz. diff. light} + \text{mz. diff. heavy}}{\text{mz. tol}}\right)/2 + (1 - 2 \times \text{int. diff})/4$$

where

$$\text{rt. diff} = \text{abs}(\text{rt. } ^{13}C\text{. peak} - \text{rt. rawdata. peak})$$

$$\text{mz. diff. light} = 1\text{E}6 \times \frac{\text{abs}\left(\text{mz. } ^{12}C\text{. peak} - \text{mz. rawdata. peak} + 2.0067\right)}{\text{mz. } ^{13}C\text{. peak}}$$

$$\text{mz. diff. heavy} = 1\text{E}6 \times \frac{\text{abs}\left(\text{mz. } ^{13}C\text{. peak} - \text{mz. rawdata. peak}\right)}{\text{mz. } ^{13}C\text{. peak}}$$

$$\text{int. diff} = \text{abs}\left(\log\left(\frac{\text{int. } ^{13}C\text{. peak}}{\text{int. rawdata. peak}}\right)\right)$$

The terms, rt.^{13}C.peak, mz.^{13}C.peak (or mz.^{12}C.peak), and int.^{13}C.peak, refer to retention time, m/z value, and intensity of the labeled peak in the metabolite-intensity table, respectively. The terms, rt.rawdata.peak, mz.rawdata.peak, and int.rawdata.peak, refer to retention time, mz value, and intensity of the labeled peak in the raw MS peak list, respectively. The value, 2.0067, comes from the mass difference of the two isotope carbons (^{13}C vs. ^{12}C labeling). The default rt tolerance (rt.tol) is 30 s and the default mz tolerance (mz.tol) is 5 ppm. These tolerance values

are instrument-dependent and can be adjusted. A different weight (divided by 2 or 4) is assigned to each of the similarity terms in the above score calculation equation. The mz value is deemed to be more important than rt and int and, therefore, given more weight. The MS peak with the maximal matching score is considered to be the correct ^{13}C-labeled peak. Once the ^{13}C-labeled peak is found in the raw peak list data, its corresponding ^{12}C-labeled peak is also identified in the same MS scan based on the mz difference of smaller than mz tolerance (default 5 ppm) from that of the ^{12}C-labeled peak in the metabolite-intensity table.

After both the ^{12}C- and ^{13}C-labeled peaks of a peak pair are identified in an MS scan, peaks in the neighboring MS scans are checked to see if the peak pair is also present. The check procedure stops once either the ^{12}C or ^{13}C peak is not found in a particular MS scan. After this procedure is completed, all the ^{12}C- and ^{13}C-labeled MS-peak intensities in these continuous MS scans over a chromatographic peak are used for chromatographic area calculation. Figure 3.3 shows how to calculate the chromatographic peak area from the sum of MS peak intensities.

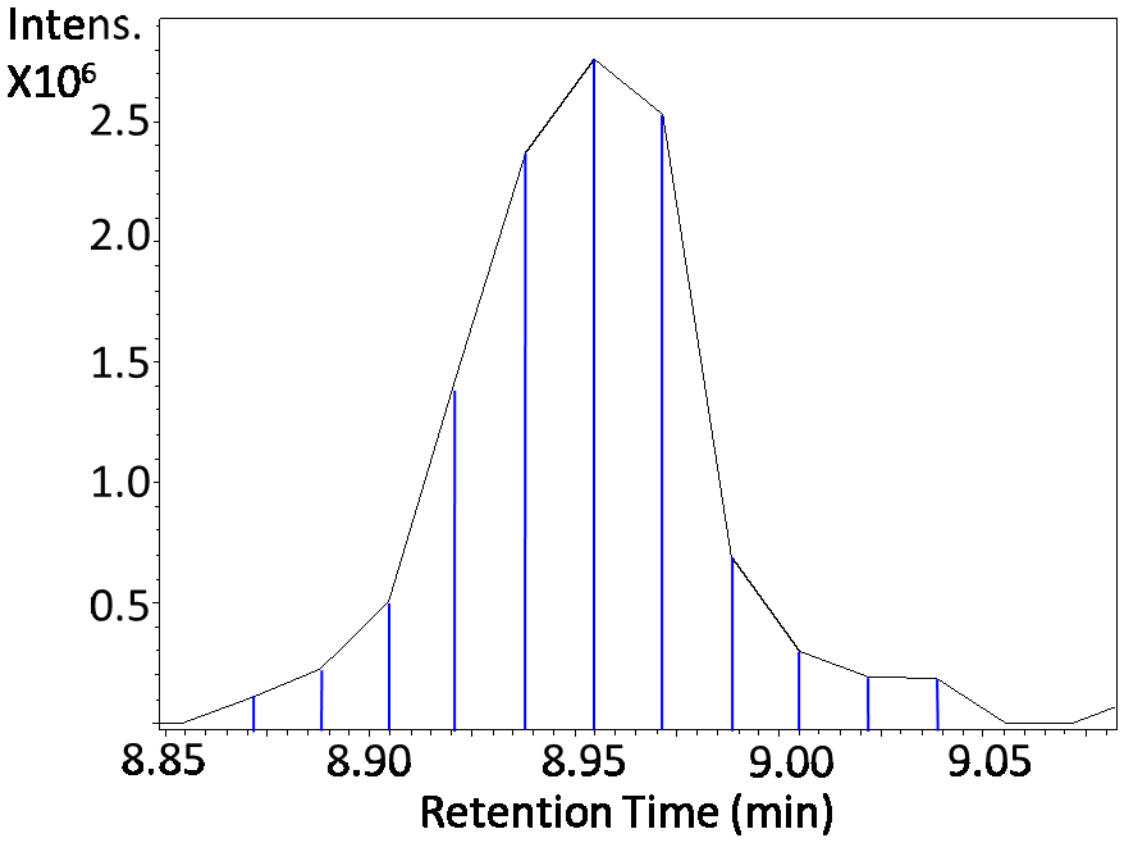


Figure 3.3 Schematic of chromatographic peak area calculation from mass spectral intensity values (blue lines)

In a typical LC-MS experiment, the MS signals are acquired at a constant time interval (e.g., at a spectral acquisition or scan rate of 1 Hz used in this work) and thus the chromatographic peak area can be calculated as the sum of all the segmented areas in trapezoids:

$$\text{Area} = \sum_{1}^{n} Area_i$$

where n is the number of scans where the same peak pair is detected. The area of each trapezoid can be described as:

$$Area_i = \frac{1}{2}(int_k + int_{k+1}) \times time.interval$$

Since the chromatographic peak area ratio of a peak pair is:

$$\text{Ratio} = \frac{Area\ of\ ^{12}C}{Area\ of\ ^{13}C}$$

by substituting the area with MS peak intensity, the ratio becomes:

$$\text{Ratio} = \frac{\sum_1^n int\ \ of\ ^{12}C}{\sum_1^n int\ of\ ^{13}C}$$

Thus the chromatographic peak area ratio of a peak pair can be determined as the sum of all the MS intensity values of the ^{12}C-labeled peaks divided by the sum of all the intensity values of the ^{13}C-labeled peaks.

After IsoMS-Quant completes the ratio calculation, it will compare the new ratio to the original intensity ratio. If the ratio difference is greater than 4-fold, the chromatographic ratio would be rejected; using manual inspections of the ratio results, we found that they belonged to less than 0.5% of the total number of peak pairs found and they were all falsely picked pairs. Otherwise, the new ratio will replace the original intensity ratio in the metabolite-intensity table which can be exported for statistical analysis or other uses.

3.3 Results and Discussion

CIL LC-MS is a platform that allows in-depth profiling of chemical-group-based submetabolomes using different labeling reagents targeting different classes of metabolites (e.g.,^{13}C-/^{12}C-dansylation labeling for quantifying amine- and phenol-containing metabolites.[125, 138-140]). The major difference between CIL LC-MS and conventional LC-

MS is that in CIL LC-MS all the true metabolites show up in the mass spectra as peak pairs which can be readily differentiated from the singlet peaks originated from background or noise. Thus it is much easier and more reliable to detect the true metabolite peaks. Based on this unique feature of peak pair detection, we have developed two software modules, IsoMS and Zero-fill, to process CIL LC-MS data to produce a metabolite-intensity table.[123, 137] The peak ratio in the table was calculated from the highest intensity peak pair found in multiple mass spectra. This way of calculation, while it is simple to implement, does not use the intensity ratio information in other neighboring mass spectra. In contrast, IsoMS-Quant utilizes all the mass spectral peak pair information to calculate a peak ratio.

Comparing the performance of using MS peak intensity ratio vs. chromatographic peak area ratio, three cases can be considered. Figure 3.4A shows an example of good chromatographic peaks where the peak ratios are basically the same: 0.51 from the chromatographic peak area calculation vs. 0.50 from the mass spectral intensity calculation. In this case, the overall mass spectral signals are strong (Figure 3.4B), representing a high abundance or readily ionizable metabolite found in a ^{12}C-labeled human serum sample mixed with a ^{13}C-labeled pooled sample. However, because of a wide concentration dynamic range of metabolites present in a sample such as human serum, there are many low-intensity peaks detected in LC-MS. For these peaks, the highest intensity peak pair may not be representative of the concentration ratio of the labeled metabolite. Figure 3.4C shows an example of relatively poor chromatographic peaks for both the ^{12}C- and ^{13}C-labeled mass spectral peaks (Figure 3.4D). Poor peak shape is likely due to the effects of other co-eluting components or background ions present in a complex sample along with the analyte during the analyte elution; these peaks show up randomly and unpredictably and cannot be mimicked using simple standards. The ratio

calculated using the highest mass spectral peak intensities (1.26) does not match well with the ratio determined from the chromatographic peak areas (0.98).

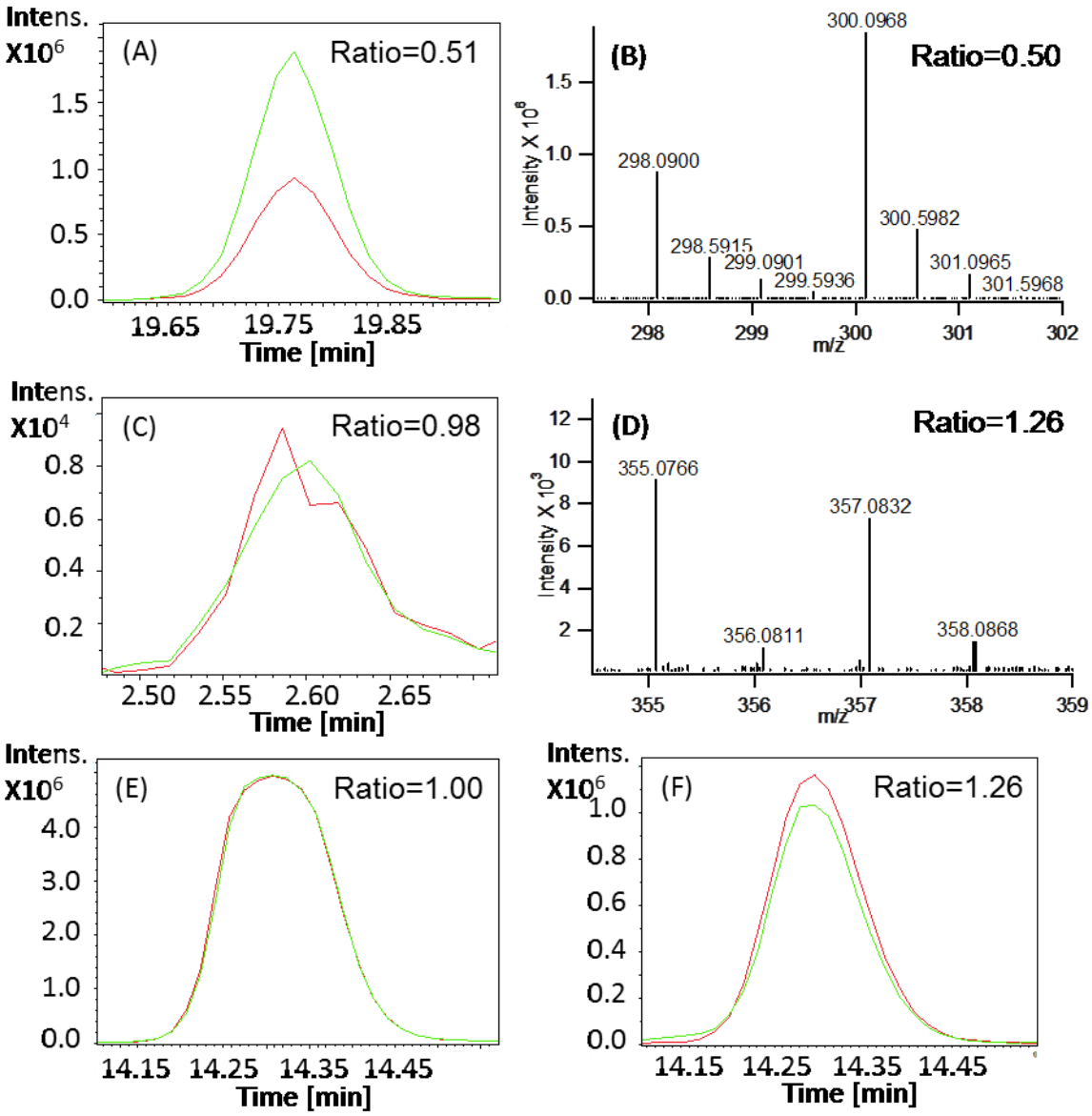
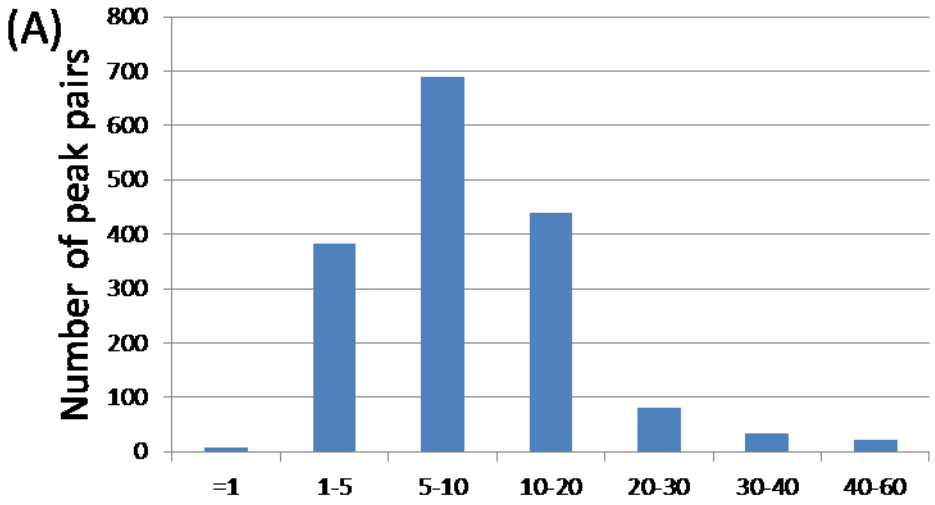


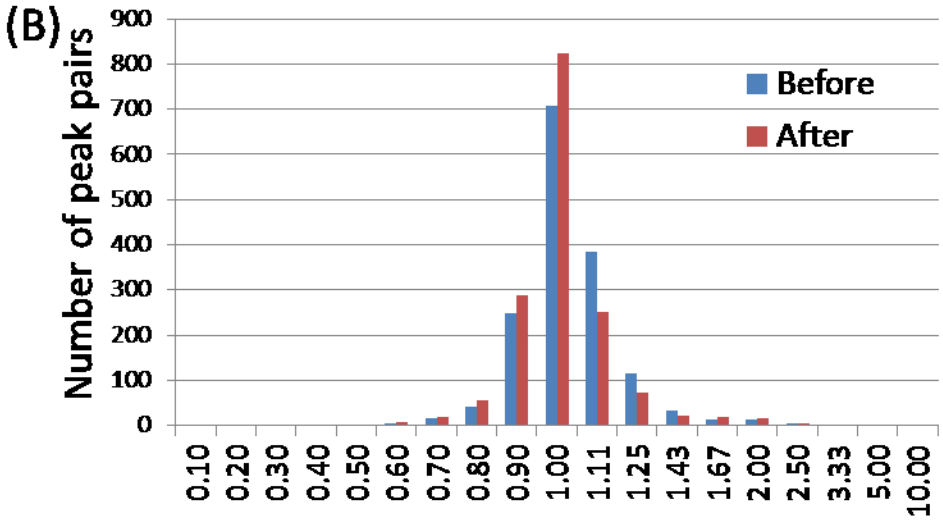
Figure 3.4 (A) Extracted ion chromatograms (EICs) of a relatively high abundance or easily ionizable ^{13}C-/^{12}C-labeled peak pair (green: ^{12}C-labeled metabolite; red: ^{13}C-labeled metabolite) found in a mixture of a ^{12}C-labeled individual human serum and a ^{12}C-labeled pooled serum prepared from 100 healthy individuals and (B) the highest intensity mass spectrum of the pair.

(C) EICs of a relatively low abundance or not readily ionizable peak pair and (D) the high intensity mass spectrum of the pair. (E) EICs of a saturated peak pair and (F) EICs of the corresponding pair plotted using their ^{13}C natural abundance peaks.

Another case is related to the saturation of MS detection which can lead to distorted peak shapes. Figure 3.4E shows an example where the MS signals are saturated and the mass spectral peak intensity no longer reflects the real metabolite concentration. IsoMS-Quant addresses this issue by automatically finding the ^{13}C natural isotope peaks of both the ^{12}C- and ^{13}C-labeled metabolite peaks and then using these peaks to reconstruct the chromatographic peaks (Figure 3.2F) for ratio calculation. Since the ^{13}C natural isotope peak is much lower in intensity, they are less likely to be saturated in MS detection and thus can be used for more accurate quantification. In the Impact QTOF instrument, we rarely observed the saturation of the ^{13}C natural isotope peak; electrospray ionization saturation often occurred before detection saturation. In IsoMS-Quant, a user can enter a threshold above which saturation occurs, depending on the MS instrument used.

The overall performance improvement for quantitative metabolomics can be demonstrated using the results of triplicate analysis of a ^{13}C-/^{12}C-labeled human urine sample. In this experiment, an equal amount of ^{12}C-labeled and ^{13}C-labeled same urine was mixed for analysis and thus the peak ratios for all the metabolite peak pairs should be equal to 1. Figure 3.5A plots the number of peak pairs detected in multiple neighboring mass spectral scans as a function of the scan number. Out of the 1660 peak pairs detected, only 7 pairs (<1%) were detected in a single mass spectrum. The highest percentage of peak pairs belongs to those detected over 6 to 10 mass spectra or chromatographic peaks of 6 to 10 s. Thus, for most of the peak pairs detected, they appear in multiple scans and integration of peak pair intensities over these scans should improve quantification.


(A)
Number of peak pairs
800
700
600
500
400
300
200
100
0
=1
1-5
5-10
10-20
20-30
30-40
40-60


(B)
Number of peak pairs
900
800
700
600
500
400
300
200
100
0
Before
After
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
1.11
1.25
1.43
1.67
2.00
2.50
3.33
5.00
10.00

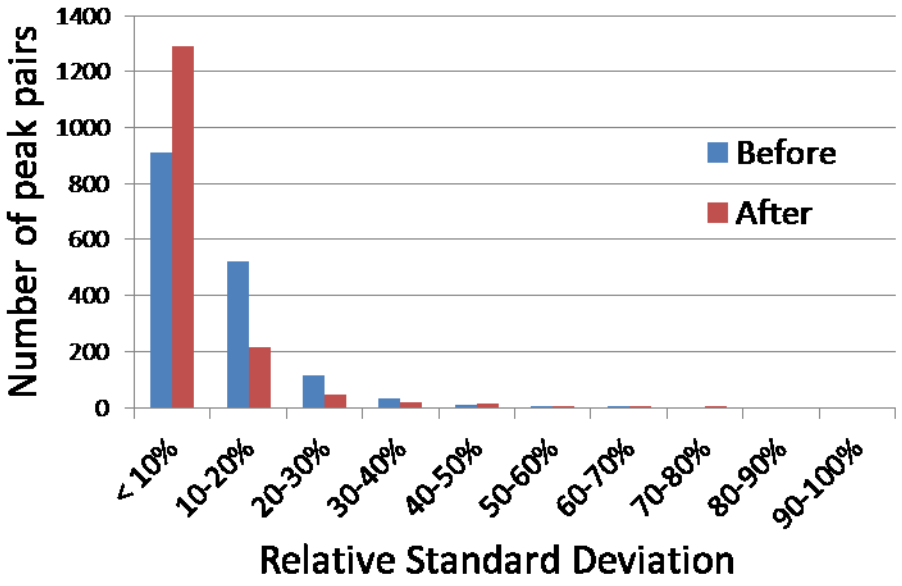


Figure 3.5 Distributions of the number of peak pairs detected in a 1:1 13C-/12C-labeled human urine sample as a function of (A) number of neighboring MS scans where a peak pair is detected, (B) peak ratios calculated before and after applying IsoMS-Quant, and (C) relative standard deviation of peak ratios from triplicate experiments (n=3)

Figure 3.5B shows a distribution of the number of peak pairs as a function of the peak ratio determined with and without applying IsoMS-Quant. The peak ratio distribution becomes more symmetric after using IsoMS-Quant and there are more peak pairs with peak ratio values close to 1. Figure 3.5C shows a distribution of the number of peak pairs as a function of the relative standard deviation (RSD) of the peak ratio from the mean from experimental triplicate runs. More peak pairs have their ratios close to the mean after using IsoMS-Quant. Without using IsoMS-Quant, the $^{12}C/^{13}C$ ratios have an averaged RSD of 10.4%, and with IsoMS-Quant, the averaged RSD is reduced to 6.7%. The results shown in Figure 3.5B, C illustrate that the use of IsoMS-Quant can improve the accuracy and precision for quantitative metabolomics by CIL LC-MS. We note that we have not studied how the retention time precision or mass accuracy of

different instruments would affect the degree of improvement achievable by IsoMS-Quant. In our studies, we usually used an LC instrument that can readily provide a retention time precision of better than 30 s and a mass spectrometer that can provide a mass accuracy of better than 5 ppm for CIL LC-MS.

We have used the IsoMS-Quant program for a number of metabolomics research projects and observed improvement in quantitative results that lead to better statistical analysis of the metabolomic data. One example is in a metabolomics study where a set of 109 LC-MS runs of dansyl labeled urine samples collected from 55 bladder cancer patients and 54 controls were processed to search for potential metabolite biomarkers for diagnosis of bladder cancer.[125, 137] The two groups could be readily separated using PLS-DA or volcano plots based on concentration variations of a number of significant metabolites.[125, 137] Figure 3.6 shows a plot of the p-values of three representative significant metabolites obtained before and after applying IsoMS-Quant. The p-values increase by more than 10-fold after using IsoMS-Quant. This level of improvement can be attributed to the fact that IsoMS-Quant generates more precise and accurate peak ratio values, allowing the reduction of intra-group variations and better separation of inter-group differences.

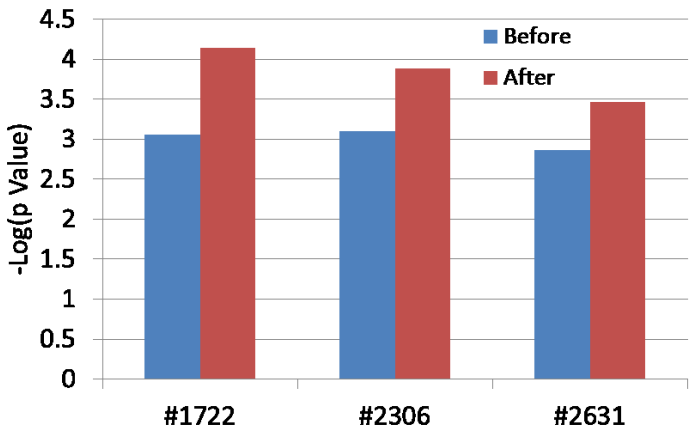


Figure 3.6 P-values of three significant metabolites differentiating bladder cancer and control groups in a metabolomics study of 109 bladder cancer and control samples. Metabolite #1722 with molecular mass of 323.1703 was putatively identified as a derivative of 1,3-diaminopropane with the addition of adenosine, Metabolite #2306 with molecular mass of 290.1474 was putatively identified as a derivative of glutamic acid with the addition of carnitine, and Metabolite #2631 with molecular ion mass of 144.1017 was putatively identified as proline betaine.

Finally, the use of IsoMS-Quant can also improve the analytical performance of targeted metabolite quantification. For targeted metabolite quantification using CIL LC-MS, a reference sample such as a pooled sample is ^{13}C-labeled, following by spiking ^{12}C-labeled metabolite standards with known concentrations to determine the absolute concentrations of all the metabolites of interest. This reference sample can then be used to quantify the metabolites in an individual sample by measuring the peak ratio of a metabolite peak pair from a mixture of ^{12}C-labeled individual sample and ^{13}C-labeled reference sample. As an example, we performed absolute quantification of 20 metabolites in an individual human serum sample. Standard addition method using ^{12}C-dansyl labeled metabolite standards was used to determine the

absolute concentrations of these 20 metabolites in a ^{13}C-labeled pooled sample generated from mixing serums of 100 healthy individuals. An aliquot of the ^{13}C-labeled pooled sample was spiked into a ^{12}C-labeled individual sample in 1:1 volume ratio. The mixture was analyzed by LC-MS. The absolute concentrations of the 20 metabolites in the individual sample were determined by using the peak ratio of a metabolite peak pair and the absolute concentration of the metabolite in the pooled sample. Peak ratio was calculated with and without using IsoMS-Quant.

Table 3.1 lists the concentrations of 20 metabolites found in the individual serum sample using data processing with and without applying IsoMS-Quant. As Table 3.1 shows, the percentage of concentration difference for an individual metabolite by the two processing methods or relative error can be up to 32% (for serine) and as high as over 55% (for glycine where mass spectral peaks were saturated); the average difference was 13%. Manual inspection of the concentration data generated from the IsoMS-Quant method indicated that these concentration values were much more reliable, as the chromatographic peak shapes of a peak pair were well represented. Better precision is also achieved using IsoMS-Quant (mean RSD of 4.2% vs. 6.7% without IsoMS-Quant from triplicate experiments). This example demonstrates that by using IsoMS-Quant better accuracy and precision can be achieved for targeted quantification of metabolites of interest using CIL LC-MS.

Table 3.1 Results of targeted quantification of 20 metabolites in a human serum sample by LC-MS analysis of a mixture of the 12C-labeled sample and the 13C-labeled pooled serum standard with known concentrations of these metabolites.

Name	Absolute conc. from peak area (μM)	Absolute conc. from MS intensity (μM)	Absolute conc. Relative error

Taurine	68 ± 1	69 ± 5	1%
Arginine	364.6 ± 0.3	371 ± 1	2%
Asparagine	10 ± 4	10 ± 4	0%
Glutamine	89 ± 1	96 ± 4	8%
Homoserine	0.98 ± 0.02	0.8 ± 0.1	18%
Serine	428 ± 1	293 ± 6	32%
Aspartic Acid	1289 ± 4	1310 ± 30	2%
Trans-4-Hydroxyl-L-Proline	20.2 ± 0.2	19.1 ± 0.5	5%
Threonine	220.6 ± 0.5	169 ± 3	23%
Aminoadipic acid	2.7 ± 0.2	2.6 ± 0.4	4%
Glycine	411 ± 2	183 ± 4	55%
Glycylproline	0.56 ± 0.04	0.58 ± 0.06	4%
Tryptophan	182 ± 1	206 ± 4	13%
Phenylalanine	330 ± 4	230 ± 19	30%
Isoleucine	128 ± 2	99 ± 1	23%
Lysine	430 ± 8	365 ± 6	15%
4-Hydroxybenzoic acid	0.64 ± 0.01	0.65 ± 0.01	2%
Desaminotyrosine	0.25 ± 0.02	0.24 ± 0.01	4%
Histidine	254 ± 2	211 ± 4	17%
Pyrocatechol	0.0041 ± 0.0005	0.0041 ± 0.0005	0%

3.4 Conclusions

We have developed a method of generating quantitative peak ratio data using chromatographic peak areas of a peak pair in CIL LC-MS. A mass spectral peak pair found in the metabolite-intensity table generated by IsoMS and Zero-fill is searched against the raw LC-MS data to find all neighboring mass spectra where the same peak pair is continuously detected. The chromatographic peaks of the light-labeled and heavy-labeled metabolites in the pair are constructed and their peak areas are determined for peak ratio measurement. We implemented this method by developing a software program, IsoMS-Quant, for automatic peak ratio calculation. IsoMS-Quant is demonstrated to provide better precision and accuracy for both untargeted and targeted metabolic profiling work using CIL LC-MS. IsoMS-Quant, along with IsoMS and Zero-fill, forms a complete workflow for rapid processing of raw LC-MS data generated by CIL LC-MS.

Chapter 4

DnsID in MyCompoundID for Rapid Identification of Dansylated Amine- and Phenol-Containing Metabolites in LC-MS-Based Metabolomics

4.1 Introduction

Metabolite identification in liquid chromatography mass spectrometry (LC-MS)-based metabolomics remains to be one of the major analytical challenges. The first path of metabolite identification often involves mass search and/or MS/MS spectral search of a given peak against a compound library for possible match. Several compound libraries containing accurate masses and MS/MS spectra of standards have been developed.[141-143] However, there are a limited number of MS/MS spectra of metabolites available. Moreover, not all metabolites can produce a sufficient number of fragment ions for library search. Using mass search alone can lead to many possible structure candidates.[144-146] On the other hand, retention time (RT) of metabolites can be another important piece of information.[120, 147-151] However, RT can vary greatly, depending on a number of factors including instrumental setup, column type and elution conditions used. Thus, RT is not commonly used as a search parameter in a publicly available compound library. Instead, RT match is often performed at the final stage of confirming a metabolite identity using an authentic standard. By spiking a standard to a sample or running identical LC-MS conditions for the standard and sample, retention time can then be compared.[152]

We have been developing a high-performance chemical isotope labeling (CIL) LC-MS platform for metabolomics where a mass-coded chemical labeling reagent (i.e., an isotope reagent) is rationally designed to label a submetabolome of the same functional group in order to

improve LC separation, increase MS sensitivity and enhance quantification simultaneously.[45, 120, 123, 130, 138, 153, 154] For example, $^{12}C_2/^{13}C_2$-dansylation (Dns) labeling LC-MS can be used to profile amine- and phenol-containing submetabolome with much improved metabolite coverage.[120] Other labeling reagents for targeted or untargeted metabolomic profiling have also been reported from a number of research groups.[47-64] In the case of dansylation, it offers a means of reducing the great diversity of physicochemical properties of many different metabolites in a metabolome so that the labeled metabolites can be efficiently separated using a reversed phase (RP) LC column alone.[120, 155] This averts the use of different columns to analyze different groups of metabolites (e.g., ionic, hydrophilic, hydrophobic, etc).[156] Since only one type of column is used, it should be relatively easier to use the retention time information of metabolites for comparison.

However, even in RPLC-MS, retention time is sensitive to the experimental conditions used. Using the same LC-MS setup, slight changes in elution conditions can vary the retention time significantly. In this work, we report our study of applying a retention time calibration method to correct the RT drifts from one LC-MS dataset to another. This method is shown to be robust to generate a normalized RT for each dansyl labeled standard that can be used for library search. In addition to normalized RT information, MS and MS/MS spectra were obtained for individual standards. A search program, DnsID, was developed to match MS, MS/MS and RT of an unknown metabolite to those in the library for identification of labeled metabolites in dansylation LC-MS. To allow other researchers to use this resource for metabolite identification, we implemented the RT calibration and library search algorithms in a web-based interface that are freely available at the www.MyCompoundID.org website.

4.2 Experimental Section

4.2.1 Construction of Dns-library

The current Dns-library contains a total of 315 Dns-compounds from 273 unique metabolites (see Table T 4.1 for the complete list). To build this library, each metabolite standard was individually labeled using the dansylation labeling protocol previously published.[130] The final concentration of the dansylated standard for LC-MS analysis was about 10 μM after diluting by 0.1% (v/v) formic acid in 9:1 (v/v) H_2O/ACN. All the dansylated standards were analyzed by LC-QTOF (see below) to produce MS, RT and MS/MS information. A mixture of RT calibrants (see below) was run every 10 injections of different Dns-standards to correct for any significant RT drifts during the data collection process.

Table 4.1 The complete Dns-compounds list in the Dns-library

ID	HMDB No.	Name	Acc. mass	mz_light	Normalized RT (min)	Tag No.	Charge No.
1	HMDB00001	1-Methylhistidine	169.0851	403.1434	2.17	1	1
2	HMDB00002	1,3-Diaminopropane	74.0844	308.1427	2.63	1	1
3	HMDB00002_2	1,3-Diaminopropane - multi-tags	74.0844	271.0583	20.49	2	2
4	HMDB00020	p-Hydroxyphenylacetic acid	152.0473	386.1057	16.91	1	1
5	HMDB00021	Iodotyrosine	306.9705	387.5436	23.88	2	2
6	HMDB00022	3-Methoxytyramine	167.0946	317.6056	25.49	2	2
7	HMDB00045	Adenosine monophosphate	347.0631	581.1214	1.75	1	1
8	HMDB00050	Adenosine	267.0968	501.1551	3.94	1	1
9	HMDB00051	Ammonia	17.0266	251.0849	5.82	1	1
10	HMDB00056	Beta-Alanine	89.0477	323.1060	7.24	1	1
11	HMDB00064	Creatine	131.0695	365.1278	3.02	1	1
12	HMDB00070	D-Pipecolic acid	129.0790	363.1373	13.23	1	1
13	HMDB00085	Deoxyguanosine	267.0968	501.1551	8.49	1	1
14	HMDB00087	Dimethylamine	45.0578	279.1162	15.07	1	1
15	HMDB00089	Cytidine	243.0855	477.1438	5.87	1	1
16	HMDB00089_2	Cytidine - H2O	243.0855	459.1333	7.38	1	1
17	HMDB00095	Cytidine monophosphate	323.0519	557.1102	1.88	1	1
18	HMDB00095_2	Cytidine monophosphate - Isomer	323.0519	557.1102	2.87	1	1
19	HMDB00099	L-Cystathionine	222.0674	345.0920	13.34	2	2
20	HMDB00099_2	L-Cystathionine - Isomer	222.0674	345.0920	13.69	2	2
21	HMDB00101	Deoxyadenosine	251.1018	485.1602	8.72	1	1
22	HMDB00112	Gamma-Aminobutyric acid	103.0633	337.1216	7.79	1	1
23	HMDB00112_2	Gamma-Aminobutyric acid - H2O	103.0633	319.1144	13.57	1	1
24	HMDB00118	Homovanillic acid	182.0579	416.1162	16.51	1	1
25	HMDB00123	Glycine	75.0320	309.0903	6.59	1	1
26	HMDB00128	Guanidoacetic acid	117.0538	351.1121	2.74	1	1

27	HMDB00130	Homogentisic acid	168.0423	318.0794	24.84	2	2
28	HMDB00133	Guanosine	283.0917	517.1500	2.22	1	1
29	HMDB00148	L-Glutamic Acid	147.0532	381.1115	5.05	1	1
30	HMDB00148_2	L-Glutamic Acid - H2O	147.0532	363.1009	9.46	1	1
31	HMDB00149	Ethanolamine	61.0528	295.1111	6.00	1	1
32	HMDB00152	Gentisic acid	154.0266	388.0849	17.11	1	1
33	HMDB00152_2	Gentisic acid - multi-tags	154.0266	311.0716	24.69	2	2
34	HMDB00153	Estriol	288.1725	522.2309	20.36	1	1
35	HMDB00157	Hypoxanthine + H2O	136.0385	388.1098	2.12	1	1
36	HMDB00157_2	Hypoxanthine - multi-tags	136.0385	370.0968	8.73	1	1
37	HMDB00157_3	Hypoxanthine - Isomer	136.0385	370.0968	9.65	1	1
38	HMDB00158	L-Tyrosine	181.0739	324.5953	22.65	2	2
39	HMDB00159	L-Phenylalanine	165.0790	399.1373	12.74	1	1
40	HMDB00161	L-Alanine	89.0477	323.1060	7.57	1	1
41	HMDB00162	L-Proline	115.0633	349.1216	10.18	1	1
42	HMDB00164	Methylamine	31.0422	265.1005	9.82	1	1
43	HMDB00167	L-Threonine	119.0582	353.1166	5.79	1	1
44	HMDB00168	L-Asparagine	132.0535	366.1118	3.00	1	1
45	HMDB00168_2	L-Asparagine - H2O	132.0535	348.1070	6.40	1	1
46	HMDB00172	L-Isoleucine	131.0946	365.1529	13.06	1	1
47	HMDB00177	L-Histidine	155.0695	389.1278	18.09	1	1
48	HMDB00182	L-Lysine	146.1055	307.1111	17.47	2	2
49	HMDB00187	L-Serine	105.0426	339.1009	4.40	1	1
50	HMDB00191	L-Aspartic Acid	133.0375	367.0958	5.16	1	1
51	HMDB00192	L-Cystine	240.0238	354.0702	14.11	2	2
52	HMDB00206	N6-Acetyl-L-Lysine	188.1161	422.1744	5.71	1	1
53	HMDB00210	Pantothenic acid	219.1107	453.1690	8.37	1	1
54	HMDB00214	Ornithine	132.0899	300.1033	16.58	2	2
55	HMDB00224	O-Phosphoethanolamine	141.0191	375.0774	2.02	1	1
56	HMDB00228	Phenol	94.0419	328.1002	23.16	1	1

57	HMDB00238	Sepiapterin	237.0862	471.1445	10.14	1	1
58	HMDB00239	Pyridoxine	169.0739	403.1322	10.12	1	1
59	HMDB00239_2	Pyridoxine - H2O	169.0739	385.1243	18.01	1	1
60	HMDB00251	Taurine	125.0147	359.0730	2.24	1	1
61	HMDB00259	Serotonin	176.0950	322.1058	24.65	2	2
62	HMDB00262	Thymine	126.0429	360.1012	13.21	1	1
63	HMDB00265	Liothyronine	650.7900	884.8484	19.14	1	1
64	HMDB00271	Sarcosine	89.0477	323.1060	9.34	1	1
65	HMDB00279	Saccharopine	276.1321	510.1905	2.26	1	1
66	HMDB00279_2	Saccharopine - H2O	276.1321	492.1799	5.65	1	1
67	HMDB00291	Vanillylmandelic acid	198.0528	432.1111	12.81	1	1
68	HMDB00291_2	Vanillylmandelic acid - H2O	198.0528	414.1005	21.31	1	1
69	HMDB00292	Xanthine	152.0334	386.0917	8.95	1	1
70	HMDB00296	Uridine	244.0695	478.1279	7.84	1	1
71	HMDB00296_2	Uridine - H2O	244.0695	460.1173	8.67	1	1
72	HMDB00300	Uracil	112.0273	346.0856	11.34	1	1
73	HMDB00301	Urocanic acid	138.0429	372.1012	13.52	1	1
74	HMDB00303	Tryptamine	160.1000	394.1584	18.03	1	1
75	HMDB00306	Tyramine	137.0841	302.6004	25.83	2	2
76	HMDB00356	17-Epiestriol	288.1725	522.2309	23.93	1	1
77	HMDB00370	2-Amino-3-phosphonopropionic acid	169.0140	403.0723	1.69	1	1
78	HMDB00397	2-Pyrocatechuic acid	154.0266	388.0849	16.31	1	1
79	HMDB00440	3-Hydroxyphenylacetic acid	152.0473	386.1057	16.72	1	1
80	HMDB00446	N-Alpha-acetyllysine	188.1161	422.1744	6.79	1	1
81	HMDB00450	5-Hydroxylysine	162.1004	315.1085	13.88	2	2
82	HMDB00452	L-Alpha-aminobutyric acid	103.0633	337.1216	9.13	1	1
83	HMDB00455	Allocystathionine	222.0674	345.0920	13.33	2	2
84	HMDB00455_2	Allocystathionine - Isomer	222.0674	345.0920	13.61	2	2
85	HMDB00468	Biopterin	237.0862	471.1445	6.03	1	1
86	HMDB00469	5-Hydroxymethyluracil	142.0378	376.0962	8.87	1	1

87	HMDB00473	6-Dimethylaminopurine	163.0858	397.1441	18.56	1	1
88	HMDB00479	3-methyl-histidine	169.0851	403.1434	2.01	1	1
89	HMDB00484	Vanillic acid	168.0423	402.1006	17.34	1	1
90	HMDB00500	4-Hydroxybenzoic acid	138.0317	372.0900	17.57	1	1
91	HMDB00504	5-Hydroxy-L-tryptophan	220.0848	344.1007	20.29	1	1
92	HMDB00510	Aminoadipic acid	161.0688	395.1271	5.97	1	1
93	HMDB00517	L-Arginine	174.1117	408.1700	2.44	1	1
94	HMDB00557	L-Alloisoleucine	131.0946	365.1529	13.20	1	1
95	HMDB00574	Cysteine	121.0197	355.0781	14.12	1	1
96	HMDB00592	Glucosamine 6-sulfate	259.0362	493.0945	1.79	1	1
97	HMDB00615	Epinephrine	183.0895	417.1479	6.19	1	1
98	HMDB00615_2	Epinephrine - Isomer	183.0895	417.1479	7.20	1	1
99	HMDB00615_3	Epinephrine - Isomer	183.0895	417.1479	8.59	1	1
100	HMDB00630	Cytosine	111.0433	345.1016	7.58	1	1
101	HMDB00641	L-Glutamine	146.0691	380.1275	3.32	1	1
102	HMDB00650	D-Alpha-aminobutyric acid	103.0633	337.1216	9.23	1	1
103	HMDB00667	L-Thyroine	273.1001	370.6084	25.44	2	2
104	HMDB00669	Ortho-Hydroxyphenylacetic acid	152.0473	386.1057	16.42	1	1
105	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.00	1	1
106	HMDB00676	L-Homocystine	268.0551	368.0859	15.82	2	2
107	HMDB00679	Homocitrulline	189.1113	423.1697	4.47	1	1
108	HMDB00684	L-Kynurenine	208.0848	442.1431	11.44	1	1
109	HMDB00684_2	L-Kynurenine - H2O	208.0848	424.1325	11.97	1	1
110	HMDB00687	L-leucine	131.0946	365.1529	13.36	1	1
111	HMDB00696	L-Methionine	149.0510	383.1094	10.89	1	1
112	HMDB00704	Isoxanthopterin	179.0443	413.1026	9.55	1	1
113	HMDB00704_2	Isoxanthopterin - Isomer	179.0443	413.1026	10.82	1	1
114	HMDB00706	L-Aspartyl-L-phenylalanine	280.1059	514.1642	10.07	1	1
115	HMDB00714	Hippuric acid	179.0582	413.1166	7.07	1	1
116	HMDB00716	L-Pipecolic acid	129.0790	363.1373	13.45	1	1

117	HMDB00719	L-Homoserine	119.0582	353.1166	4.05	1	1
118	HMDB00719_2	L-Homoserine - H2O	119.0582	335.1060	9.26	1	1
119	HMDB00721	Glycylproline	172.0848	406.1431	7.17	1	1
120	HMDB00725	Trans-4-Hydroxyl-L-Proline	131.0582	365.1166	5.17	1	1
121	HMDB00734	Indoleacrylic acid	187.0633	421.1216	20.61	1	1
122	HMDB00750	3-Hydroxymandelic acid	168.0423	402.1006	12.94	1	1
123	HMDB00750_2	3-Hydroxymandelic acid - COOH	168.0423	356.0951	21.64	1	1
124	HMDB00755	Hydroxyphenyllactici acid	182.0579	416.1162	14.39	1	1
125	HMDB00759	Glycyl-L-Leucine	188.1161	422.1744	11.22	1	1
126	HMDB00763	5-Hydroxyindoleacetic acid	191.0582	425.1166	15.09	1	1
127	HMDB00819	Normetanephrine	183.0895	325.6031	23.41	2	2
128	HMDB00840	Salicyluric acid	195.0532	429.1115	11.05	1	1
129	HMDB00881	Xanthurenic acid	205.0375	439.0958	9.06	1	1
130	HMDB00881_2	Xanthurenic acid - multi-tags	205.0375	629.0732	26.34	2	1
131	HMDB00883	L-Valine	117.0790	351.1373	10.81	1	1
132	HMDB00884	Ribothymidine	258.0852	492.1435	5.85	1	1
133	HMDB00884_2	Ribothymidine - Isomer	258.0852	492.1435	8.54	1	1
134	HMDB00884_3	Ribothymidine - H2O	258.0852	474.1329	9.39	1	1
135	HMDB00897	7-Methylguanine	165.0651	399.1234	10.32	1	1
136	HMDB00904	Citrulline	175.0957	409.1540	3.74	1	1
137	HMDB00905	Deoxyadenosine monophosphate	331.0682	565.1265	4.58	1	1
138	HMDB00929	L-Tryptophan	204.0899	438.1482	11.44	1	1
139	HMDB00939	S-Adenosylhomocysteine	384.1216	426.1191	10.52	2	2
140	HMDB00954	trans-Ferulic acid	194.0579	428.1162	18.47	1	1
141	HMDB00955	Isoferulic acid	194.0579	428.1162	17.49	1	1
142	HMDB00957	pyrocatechol	110.0368	289.0767	26.70	2	2
143	HMDB00965	Hypotaurine	109.0197	343.0781	2.47	1	1
144	HMDB00982	5-Methylcytidine	257.1012	491.1595	2.94	1	1
145	HMDB00982_2	5-Methylcytidine - Isomer	257.1012	491.1595	5.98	1	1
146	HMDB00991	2-aminooctanoic acid	159.1259	393.1842	19.20	1	1

147	HMDB01044	2'-Deoxyguanosine 5'-monophosphate	347.0631	581.1214	5.57	1	1
148	HMDB01049	Gamma-Glutamylcysteine	250.0623	483.1211	8.62	1	1
149	HMDB01065	2-Hydroxyphenethlamine	137.0841	371.1424	10.77	1	1
150	HMDB01065_2	2-Hydroxyphenethlamine - Isomer	137.0841	371.1424	13.77	1	1
151	HMDB01069	2-Phenylaminoadenosine	358.1390	592.1973	8.73	1	1
152	HMDB01123	2-Aminobenzoic acid	137.0477	371.1060	16.62	1	1
153	HMDB01149	5-Aminolevulinic acid	131.0582	365.1166	7.59	1	1
154	HMDB01169	4-Aminophenol	109.0528	343.1111	16.30	1	1
155	HMDB01169_2	4-Aminophenol - multi-tags	109.0528	288.5847	25.04	2	2
156	HMDB01173	5'-Methylthioadenosine	297.0896	531.1479	6.97	1	1
157	HMDB01202	dCMP	307.0569	541.1153	4.66	1	1
158	HMDB01232	4-Nitrophenol	139.0269	373.0853	23.45	1	1
159	HMDB01238	N-Acetylserotonin	218.1055	452.1638	14.32	1	1
160	HMDB01254	Glucosamine 6-phosphate	259.0457	493.1040	1.60	1	1
161	HMDB01257	Spermidine	145.1579	306.6373	10.54	2	2
162	HMDB01336	3,4-Dihydroxybenzeneacetic acid	168.0423	318.0794	23.90	2	2
163	HMDB01341	ADP	427.0294	661.0877	1.49	1	1
164	HMDB01370	Diaminopimelic acid	190.0954	329.1060	12.30	2	2
165	HMDB01370_2	Diaminopimelic acid - Isomer	190.0954	329.1060	12.96	2	2
166	HMDB01392	p-Aminobenzoic acid	137.0477	371.1060	11.52	1	1
167	HMDB01397	Guanosine monophosphate	363.0580	597.1163	1.15	1	1
168	HMDB01398	Guaiacol	124.0524	358.1107	22.54	1	1
169	HMDB01413	Citicoline	488.1073	722.1657	1.48	1	1
170	HMDB01413_2	Citicoline - Isomer	488.1073	722.1657	1.28	1	1
171	HMDB01414	1,4-diaminobutane	88.1000	278.1083	21.27	2	2
172	HMDB01431	Pyridoxamine	168.0899	318.1033	19.47	2	2
173	HMDB01432	Agmatine	130.1218	364.1802	4.52	1	1
174	HMDB01432_2	Agmatine - multi-tags	130.1218	299.1192	15.28	2	2
175	HMDB01476	3-Hydroxyanthranilic acid	153.0426	387.1009	18.14	1	1
176	HMDB01522	Methylguanidine	73.0640	307.1223	3.84	1	1

177	HMDB01522_2	Methylguanidine - multi-tags	73.0640	270.5903	22.52	2	2
178	HMDB01525	Imidazole	68.0374	302.0958	14.29	1	1
179	HMDB01545	Pyridoxal	167.0582	401.1166	12.01	1	1
180	HMDB01645	L-Norleucine	131.0946	365.1529	14.11	1	1
181	HMDB01713	m-Coumaric acid	164.0473	398.1057	18.51	1	1
182	HMDB01833	Aminopterin	440.1557	674.2140	6.06	1	1
183	HMDB01842	Guanidine	59.0483	293.1067	3.00	1	1
184	HMDB01849	Propranolol	259.1572	493.2155	24.95	1	1
185	HMDB01855	5-Hydroxytryptophol	177.0790	411.1373	15.55	1	1
186	HMDB01856	Protocatechuic acid	154.02661	311.0716	24.51	2	2
187	HMDB01858	p-Cresol	108.0575	342.1158	24.54	1	1
188	HMDB01859	Acetaminophen	151.0633	385.1216	16.35	1	1
189	HMDB01861	3-Methylhistamine	125.0953	359.1536	3.27	1	1
190	HMDB01866	3,4-Dihydroxymandelic acid	184.0372	326.0769	21.73	2	2
191	HMDB01867	4-Aminohippuric acid	194.0691	428.1275	8.38	1	1
192	HMDB01868	5-Methoxysalicylic acid	168.0423	402.1006	16.38	1	1
193	HMDB01885	3-Chlorotyrosine	215.0349	341.5758	23.57	2	2
194	HMDB01889	Theophylline	180.0647	414.1230	15.42	1	1
195	HMDB01891	m-Aminobenzoic acid	137.0477	371.1060	11.76	1	1
196	HMDB01894	Aspartame	294.1216	528.1799	13.72	1	1
197	HMDB01895	Salicylic acid	138.0317	372.0900	15.62	1	1
198	HMDB01901	Aminocaproic acid	131.0946	365.1529	10.21	1	1
199	HMDB01904	3-Nitrotyrosine	226.0590	335.1080	22.61	2	2
200	HMDB01904_2	3-Nitrotyrosine - H2O	226.0590	347.0878	22.61	2	2
201	HMDB01906	2-Aminoisobutyric acid	103.0633	337.1216	8.91	1	1
202	HMDB01915	Alendronic acid	249.0167	483.0750	1.59	1	1
203	HMDB01918	Thyroxine	776.6867	622.4017	27.74	2	2
204	HMDB01924	Atenolol	266.1630	500.2214	11.91	1	1
205	HMDB01932	Metoprolol	267.1834	501.2418	22.09	1	1
206	HMDB01937	Salbutamol	239.1521	473.2105	8.27	1	1

207	HMDB01937_2	Salbutamol - H2O	239.1521	455.1999	8.42	1	1
208	HMDB01938	Lisinopril	405.2264	320.1460	7.68	1	2
209	HMDB01942	Phenylpropanolamine	151.0997	385.1580	15.13	1	1
210	HMDB01943	Pseudoephedrine	165.1154	399.1737	19.38	1	1
211	HMDB01964	Caffeic acid	180.0423	324.0800	24.64	1	1
212	HMDB01972	3-Aminosalicylic acid	153.0426	387.1009	12.22	1	1
213	HMDB02005	Methionine Sulfoxide	165.0460	399.1043	3.72	1	1
214	HMDB02005_2	Methionine Sulfoxide - Isomer	165.0460	399.1043	4.20	1	1
215	HMDB02006	2,3-Diaminoproprionic acid	104.0586	286.0876	15.65	2	2
216	HMDB02017	1-Phenylethylamine	121.0891	355.1475	18.63	1	1
217	HMDB02024	Imidazoleacetic acid	126.0429	360.1012	11.12	1	1
218	HMDB02048	m-Cresol	108.0575	342.1158	24.44	1	1
219	HMDB02055	o-Cresol	108.0575	342.1158	24.60	1	1
220	HMDB02064	N-Acetylputrescine	130.1106	364.1689	7.25	1	1
221	HMDB02085	Syringic acid	198.0528	432.1111	18.10	1	1
222	HMDB02099	6-Methyladenine	149.0701	383.1285	12.22	1	1
223	HMDB02099_2	6-Methyladenine - Isomer	149.0701	383.1285	12.74	1	1
224	HMDB02108	Methylcysteine	135.0354	369.0937	9.37	1	1
225	HMDB02128	2,4-Diamino-6-hydroxypyrimidine	126.0542	360.1125	8.95	1	1
226	HMDB02141	N-Methyl-a-aminoisobutyric acid	117.0790	351.1373	13.65	1	1
227	HMDB02182	Phenylephrine	167.0946	317.6056	25.39	2	2
228	HMDB02199	Desaminotyrosine	166.0630	400.1213	18.04	1	1
229	HMDB02205	L-Homocysteic acid	183.0201	417.0785	1.64	1	1
230	HMDB02210	2-Phenylglycine	151.0633	385.1216	11.69	1	1
231	HMDB02322	Cadaverine	102.1157	285.1162	22.39	2	2
232	HMDB02339	5-Methoxytryptophan	234.1004	468.1588	9.79	1	1
233	HMDB02362	2,4-Diaminobutyric acid	118.0742	293.0954	15.80	2	2
234	HMDB02390	3-Cresotinic acid	152.0473	386.1057	16.80	1	1
235	HMDB02393	N-methyl-D-aspartic acid	147.0532	381.1115	7.53	1	1
236	HMDB02658	6-Hydroxynicotinic acid	139.0269	373.0853	12.21	1	1

237	HMDB02658_2	6-Hydroxynicotinic acid - Isomer	139.0269	373.0853	15.32	1	1
238	HMDB02670	Naringenin	272.0685	370.0926	15.29	2	2
239	HMDB02706	Canavanine	176.0909	322.1038	11.37	2	2
240	HMDB02706_2	Canavanine - Isomer	176.0909	322.1038	11.59	2	2
241	HMDB02991	Cysteamine	77.0299	311.0882	15.08	1	1
242	HMDB03012	Aniline	93.0578	327.1162	17.32	1	1
243	HMDB03134	Biocytin	372.1831	606.2414	6.72	1	1
244	HMDB03157	Guanidinosuccinic acid	175.0593	409.1176	2.81	1	1
245	HMDB03164	Chlorogenic acid	354.0951	411.1059	19.45	2	2
246	HMDB03164_2	Chlorogenic acid - Isomer	354.0951	411.1059	21.24	2	2
247	HMDB03282	1-Methylguanine	165.0651	399.1234	9.57	1	1
248	HMDB03320	Indole-3-carboxylic acid	161.0477	395.1060	19.27	1	1
249	HMDB03334	Symmetric dimethylarginine	202.1430	436.2013	3.05	1	1
250	HMDB03337	Oxidized glutathione	612.1520	540.1343	8.07	2	2
251	HMDB03338	Hydroxylamine	33.0215	267.0798	12.02	1	1
252	HMDB03355	5-Aminopentanoic acid	117.0790	351.1373	8.68	1	1
253	HMDB03423	D-Glutamine	146.0691	380.1275	3.32	1	1
254	HMDB03431	L-Histidinol	141.0902	375.1485	1.44	1	1
255	HMDB03464	4-Guanidinobutanoic acid	145.0851	379.1434	3.65	1	1
256	HMDB03464_2	4-Guanidinobutanoic acid - H2O	145.0851	361.1329	11.00	1	1
257	HMDB03474	3,5-Diiodo-L-tyrosine	432.8672	666.9255	11.82	1	1
258	HMDB03474_2	3,5-Diiodo-L-tyrosine - multi-tags	432.8672	450.4919	24.18	2	2
259	HMDB03640	Beta-Leucine	131.0946	365.1529	10.78	1	1
260	HMDB03640_2	Beta-Leucine - H2O	131.0946	347.1424	20.77	1	1
261	HMDB03911	3-Aminoisobutanoic acid	103.0633	337.1216	8.67	1	1
262	HMDB03911_2	3-Aminoisobutanoic acid - H2O	103.0633	319.1110	16.29	1	1
263	HMDB04095	5-Methoxytryptamine	190.1106	424.1689	16.52	1	1
264	HMDB04122	Selenocystine	335.9128	402.0147	15.09	2	2
265	HMDB04437	Diethanolamine	105.0790	339.1373	5.49	1	1
266	HMDB04811	2,4-Dichlorophenol	161.9639	396.0222	26.30	1	1

267	HMDB04815	4-Hydroxy-3-methylbenzoic acid	152.0473	386.1057	19.43	1	1
268	HMDB04987	Alpha-Aspartyl-lysine	261.1325	364.6246	13.61	2	2
269	HMDB04992	Benzocaine	165.0790	399.1373	20.08	1	1
270	HMDB06050	o-Tyrosine	181.0739	324.5953	22.38	2	2
271	HMDB11177	L-phenylalanyl-L-proline	262.1317	496.1901	13.13	1	1
272	HMDB11737	Gamma Glutamylglutamic acid	276.0958	510.1541	3.44	1	1
273	HMDB13243	Leucyl-phenylalanine	278.1630	512.2214	16.59	1	1
274	HMDB13302	Phenylalanylphenylalanine	312.1474	546.2057	16.55	1	1
275	HMDB28689	Alanyl-Histidine	226.1066	460.1649	17.62	1	1
276	HMDB28691	Alanyl-Leucine	202.1317	436.1901	11.36	1	1
277	HMDB28694	Alanyl-Phenylalanine	236.1161	470.1744	12.11	1	1
278	HMDB28698	Alanyl-Tryptophan	275.1270	509.1853	11.09	1	1
279	HMDB28699	Alanyl-Tyrosine	252.1110	360.1138	21.85	2	2
280	HMDB28716	Arginyl-Phenylalanine	321.1801	555.2384	6.80	1	1
281	HMDB28844	Glycyl-Isoleucine	188.1161	422.1744	10.78	1	1
282	HMDB28848	Glycyl-Phenylalanine	222.1004	456.1588	11.65	1	1
283	HMDB28852	Glycyl-Tryptophan	261.1113	495.1697	11.19	1	1
284	HMDB28853	Glycyl-Tyrosine	238.0954	353.1060	21.63	2	2
285	HMDB28854	Glycyl-Valine	174.1004	408.1588	9.19	1	1
286	HMDB28878	Histidinyl-Alanine	226.1066	460.1649	16.69	1	1
287	HMDB28937	Leucyl-Proline	228.1474	462.2057	12.99	1	1
288	HMDB28940	Leucyl-Tryptophan	317.1739	551.2323	15.77	1	1
289	HMDB28941	Leucyl-Tyrosine	294.1580	381.1373	23.98	1	1
290	HMDB28988	Phenylalanyl-Alanine	236.1161	470.1744	10.58	1	1
291	HMDB28995	Phenylalanyl-Glycine	222.1004	456.1588	9.43	1	1
292	HMDB29001	Phenylalanyl-Methionine	296.1195	530.1778	13.87	1	1
293	HMDB29007	Phenylalanyl-Tyrosine	328.1423	398.1295	24.22	2	2
294	HMDB29008	Phenylalanyl-Valine	264.1474	498.2057	13.62	1	1
295	HMDB29043	Serinyl-Leucine	218.1267	452.1850	8.90	1	1
296	HMDB29046	Serinyl-Phenylalanine	252.1110	486.1693	9.38	1	1

297	HMDB29065	Threoninyl-Leucine	232.1423	466.2006	10.18	1	1
298	HMDB29082	Tryptophyl-Glutamate	332.1246	566.1830	8.29	1	1
299	HMDB29087	Tryptophyl-Leucine	317.1739	551.2323	14.60	1	1
300	HMDB29090	Tryptophyl-Phenylalanine	351.1583	585.2166	15.36	1	1
301	HMDB29095	Tryptophyl-Tyrosine	367.1532	417.6349	23.25	1	1
302	HMDB29098	Tyrosyl-Alanine	252.1110	360.1138	20.86	2	2
303	HMDB29105	Tyrosyl-Glycine	238.0954	353.1060	20.19	2	2
304	HMDB29109	Tyrosyl-Leucine	294.1580	381.1373	23.77	2	2
305	HMDB29118	Tyrosyl-Valine	280.1423	374.1295	22.83	2	2
306	HMDB29306	4-Ethylphenol	122.0732	356.1315	25.63	1	1
307	HMDB59964	2,3,4-Trihydroxybenzoic acid	170.0215	319.0691	24.10	2	2
308	HMDB59966	3,5-Dimethoxyphenol	154.0630	388.1213	23.75	1	1
309	HMDB60003	Isovanillic acid	168.0423	402.1006	15.69	1	1
310	312	Gly-Gly-Gly-Gly	246.0964	480.1548	3.39	1	1
311	313	Trp-Gly-Gly	318.1328	552.1911	7.60	1	1
312	314	Gly-Norvaline	174.1005	408.1588	9.51	1	1
313	315	Gly-Norleucine	188.1161	422.6993	11.36	1	1
314	316	Phenyl-Leucine	278.1631	512.2214	15.90	1	1
315	317	Phe-Phe-Phe	459.2158	693.2741	19.95	1	1

4.2.2 LC-MS and MS/MS

Several LC-MS setups were used in this work. Detailed information on the LC and MS settings is provided in "Supplemental Note for Dns-lib Instrumental Settings" in the Appendix.

To construct the dansyl library, an individual ^{12}C-dansyl labeled standard was injected into a Bruker HD Impact QTOF mass spectrometer (Billerica, MA, USA) with electrospray ionization (ESI) linked to an Agilent 1100 HPLC system (Palo Alto, CA, USA). Reversed-phase Zorbax Eclipse C18 column (2.0 mm × 100 mm, 1.7 μm particle size, 95 Å pore size) from Agilent was used.

For human urine sample analysis, two equal aliquots were taken from the same urine collected from a healthy individual with the approval of the University of Alberta Ethics Board and labeled separately by ^{12}C- and ^{13}C-dansylation. The labeled samples were mixed in 1:1 and centrifuged at 20,800 *g* for 10 min before injecting into LC-QTOF-MS for analysis. IsoMS was used to process the data including adduct ion filtering to retain only $[M+H]^+$ ions to generate unique peak pairs from individual labeled metabolites.[123] While it is not the focus of this work, both relative and absolute quantification of dansyl labeled metabolites can be performed using differential isotope labeling.[157]

To validate the performance of the RT correction algorithm, a mixture of Dns-metabolites was prepared using 20 Dns-standards (see Table T 4.3 for the list) and analyzed using the Bruker LC-QTOF-MS as well as the Bruker 9.4-T Fourier-transform Ion Cyclotron Resonance (FTICR) MS.

To examine the differences of fragmentation patterns obtained using different tandem mass spectrometers, MS/MS spectra of several Dns-metabolites were also collected using a QTRAP 4000 mass spectrometer (AB Sciex, Foster City, CA).

4.2.3 Retention Time Calibration

Table 4.1 lists the 22 Dns-standards selected according to their elution times in a typical dansylation RPLC-MS run. They were mixed in equal moles and served as the retention time calibrants (RTcal). They are generally very stable and can be stored at -20 °C for a year or longer without degradation.

4.2.4 DnsID for Metabolite Identification

Identification of labeled metabolites in a sample in a user's laboratory is done in two steps. The first step is to run the RTcal mixture in LC-MS to produce the retention time information for all the calibrants. The next step is to run a real sample under the same LC-MS conditions as those used for running the RTcal. The two data files are then uploaded to the DnsID program which is hosted at the MyCompoundID website (www.mycompoundid.org).[144] In DnsID, the retention times of all the labeled metabolites detected in the sample are first corrected using the retention time information obtained from the RTcal. The program then compares the mass and the corrected retention time of an individual unknown metabolite to those in the Dns-library for possible match. If a tandem mass spectrometer is available, MS/MS spectrum of a matched metabolite can be generated and searched against the standard MS/MS spectra in the Dns-library for further confirmation of the metabolite identity. The scoring system used for MS/MS search is the same as the one described previously for MS/MS-based di/tripeptide identification.[158]

A user manual of the DnsID program and an example of search are shown in the Appendix, respectively. These documents are also provided at the website.

4.3 Results and Discussion

4.3.1 Retention Time Calibration.

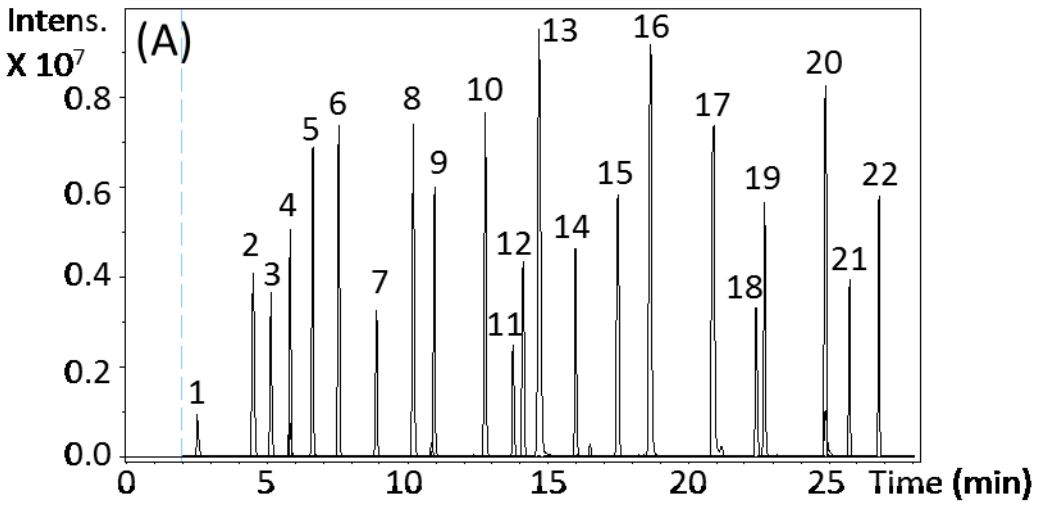
Table 4.2 lists the composition of the RTcal mixture which was carefully developed to space the individual calibrants at similar RT intervals over the whole metabolite elution window. Figure 4.1A shows the base-peak ion chromatogram of the RTcal mixture obtained using a typical dansylation LC-MS running condition. In most cases, the interval is less than 2 min, allowing the use of a simple linear calibration equation to correct the retention times of any metabolite peaks falling within a given short interval. There are a few unlabeled low-intensity peaks in Figure 4.1A from the impurities (e.g., a peak at RT 10.9 with m/z 355.1458 which was a natural isotopic peak of an unknown peak pair at m/z 351.1396 and 353.1461).

Table 4.2 A list of dansyl labeled standards used for retention time calibration (i.e., RT calibrants).

	HMDB No.	Name	mz_light	mz_heavy	RT (min)
1	HMDB00517	Dns-Arginine	408.1700	410.1767	2.44
2	HMDB00187	Dns-Serine	339.1009	341.1076	4.40
3	HMDB00148	Dns-Glutamic acid	381.1115	383.1182	5.05
4	HMDB00167	Dns-Threonine	353.1166	355.1233	5.79
5	HMDB00123	Dns-Glycine	309.0903	311.0970	6.59
6	HMDB00161	Dns-Alanine	323.1060	325.1127	7.57
7	HMDB01906	Dns-2-Aminoisobutyric acid	337.1216	339.1283	8.91
8	HMDB00162	Dns-Proline	349.1216	351.1283	10.18
9	HMDB00696	Dns-Methionine	383.1094	385.1161	10.89
10	HMDB00159	Dns-Phenylalanine	399.1373	401.1440	12.74
11	HMDB01894	Dns-Aspartame	528.1799	530.1866	13.72

12	HMDB00192	Dns-Cystine	354.0702	356.0769	14.11
13	HMDB00087	Dns-Dimethylamine	279.1162	283.1308	14.67
14	HMDB00676	Dns-Homocystine	368.0859	370.0926	15.82
15	HMDB00182	Dns-Lysine	307.1111	309.1178	17.47
16	HMDB02017	Dns-1-phenylethylamine	355.1475	357.1542	18.63
17	HMDB29098	Dns-Tyrosyl-Alanine	360.1138	362.1205	20.86
18	HMDB02322	Dns-Cadaverine	285.1162	287.1229	22.39
19	HMDB00158	Dns-Tyrosine	324.5953	326.6020	22.65
20	HMDB00259	Dns-Serotonin	322.1058	324.1125	24.65
21	HMDB29306	Dns-4-Ethylphenol	356.1315	358.1382	25.63
22	HMDB00957	Dns-Pyrocatechol	289.0767	291.0834	26.70

Figure 4.1B shows the schematic display of the RT calibration method. It works by dividing the whole LC chromatogram into 23 time intervals. Except the first and last time intervals, all the other 21 intervals have each of them bracketed by two reference standards from the RTcal. Within each interval, the RT differences of two pairs of standards between the user's RTcal run and the library RTcal data are calculated (e.g., Δt_1 and Δt_2 for the time shifts of standards 1 and 2, respectively). Then, a linear RT correction is applied to calculate the RT shift (Δt) (see Figure 1B for the equation). To correct the RT shift of any peak within the interval for a real sample, the measured RT of the metabolite in the LC-MS run of the sample (t_i in Figure 1B) is subtracted by the RT shift to generate a corrected RT ($t_{i\_corrected}$). For any metabolite peaks present in the first and last time intervals, only one pair of standard is available (either the first or the last reference standard). Thus, only one RT difference ($\Delta t_{first\_standard}$ or $\Delta t_{last\_standard}$) is calculated between the user's RTcal run and the library RTcal data. The measured RT of a labeled metabolite in a sample is corrected by subtracting $\Delta t_{first\_standard}$ or $\Delta t_{last\_standard}$. After processing the chromatographic peaks at all the intervals in the sample LC-MS run, a CSV file containing the corrected RT data is created for the sample.



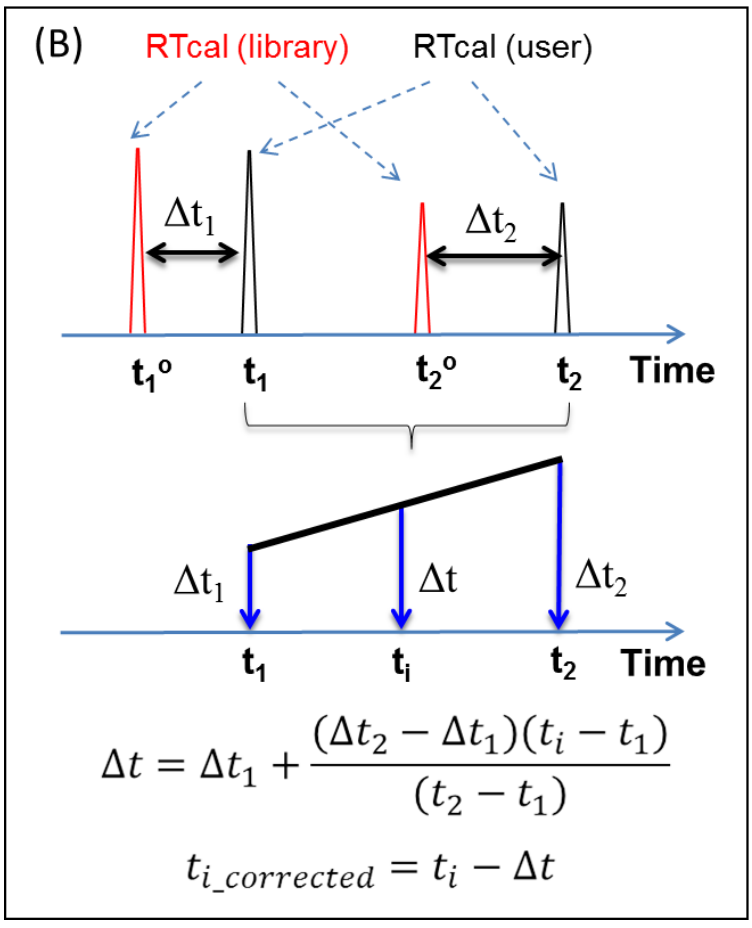


Figure 4.1(A) Base peak ion chromatogram of a mixture of 22 RT calibration standards (RTcal) (see Table 1 for the list) obtained by using RPLC-MS with a linear gradient elution. (B)

Schematic of the retention time calibration method where t_1^0 (or t_2^0) and t_1 (or t_2) refer to retention time of standard 1 (or standard 2) found in the library and the user's RTcal chromatogram, respectively, Δt_1 and Δt_2 refer to the retention time shift of the user's chromatogram from the library data for standard 1 and 2, respectively, t_i refers to the retention time of any metabolite peak resided in between the retention times of standards 1 and 2, and Δt is the retention time correction for t_i .

The performance of the RT calibration method is illustrated in Figure 4.2 where retention time correlations of different LC-MS experiments (LC-FTICR-MS and LC-QTOF-MS) before and after applying RT calibration are shown. In this case, 20 standards have been selected from the library with retention time span over the entire metabolite elution window. Two different HPLC systems with different batches of the same type of column were used. Also, the connection tubing lengths and the interfaces of LC-FTICR-MS and LC-QTOF-MS were different. Figure 4.2 shows the RT correlation plots of the 20 standards from the data obtained by LC-FTICR-MS and those in the Dns-library. Before applying the RT calibration, there is a shift to a higher RT for the LC-FTICR-MS data. The RT shift can be as large as 4.8 min (see Table T 4.3). Although the RT shift becomes smaller at the high organic elution region, the shift is still greater than 0.5 min. However, even with these large and non-linear RT variations, after applying the RT calibration, an excellent linear correlation between the corrected RT and the library RT can be obtained (R^2=0.9996). As it is shown in Table T 4.3, the RT shift after calibration for all the metabolites is below 15 s, which is the RT tolerance threshold we typically use for performing DnsID M-RT search (see below). This example illustrates that the RT calibration method is able to correct for RT shifts found in different LC-MS setups.

Table 4.3 The RT shifts of 20 standards before and after calibration

Test std	HMDB No.	Name (Dns labeled)	mz_light	Library RT	LC-FTICR-MS original RT	LC-FTICR-MS corrected RT	RT shift before calibration	RT shift after calibration
1	HMDB00251	Taurine	359.0730	2.24	7.09	2.36	4.85	0.12
2	HMDB00168	Asparagine	366.1117	3.00	7.69	2.96	4.69	-0.04
3	HMDB00641	Glutamine	380.1275	3.32	7.99	3.39	4.67	0.07
4	HMDB00679	Homocitrulline	423.1699	4.47	8.77	4.49	4.30	0.02
5	HMDB00191	Aspartic Acid	367.0958	5.16	9.31	5.14	4.15	-0.02
6	HMDB00510	Aminoadipic acid	395.1271	5.97	10.04	5.96	4.07	-0.01
7	HMDB00721	Glycylproline	406.1432	7.17	11.32	7.08	4.15	-0.09
8	HMDB00210	Pantothenic acid	453.1694	8.37	12.58	8.19	4.21	-0.18
9	HMDB00883	Valine	351.1364	10.81	15.58	10.89	4.77	0.08
10	HMDB00929	Tryptophan	438.1476	11.44	16.18	11.53	4.74	0.09
11	HMDB00172	Isoleucine	365.1512	13.06	17.75	13.28	4.69	0.22
12	HMDB00687	Leucine	365.1512	13.36	17.99	13.58	4.63	0.22
13	HMDB29087	Tryptophyl-Leucine	551.2323	14.60	19.25	14.63	4.23	0.03
14	HMDB00214	Ornithine	300.1034	16.58	20.57	16.46	3.99	-0.12
15	HMDB00177	Histidine	389.1278	18.09	21.89	18.01	3.80	-0.08
16	HMDB00954	trans-ferulic acid	428.1162	18.47	22.25	18.43	3.78	-0.04
17	HMDB01414	1,4-diaminobutane	278.1083	21.27	23.75	21.47	2.48	0.20
18	HMDB00228	Phenol	328.0993	23.16	24.84	23.36	1.73	0.20
19	HMDB00130	Homogenstic acid	318.0794	24.84	25.73	24.65	0.89	-0.19
20	HMDB00881	Xanthurenic acid	629.0730	26.34	26.87	26.37	0.53	0.03

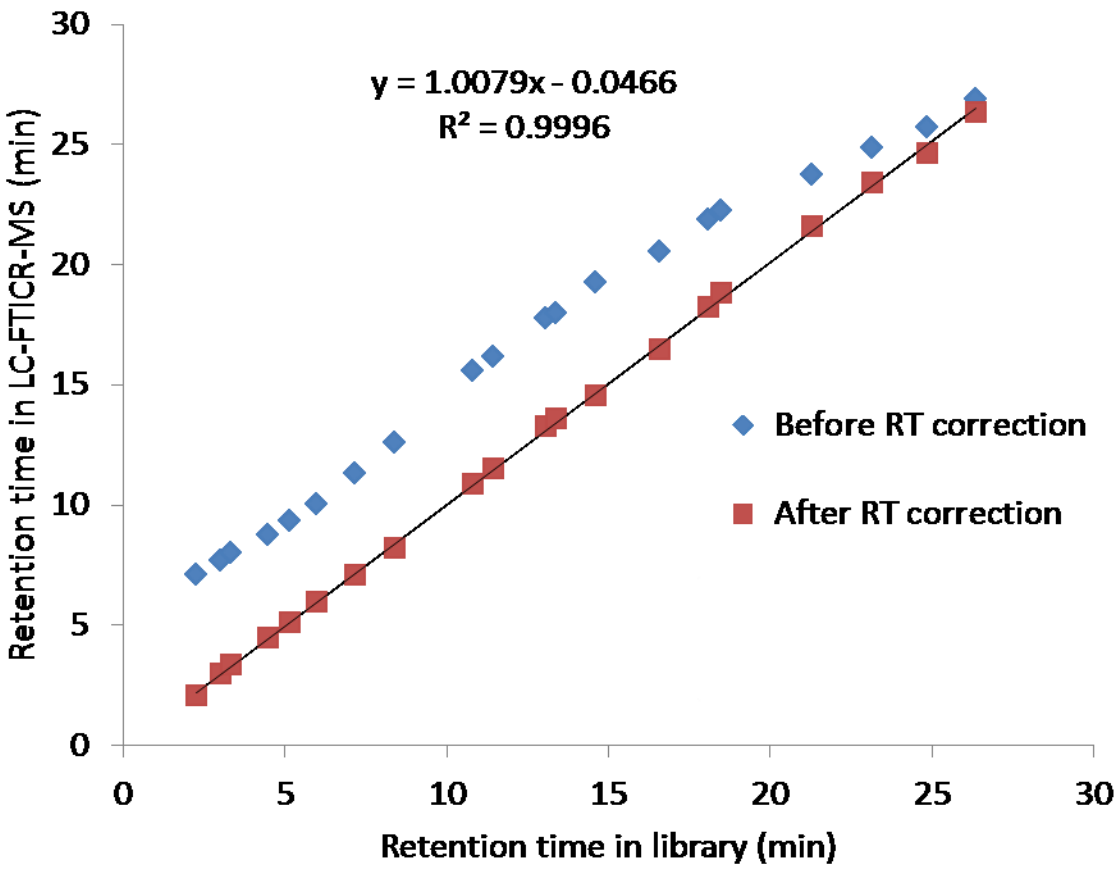


Figure 4.2 Correlation plots of the retention times of RTcal obtained by LC-FTICR-MS vs. those in the Dns-library before (in blue) and after (in Red) applying the RT calibration method.

We have also examined whether RT changes with sample matrix. In this case, we injected a RTcal mixture followed by three injections of the dansylated urine. Many of the labeled amino acids in RTcal were detected in the samples. Table T 4.4 shows the measured RTs for 10 labeled amino acids in RTcal and urine samples. The RT differences between those in urine and standard mixture are within 2.6 s, indicating that RTs were not affected by sample matrix.

Table 4.4 RTs of 10 labeled amino acids in stds and urine

	Name	mz_light	Std_1 rt (min)*	Urine_1 rt (min)*	Urine_2 rt (min)*	Urine_3 rt (min)*	Averaged urine rt (min)	Averaged rt difference (second)
1	Dns-Arginine	408.1700	2.58	2.60	2.60	2.60	2.60	1.2
2	Dns-Serine	339.1009	4.56	4.56	4.57	4.46	4.53	-1.8
3	Dns-Threonine	353.1166	5.87	5.83	5.83	5.85	5.84	-2.0
4	Dns-Alanine	323.1060	7.60	7.55	7.54	7.58	7.56	-2.6
5	Dns-Proline	349.1217	10.24	10.24	10.24	10.22	10.23	-0.4
6	Dns-Methionine	383.1094	10.99	10.99	11.01	10.98	10.99	0.2
7	Dns-Phenylalanine	399.1373	12.79	12.81	12.80	12.78	12.80	0.4
8	Dns-Cystine	354.0703	14.14	14.16	14.14	14.15	14.15	0.6
9	Dns-Lysine	307.1111	17.51	17.54	17.53	17.53	17.53	1.4
10	Dns-Tyrosine	324.5953	22.70	22.74	22.70	22.73	22.72	1.4

*un-normalized raw RT data.

4.3.2 Dns-library

The current Dns-library consists of 273 unique metabolites that have been found in biological samples related to human, according to the Human Metabolome Database (HMDB). These are mainly amines and phenols with a few other types of metabolites that can be labeled by dansylation (see Table T 4.1). They also include 39 dipeptides and 2 tripeptides; di/tripeptides are products of enzymatic reactions and degradation products of proteins and some of them have significant biological activities and functions.[158] Table T 4.5 shows a list of metabolic pathways where one or more of these metabolites belong to. These metabolites cover more than 42 metabolic pathways, offering the possibility of probing their perturbations in metabolomics studies using dansylation LC-MS. Due to multiple products that could be formed after dansylation of a metabolite containing more than one amine or phenol group, there are 314 different labeled compounds in the current library from the 273 metabolites (i.e., two products each from 8 di-amines, 27 amine-phenols and three products from 3 amine-phenols).

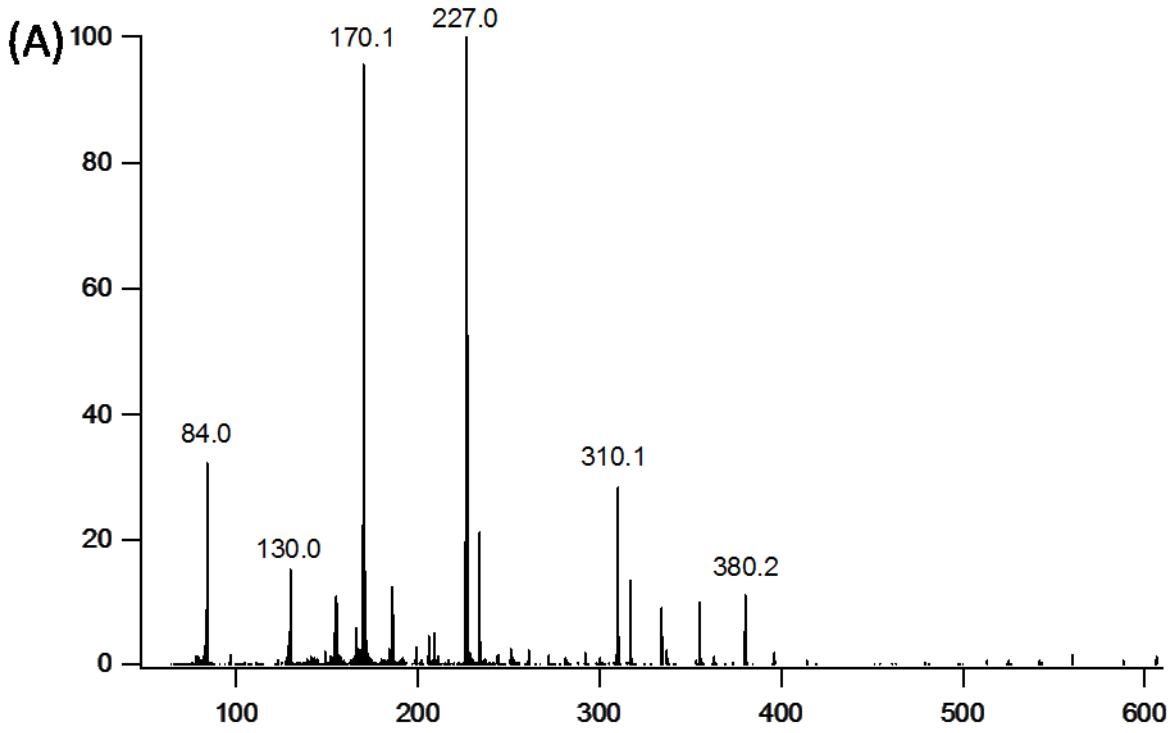
Table 4.5 List of Pathways involved

	Pathway Name
1	Nitrogen metabolism
2	Aminoacyl-tRNA biosynthesis
3	Tyrosine metabolism
4	Arginine and proline metabolism
5	Glycine, serine and threonine metabolism
6	Alanine, aspartate and glutamate metabolism
7	beta-Alanine metabolism
8	Phenylalanine metabolism
9	Glutathione metabolism
10	Cysteine and methionine metabolism
11	Pyrimidine metabolism
12	Purine metabolism

13	Lysine degradation
14	Cyanoamino acid metabolism
15	Lysine biosynthesis
16	Tryptophan metabolism
17	D-Glutamine and D-glutamate metabolism
18	Histidine metabolism
19	Phenylalanine, tyrosine and tryptophan biosynthesis
20	Pantothenate and CoA biosynthesis
21	Taurine and hypotaurine metabolism
22	D-Arginine and D-ornithine metabolism
23	Valine, leucine and isoleucine biosynthesis
24	Sulfur metabolism
25	Biotin metabolism
26	Methane metabolism
27	Glycerophospholipid metabolism
28	Valine, leucine and isoleucine degradation
29	Vitamin B6 metabolism
30	Caffeine metabolism
31	Ubiquinone and other terpenoid-quinone biosynthesis
32	Thiamine metabolism
33	Sphingolipid metabolism
34	Propanoate metabolism
35	Selenoamino acid metabolism
36	Butanoate metabolism
37	Nicotinate and nicotinamide metabolism
38	Primary bile acid biosynthesis
39	Folate biosynthesis
40	Porphyrin and chlorophyll metabolism
41	Amino sugar and nucleotide sugar metabolism
42	Steroid hormone biosynthesis

In addition to mass and normalized RT, the Dns-library contains high-resolution QTOF MS/MS spectra of individual metabolites. We recognize that low-resolution tandem mass spectrometers such as triple quadrupole (QqQ), Qtrap and ion trap are currently more readily available than QTOF as they are more widely used for targeted metabolite analysis.[159-161] The MS/MS spectra generated using these instruments should still be useful for comparison with the QTOF spectra in the library. As an example, MS/MS spectra of Dns-biocytin obtained using

Qtrap 4000 and Impact HD QTOF are shown in Figure 4.3. In this case, the fragmentation patterns in terms of relative peak intensities of the fragment ions are somewhat different, but the types of fragment ions generated are almost the same. The pattern difference is understandable for CID MS/MS where several parameters including collision cell design, collision voltage, etc. can influence the spectral pattern.



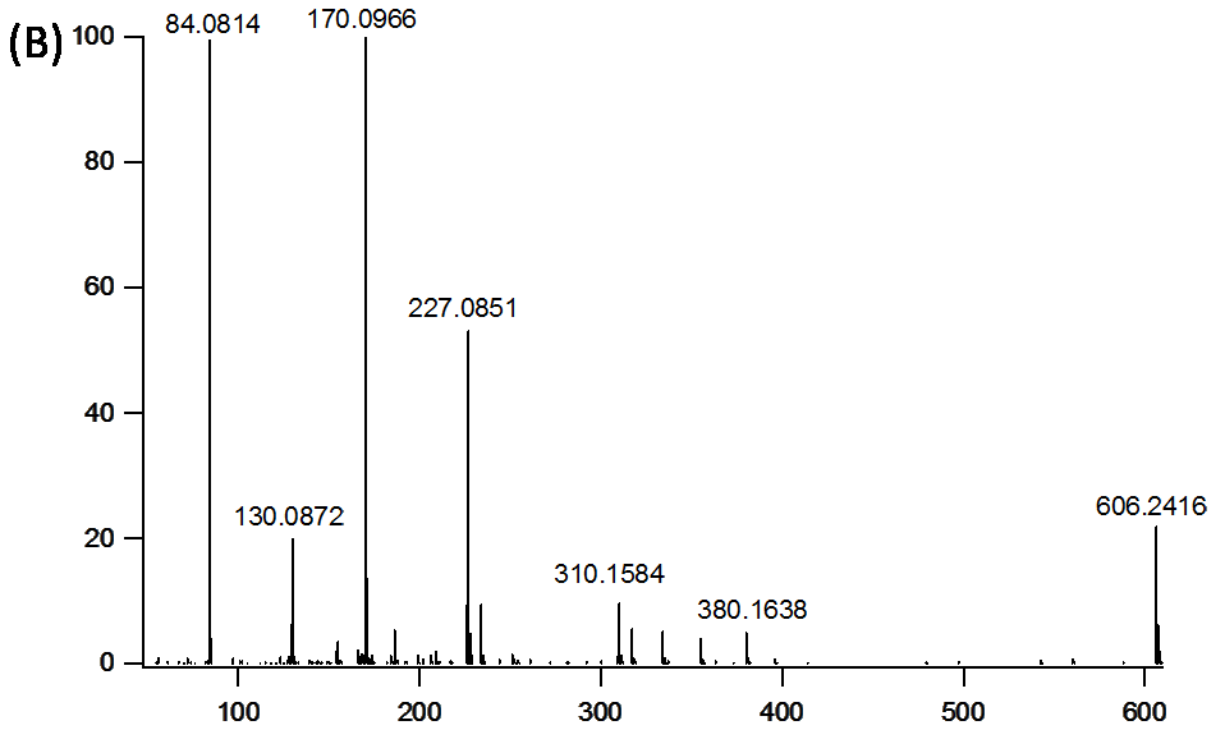


Figure 4.3 MS/MS spectra of Dns-biocytin obtained by (A) QTOF MS (library spectrum) and (B) QTrap MS (see Figure 4.5 for structure assignments of major fragment ions).

For constructing the library MS/MS spectra in QTOF, half of the data acquisition time was spent at collision energy of 20 eV and another half at 50 eV. Therefore, each MS/MS spectrum actually represents an averaged fragmentation pattern at these two collision energies. However, when using a tandem MS such as QqQ or Qtrap for targeted analysis, sometimes only one collision energy is used. As an example, panels A and B in Figure 4.4. show the Qtrap MS/MS spectra of Dns-histidinyl-alanine collected at 25 eV and 55 eV, respectively. Each spectrum matches partially with the corresponding MS/MS spectrum in the library (Figure 4.4C). Nevertheless, within this partial match, similar fragmentation patterns were observed. In addition, it is also possible to overlay the two Qtrap MS/MS spectra (Figure 4.4D) to obtain a more

complete match with the library spectrum, thereby providing higher confidence of correct identification.

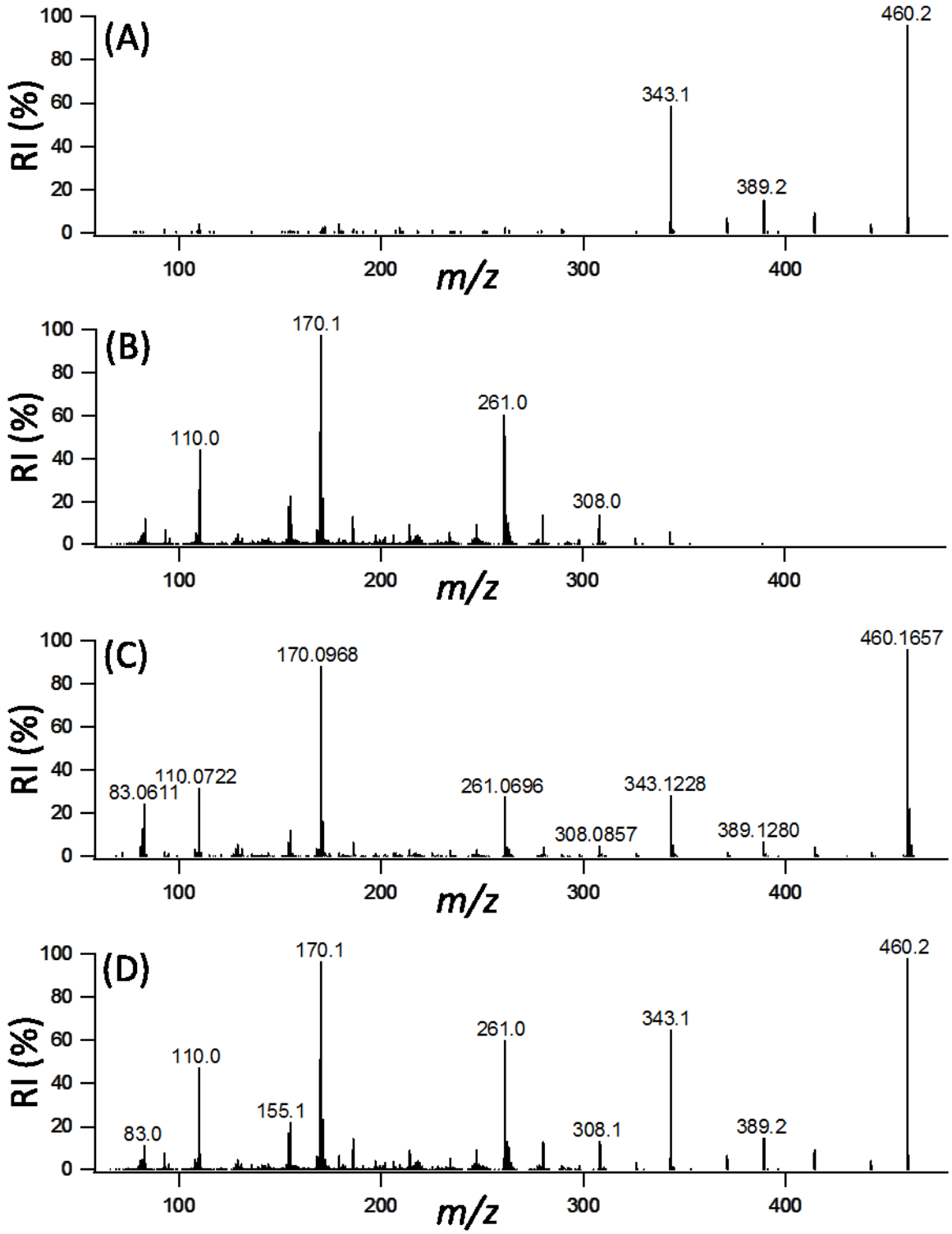


Figure 4.4 MS/MS spectra of Dns-histidinyl-alanine.

Comparisons of the MS/MS spectra shown in Figure 4.3 and Figure 4.4 indicate that there are several common fragment ions detected. They are from the fragmentation of the Dns group. Figures 4.5 and 4.6 show the interpretation of the major fragment ions present in the MS/MS spectra of Dns-biocytin and Dns-histidinyl-alanine, respectively.

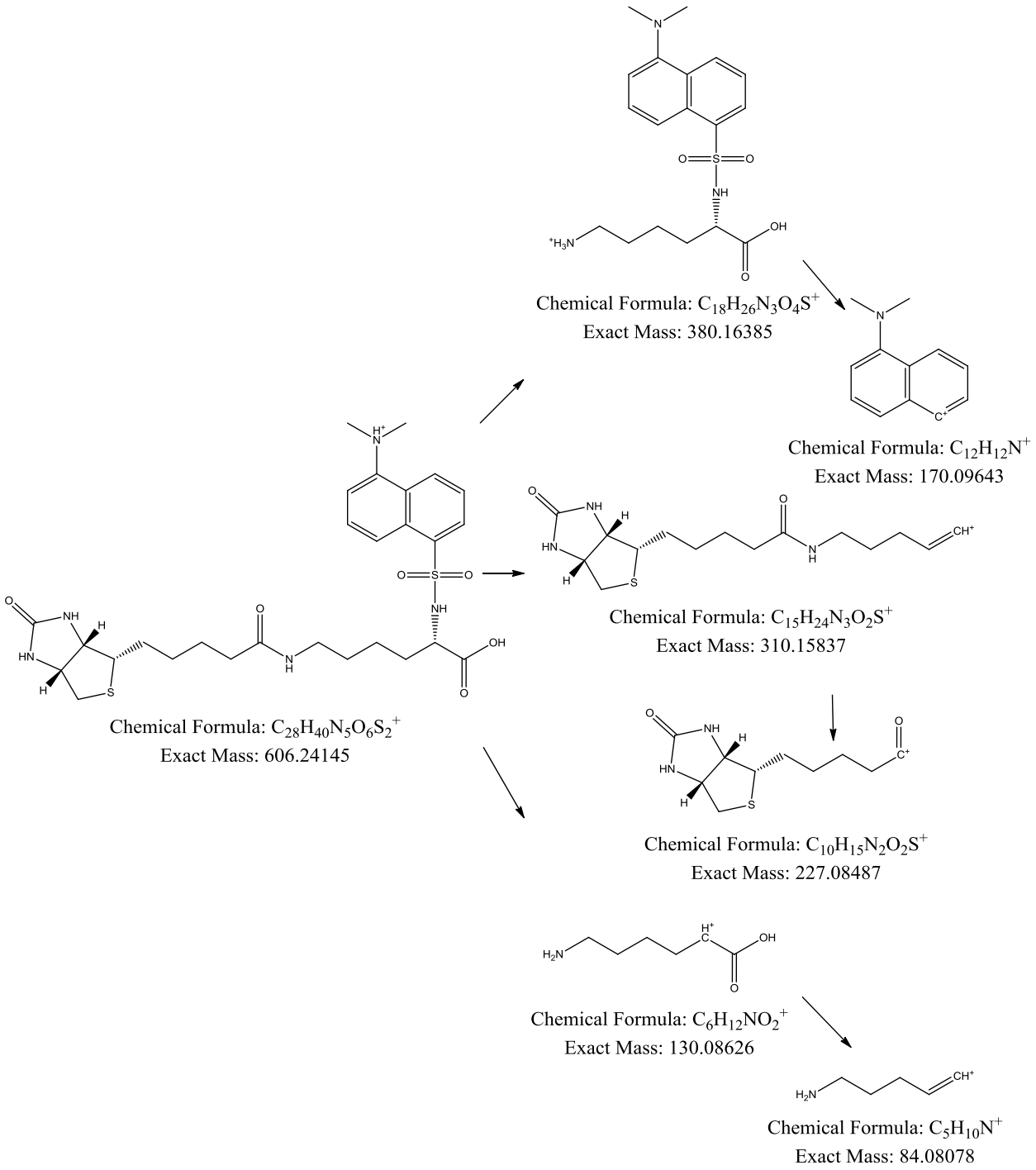


Figure 4.5 Interpretation of the major fragment ions present in the MS/MS spectra of Dns-biocytin.

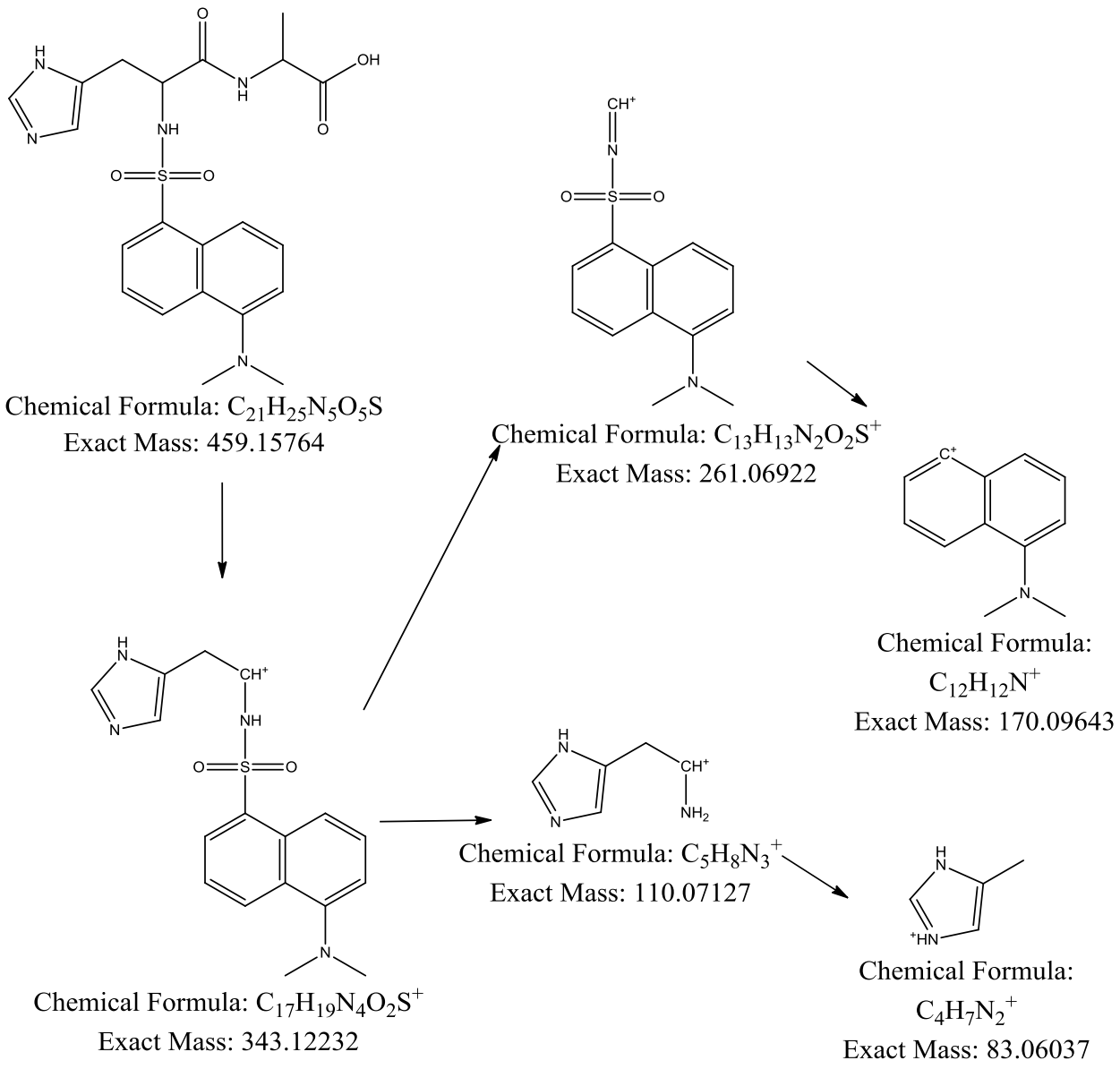


Figure 4.6 Interpretation of the major fragment ion present in the MS/MS spectra of Dns-histidinyl-alanine.

It should be noted that, although the 273 Dns-standards already cover many metabolic pathways, the size of the current library is still relatively small, compared to thousands of Dns-metabolites detectable in a biological sample using LC-MS.[137, 162] There is clearly a need to expand the library in the future through acquisition of more commercially available standards as

well as synthesis of key metabolic pathway related amines and phenols. In addition, we hope that other users will contribute to the expansion of this public library by providing us any standards that are not readily available.

4.3.3 DnsID M-RT Search

Out of the three searchable parameters for each Dns-metabolite in the Dns-library, mass and RT matches by M-RT search can result in confident identification of a Dns-metabolite. Figure 4.7 shows a screenshot of the DnsID program depicting the search process. In a user's laboratory, the RTcal mixture, which can be prepared by the user and is also available by contacting the corresponding author, is first run by LC-MS. The resultant data are processed by the user to generate a file containing accurate masses of RTcal along with their corresponding measured RT. This calibration file is uploaded to DnsID (see Figure 4.7). An LC-MS run of a sample differentially labeled by ^{12}C- and ^{13}C-dansylation from a metabolomic profiling work is then chosen for metabolite identification using DnsID.

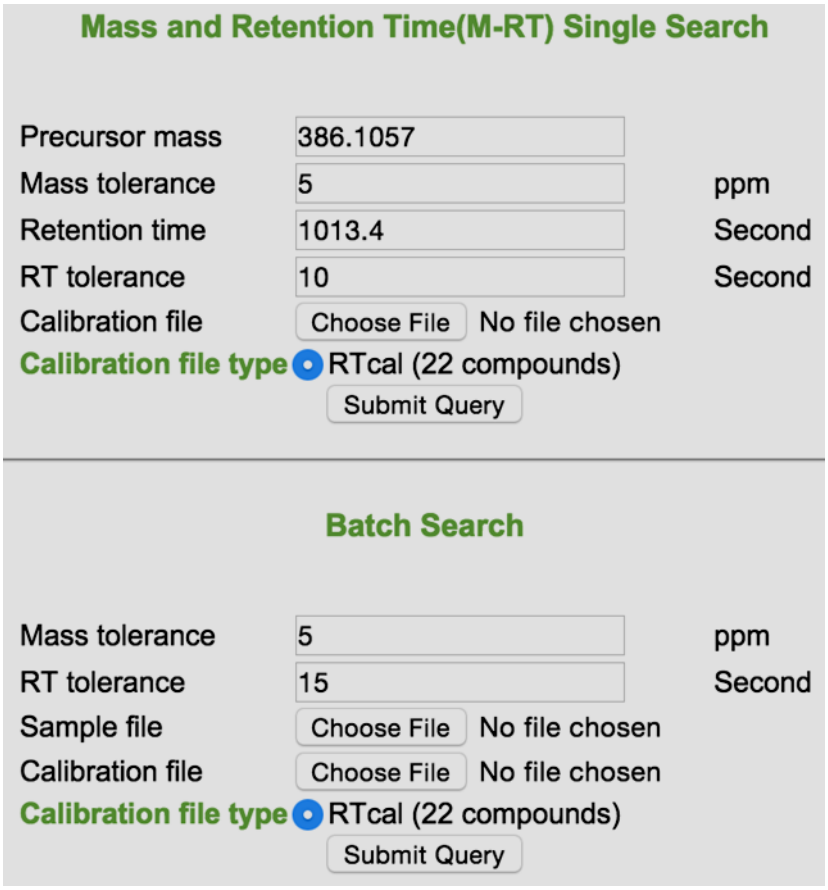


Figure 4.7 Screenshot of the mass and RT (M-RT) search interface of DnsID in www.mycompoindid.org

As Figure 4.7 shows, there are two modes of M-RT search. In the single search mode, the calibration file is first uploaded. The measured mass of a Dns-metabolite of interest found in a sample (e.g., a significant metabolite differentiating two groups of samples in a metabolomics study) is entered along with the mass tolerance which is dependent on the instrument used. The measured RT of the Dns-metabolite and its tolerance are then entered. The RT tolerance should be within the limit of RT variation encountered in LC-MS. In our LC-QTOF-MS and LC-FTICR-MS setups, the RT tolerance is typically within 15 s and the mass error is within 10 ppm. However, for a lower abundance chromatographic peak, the peak shape may not be perfectly

symmetric and thus increasing the RT tolerance to some extent (e.g., using 30 s) is warranted. While mass error for most of the metabolite peaks detected is less than 2 ppm, low abundance peaks can have mass error of up to 10 ppm due to relatively poor peak shapes.

For untargeted metabolite identification, the batch mode M-RT search can be used. In this case, both the calibration file and the IsoMS-processed CSV file of a sample LC-MS run are uploaded (see Figure 4.7). The mass tolerance and retention time tolerance are also entered. DnsID automatically performs RT calibration using the calibration file against the Dns-library data, followed by applying the RT calibration to the uploaded sample CSV file to correct any RT shifts caused by the differences in the user's LC-MS setup and the library LC-MS setup. The search result page displays all the matches that can be sorted according to an individual parameter (e.g., RT) (see Figure 4.8). The displayed information includes the name of matched metabolite, HMDB number (or Dns-library number if HMDB number is not available for the standard), several numeric parameters as well as external links to HMDB and KEGG. These external links are useful to extract biological information about the matched metabolite. On the summary page, there is also a "Show Detail" column which provides a link to the ion chromatogram and MS/MS spectrum of the Dns-standard (see an example shown in Figure 4.9). The standard's chromatogram is particularly useful for manual inspection of a match where a larger RT difference or error (i.e., between 15 and 30 s) is found. A larger RT error is acceptable if this is due to relatively poor chromatographic peak shape. Otherwise, the match may be false.

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	Show detail
1	375.0785	2.05	1.94	HMDB00224	O-Phosphoethanolamine	141.0191	375.0774	2.02	0.0011	0.08	Link	Link	Detail
2	359.0743	2.31	2.18	HMDB00251	Taurine	125.0147	359.0730	2.24	0.0013	0.06	Link	Link	Detail
3	403.1443	2.32	2.19	HMDB00001	1_Methylhistidine	169.0851	403.1434	2.17	0.0009	0.02	Link	Link	Detail
4	403.1443	2.32	2.19	HMDB00479	3_methyl-histidine	169.0851	403.1434	2.01	0.0009	0.18	Link	Link	Detail
5	408.1708	2.61	2.47	HMDB00517	L-Arginine	174.1117	408.1700	2.44	0.0008	0.03	Link	Link	Detail

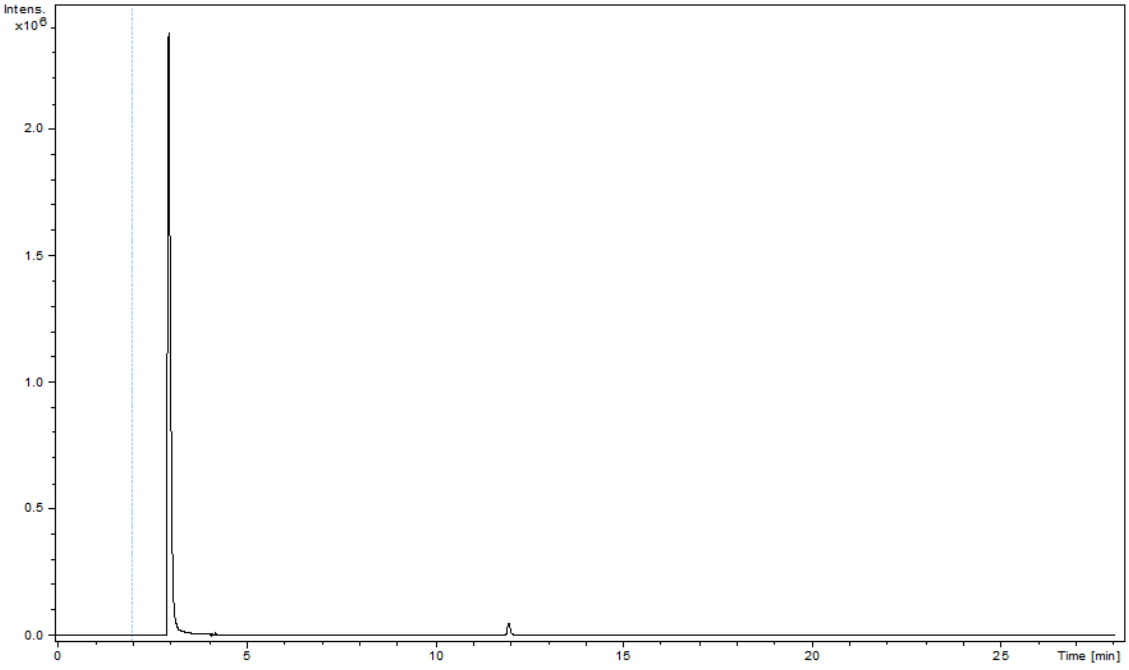
Figure 4.8 Partial list of the search result from LC-MS analysis of a 12C-/13C-dansyl labeled human urine sample.

(A)

Detail Information

HMDB No.:	HMDB00670	
Common name:	Homo-L-arginine	
Monoisotopic molecular mass:	188.1273	Da
Corrected RT:	3.0	Min
mz_light:	422.1856	
Tag number:	1	
Charge number:	1	

(B)



(C)

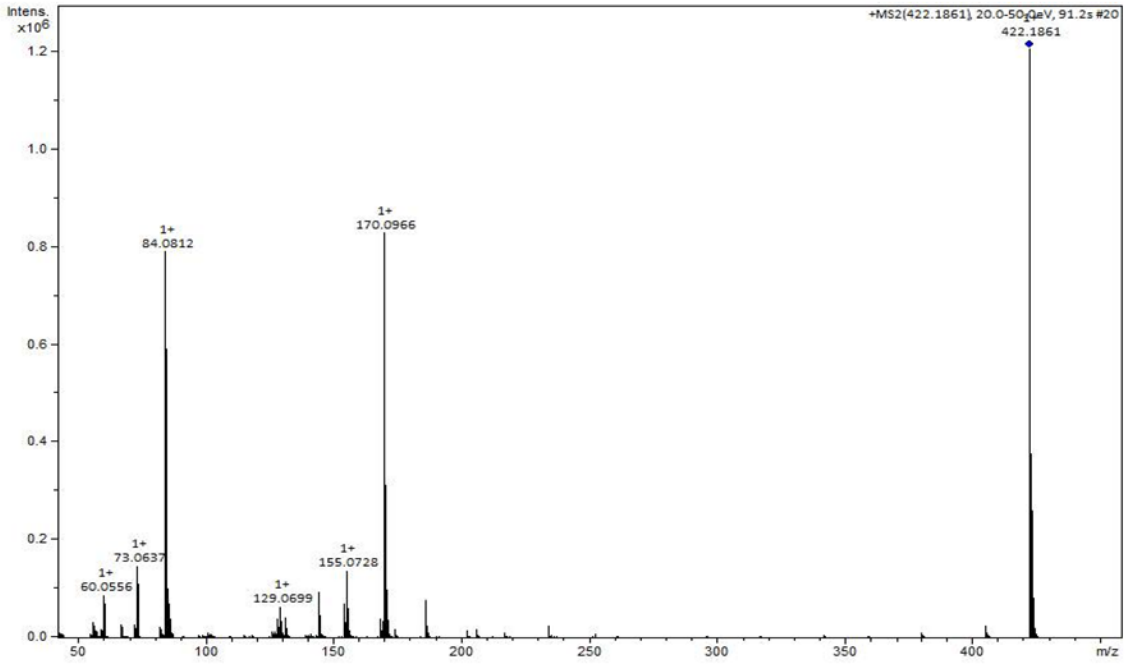


Figure 4.9 Detailed information of a Dns-standard.

While Figure 4.8 shows only a partial list of matches from the analysis of a dansyl labeled human urine sample, the complete list is shown in Table T 4.6. The mass error and RT error of each match are shown in the search result. Again, a larger RT error should trigger a manual inspection of the sample chromatogram to compare the chromatographic peak of the match with that of the Dns-standard. As Table T 4.6 shows, a total of 105 metabolites were matched. Manual inspection of all these M-RT matches did not find any obvious mistake in the match result. On the search result page, there is an option of saving the search results as a CSV file to the user's computer. This file can be opened locally by Excel or other program for presentation or further processing. No result files or any original data from a user are saved in the server.

Table 4.6 The entire urine search results

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	Show detail	Matched with MS/MS search
1	375.0785	2.05	1.94	HMDB00224	O-Phosphoethanolamine	141.0191	375.0774	2.02	0.0011	0.08	Link	Link	Detail	Yes
2	359.0743	2.31	2.18	HMDB00251	Taurine	125.0147	359.073	2.24	0.0013	0.06	Link	Link	Detail	Yes
3	403.1443	2.32	2.19	HMDB00001	1-Methylhistidine	169.0851	403.1434	2.17	0.0009	0.02	Link	Link	Detail	Yes
	403.1443	2.32	2.19	HMDB00479	3-methyl-histidine	169.0851	403.1434	2.01	0.0009	0.18	Link	Link	Detail	No
4	408.1708	2.61	2.47	HMDB00517	L-Arginine	174.1117	408.17	2.44	0.0008	0.03	Link	Link	Detail	Yes
5	343.0781	2.64	2.5	HMDB00965	Hypotaurine	109.0197	343.0781	2.47	0	0.03	Link	Link	Detail	Yes
6	351.1124	2.81	2.67	HMDB00128	Guanidoacetic acid	117.0538	351.1121	2.74	0.0003	0.07	Link	Link	Detail	Yes
7	366.1132	3.09	2.94	HMDB00168	L-Asparagine	132.0535	366.1118	3.00	0.0014	0.06	Link	Link	Detail	Yes
8	422.1861	3.21	3.06	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.00	0.0005	0.06	Link	Link	Detail	Yes
9	359.1547	3.28	3.14	HMDB01861	3-Methylhistamine	125.0953	359.1536	3.27	0.0011	0.13	Link	Link	Detail	Yes
10	436.2014	3.44	3.29	HMDB03334	Symmetric dimethylarginine	202.143	436.2013	3.05	0.0001	0.24	Link	Link	Detail	Yes
11	380.1288	3.44	3.3	HMDB00641	L-Glutamine	146.0691	380.1275	3.32	0.0013	0.02	Link	Link	Detail	Yes
	380.1288	3.44	3.3	HMDB03423	D-Glutamine	146.0691	380.1275	3.32	0.0013	0.02	Link	Link	Detail	No
12	359.1538	3.49	3.34	HMDB01861	3_Methylhistamine	125.0953	359.1536	3.27	0.0002	0.07	Link	Link	Detail	Yes
13	409.1551	3.65	3.5	HMDB00904	Citrulline	175.0957	409.154	3.74	0.0011	0.24	Link	Link	Detail	Yes
14	399.1049	3.83	3.68	HMDB02005	Methionine Sulfoxide	165.046	399.1043	3.72	0.0006	0.04	Link	Link	Detail	Yes

15	307.1223	3.93	3.78	HMDB01522	Methylguanidine	73.064	307.1223	3.84	0	0.06	Link	Link	Detail	Yes
16	353.1167	4.19	4.03	HMDB00719	L-Homoserine	119.0582	353.1166	4.05	0.0001	0.02	Link	Link	Detail	Yes
17	399.1047	4.34	4.18	HMDB02005	Methionine Sulfoxide - Isomer	165.046	399.1043	4.20	0.0004	0.02	Link	Link	Detail	Yes
18	339.1015	4.55	4.39	HMDB00187	L-Serine	105.0426	339.1009	4.40	0.0006	0.01	Link	Link	Detail	Yes
19	423.1702	4.61	4.45	HMDB00679	Homocitrulline	189.1113	423.1697	4.47	0.0005	0.02	Link	Link	Detail	Yes
20	381.1125	5.15	5.01	HMDB00148	L-Glutamic Acid	147.0532	381.1115	5.05	0.001	0.04	Link	Link	Detail	Yes
21	492.1793	5.71	5.62	HMDB00279	Saccharopine - H2O	276.1321	492.1799	5.65	0.0006	0.03	Link	Link	Detail	Yes
22	422.1753	5.76	5.67	HMDB00206	N6-Acetyl-L-Lysine	188.1161	422.1744	5.71	0.0009	0.04	Link	Link	Detail	Yes
23	353.1173	5.85	5.77	HMDB00167	L-Threonine	119.0582	353.1166	5.79	0.0007	0.02	Link	Link	Detail	Yes
24	395.1281	6.02	5.94	HMDB00510	Aminoadipic acid	161.0688	395.1271	5.97	0.001	0.03	Link	Link	Detail	Yes
25	295.1112	6.06	5.98	HMDB00149	Ethanolamine	61.0528	295.1111	6.00	0.0001	0.02	Link	Link	Detail	Yes
26	309.0903	6.63	6.54	HMDB00123	Glycine	75.032	309.0903	6.59	0	0.05	Link	Link	Detail	Yes
27	422.1752	6.89	6.82	HMDB00446	N-Alpha-acetyllysine	188.1161	422.1744	6.79	0.0008	0.03	Link	Link	Detail	Yes
28	531.1480	7.09	7.03	HMDB01173	5'-Methylthioadenosine	297.0896	531.1479	6.97	0.0001	0.06	Link	Link	Detail	Yes
29	406.1432	7.19	7.14	HMDB00721	Glycylproline	172.0848	406.1431	7.17	0.0001	0.03	Link	Link	Detail	Yes
30	323.1065	7.27	7.22	HMDB00056	Beta-Alanine	89.0477	323.106	7.24	0.0005	0.02	Link	Link	Detail	Yes
31	323.1060	7.56	7.53	HMDB00161	L-Alanine	89.0477	323.106	7.57	0	0.04	Link	Link	Detail	Yes
32	365.1173	7.64	7.61	HMDB01149	5-Aminolevulinic acid	131.0582	365.1166	7.59	0.0007	0.02	Link	Link	Detail	Yes
33	337.1218	7.72	7.69	HMDB00112	Gamma-Aminobutyric acid	103.0633	337.1216	7.79	0.0002	0.1	Link	Link	Detail	Yes
34	453.168	8.24	8.22	HMDB0021	Pantothenic acid	219.1107	453.169	8.37	0.000	0.15	Link	Link	Detai	Yes

	9			0					1				l	
35	337.1220	8.71	8.69	HMDB01906	2-Aminoisobutyric acid	103.0633	337.1216	8.91	0.0004	0.22	Link	Link	Detail	Yes
	337.1220	8.71	8.69	HMDB03911	3-Aminoisobutanoic acid	103.0633	337.1216	8.67	0.0004	0.02	Link	Link	Detail	Yes
36	370.0976	8.76	8.74	HMDB00157	Hypoxanthine - multi-tags	136.0385	370.0968	8.73	0.0008	0.01	Link	Link	Detail	Yes
37	351.1375	8.85	8.83	HMDB03355	5-Aminopentanoic acid	117.079	351.1373	8.68	0.0002	0.15	Link	Link	Detail	Yes
38	376.0960	8.9	8.88	HMDB00469	5-Hydroxymethyluracil	142.0378	376.0962	8.87	0.0002	0.01	Link	Link	Detail	Yes
39	386.0923	9.06	9.03	HMDB00292	Xanthine	152.0334	386.0917	8.95	0.0006	0.08	Link	Link	Detail	Yes
40	337.1222	9.15	9.12	HMDB00452	L-Alpha-aminobutyric acid	103.0633	337.1216	9.13	0.0006	0.01	Link	Link	Detail	Yes
	337.1222	9.15	9.12	HMDB00650	D-Alpha-aminobutyric acid	103.0633	337.1216	9.23	0.0006	0.11	Link	Link	Detail	No
	337.1222	9.15	9.12	HMDB01906	2-Aminoisobutyric acid	103.0633	337.1216	8.91	0.0006	0.21	Link	Link	Detail	No
41	335.1053	9.34	9.31	HMDB00719	L-Homoserine - H2O	119.0582	335.106	9.26	0.0007	0.05	Link	Link	Detail	Yes
42	399.1244	9.44	9.41	HMDB03282	1-Methylguanine	165.0651	399.1234	9.57	0.001	0.16	Link	Link	Detail	Yes
43	456.1582	9.46	9.42	HMDB28995	Phenylalanyl-Glycine	222.1004	456.1588	9.43	0.0006	0.01	Link	Link	Detail	Yes
44	363.1011	9.55	9.51	HMDB00148	L-Glutamic Acid - H2O	147.0532	363.1009	9.46	0.0002	0.05	Link	Link	Detail	Yes
45	413.1030	9.56	9.52	HMDB00704	Isoxanthopterin	179.0443	413.1026	9.55	0.0004	0.03	Link	Link	Detail	Yes
46	369.0939	9.58	9.54	HMDB02108	Methylcysteine	135.0354	369.0937	9.37	0.0002	0.17	Link	Link	Detail	Yes
47	265.1007	9.69	9.64	HMDB00164	Methylamine	31.0422	265.1005	9.82	0.0002	0.18	Link	Link	Detail	Yes
48	370.0972	9.77	9.73	HMDB00157	Hypoxanthine - Isomer	136.0385	370.0968	9.65	0.0004	0.08	Link	Link	Detail	Yes
49	349.1217	10.23	10.17	HMDB00162	L-Proline	115.0633	349.1216	10.18	0.0001	0.01	Link	Link	Detail	Yes
50	399.1228	10.34	10.28	HMDB00897	7-Methylguanine	165.0651	399.1234	10.32	0.0006	0.04	Link	Link	Detail	Yes

51	351.1379	10.88	10.78	HMDB00883	L-Valine	117.079	351.1373	10.81	0.0006	0.03	Link	Link	Detail	Yes
52	383.1103	10.98	10.88	HMDB00696	L-Methionine	149.051	383.1094	10.89	0.0009	0.01	Link	Link	Detail	Yes
53	429.1112	11.17	11.07	HMDB00840	Salicyluric acid	195.0532	429.1115	11.05	0.0003	0.02	Link	Link	Detail	Yes
54	360.1015	11.24	11.15	HMDB02024	Imidazoleacetic acid	126.0429	360.1012	11.12	0.0003	0.03	Link	Link	Detail	Yes
55	346.0862	11.25	11.15	HMDB00300	Uracil	112.0273	346.0856	11.34	0.0006	0.19	Link	Link	Detail	Yes
56	438.1479	11.53	11.44	HMDB00929	L-Tryptophan	204.0899	438.1482	11.44	0.0003	0	Link	Link	Detail	Yes
57	346.0862	11.54	11.46	HMDB00300	Uracil	112.0273	346.0856	11.34	0.0006	0.12	Link	Link	Detail	Yes
58	442.1449	11.55	11.47	HMDB00684	L-Kynurenine	208.0848	442.1431	11.44	0.0018	0.03	Link	Link	Detail	Yes
59	424.1327	12.21	12.14	HMDB00684	L-Kynurenine - H2O	208.0848	424.1325	11.97	0.0002	0.17	Link	Link	Detail	Yes
60	399.1372	12.79	12.74	HMDB00159	L-Phenylalanine	165.079	399.1373	12.74	0.0001	0	Link	Link	Detail	Yes
61	432.1104	12.97	12.91	HMDB00291	Vanillylmandelic acid	198.0528	432.1111	12.81	0.0007	0.1	Link	Link	Detail	Yes
62	462.2048	12.97	12.92	HMDB28937	Leucyl-Proline	228.1474	462.2057	12.99	0.0009	0.07	Link	Link	Detail	Yes
63	365.1535	13.12	13.06	HMDB00172	L-Isoleucine	131.0946	365.1529	13.06	0.0006	0	Link	Link	Detail	Yes
	365.1535	13.12	13.06	HMDB00557	L-Alloisoleucine	131.0946	365.1529	13.20	0.0006	0.14	Link	Link	Detail	No
64	462.2047	13.15	13.09	HMDB28937	Leucyl-Proline	228.1474	462.2057	12.99	0.001	0.1	Link	Link	Detail	Yes
65	363.1376	13.31	13.24	HMDB00070	D-Pipecolic acid	129.079	363.1373	13.23	0.0003	0.01	Link	Link	Detail	Yes
	363.1376	13.31	13.24	HMDB00716	L-Pipecolic acid	129.079	363.1373	13.45	0.0003	0.21	Link	Link	Detail	No
66	360.1013	13.36	13.29	HMDB00262	Thymine	126.0429	360.1012	13.21	0.0001	0.08	Link	Link	Detail	Yes
67	365.1537	13.42	13.36	HMDB00557	L-Alloisoleucine	131.0946	365.1529	13.20	0.0008	0.16	Link	Link	Detail	No
	365.153	13.42	13.36	HMDB0068	L-leucine	131.0946	365.152	13.36	0.000	0	Link	Link	Detai	Yes

	7			7			9		8				l	
68	372.1011	13.71	13.63	HMDB00301	Urocanic acid	138.0429	372.1012	13.52	0.0001	0.11	Link	Link	Detail	Yes
69	345.0921	13.72	13.64	HMDB00099	L-Cystathionine - Isomer	222.0674	345.092	13.69	0.0001	0.05	Link	Link	Detail	No
	345.0921	13.72	13.64	HMDB00455	Allocystathionine - Isomer	222.0674	345.092	13.61	0.0001	0.03	Link	Link	Detail	Yes
70	315.1078	13.83	13.75	HMDB00450	5-Hydroxylysine	162.1004	315.1085	13.88	0.0007	0.13	Link	Link	Detail	Yes
71	315.1073	14.09	14.05	HMDB00450	5-Hydroxylysine	162.1004	315.1085	13.88	0.0012	0.17	Link	Link	Detail	Yes
72	354.0701	14.14	14.1	HMDB00192	L-Cystine	240.0238	354.0702	14.11	0.0001	0.01	Link	Link	Detail	Yes
73	425.1161	15.21	15.11	HMDB00763	5-Hydroxyindoleacetic acid	191.0582	425.1166	15.09	0.0005	0.02	Link	Link	Detail	Yes
74	414.1224	15.58	15.43	HMDB01889	Theophylline	180.0647	414.123	15.42	0.0006	0.01	Link	Link	Detail	Yes
75	368.0856	16.02	15.84	HMDB00676	L-Homocystine	268.0551	368.0859	15.82	0.0003	0.02	Link	Link	Detail	Yes
76	319.1109	16.46	16.32	HMDB03911	3_Aminoisobutanoic acid - H2O	103.0633	319.111	16.29	0.0001	0.03	Link	Link	Detail	Yes
77	388.0857	16.52	16.38	HMDB00397	2-Pyrocatechuic acid	154.0266	388.0849	16.31	0.0008	0.07	Link	Link	Detail	Yes
78	371.1060	16.57	16.44	HMDB01123	2-Aminobenzoic acid	137.0477	371.106	16.62	0	0.18	Link	Link	Detail	Yes
79	402.0994	16.58	16.45	HMDB01868	5-Methoxysalicylic acid	168.0423	402.1006	16.38	0.0012	0.07	Link	Link	Detail	Yes
80	386.1053	16.62	16.49	HMDB00440	3-Hydroxyphenylacetic acid	152.0473	386.1057	16.72	0.0004	0.23	Link	Link	Detail	No
	386.1053	16.62	16.49	HMDB00669	Ortho-Hydroxyphenylacetic acid	152.0473	386.1057	16.42	0.0004	0.07	Link	Link	Detail	Yes
81	416.1162	16.68	16.56	HMDB00118	Homovanillic acid	182.0579	416.1162	16.51	0	0.05	Link	Link	Detail	Yes
82	386.1061	17.1	17.02	HMDB00020	p-Hydroxyphenylacetic acid	152.0473	386.1057	16.91	0.0004	0.11	Link	Link	Detail	Yes
	386.1061	17.1	17.02	HMDB02390	3_Cresotinic acid	152.0473	386.1057	16.80	0.0004	0.22	Link	Link	Detail	No
83	388.0848	17.39	17.34	HMDB00152	Gentisic acid	154.0266	388.0849	17.11	0.0001	0.23	Link	Link	Detail	Yes

84	307.1098	17.52	17.48	HMDB00182	L-Lysine	146.1055	307.1111	17.47	0.0013	0.01	Link	Link	Detail	Yes
85	402.1002	17.59	17.55	HMDB00484	Vanillic acid	168.0423	402.1006	17.34	0.0004	0.21	Link	Link	Detail	Yes
86	428.1155	17.68	17.65	HMDB00955	Isoferulic acid	194.0579	428.1162	17.49	0.0007	0.16	Link	Link	Detail	Yes
87	372.0901	17.81	17.77	HMDB00500	4-Hydroxybenzoic acid	138.0317	372.09	17.57	0.0001	0.2	Link	Link	Detail	Yes
88	389.1273	18.07	18.04	HMDB00177	L-Histidine	155.0695	389.1278	18.09	0.0005	0.05	Link	Link	Detail	Yes
89	400.1207	18.14	18.11	HMDB02199	Desaminotyrosine	166.063	400.1213	18.04	0.0006	0.07	Link	Link	Detail	Yes
90	394.1572	18.18	18.15	HMDB00303	Tryptamine	160.1	394.1584	18.03	0.0012	0.12	Link	Link	Detail	Yes
91	398.1051	18.57	18.55	HMDB01713	m-Coumaric acid	164.0473	398.1057	18.51	0.0006	0.04	Link	Link	Detail	Yes
92	428.1152	18.68	18.66	HMDB00954	trans-Ferulic acid	194.0579	428.1162	18.47	0.001	0.19	Link	Link	Detail	Yes
93	393.1826	19.15	19.13	HMDB00991	2-aminooctanoic acid	159.1259	393.1842	19.20	0.0016	0.07	Link	Link	Detail	Yes
94	411.1043	19.55	19.52	HMDB03164	Chlorogenic acid	354.0951	411.1059	19.45	0.0016	0.07	Link	Link	Detail	Yes
95	278.1069	21.22	21.18	HMDB01414	1,4-diaminobutane	88.1	278.1083	21.27	0.0014	0.09	Link	Link	Detail	Yes
96	411.1044	21.23	21.19	HMDB03164	Chlorogenic acid - Isomer	354.0951	411.1059	21.24	0.0015	0.05	Link	Link	Detail	Yes
97	356.0945	21.62	21.59	HMDB00750	3-Hydroxymandelic acid - COOH	168.0423	356.0951	21.64	0.0006	0.05	Link	Link	Detail	Yes
98	326.0753	21.93	21.91	HMDB01866	3,4-Dihydroxymandelic acid	184.0372	326.0769	21.73	0.0016	0.18	Link	Link	Detail	Yes
99	324.5942	22.73	22.68	HMDB00158	L-Tyrosine	181.0739	324.5953	22.65	0.0011	0.03	Link	Link	Detail	Yes
100	328.0990	23.29	23.2	HMDB00228	Phenol	94.0419	328.1002	23.16	0.0012	0.04	Link	Link	Detail	Yes
101	311.0703	24.64	24.44	HMDB00152	Gentisic acid - multi-tags	154.0266	311.0716	24.69	0.0013	0.25	Link	Link	Detail	No
	311.0703	24.64	24.44	HMDB01856	Protocatechuic acid	154.0266	311.0716	24.51	0.0013	0.07	Link	Link	Detail	Yes
102	322.104	24.88	24.67	HMDB0025	Serotonin	176.095	322.105	24.65	0.001	0.02	Link	Link	Detai	Yes

	5			9			8		3				l	
103	318.0783	24.94	24.74	HMDB00130	Homogentisic acid	168.0423	318.0794	24.84	0.0011	0.1	Link	Link	Detail	Yes
104	317.6047	25.56	25.46	HMDB00022	3_Methoxytyramine	167.0946	317.6056	25.49	0.0009	0.03	Link	Link	Detail	Yes
	317.6047	25.56	25.46	HMDB02182	Phenylephrine	167.0946	317.6056	25.39	0.0009	0.07	Link	Link	Detail	No
105	302.5998	25.88	25.8	HMDB00306	Tyramine	137.0841	302.6004	25.83	0.0006	0.03	Link	Link	Detail	Yes

4.3.4 DnsID MS/MS Search

In DnsID, the MS/MS spectral library can be searched using an acquired MS/MS spectrum from a sample. Figure 4.10 shows the screenshot of the MS/MS search interface. The fragment ion masses and intensities along with their mass tolerance are entered. Because structurally similar metabolites may have different molecular ion masses, but similar MS/MS spectra (e.g., a methylated standard with a methyl group added to the core structure of a standard), DnsID MS/MS search has the option of not specifying the precursor ion mass for spectral match. This option is useful to find related metabolites. An example is shown in Figure 4.11. In this case, the unknown metabolite matches to Dns-arginine in the library based on the fragment ions only, but the precursor ion mass differs by 14.0156 Da. By searching the mass of the unknown using MyCompoundID against the predicted human metabolite library, there were 4 matches (Figure 4.11). This unknown was thought to be likely a metabolite of Dns-arginine with the addition of CH_2 group, possibly Dns-homo-arginine. Subsequently we obtained homo-arginine, produced Dns-homo-arginine and generated the MS/MS of the labeled standard. The MS/MS spectrum of Dns-homo-arginine matched perfectly with the MS/MS spectrum of the unknown. While in this case it was fortunate that we could obtain the standard to confirm the identity of the unknown, in many other cases we could not obtain standards for metabolite confirmation. Nevertheless, using this strategy to identify structurally related metabolites, albeit putatively, can still be useful as one can infer some biological relevance of these metabolites, potentially helpful for studying metabolic mechanism related to a biological event.

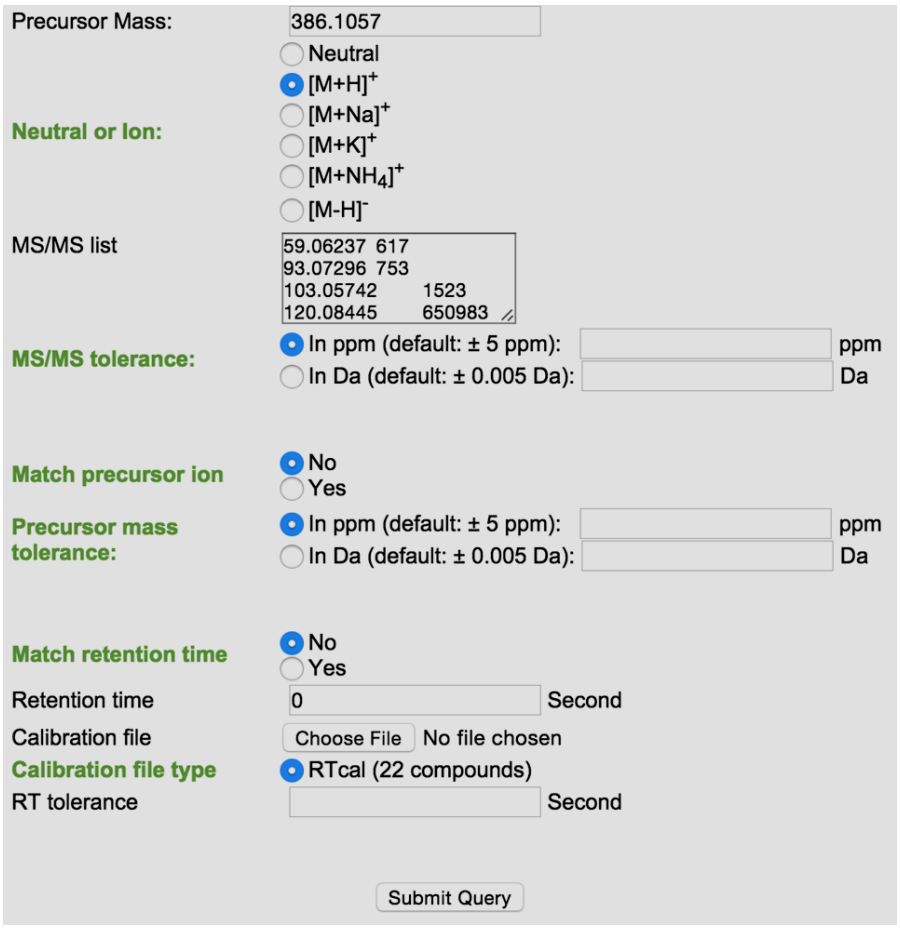
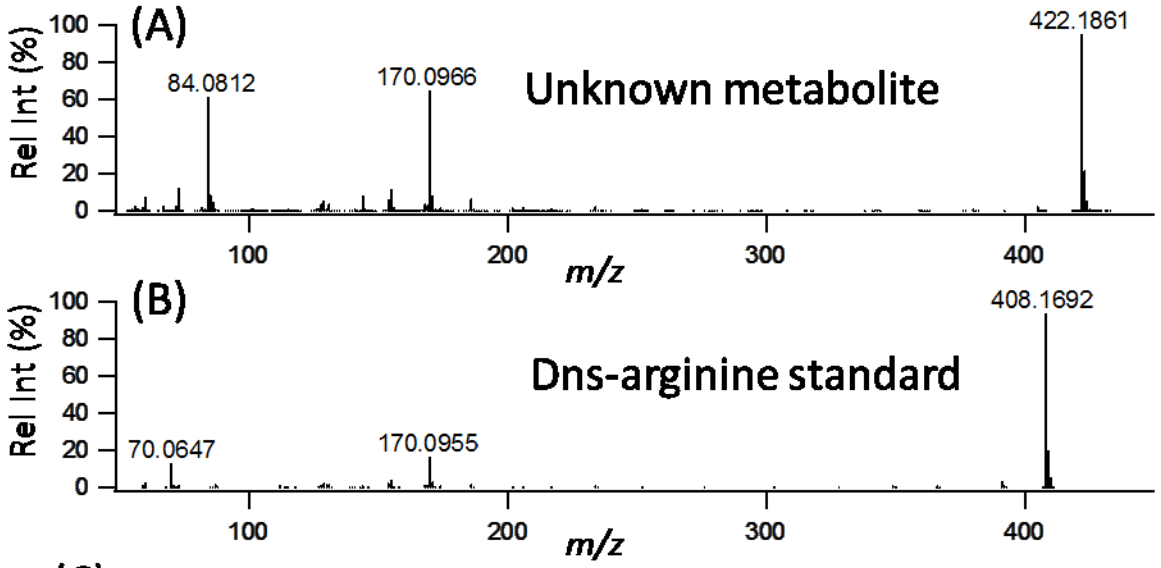


Figure 4.10 Screenshot of the MS/MS search interface of DnsID in www.mycompoindid.org.



(C)

#	HMDB ID	Common Name	Mass (Da)	Formula	Chemical Structure	Explore (for Firefox)	Possible Reactions	Reaction Offset (Da)	Mass Error (ppm)
1	HMDB00517	L-Arginine	174.111676	$C_6H_{14}N_4O_2$		ChemDraw Pro ChemDraw Plugin	[+CH_2]	14.0156500	2.413253
2	HMDB03416	D-Arginine	174.111676	$C_6H_{14}N_4O_2$		ChemDraw Pro ChemDraw Plugin	[+CH_2]	14.0156500	2.413253
3	HMDB01539	Dimethyl-L-arginine	202.142976	$C_8H_{18}N_4O_2$		ChemDraw Pro ChemDraw Plugin	[-CH_2]	-14.0156500	2.413253
4	HMDB03334	Symmetric dimethylarginine	202.142976	$C_8H_{18}N_4O_2$		ChemDraw Pro ChemDraw Plugin	[-CH_2]	-14.0156500	2.413253

Homo-Arginine

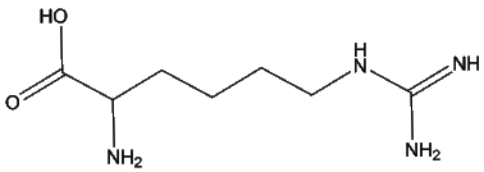


Figure 4.11 MS/MS search without specifying the precursor ion of a Dns-metabolite found in a labeled human urine sample. MS/MS spectra of (A) the unknown metabolite and (B) Dns-arginine from the Dns-library. (C) Screenshot of the result obtained from searching the precursor

ion mass of the unknown against the predicted human metabolite library in www.mycompoundid.org. The unknown was confirmed to be Dns-homo-arginine.

Another option of MS/MS search is to include RT information during the search. If precursor ion mass is also entered, this option allows the matches of all three searchable parameters. One unique application of this option is to distinguish isomers of Dns-standards in the library. Some positional isomers have different retention times, but the same or similar fragmentation patterns. One example is Dns-leucine and Dns-isoleucine (Figure 4.12). In the labeled human urine sample, two peaks with the same ion mass (m/z 365.1529) within a mass tolerance of 10 ppm were detected at two different retention times. When the MS/MS library was searched without the retention time information, there were two matches with similar matching scores. However, by including the RT information, we could clearly rule out Dns-isoleucine and confirm that the metabolite detected in urine was Dns-leucine.

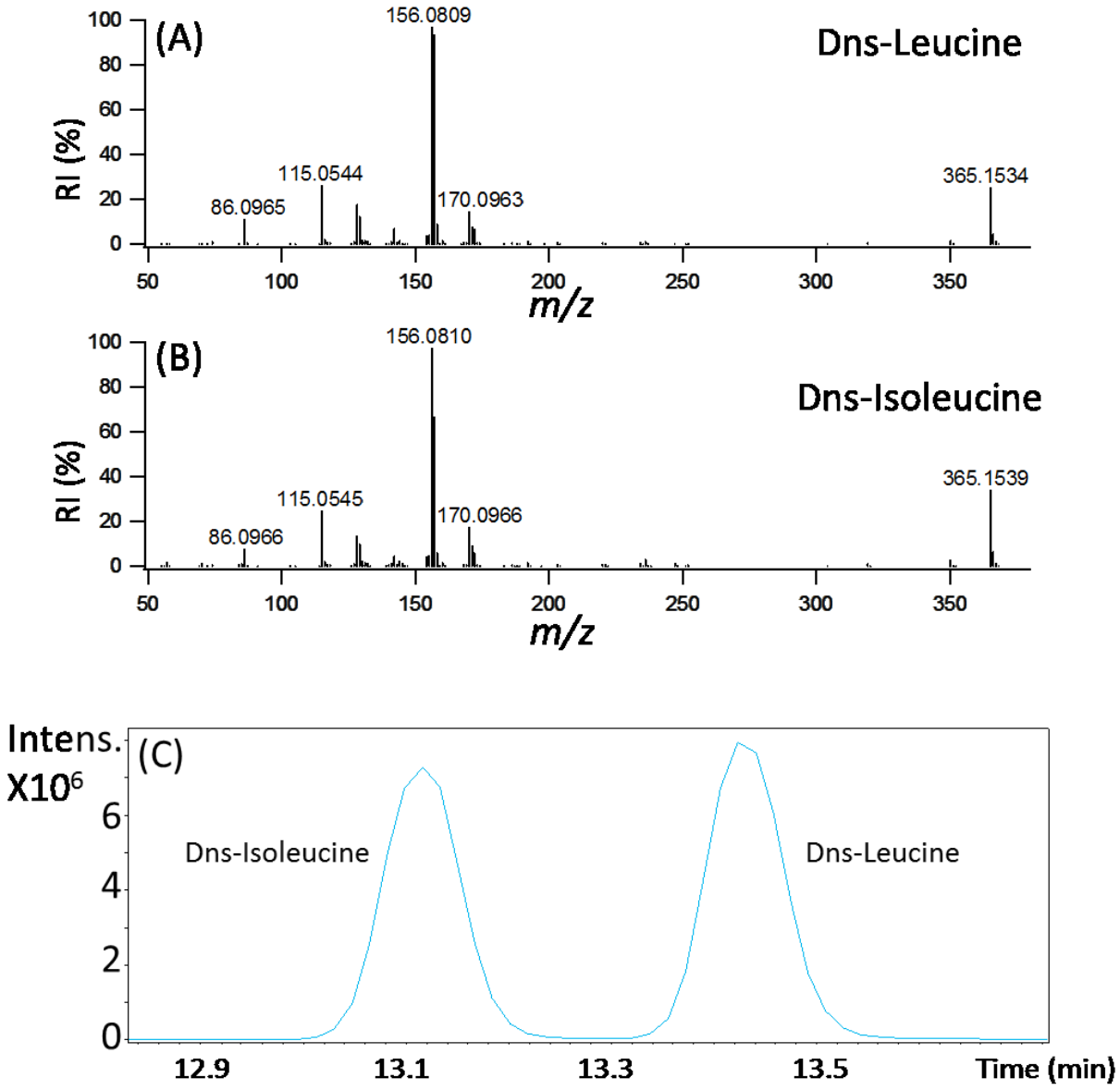


Figure 4.12 MS/MS spectra of Dns-leucine (A) and Dns-isoleucine (B). (C) Ion chromatogram showing the retention time difference of the dansyl labeled isomers.

4.3.5 Application of DnsID

For the human urine sample analyzed in this work, 105 metabolite matches were obtained using M-RT search (Table T 4.6). To validate the M-RT match results, an automatic LC-MS/MS experiment was performed on the same urine sample to generate the MS/MS spectra of the

labeled metabolites. Out of the 105 M-RT matches, 77 matched metabolites (73.3%) had their corresponding MS/MS spectra generated with relatively high quality (i.e., the precursor ion intensity was sufficiently high to produce a number of fragment ions in MS/MS) (see last column in Table T 4.6). Based on MS/MS spectral search, all these 77 M-RT matches could be confirmed to have the correct structures. The other 28 M-RT matched metabolites (26.7%) had low precursor ion signal intensities and thus their MS peaks were not selected for the auto MS/MS experiment.

The above example illustrates that out of the three searchable parameters for each Dns-metabolite, M-RT matches can result in confident identification of a Dns-metabolite. In principle, metabolite identification can also be done by using mass and MS/MS matches (M-MS/MS matches). In each case, an authentic standard is required to confirm the identity of a match. However, M-RT search does not require the use of a tandem MS, while M-MS/MS search does. This difference is significant and may play an important role in deciding the initial infrastructure investment and subsequent usage of the MS equipment. A simple MS instrument such as LC-TOF-MS that provides adequate mass resolution and mass measurement accuracy for CIL MS metabolome profiling is a relatively inexpensive capital investment.[162] Moreover, all the time-consuming profiling work is based on MS detection, which does not require MS/MS. After the MS profiling work, one can just use M-RT search to identify the metabolites of interest and then use authentic standards to confirm the metabolite identities. In future works, we will continue to examine the robustness of the M-RT search approach for metabolite identification in various metabolomics applications.

It should be noted that, for the ^{13}C-/^{12}C-dansyl labeled human urine, we detected 1552 peak pairs in triplicate runs. Using accurate mass search (5 ppm tolerance) against HMDB and

the MCID library with one reaction,[144] we matched 378 and 600 metabolites, respectively. Compared to 105 metabolites identified using DnsID, many dansyl metabolites still remain to be positively identified. Some compounds such as drugs or drug metabolites in urine could not be detected using DnsID as it does not contain these compounds in the library. Expanding the dansyl standard library is clearly needed.

4.4 Conclusions

We have developed a dansyl standards library and a library search program, DnsID, for rapid identification of metabolites in dansylation LC-MS targeting the analysis of the amine/phenol submetabolome. The current Dns-library consists of 273 unique metabolites and should be expandable in the future. Construction of other types of labeled standards, such as DmPA labeled acids for profiling the acid submetabolome, is currently on the way. As CIL LC-MS technology further advances, we envisage a broad use of this resource for rapid identification of labeled metabolites for metabolomics research.

Chapter 5

MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites

5.1 Introduction

Mass spectrometry (MS)-based metabolomics has been developed rapidly in the past decade or so. However, metabolite identification from the MS data is still a challenge. Accurate mass search alone against a chemical database can result in many possible matches. To generate structural information of a metabolite, MS/MS or fragment ion spectrum can be produced using a tandem mass spectrometer. The fragmentation pattern can be manually interpreted, often against a probable chemical structure found using accurate mass search, to confirm or disapprove a structure.[144] Considering that manual spectral interpretation is a time-consuming process, spectral search using an MS/MS spectral library of metabolite standards has been developed for rapid metabolite identification.[163, 164] Besides in-house and commercial libraries,[165, 166] several public libraries have been developed as a very useful resource. For example, our laboratory constructed the HMDB MS/MS spectral library using 800 endogenous human metabolites.[1] Other libraries such as Metlin[141, 167] and MassBank[143] contain MS/MS spectra of metabolites as well as other synthetic compounds such as common drugs. However, the number of metabolites with reference spectra available is still very small, due to the lack of standards.

In the absence of a standard, a predicted MS/MS spectrum of a given structure can be helpful in manual spectral interpretation as well as in spectral match. There are several approaches for generating predicted MS/MS spectra (more precisely a list of fragment ions with

unit intensity), depending on the chemical bond breakage rules used and the number or level of fragment ions included in a predicted fragment ion spectrum.[80-91] Commercial products (e.g., Mass Frontier from Thermo Scientific, Waltham, US and ACD/MS Fragmenter from Advanced Chemistry Labs, Toronto, Canada) and published tools (e.g., Metfrag[82], Fragment Identificator or FiD[81] and MIDAS[88]) are available for generating predicted MS/MS spectra with varying degrees of success.

Our approach is to develop a web-based online tool for metabolite identification based on integrated MS and MS/MS search using a comprehensive library of predicted spectra of all metabolites in MyCompoundID.org (MCID).[144] The current MCID compound library includes 8,021 known endogenous human metabolites in the Human Metabolome Database (HMDB) and 375,809 predicted human metabolites in the Evidence-based Metabolome Library (EML) with one metabolic reaction. We developed an *in silico* method of predicting fragment ions using heteroatom-initiated bond breakage rules and applied it to all MCID metabolites to generate a predicted MS/MS spectral library. An automated MS/MS search program was developed that allows a user to search an experimental MS/MS spectrum, in single or batch search mode, against the library for spectral match. In this paper, we describe the MCID spectral library and MS/MS search tool and demonstrate its performance using MS/MS spectra of metabolite standards and those acquired from a human urine sample.

5.2 Experimental Section

5.2.1 Overall Workflow

In the MCID MS/MS search method, a precursor ion mass of a metabolite is first used to search against the MCID library to generate a list of candidate compounds with matched molecular ion masses. The fragment ion masses from an experimental MS/MS spectrum are then compared to the predicted fragment ion masses of each candidate compound in the list. A fit score is assigned to each comparison to measure the similarity between the experimental and predicted fragment ions. Once all the comparisons are done, the candidates in the list are ranked by the fit scores.

5.2.2 Predicting MS/MS Fragment Ions

The MCID spectral library contains the predicted MS/MS spectra of 383,830 known and potentially existing human metabolites.[144] Each predicted spectrum was generated using a "chopping" program following a series of *in silico* fragmentation rules. A .mol file of a compound structure is used by the program. The algorithm in the chopping program involves two steps. The first step is the heteroatom-initiated bond breakage or chopping. Heteroatoms in a compound such as O and N are identified and the bonds connecting to the heteroatoms are broken to create possible fragments. The second step is the splittable-bond chopping. Splittable-bonds are linear single bonds and double bonds in aromatic structures. If there are less than 40 splittable-bonds in the chemical structure, four layers of chopping are done. In cases that there are 40-60 splittable-bonds, three layers of chopping are done to avoid generating too many fragment ions. For a very large compound with > 60 splittable-bonds, only two layers of chopping are carried out. After applying these two steps of chopping to a compound structure, a mass redundancy check is performed to combine the same fragment ion masses. A list of

fragment ion masses are then compiled for the compound and stored as a predicted MS/MS spectrum. All predicted spectra are stored in a local MySQL database in the MCID web server.

5.2.3 Match Algorithm

Two layers of scoring have been developed to gauge the similarity between the experimental MS/MS data and the predicted MS/MS data. At first, we calculate an initial match score, according to:

$$score_i = \frac{1}{\max(weight)} weight_i$$

where

$$weight_i = < \overrightarrow{m/z}(matched) > \cdot < \overrightarrow{Int}(matched) >$$

$$< \overrightarrow{\mathrm{m/z}}(\mathrm{matched}) >: the\ matched\ list\ of\ m/z$$

$$< \overrightarrow{\mathrm{Int}}(\mathrm{matched}) >: the\ matched\ list\ of\ intensities$$

Using the above equation, a weight is calculated for each comparison by the dot product of the matched m/z's and intensities. An m/z tolerance is set to determine if the experimental m/z is matched with the predicted m/z. The initial score is calculated by normalization against the maximum weight in all the candidates. For the candidates with no-zero initial scores, a fit score is then used to quantify and rank how well the experimental spectrum is matched to the predicted spectrum. The fit score is defined as:

$$\mathrm{fit.score} = \frac{< \overrightarrow{m/z}(matched) > \cdot < \overrightarrow{Int}(matched) >}{< \overrightarrow{m/z}(experimental) > \cdot < \overrightarrow{Int}(experimental) >}$$

where

$$< \overrightarrow{m/z}(matched) >: the\ matched\ list\ of\ m/z$$

$$< \overrightarrow{Int}(matched) > : the\ matched\ list\ of\ intensities$$

$$< \overrightarrow{m/z}(experimental) > : the\ entire\ list\ of\ experimental\ m/z$$

$$< \overrightarrow{Int}(experimental) > : the\ entire\ list\ of\ experimental\ intensities$$

5.2.4 MS/MS of Standards

35 metabolites were selected to generate the MS/MS spectra. An individual standard was used to produce a final concentration of 10 μM. A Bruker Impact HD QTOF mass spectrometer (Billerica, MA) was used to generate the MS/MS spectra using direct infusion with collision energy of 20-50 eV.

5.2.5 LC-MS/MS of Urine

A human urine sample was collected from an healthy individual and filtered using 0.22 μM-pore-size filter (Millipore Corp., MA) twice. LC-MS/MS analysis was performed on the Bruker QTOF-MS equipped with an Agilent 1100 HPLC system (Palo Alto, CA, USA). 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) LC-MS grade formic acid in 2% (v/v) grade ACN, and solvent B was 0.1% (v/v) LC-MS grade formic acid in LC-MS grade 98% ACN. The gradient elution profile was as follows: t = 0.0 min 0% B, t = 10 min, 0% B, t = 50.0 min, 80% B, t = 55 min, 100%B, t = 60 min, 100% B, t = 60.1 min, 0% B, t = 80 min, 0% B. The flow rate was 100 μL/min. The sample injection volume was 20 μL.

5.3 Results and Discussion

5.3.1 MCID MS/MS Search

There are two search modes available (see Figure 5.1 for a screenshot of the web interface and Appendix for a user manual for MCID MS/MS search). In the single-spectrum search which is useful for targeted metabolite identification, a user selects either the zero-reaction library containing all the known metabolites in HMDB or the one-reaction library containing all the predicted metabolites in EML. The precursor ion type, mass and mass tolerance are entered. The fragment ion masses and their relative intensities, and the m/z tolerance value for the fragment ions are also entered. "Deisotope" is selected as a default to remove the ^{13}C natural abundance isotopic peak(s) of a fragment ion. Figure 1A shows an example of the search results obtained by searching the zero-reaction library. The result page lists all the mass-matched metabolites. For each candidate, the HMDB ID number with a link to the HMDB database is given along with other information.

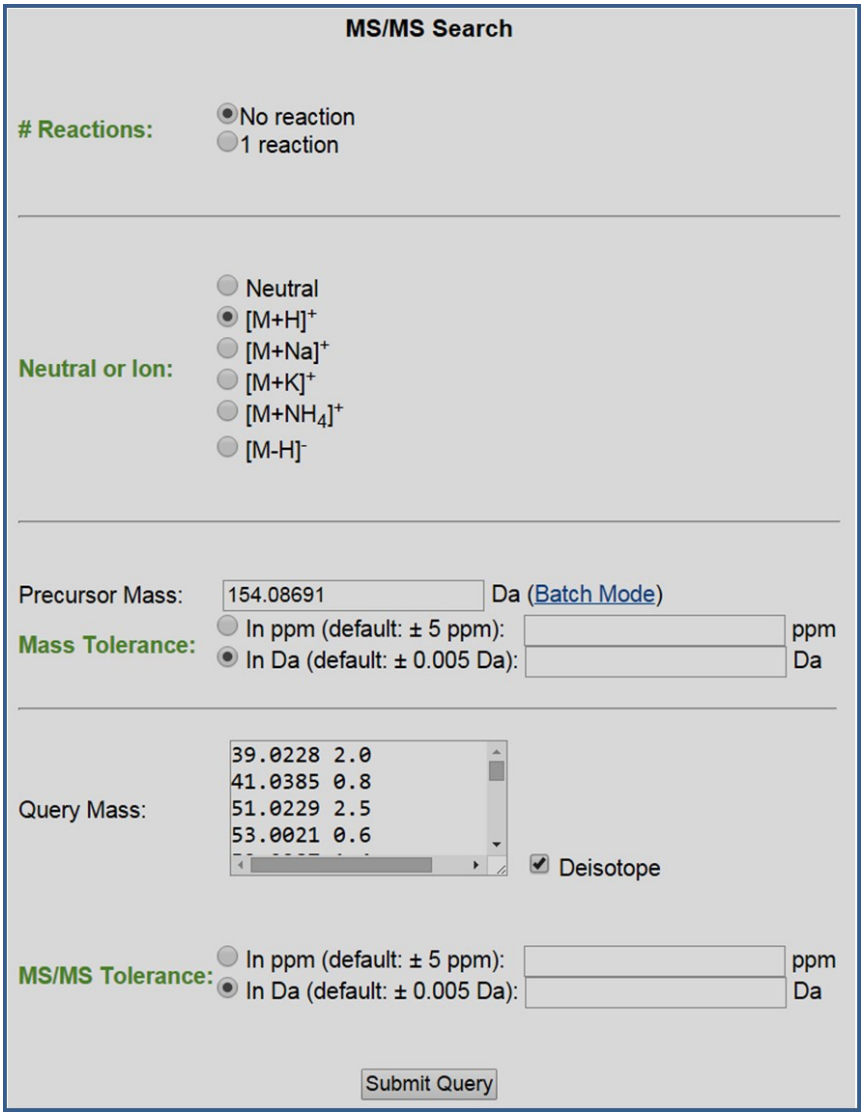
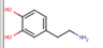
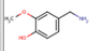


Figure 5.1 Screenshot of the web interface

For each candidate, the initial score and fit score from MS/MS spectral comparison are given. In the case shown in Figure 5.2A, the three candidates are isomers and the fit score is the same. By clicking the initial score, a new page will be displayed. Figure 5.3A shows an example where the experimental MS/MS spectrum is shown. The matched peaks to the predicted spectrum are shown in red and unmatched ones are shown in grey. It also displays a table (Figure 5.3B) containing information on the masses and intensities of the experimental fragment ions

(matched ones in red and unmatched ones in black), the number of matched fragment ion structures, and a link called "Detail". By clicking "Detail", another page will be displayed (Figure 5.3C) which provides a summary of the matched fragment ion(s) including the predicted structure(s). These multiple layers of information can be very helpful for manual confirmation of a MS/MS match. Manual interpretation may assist in determining which structure among the matches is the most probable one fitting to the MS/MS fragmentation pattern.

(A)

#	HMDB ID	Common Name	Mass (Da)	Formula	Chemical Structure	Explore (for Firefox)	Possible Reactions	Reaction Offset (Da)	Mass Error (Da)	Initial Score	Fit Score	To Del	Attachments
1	HMDB00073	Dopamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000655	1.000	0.819		Add
2	HMDB12309	Vanillylamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000655	1.000	0.819		Add
3	HMDB04825	p-Octopamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000655	0.998	0.817		Add

(B)

Filter the Result:

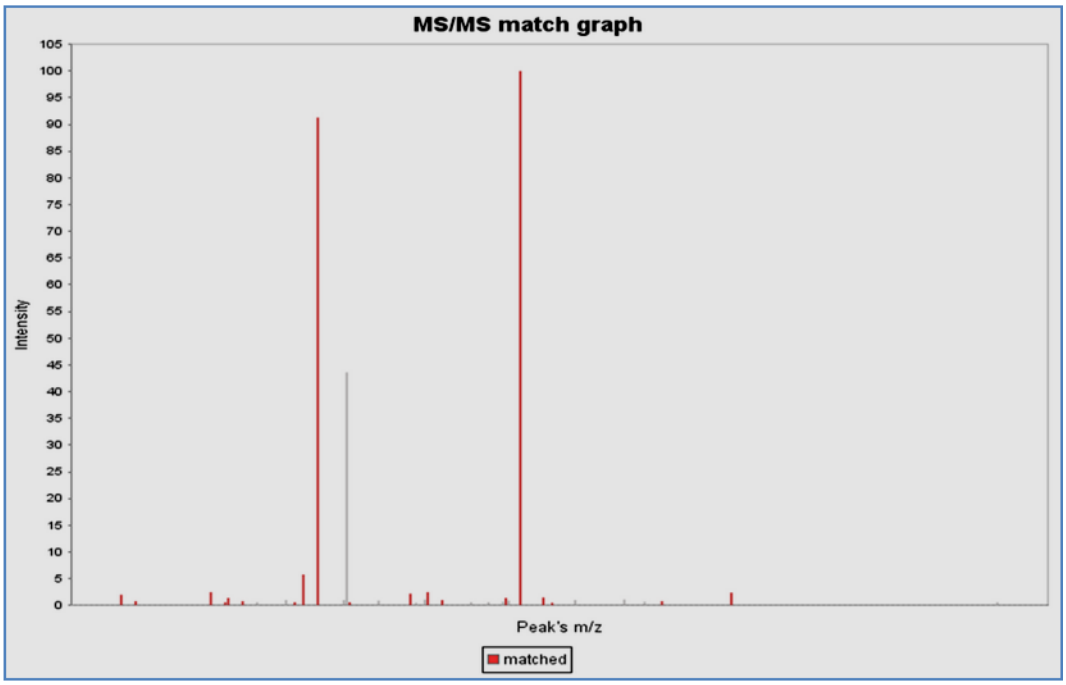
Min Precursor Mass:		Max Precursor Mass:	
Min Intensity:		Max Intensity:	
Min Fragments:		Max Fragments:	
Min Hits:		Max Hits:	
Min Fit Score:	0.7	Max Fit Score:	

Show 10 entries Search:

#	Retention Time	Precursor Mass	Precursor Intensity	No. of Fragments	No. of Hits	Max Fit Score	Show Details	Save Result
4	5.34	110.06063	114212	20	1	0.89	Show detail	CSV
6	5.44	282.12013	478864	84	4	0.98	Show detail	CSV
9	5.65	132.10302	99708	22	6	0.99	Show detail	CSV

Figure 5.2 Screenshot of search MCID MS/MS results. (A) Single-mode search. (B) Batch-mode search.

(A)



(B)

Experimental peak	Intensity	Matched simulated peaks	Detail information	Experimental peak	Intensity	Matched simulated peaks	Detail information
39.0228	2.0	2	Detail	41.0385	0.8	3	Detail
51.0229	2.5	3	Detail	53.0021	0.6	2	Detail
53.0387	1.4	2	Detail	55.018	0.8	2	Detail
56.9648	0.6	0		60.9865	1.0	0	
62.0154	0.6	2	Detail	63.023	5.8	3	Detail
65.0388	91.3	3	Detail	68.4526	1.0	0	
68.9523	43.6	0		68.9975	0.6	1	Detail

(C)

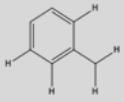
Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
89.0391	2	91.0542	91.0542	0.0000	

Figure 5.3 Screenshot of showing more details

To facilitate manual comparison of an experimental MS/MS spectrum and a predicted spectrum, there is also a function of uploading the matched metabolite structure to a local ChemDraw software or an online ChemDraw Plugin (freeware). In both programs, a built-in

"Fragmentation Tools" can be used to direct a bond breakage of a structure to show the resulting fragment ion structures and their masses. An example of how to use this tool for manual fragmentation pattern interpretation has been given in the original MCID paper.[144]

In addition to single spectrum search, a user can upload a CSV file generated from LC-MS/MS analysis of a sample for batch mode search. This is useful for examining all the possible matches in a metabolomic profiling experiment. The file format used is shown in Supplemental File F5.1. To share the computation resource in the MCID server by multiple users, the file size is limited to 100 MS/MS spectra. For a large file, a file split program can be downloaded to split the file into several small files with each limited to 100 spectra. These split files can be uploaded individually for MS/MS search. The search time for each file depends on the parameters used (e.g., a smaller precursor mass tolerance would increase the search speed as fewer candidates would need to be examined in MS/MS search) and the number of search jobs in the server. After the searches, the individual search results can be merged by a file merge program which can also be downloaded from the website to produce the final result in CSV.

Figure 5.2 B shows a screenshot of part of a search result page from MS/MS spectra of a human urine sample acquired by LC-QTOF-MS. A summary table lists information on retention time, precursor ion mass, precursor ion intensity, the number of fragment ions detected, the number of mass-matched metabolites (i.e., number of hits), fit score, show-details with links and save-result in CSV for a given match. By clicking the show-detail, several layers of information can be displayed for each MS/MS match as in the case of single spectrum search discussed earlier.

The search results can be sorted according to any parameters in the summary table. There are several parameters (see the top of Figure 5.2B) that can be used to filter the search results to

retain the matches of interest. By clicking "Download Table Result", all the filtered matches are saved to the user's computer in a CSV file (see Supplemental Table T5.1 as an example). For privacy and confidentiality, the server does not store any search file or search results. However, in the saved CSV file that can be opened in Excel, there is a link column containing long names for all the individual matches. By copying and pasting a link name of a match to the web, the user can retrieve the search result in MCID for the match. This is possible because the long name contains all the MS and MS/MS information required for a new MCID MS/MS search to generate the match result again. This feature allows a user to examine any matches in the result table without the need of repeating the batch mode search.

5.3.2 MS/MS Search of Standards

To evaluate the performance of MCID MS/MS search, we searched the MS/MS spectra of 35 standards generated by QTOF-MS against the predicted MS/MS spectral library. These metabolites were randomly picked in order to cover as many different types of compounds as possible. Table 5.1 shows the list of metabolites and their search results. The MS/MS spectra were searched using the zero- and one-reaction libraries with a normal (i.e., 0.005 Da, a typical mass accuracy from QTOF-MS). Also, a wider (i.e., 0.05 Da for zero-reaction and 0.01 Da for one-reaction) precursor ion mass tolerance was used to artificially include more possible candidates and thus evaluate the MS/MS search in terms of distinguishing the correct structure from false ones. The wider tolerance was deliberately used in order to increase the number of mass-matched metabolites including many false ones for the purpose of testing the ability of using MS/MS search to distinguish the correct structure from the false ones. The fragment ion mass tolerance was set to be 0.005 Da, according to the QTOF-MS/MS mass accuracy.

Table 5.1 Summary of MCID MS and MS/MS search results for 35 metabolite standards.

#	Name	Zero-reaction library search						One-reaction library search						Fit score for correct structure
		Precursor mass tolerance ±0.05 Da			Precursor mass tolerance ±0.005 Da			Precursor mass tolerance ±0.01 Da			Precursor mass tolerance ±0.005 Da			
		Rank	# of MS/MS match	# of MS match	Rank	# of MS/MS match	# of MS match	Rank	# of MS/MS match	# of MS match	Rank	# of MS/MS match	# of MS match	
1	Adenine	1	1	5	1	1	1	1	3	19	1	3	19	0.815
2	Androstenedione	3	10	12	1	1	1	3	28	33	2	28	32	0.896
3	Dopamine	1	3	6	1	3	3	1	19	22	1	19	22	0.819
4	Folic acid	1	1	1	1	1	1	4 (3)*	12	23	3	6	12	0.873
5	Glycine	2	3	3	1	1	1	1	15	17	1	15	17	0.993
6	Glutathione	1	1	4	1	1	1	1	5	22	1	5	6	0.892
7	L-Phenylalanine	1	4	14	1	2	4	1	28	40	1	20	40	0.838
8	L-Alanine	1	4	4	1	4	4	1	27	31	1	23	32	0.810
9	Riboflavin	1	0	1	1	0	1	1	0	13	1	0	7	0.554
10	Thymine	1	0	6	1	0	2	1	0	4	1	0	6	0.114
11	Sarcosine	1	4	4	1	4	4	1	27	31	1	27	32	0.909
12	Tryptamine	1	2	13	1	1	2	1	2	25	1	2	10	0.869
13	Tyramine	1	3	10	1	3	4	1	11	17	1	11	18	0.700
14	Chenodeoxycholic acid	3	0	18	3	0	18	7 (2)	0	44	7 (2)	0	44	0.465
15	Creatinine	1	1	3	1	1	1	1	1	1	1	1	1	0.764
16	Isovalerylcarnitine	1	3	3	1	3	3	2	17	18	2	15	16	0.980
17	L-Methionine	1	2	6	1	2	2	1	13	17	1	13	19	0.965
18	trans-Ferulic acid	1	2	13	1	2	3	1	30	39	1	30	39	0.969

19	2'-Deoxyguanosine 5'-monophosphate	1	4	4	1	4	4	1	22	34	1	22	32	0.958
20	N-Acetylmannosamine	2	1	8	3	1	7	11 (3)	28	43	11 (3)	23	37	0.407
21	Melatonin	1	1	3	1	1	1	1	7	16	1	6	12	0.986
22	Pyridoxamine	1	1	8	1	1	1	2	5	16	2	5	16	0.820
23	N-Acetylputrescine	1	2	11	1	1	1	1	10	10	2	10	10	0.971
24	Creatine	1	1	17	1	1	2	1	4	5	1	4	7	0.985
25	L-Asparagine	1	5	24	1	3	5	1	9	13	1	9	12	0.984
26	L-Cystine	1	1	1	1	1	1	1	3	6	1	3	3	0.949
27	Ornithine	1	2	23	1	2	2	1	12	12	1	12	12	0.945
28	Pyridoxine	1	4	8	1	4	4	4 (2)	20	25	4 (2)	20	25	0.935
29	Taurine	1	5	5	1	1	1	1	4	1	1	2	2	0.873
30	Uric acid	1	1	9	1	1	1	1	3	9	1	3	3	0.839
31	Xanthine	1	3	19	1	3	3	1	3	7	1	3	6	0.962
32	Xanthosine	1	3	4	1	1	1	2	11	21	1	7	10	0.971
33	DL-Homocystine	1	2	10	1	2	2	1	4	11	1	2	8	0.936
34	4-Hydroxyproline	1	4	17	1	3	8	1	8	52	1	8	52	0.998
35	Xanthurenic acid	1	4	5	1	1	1	1	9	17	1	9	17	0.958

*(x) where x=new rank after grouping isomers as one group.

With a wider precursor ion mass tolerance, for the zero-reaction library search, an average of 8.6 library compounds were mass-matched to a tested standard, while MS/MS search resulted in an average of 2.5 matched compounds with a fit score of ≥0.700 (see below). For the one-reaction search, an average of 20.4 compounds were matched to a standard if only mass search was used. With MS/MS search, an average of 11.4 compounds with a fit score of ≥0.700 was matched to a standard. Using the precursor ion mass tolerance of 0.005 Da, for the zero-reaction search, an average of 2.9 and 1.7 compounds were matched to a standard using MS search and MS/MS search, respectively. For the one-reaction library, MS search and MS/MS search resulted in an average of 18.2 and 10.4 compounds matched to a standard. These results show that the number of MS/MS matched structures with a fit score of ≥0.700 is significantly lower than the number of MS matched structures.

Since the structures of the 35 standards are known, we can examine the accuracy of the MS/MS matches in a rank according to the fit score. For the zero-reaction search using a wider mass tolerance, 31, 2 and 2 standards (88.6%, 5.7% and 5.7%) gave the correct compound as the top, 2nd and 3rd ranked match, respectively (see Table 5.1). Even for the one-reaction search, 27, 3 and 1 standards (77.1%, 8.6% and 2.9%) gave the correct structure as the top, 2nd and 3rd ranked match, respectively. Only 4 standards had the correct structure ranked below the 3rd match. The 11th ranked N-acetylmannosamine, out of 43 mass-matched compounds, have isomers ranked from top 1 to 10. Effectively this match was ranked 3rd if isomers were counted as one (see Table 1 with the new rank in brackets). Similarly, for the 7th ranked chenodeoxycholic acid, out of 44 mass-matched compounds, the top 5 matches were isomers. Counting all the isomers as one, this match was ranked 2nd. In the case of using 0.005 Da

precursor ion mass tolerance for MS/MS search, as Table 1 shows, for the zero-reaction library, 33 (94.3%), 0 (0%), 2 (5.7%) standards gave the correct structure as the top, 2nd and 3rd ranked match, respectively. For the one-reaction library, 27 (77.1%), 4 (11.4%), 1 (2.9%) standards gave the correct structure as the top, 2nd and 3rd ranked match. Only 3 (8.6%) standards were below top three. These three cases would be ranked top 3 if grouping the isomers as one.

The above results show that the correct structure of a MS/MS search belongs to one of the top three structures with majority of them as the top match. This finding would suggest that, for a MS/MS search, only the top 3 structures including isomers need to be inspected manually to confirm or disapprove a match. This should greatly improve the overall metabolite identification efficiency. For the 35 standards, after generating the top three structure matches for each metabolite in the one-reaction search results, we manually checked the matches to validate the automatic MS/MS search results. 27 top ranked metabolites could be manually confirmed. For the 2nd ranked metabolites, 3 out of 4 could be confirmed by manually eliminating the top ranked false match. Only one of the 2nd ranked metabolites (isovalerlcarnitine, #16 in Table 5.1) could not be differentiated from the top ranked structure due to the lack of characteristic fragment ions from the two structures.

Table 5.1 also lists the fit score of the correctly matched structure from MS/MS search for each standard. The fit score determines the matching quality of the experimental MS/MS data with the predicted MS/MS data. The average fit score for all the correct structures is 0.860 and 90% of the correct structures have fit scores of ≥0.700. There are 3 cases that the correct structures have a fit score of <0.700. Manual inspection of these matches shows that these spectra do not have enough high intensity and informative fragment ion peaks. For example, in the case of thymine (#10 in Table 5.1) with a fit score of 0.114, the MS/MS spectrum shows only

one fragment ion peak and this peak cannot be explained even with manual interpretation. This observation is not surprising, considering that not all the metabolites can be fragmented or produce a sufficient number of characteristic fragment ions. Nevertheless, more than 90% of the 35 standards could produce MS/MS spectra with sufficiently high quality to render a fit score of ≥0.700. Thus, a fit score of 0.700 can be used as a cut-off threshold for automated MS/MS search to produce a list of structure candidates from which manual interpretation can be carried out to approve or disapprove a structure match.

5.3.3 MS/MS Search of Urine Metabolites

To demonstrate the utility of MCID MS/MS search in real world applications, a human urine sample was analyzed by LC-QTOF-MS, followed by library search for metabolite identification. In this experiment, a precursor ion exclusion (PIE) strategy, similar to that used in shotgun proteomics work,[168] was applied to acquire as many MS/MS spectra as possible from triplicate runs of the same human urine.

In total, 5794 MS/MS spectra were generated using the PIE strategy. We used 0.005 Da mass tolerance for precursor and fragment ions in MCID MS/MS search and generated 1698 spectral matches using the zero-reaction library (see Supplemental Table T5.2 for the list). We then performed a cross-validation of some of the spectral matches using a Bruker HMDB MS/MS spectral library. This Bruker library containing 800 standards was created in the same QTOF instrument as the one used for running the urine sample. Thus, excellent fragmentation pattern match of the urine metabolite and library standard is expected, which should in turn provide high confidence for validation of the MCID MS/MS search results. One example of the validation work is shown in Figure 5.4. Figure 5.4A shows the experimental MS/MS spectrum of carnitine found in urine at the retention time of 2.55 min. Figure 5.4B shows the match of the

experimental MS/MS with the predicted MS/MS spectrum of carnitine in the MCID library. Figure 5.4C shows the standard MS/MS spectrum of carnitine in the Bruker library. The fit score for this compound using the predicted spectral library was 0.995, compared to purity score of 954 out of 1000 using the Bruker library. Thus, the MCID MS/MS search result or identification of carnitine was cross-validated.

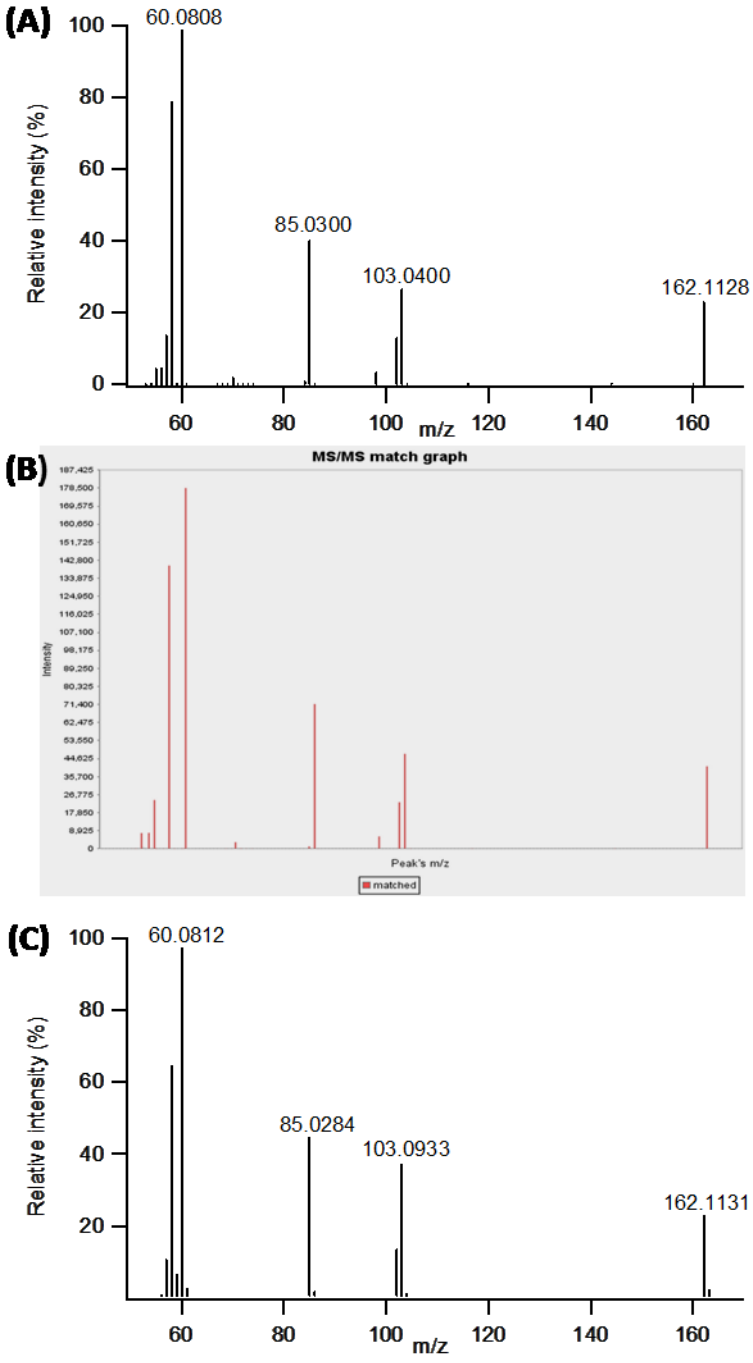


Figure 5.4 One example of the MS/MS validation

Table 5.2 lists the metabolites initially identified using MCID MS/MS search with the zero-reaction library and subsequently validated using the Bruker experimental spectral library. Out of the 77 validated spectral matches, 54, 18, 3 metabolites were correctly identified by

MCID MS/MS as the top (70.1%), 2nd (23.4%) and 3rd (3.9%) ranked structure, respectively. Two of them were ranked below top 3. However, if treating isomers as a group, only 1 spectral match had the correct structure ranked below top 3. Thus, 76 out of the 77 spectral matches (97.4%) had the correct structure belonging to one of the top 3 matched structures. The average fit score was 0.775. These results indicate that using MCID MS/MS search, almost all the correctly matched metabolites could be found as the top 3 structures in a LC-MS/MS experiment of a real biological sample. If this holds true for the other non-validated matches, only the top 3 structures from a MCID MS/MS search of an unknown metabolite would need to be inspected or confirmed for identification. Of course, more validation work is needed to generalize this finding in the future. Nevertheless, the urine results illustrate that MCID MS/MS search is capable of identifying metabolites with high confidence.

Table 5.2 Summary of zero-reaction library MS/MS spectral match results of 77 metabolites used for cross-validation in the urine sample analysis.

No.	Precursor *m/z*	RT (min)	LC-MS run #	ID	Correct structure rank	# of MS match	Fit score
1	156.078	2.07	1	L-Histidine	1	1	0.969
2	147.114	2.12	1	L-Lysine	1	1	0.957
3	118.086	2.45	2	Betaine	4	5	0.559
4	170.093	2.52	2	1-Methylhistidine	1	2	0.978
5	162.114	2.55	1	L-Carnitine	1	2	0.995
6	76.0759	2.56	1	Trimethylamine N-oxide	2	2	0.99
7	137.071	2.59	3	1-Methylnicotinamide	1	1	0.436
8	104.07	2.62	3	L-Alpha-aminobutyric acid	1	8	0.909
9	138.055	2.67	2	Trigonelline	3	5	0.454
10	189.123	2.71	3	N6-Acetyl-L-lysine	1	3	0.678
11	204.123	3.03	1	L-Acetylcarnitine	1	1	0.884
12	130.05	3.39	3	Pyrrolidonecarboxylic acid	3	5	0.634

13	189.124	3.66	1	N-Alpha-acetyllysine	1	3	0.952
14	150.06	3.97	2	L-Methionine	1	2	0.951
15	139.052	4.13	1	Urocanic acid	1	2	0.927
16	148.061	4.4	2	L-Glutamic acid	2	6	0.372
17	137.047	4.79	1	Hypoxanthine	1	3	0.955
18	282.12	5.44	1	1-Methyladensine	1	4	0.98
19	240.11	5.89	1	Dihydrobiopterin	1	3	0.598
20	132.102	5.97	3	L-Norleucine	1	6	0.992
21	132.103	6.25	2	L-Isoleucine	1	6	0.988
22	182.082	6.4	1	L-Tyrosine	2	5	0.943
23	138.092	7.56	2	Tyramine	1	4	0.884
24	385.13	8.88	2	S-Adenosylhomocysteine	1	1	0.887
25	268.104	10.41	2	Adenosine	2	3	0.921
26	330.06	11.85	3	Cyclic AMP	2	2	0.998
27	166.087	13.73	1	L-Phenylalanine	1	4	0.89
28	154.05	17.35	1	3-Hydroxyanthranilic acid	1	2	0.826
29	220.119	22.24	2	Pantothenic acid	1	2	0.931
30	137.046	22.33	1	Hypoxanthine	1	3	0.836
31	197.067	22.65	1	1,3-Dimethyluric acid	2	8	0.635
32	153.128	22.88	2	Perillyl alcohol	2	3	0.472
33	181.072	22.94	1	Theobromine	1	17	0.906
34	196.06	23.54	2	Salicyluric acid	3	6	0.65
35	158.082	23.6	1	Tiglylglycine	1	2	0.939
36	160.097	24.12	2	Isovalerylglycine	2	6	0.98
37	118.086	24.44	2	L-Valine	1	5	0.557
38	169.05	24.51	2	Isovanillic acid	4 (2)	7	0.863
39	181.072	24.68	2	Theophylline	2	17	0.975
40	206.044	24.91	2	Xanthurenic acid	1	1	0.934
41	181.061	25.02	3	Nicotinuric acid	1	1	0.994
42	205.097	25.04	1	L-Tryptophan	1	1	0.743
43	162.056	25.74	3	Indole-3-carboxylic acid	2	3	0.99
44	190.05	25.79	1	Kynurenic acid	1	1	0.642
45	153.128	26.02	3	Perillyl alcohol	2	3	0.881
46	281.113	26.04	1	L-Aspartyl-L-phenylalanine	1	2	0.884
47	164.037	26.57	1	Acetylcysteine	1	1	0.427
48	295.129	26.62	2	Aspartame	2	2	0.677
49	246.17	26.76	1	Isovalerylcarnitine	1	3	0.985

50	161.107	26.78	2	Tryptamine	1	2	0.806
51	195.088	26.94	1	Caffeine	1	1	0.954
52	180.066	27.31	1	Hippuric acid	1	2	0.994
53	111.043	27.41	3	Pyrocatechol	1	2	0.858
54	130.051	27.66	1	Pyrrolidonecarboxylic acid	2	5	0.885
55	265.119	27.69	2	Alpha-N-phenylacetyl-L-glutamine	1	2	0.968
56	377.146	28.06	1	Riboflavin	1	1	0.879
57	160.133	28.29	1	DL-2-aminooctanoic acid	1	1	0.779
58	116.071	28.45	1	L-Proline	1	2	0.878
59	194.081	29.17	2	Phenylacetylglycine	2	3	0.953
60	162.056	29.96	1	Indole-3-carboxylic acid	1	3	0.968
61	131.107	30.47	2	Heptanoic acid	1	1	0.398
62	197.082	30.55	2	Homoveratric acid	1	1	0.943
63	147.076	33.06	2	L-Glutamine	1	4	0.692
64	206.082	32.4	1	Indolelactic acid	1	3	0.877
65	231.16	32.75	1	Dodecanedioic acid	1	1	0.261
66	189.113	33.58	1	Azelaic acid	2	2	0.824
67	365.232	34.53	1	Tetrahydrocortisone	2	6	0.681
68	176.07	34.82	1	Indoleacetic acid	2	3	0.896
69	466.316	35.5	1	Glycocholic acid	1	2	0.129
70	160.134	36.82	3	DL-2-aminooctanoic acid	1	1	0.704
71	245.175	43.07	2	1,11-Undecanedicarboxylic acid	1	1	0.37
72	173.154	50.88	1	Capric acid	1	1	0.446
73	122.097	51.25	2	N,N-Dimethylaniline	1	3	0.418
74	199.171	52.15	1	5-Dodecenoic acid	1	2	0.274
75	201.185	56.89	1	Dodecanoic acid	1	1	0.531
76	283.264	57.94	2	Elaidic acid	2	3	0.716
77	283.262	58.73	2	Vaccenic acid	1	3	0.245

The fit scores of the 77 metabolites were analyzed to determine the best cut-off for high confident MS/MS match. From the study of the 35 metabolite standards, we proposed to use 0.700 as the cut-off. However, we noticed that, in the urine sample analysis, this cut-off score is too restricted in some cases. This is because, in the analysis of a complicated biological sample,

metabolites have a wide concentration range and their MS/MS signals can be affected by precursor ion intensity, background impurities and co-eluting compounds. For example, of the 77 metabolites, only 52 (67.5%) of the correctly matched structures had their fit score of >0.700. Another 18 (23.4%) of the correct structures had a fit score of between 0.700 and 0.400 and 7 (9.1%) structures even had a fit score of below 0.400. These results indicate that using a cut-off fit score of 0.700 will exclude a large fraction of the correct structures. On the other hand, of the 52 metabolites with a fit score of larger than 0.700, their correct structures were all ranked at the top 3. Thus, we can use a cut-off of 0.700 to generate a list of high confident structure matches where one of the top 3 structures is expected to be correct. For the remaining spectral matches with a fit score of below 0.700, we would still examine the top three structure matches of an experimental MS/MS spectrum to determine if one of the matches is correct; however, there is no guarantee that any of the top 3 structures is correct in these cases. We recognize that simply using a fit score cut-off of 0.700 represents a compromise between the search specificity and sensitivity. Future work will be needed to develop a more robust scoring system for MS/MS search to increase both specificity and sensitivity.

We applied the 0.700 cut-off threshold to all of the 1160 spectral matches including the 77 validated matches. We found that 636 MS/MS spectra have structure matches with a fit score of ≥0.700 for a total of 1227 structures (see Supplemental Table T5.3). Among them, 378, 126, and 54 spectral match with 1, 2, and 3 structures, respectively, and 78 spectral match with 4 or more high-score structures. While we cannot narrow down each spectral match to one structure, we can state that 636 MS/MS spectra have high confident structure matches and one of the top three structures for each spectral match is most likely correct. It is clear that MCID MS/MS search can generate many high confident, but still putative identifications from a urine sample.

Finally, we would like to illustrate the power of using MCID MS/MS to search a predicted MS/MS spectral library of one-reaction metabolites. Out of the 5794 MS/MS spectra collected from the urine sample, we took the remaining unmatched or unconfirmed spectra (i.e., 5158) from the zero-reaction library search to search the one-reaction library and the results are shown in Supplemental Table T5.4. A total of 3920 (76.0%) MS/MS spectra were matched to the one-reaction library. Among them, 1250 spectra have a total of 5966 structures matched with a fit score of ≥0.700 (see Supplemental Table T5.5). This includes 587, 380, and 123 spectra match with 1, 2, and 3 structures, respectively, and 160 spectra match with 4 or more high-score structures.

To validate some of these matches, we used the published data of 87 one-reaction metabolites that were identified based on manual interpretation of the mass-matched metabolites in MCID MS search.[144] Based on the match of retention time, precursor mass and MS/MS fragmentation pattern, 78 out of these 87 metabolites (88.5%) were identified in the current urine sample (see Table 5.3). Among the 78 metabolites, 44 (56.8%), 17 (21.8%) and 6 (7.7%) spectra had their correctly matched structures ranked at the top, 2nd and 3rd, respectively. Only 11 correct structure matches were ranked below the top 3. If treating isomers as a group, out of these 11 matches, 3 matches were ranked #2 and 2 matches were ranked #3. Thus, 72 out of the 78 metabolites (92.3%) had the correct structure belonging to one of the top 3 structure matches. Among the 78 metabolites, there were 57 spectral matches with a fit score of ≥0.700. For these matches, 56 of them (98.2%) had a correct structure listed at the top 3 matches (treating isomers as a group). These results demonstrate that a fit score cut-off threshold of 0.700 can narrow down the correct structure to one of the top 3 structure matches, even for the one-reaction library search. If this holds true, one of the top 3 matches for each of the 1250 MS/MS spectra having

one-reaction library metabolite matches with a fit score of ≥0.700 in Supplemental Table T5.5 should be the correct structure.

Table 5.3 Summary of one-reaction library MS/MS spectral match results of 78 metabolites used for cross-validation in the urine sample analysis.

No.	Precursor m/z	RT (min)	LC-MS run #	ID	Correction structure rank	# of MS match	Fit score
1	137.0467	4.79	1	8-Hydroxypurine - H2 or isomers	1	16	0.960
2	169.0601	4.12	2	3-Hydroxyanthranilic acid + NH or isomers	1	20	0.537
3	171.088	4.88	1	L-Histidine + NH or isomers	1	2	0.689
4	188.1035	4.1	1	Homocitrulline - H2 or isomers	1	9	0.893
5	190.1067	26.74	2	Suberic acid + NH or isomers	5 (2)	25	0.900
6	197.1279	25.97	1	L-prolyl-L-proline – O or isomers	4	7	0.910
7	203.0802	8.71	1	Indoleacrylic acid + NH or isomers	1	9	0.731
8	209.1168	31.75	3	3-Methoxybenzenepropanoic acid + C2H4 or isomers	1	2	0.700
9	220.0601	28.04	1	N'-Formylkynurenine – NH3 or isomers	8	13	0.654
10	222.0787	7.51	1	3-Hydroxyvaleric acid + C3H5NOS or isomers	6	26	0.617
11	224.1278	39.31	1	Perillic acid + C2H3NO or isomers	1	1	0.930
12	226.0823	4.25	1	Deoxycytidine – H2 or isomers	1	22	0.824
13	257.1485	30.62	1	Dethiobiotin + C2H2O	3	5	0.685
14	257.2263	47.61	1	Retinoic acid or isomers - CO2	1	6	0.623
15	257.2261	49.04	1	Retinoic acid or isomers - CO2	1	6	0.969
16	259.165	31.86	2	Capryloylglycine + C2H3NO	4	6	0.328
17	262.1652	9.95	1	Hydroxyvalerylcarnitine or isomers – H2	1	22	0.891
18	262.1658	11.42	1	Hydroxyvalerylcarnitine or isomers – H2	1	22	0.956
19	263.1382	28.31	2	L-phenylalanyl-L-hydroxyproline - O	1	8	0.883
20	266.102	30.16	1	(R)-2-Benzylsuccinate + C2H3NO	2	19	0.883
21	269.1231	3.69	1	Homocarnosine + CO	2	10	0.969
22	272.1845	31.99	1	Heptanoylcarnitine - H2 or isomers	1	7	0.802

23	272.1852	31.29	1	Heptanoylcarnitine - H2 or isomers	1	7	0.968
24	272.1852	30.51	1	Heptanoylcarnitine - H2 or isomers	1	7	0.979
25	273.2207	41.63	1	Androstenol - H2	3	23	0.733
26	273.2214	41.63	1	Androstenol - H2	4(2)	23	0.733
27	284.1851	31.54	1	2-Octenoylcarnitine - H2 or isomers	2	3	0.964
28	284.1856	30.47	2	2-Octenoylcarnitine - H2 or isomers	2	3	0.975
29	285.2573	40.79	1	Docosahexaenoic acid - CO2	1	1	0.894
30	286.1272	12.45	3	trans-trans-Muconic acid + C7H13NO2 or isomers	1	13	0.946
31	287.1996	35.42	2	Testosterone - H2 or isomers	6 (2)	32	0.412
32	300.2166	35.37	1	9-Decenoylcarnitine - CH2 or isomers	1	4	0.989
33	300.2169	46.77	2	9-Decenoylcarnitine - CH2 or isomers	1	4	0.991
34	302.1608	23.74	1	Pimelylcarnitine - H2 or isomers	2	5	0.803
35	302.1959	39.65	2	Heptanoylcarnitine + CO or isomers	2	9	0.522
36	302.1953	29.04	1	Heptanoylcarnitine + CO or isomers	2	9	0.730
37	302.1955	34.78	2	Heptanoylcarnitine + CO or isomers	3	9	0.634
38	302.1958	30.37	2	Heptanoylcarnitine + CO or isomers	1	9	0.612
39	303.1004	28.13	1	4-Hydroxy tolbutamide + O	4 (3)	5	0.446
40	304.2109	31.51	1	3-Hydroxyoctanoic acid + C7H13NO2 or isomers	1	10	0.978
41	304.2119	31.12	1	3-Hydroxyoctanoic acid + C7H13NO2 or isomers	1	19	0.991
42	310.2008	35.02	1	2-trans,4-cis-Decadienoylcarnitine - H2 or isomers	2	4	0.976
43	310.2012	36.32	1	2-trans,4-cis-Decadienoylcarnitine - H2 or isomers	2	4	0.983
44	316.175	26.19	1	2-Octenedioic acid + C7H13NO2 or isomers	1	2	0.952
45	316.211	30.76	2	6-Keto-decanoylcarnitine - CH2 or isomer	3	5	0.901
46	318.1908	27.94	1	Hydroxyhexanoycarnitine + C2H2O or isomers	2	6	0.877
47	318.2068	35.16	2	16a-Hydroxyandrost-4-ene-3,17-dione + NH3 or isomers	1	10	0.188
48	319.1651	30.09	1	Indoleacetic acid + C7H13NO2	1	5	0.981
49	326.0869	10.47	2	Inodxyl glucuronide + O or isomers	6 (3)	19	0.953
50	328.2108	30.28	1	2-trans,4-cis-Decadienoylcarnitine + O or isomers	1	3	0.681

51	328.2111	31.78	1	2-trans,4-cis-Decadienoylcarnitine + O or isomers	1	3	0.972
52	328.2112	33.09	1	2-trans,4-cis-Decadienoylcarnitine + O or isomers	1	3	0.986
53	328.248	39.89	1	4,8 dimethylnonanoyl carnitine - H2 or isomers	2	12	0.992
54	328.2472	39.82	2	4,8 dimethylnonanoyl carnitine - H2 or isomers	2	12	0.993
55	330.1907	28.51	1	2-Octenoylcarnitine + CO2	1	2	0.971
56	330.2267	30.07	1	9-Decenoylcarnitine + O	2	6	0.847
57	332.2061	29.58	1	Nonate + C7H13NO2	1	8	0.939
58	332.2422	33.5	1	(R)-3-Hydroxydecanoic acid + C7H13NO2 or isomers	1	7	0.659
59	332.2422	36.09	1	(R)-3-Hydroxydecanoic acid + C7H13NO2 or isomers	1	7	0.788
60	337.1754	30.2	1	Phenylacetylglycine + C7H13NO2 or isomers	2	6	0.899
61	341.1697	20.5	2	L-Dopa + C7H13NO2	1	9	0.982
62	342.2272	31.98	1	9-Decenoylcarnitine + CO	2	3	0.861
63	344.2055	25.73	1	Decenedioic acid + C7H13NO2 or isomers	1	1	0.447
64	344.2059	30.79	3	Decenedioic acid + C7H13NO2 or isomers	1	1	0.705
65	346.1256	5.26	1	Muramic acid + C4H2N2O	5 (4)	25	0.598
66	346.2207	31.43	1	Nonanoylcarnitine + CO2 or isomers	1	4	0.862
67	356.2428	37.22	1	9-Decenoylcarnitine + C2H2O	3	3	0.889
68	358.258	38.66	1	2-Hydroxylauroylcarnitine – H2	3	10	0.978
69	384.115	23.66	1	Adenylsuccinic acid -HPO3	1	15	0.901
70	384.2738	38.32	1	3, 5-Tetradecadiencarnitine - O	1	11	0.746
71	402.2832	34.54	1	Dodecanedioylcarnitine + +C2H4	2	6	0.898
72	432.3101	40.89	1	Lithocholic acid glycine conjugate - H2 or isomer	1	9	0.736
73	464.1897	34.77	1	Tetrahydrofolic acid + H2O	7 (5)	11	0.355
74	523.2527	32.21	2	Cortexolone + C6H8O6	1	11	0.446
75	593.3332	44.33	1	D-Urobilinogen + H2 or isomers	1	7	0.987
76	595.3481	45.12	1	L-Urobilinogen - H2	2	5	0.971
77	626.2066	31.06	3	Hesperidin + NH	1	2	0.585
78	642.3468	33.03	1	Glycochenodeoxycholic acid 3-glucuronide + O	1	10	0.636

Taken together, the urine sample analysis results indicate that, in most cases, only the top 3 structure matches from the MS/MS search of an experimental MS/MS spectrum with a fit score of ≥0.700 needs to be manually inspected for confirming or disapproving a match. Of course, for positive metabolite identification, an authentic standard is needed to confirm a structure match. In this regard, using MCID MS/MS search, standards of only a few top ranked candidates need to be acquired or synthesized, which should greatly reduce the time and efforts needed for metabolite identification.

5.4 Conclusions

We have developed a web-based MS/MS spectral search tool for improving metabolite identification based on the use of a large library of predicted fragment-ion-spectra of over 383,830 possible human metabolites. This tool is freely accessible at www.MyCompoundID.org, allowing a user to search a MS/MS spectrum or a batch of spectra against the library for possible structure matches. Using MS/MS spectra collected from 35 standards and a human urine sample, we demonstrated that one of the top 3 matches from a MS/MS spectrum with a fit score of greater than 0.700 (out of 1.000) is a correct structure. While MCID MS/MS spectral search cannot always produce one unique structure match, narrowing down the possible matches to the top 3 candidates should save the time and efforts to find or synthesize authentic compound standards for positive identification.

Chapter 6

Saliva Metabolomic Changes Associated with Mild Cognitive Impairment and Alzheimer's Disease Revealed by Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry

6.1 Introduction

Mild cognitive impairment (MCI) is defined as a clinical state characterized by significant cognitive impairment in the absence of dementia. MCI is known as the transition between normal aging and the prodromal phase of Alzheimer's disease (AD).[169] Early diagnosis of MCI and AD can assist in management and treatment of the diseases.[170, 171] There is a pressing need for an improved, rapid and sensitive method for diagnosing MCI patients, as the preclinical period of AD can reveal valuable information to develop interventions that may delay or prevent AD.[172] While discovering protein biomarkers of MCI and AD is an active area of research,[173-176] applying metabolomics to reveal metabolic perturbations in many biochemical pathways related to MCI and AD may allow us to discover a panel of metabolite biomarkers specific to MCI and AD.[177]

Over the past few years, a number of metabolomics studies on AD and a few on MCI have been reported. These studies used human CSF,[178, 179] plasma/serum,[180-182] or brain tissue samples.[183] Human saliva is another attractive medium because it is inexpensive and easy to collect and non-invasive. Saliva, often considered as "the mirror of the body", is secreted from three pairs of major salivary glands and many salivary glands lying beneath the oral mucosa.[184] Saliva metabolomics is an attractive approach to screen for potential diagnostic and prognostic biomarkers to distinguish different states of diseases.[185] For example, saliva

metabolomic profiles of healthy controls and of patients with oral, breast or pancreatic cancer have been investigated to differentiate different groups.[186]

To search for potential biomarkers of MCI and AD, an analytical technique that allows accurate quantification and high metabolome coverage is needed. To this end, we have recently reported a quantitative technique based on ^{13}C-/^{12}C-isotope dansylation labeling combined with liquid chromatography Fourier-transform ion cyclotron resonance (FTICR) mass spectrometry (MS) for in-depth profiling of the amine/phenol submetabolome in human saliva sample.[187] In this work, we report a workflow with improved submetabolome coverage for saliva metabolomics and the application of this workflow for discovering potential metabolite biomarkers to differentiate MCI, AD and normal controls (NA) using 82 saliva samples in the training set and 27 independent samples in the validation set.

6.2 Experimental Section

6.2.1 Subjects

Ethics approval of this work was obtained from the University of Alberta according to the university's health research policy. Saliva samples were collected from 109 participants including 82 samples for the biomarker discovery work (i.e., the training set) and 27 samples for the biomarker validation work (i.e., the validation set). The training set (TS) consisted of n = 35 NA adults (age = 64-75 years; 62.9% female), n = 25 MCI adults (age: 64-75 years; 60% female), and n = 22 AD patients (age = 52-91 years; 72.7 % female). The validation set consisted of n = 10 NA adults, n = 10 MCI adults, and n = 7 AD adults. The demographic characteristics

and clinical information of the participants are shown in Table 6.1. Each individual sample was analyzed in experimental triplicates.

Table 6.1 Baseline characteristics for discovery and validation samples

	Discovery			**Validation**		
Characteristics	NA	MCI	AD	NA	MCI	AD
n	35	25	22	10	10	7
Age (years)	69.94 (3.80)	70.40 (3.38)	77.09 (11.20)	71.40 (2.84)	71.50 (2.51)	70.11 (16.60)
Gender (M/F)	13/22	10/15	6/16	5/5	5/5	2/5
Education, years	15.69 (2.69)	14.68 (2.94)	11.59 (3.23)	14.80 (3.26)	15.40 (3.03)	14.00 (2.08)
Mini-Mental State Exam	28.46 (1.42)	27.39 (3.14)	21.32 (4.76)	28.70 (1.06)	27.70 (1.89)	19.57 (6.58)

6.2.2 Sample Collection and Storage

Salivary samples were collected using Oragene®•DNA Self-Collection Kit OG-500 DNA Genotek, Inc., Ottawa, Ontario, Canada). Whole saliva was collected according to the manufacturer's instructions and was placed inside the kit which also contained an Oragene DNA-preserving solution. The ingredients of Oragene solution include ethyl alcohol (<24 %) and Tris-HCl buffer (pH 8). As provided by established procedures, samples were stored at room temperature before dansylation labeling experiments and were preserved in -20°C or -80°C after the labeling experiments for long-term storage and follow-up studies.

6.2.3 Chemicals and reagents

^{13}C-dansyl chloride (DnsCl) was synthesized in-house as described by Guo and Li.[44] ^{12}C-dansyl chloride was purchased from Sigma-Aldrich (Milwaukee, WI). All reagents were of ACS grade or higher with water and organic solvents being of MS grade.

6.2.4 Metabolite extraction and isotope labelling

Based on our earlier work on saliva metabolome profiling,[187] we have developed an improved workflow in this work. Specifically, an aliquot of 5 μL of saliva sample was dissolved in 20 μL ACN/ H_2O (50/50) in a screw cap vial. 12.5 μL of $NaHCO_3/NaH_2CO_3$ buffer solution (500 mM, 1:1, v/v) was mixed in the solution and the vial was vortexed and then spun down. 36.6 μL of freshly prepared ^{12}C-DnsCl or ^{13}C-DnsCl in acetonitrile (12 mg/mL) was added into the vial. The solution was vortexed, spun down again, and then let to react for 60 min in an oven at 60 ^{o}C. 5 μL of NaOH (250 mM) was added to quench the excess DnsCl. After another 10 min incubation in the 60 ^{o}C oven, 25 μL of formic acid in ACN/ H_2O (425 mM, 50/50) was added to neutralize the solution. Each individual saliva sample was directly labeled with ^{12}C-DnsCl. A pooled saliva sample was prepared by pooling 10 μL of each from all 82 saliva samples together. This pooled sample was labeled with ^{13}C-DnsCl under the exact same reaction condition. Both the training sample set and validation sample set used the same ^{13}C-labeled pooled sample to mix with their ^{12}C-labeled individual saliva samples.

6.2.5 UPLC-UV

After being labeled with ^{12}C- or ^{13}C-DnsCl, the total concentration of the labeled submetabolome in a sample was quantified by a step-gradient ultra-high pressure liquid chromatography (UPLC) with UV detection at 338 nm.[188] An ACQUITY UPLC system (Waters Corporation, Milford, MA) equipped with photo diode array (PDA) detector, and a

Waters ACQUITY UPLC BEH (Ethylene Bridged Hybrid) C18 column (2.1 mm × 50 mm, 1.7 μm particle size, 130 Å) were used for online LC-UV analysis. LC solvent A was 0.1% (v/v) in 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 min, 0% B; t = 1.00 min, 0% B; t = 1.01 min, 95% B; t = 2.50 min, 95% B; t = 3.00 min, 0% B; t = 6.00 min, 0% B. The flow rate was 450 μL/min, and the sample injection volume was 2 μL.

6.2.6 LC-FTICR-MS

Metabolomic analyses were performed using an Agilent 1100 series capillary HPLC system (Agilent, Palo Alto, CA, USA) connected to a Bruker 9.4 T Apex-Qe Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker, Billerica, MA, USA) equipped with an electrospray ionization (ESI) interface operating in positive mode. Reversed phase (RP) chromatographic separation was carried out on an Eclipse C18 column (2.1 mm ×100 mm, 1.8 μm, 95 Å), with solvent A being water with 0.1% (v/v) formic acid and 5% acetonitrile (ACN) (v/v), and solvent B being ACN with 0.1% (v/v) formic acid. The LC flow rate was 180 μL/min and running time was 26.50 min. The gradient was: t = 0 min, 20% B; t = 3.50 min, 35% B; t = 18.00 min, 65% B; t = 21.00 min, 95% B; t = 21.50 min, 95% B; t = 23.00 min, 98% B; t =24.00 min, 98% B; t = 26.50 min, 99% B. The sample injection volume was 6 μL and the flow was split 1:2 and 60 μL/min of the LC eluate entered the ESI-MS system.

6.2.7 Data processing and statistical analysis

The $^{12}C/^{13}C$ ion pairs were extracted from raw LC-MS data by a peak pair picking program, IsoMS.[27] The peak pairs of all individual samples were then aligned by retention time and accurate mass to produce a metabolite-intensity table. The alignment parameters were set as retention time tolerance of 30 seconds and accurate mass tolerance of 8 ppm. A Zero-fill

program[65] was then applied to all the peak pairs in the table to retrieve the missing values from the raw LC-MS data.

Multivariate statistical analysis of the LC-MS data was carried out using SIMCA-P+ 12.0 (Umetrics, Umea, Sweden). Principle component analysis (PCA) and orthogonal partial least squares - discriminant analysis (OPLS-DA) were used to analyze the data. Receiver operating characteristic (ROC) analysis and linear SVM model was performed using MetaboAnalyst[26] (http://www.metaboanalyst.ca/). Mean center and autoscaling were used to normalize all the peak ratio values prior to the statistical analysis.

6.2.8 Metabolite identification

For positive or definitive metabolite identification, the peak pairs were matched against a Dns-standards library by retention time and accurate mass. In addition, putative metabolite identification was performed based on accurate mass match of the peak pairs found to the metabolites in the Human Metabolome Database (HMDB)[189] and the Evidence-based Metabolome Library (EML) using MyCompoundID,[46] with a mass tolerance of 5 ppm.

6.3 Results and Discussion

The major objective of this work is to identify and verify a set of metabolites that can help diagnosis of AD and MCI patients. Figure 6.1 shows the entire workflow of this study. 5 μL saliva sample was aliquoted out from each individual sample and labeled with ^{12}C-DnsCl. A pooled sample was prepared by mixing small aliquots of individual samples and then labeled with ^{13}C-DnsCl. The ^{12}C-labeled individual sample was then mixed with ^{13}C-labeled pooled sample in a 1:1 amount ratio after the total concentration of the labeled metabolites was

determined by LC-UV. After the LC-MS analysis, automatic data preprocessing was performed to extract the peak pairs belonging to the labeled amine/phenol submetabolome. To discover the metabolites that contribute to the statistical difference of AD, MCI and NA, we performed pair-wise comparison using OPLS-DA and volcano plot analysis. The diagnostic power of the common metabolites that were highly ranked in both statistical tools was then evaluated by ROC analysis. Top ranked metabolites were identified and externally validated by another set of samples. These metabolites could potentially be served as biomarkers for the diagnosis of AD and MCI.

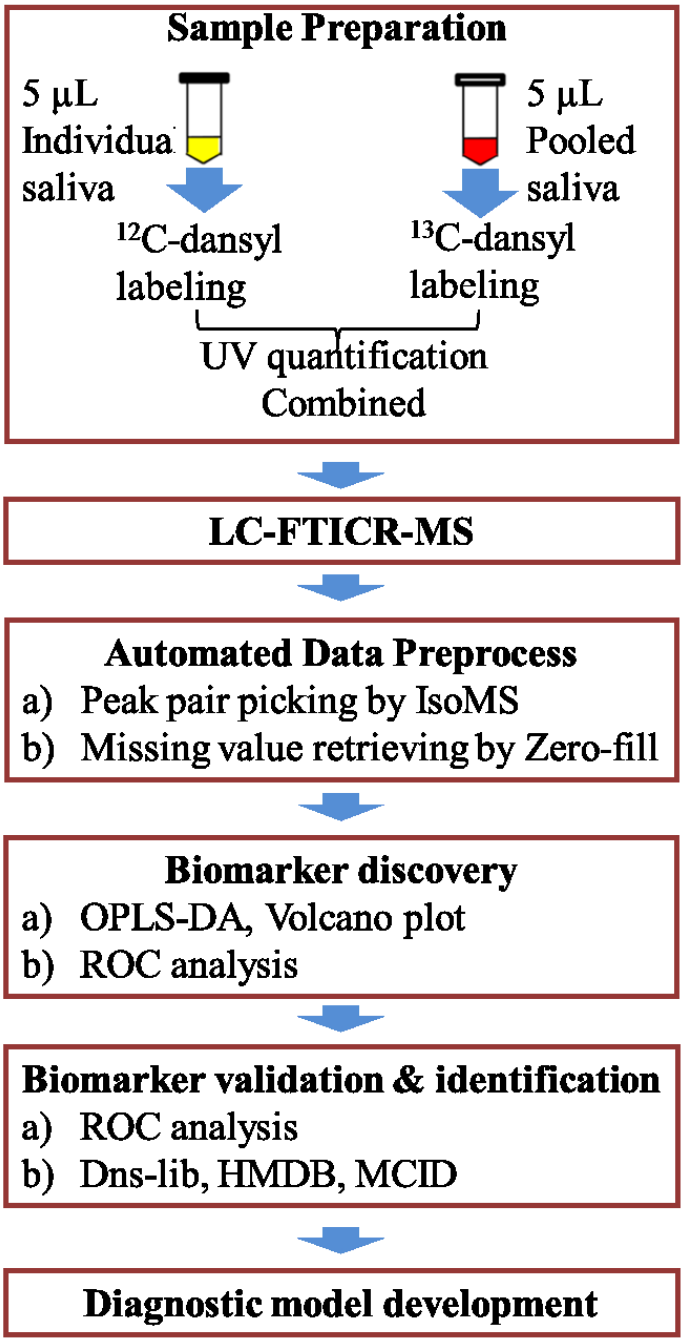


Figure 6.1 Experimental workflow

6.3.1 Improved workflow for saliva metabolome profiling

In our previous study, cold acetone (-20 ˚C) was used to remove the proteins before the dansylation labeling.[187] However, it was noted that most of the proteins have already been

rapidly precipitated by the presence of ethyl alcohol in Oragene·DNA saliva collection kits. Therefore, a set of experiments were carried out to determine the best solvent to dissolve the saliva sample without subsequent protein precipitation and concentration steps in order to prevent any progressive loss of samples. A 5 μL starting saliva sample was aliquoted out and dissolved in 20 μL acetone (ACE), 20 μL H_2O, or 20 μL ACN/H_2O (50/50) at room temperature. Another 5 μL starting saliva sample was dissolved in acetone and stored at -20 ^{o}C overnight to serve as the control sample. Each solution was then subjected to dansyl labeling, followed by LC-UV quantification. The peak areas of the labeled metabolites were measured and compared among the three solvent system experiments. As shown in Figure 6.2, the labeling efficiencies of the experiments with 3 chosen solvents are higher than that of the control. The ACN/H_2O (50/50) system is shown to have the best performance.

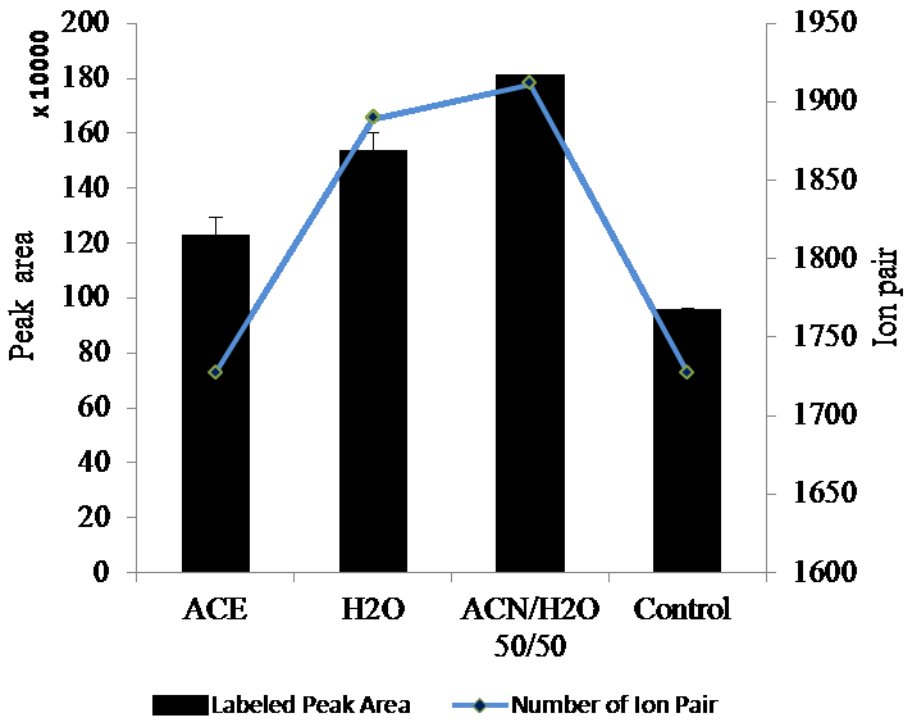


Figure 6.2 UV peak area and peak pair numbers detected in different dissolving solvent systems

The number of peak pairs detected in each of the dissolving solvent systems was extracted using IsoMS. The right x-axis in Figure 6.2 shows the number of peak pairs detected. There are 1911±15 (n=3) peak pairs detected in the ACN/H_2O (50/50) experiment, showing a marked improvement over the old protocol as described in the control experiment (i.e., 1727±16 peak pairs). These results indicate that by omitting the protein precipitation step more metabolites could be detected. Therefore, we used ACN/H_2O (50/50) as the solvent to dissolve the 5 μL saliva sample for all the subsequent experiments.

Another major improvement of the current workflow over the original protocol is on data processing. An optimized IsoMS was used for peak picking with high sensitivity and specificity. The newly developed Zero-fill program[65] was applied to all the metabolome data produced which retrieved many missing peak ratio values in the initial metabolite-intensity table generated

by IsoMS. Thus, the overall metabolome coverage with consistent peak ratio values was more than doubled, compared to the data set produced in our original work (see below).

6.3.2 The saliva metabolome

The optimized sample and data processing workflow was used to profile the 82 saliva samples in the training set. Relative standard deviations (RSD) for peak pair ratios ranged from 0.1% to 15%, with an average of 2% for the triplicate experiments. A set of data that fell within the range of 10-15% was subjected to Grubbs test at 99% confidence level to detect any statistical outlier. A total of 6230 unique pairs or metabolites (defined by molecular ion m/z coupled with its retention time) were obtained from the LC-FTICR-MS analysis with an average of 3669 peak pairs detected from each sample. Among them, 3801 peak pairs were commonly detected in more than 50% of the samples. By searching these peak pairs against the Dns-library composed of 273 labeled standards, using mass tolerance of 5 ppm and RT tolerance of 15 s, 79 metabolites were positively identified based on mass and RT matches (see Table T6.2 for the list). Using MyCompoundID MS search based on accurate mass of peak pairs with mass tolerance of 5 ppm, 616 (9.9%) metabolites were putatively identified using the HMDB library and 2972 (47.7%) were identified using the predicted human metabolite library with one reaction (see Supplemental Table T6.1, T6.2). These results demonstrate a high coverage of the amine/phenol submetabolome in saliva using the optimized dansylation LC-MS workflow and also show the complexity and great diversity of the salivary metabolites present in a sample.

Table 6.2 Dns-lib search results

#	Input mass	Input rt	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT
1	408.1707	1.40	2.43	HMDB00517	L-Arginine	174.1117	408.1700	2.44
2	366.1118	2.22	2.94	HMDB00168	L-Asparagine	132.0535	366.1118	3.00
3	399.1054	3.88	3.94	HMDB02005	Methionine Sulfoxide	165.0460	399.1043	3.72
4	339.1011	4.58	4.37	HMDB00187	L-Serine	105.0426	339.1009	4.40
5	381.1119	5.24	4.85	HMDB00148	L-Glutamic Acid	147.0532	381.1115	5.05
6	339.1372	6.05	5.46	HMDB04437	Diethanolamine	105.0790	339.1373	5.49
7	353.1167	6.51	5.81	HMDB00167	L-Threonine	119.0582	353.1166	5.79
8	295.1111	6.87	6.13	HMDB00149	Ethanolamine	61.0528	295.1111	6.00
9	309.0914	7.46	6.68	HMDB00123	Glycine	75.0320	309.0903	6.59
10	406.1437	7.74	6.93	HMDB00721	Glycylproline	172.0848	406.1431	7.17
11	364.1677	7.97	7.14	HMDB02064	N-Acetylputrescine	130.1106	364.1689	7.25
12	323.1060	8.14	7.30	HMDB00056	Beta-Alanine	89.0477	323.1060	7.24
13	323.1055	8.47	7.59	HMDB00161	L-Alanine	89.0477	323.1060	7.57
14	381.1118	8.52	7.64	HMDB02393	N-methyl-D-aspartic acid	147.0532	381.1115	7.53
15	337.1217	8.78	7.88	HMDB00112	Gamma-Aminobutyric acid	103.0633	337.1216	7.79
16	337.1218	9.39	8.43	HMDB03911	3-Aminoisobutanoic acid	103.0633	337.1216	8.67
17	337.1220	9.77	8.77	HMDB01906	2-Aminoisobutyric acid	103.0633	337.1216	8.91
	337.1220	9.77	8.77	HMDB03911	3-Aminoisobutanoic acid	103.0633	337.1216	8.67
18	351.1376	9.90	8.89	HMDB03355	5-Aminopentanoic acid	117.0790	351.1373	8.68
19	452.1863	9.97	8.96	HMDB29043	Serinyl-Leucine	218.1267	452.1850	8.90
20	408.1594	10.16	9.13	HMDB28854	Glycyl-Valine	174.1004	408.1588	9.19
21	337.1219	10.37	9.32	HMDB00650	D-Alpha-aminobutyric acid	103.0633	337.1216	9.23
	337.1219	10.37	9.32	HMDB00452	L-Alpha-aminobutyric acid	103.0633	337.1216	9.13
22	456.1599	10.48	9.42	HMDB28995	Phenylalanyl-Glycine	222.1004	456.1588	9.43

23	408.1594	10.49	9.43		Gly-Norvaline	174.1005	408.1588	9.51
24	323.1060	10.52	9.45	HMDB00271	Sarcosine	89.0477	323.1060	9.34
25	363.1009	10.53	9.47	HMDB00148	L-Glutamic Acid - H2O	147.0532	363.1009	9.46
26	466.2014	11.30	10.17	HMDB29065	Threoninyl-Leucine	232.1423	466.2006	10.18
27	514.1650	11.33	10.19	HMDB00706	L-Aspartyl-L-phenylalanine	280.1059	514.1642	10.07
28	349.1218	11.35	10.21	HMDB00162	L-Proline	115.0633	349.1216	10.18
29	470.1758	11.73	10.47	HMDB28988	Phenylalanyl-Alanine	236.1161	470.1744	10.58
30	365.1533	11.90	10.59	HMDB03640	Beta-Leucine	131.0946	365.1529	10.78
31	351.1372	12.08	10.71	HMDB00883	L-Valine	117.0790	351.1373	10.81
32	383.1095	12.14	10.75	HMDB00696	L-Methionine	149.0510	383.1094	10.89
33	422.1747	12.17	10.77	HMDB28844	Glycyl-Isoleucine	188.1161	422.1744	10.78
34	371.1409	12.37	10.91	HMDB01065	2-Hydroxyphenethlamine	137.0841	371.1424	10.77
35	360.1013	12.38	10.92	HMDB02024	Imidazoleacetic acid	126.0429	360.1012	11.12
36	361.1330	12.40	10.94	HMDB03464	4-Guanidinobutanoic acid - H2O	145.0851	361.1329	11.00
37	495.1691	12.62	11.18	HMDB28852	Glycyl-Tryptophan	261.1113	495.1697	11.19
38	422.1747	12.67	11.24	HMDB00759	Glycyl-L-Leucine	188.1161	422.1744	11.22
39	438.1494	12.72	11.29	HMDB00929	L-Tryptophan	204.0899	438.1482	11.44
40	436.1906	12.75	11.32	HMDB28691	Alanyl-Leucine	202.1317	436.1901	11.36
41	456.1588	13.06	11.66	HMDB28848	Glycyl-Phenylalanine	222.1004	456.1588	11.65
42	470.1750	13.50	12.14	HMDB28694	Alanyl-Phenylalanine	236.1161	470.1744	12.11
43	399.1377	14.07	12.78	HMDB00159	L-Phenylalanine	165.0790	399.1373	12.74
44	462.2066	14.25	12.97	HMDB28937	Leucyl-Proline	228.1474	462.2057	12.99
45	365.1522	14.29	13.01	HMDB00172	L-Isoleucine	131.0946	365.1529	13.06
46	496.1912	14.29	13.01	HMDB11177	L-phenylalanyl-L-proline	262.1317	496.1901	13.13
47	365.1532	14.57	13.31	HMDB00557	L-Alloisoleucine	131.0946	365.1529	13.20
48	365.1529	14.70	13.44	HMDB00687	L-leucine	131.0946	365.1529	13.36
49	372.1014	14.87	13.62	HMDB00301	Urocanic acid	138.0429	372.1012	13.52
50	364.6247	14.88	13.63	HMDB04987	Alpha-Aspartyl-lysine	261.1325	364.6246	13.61

51	498.2060	14.95	13.70	HMDB29008	Phenylalanyl-Valine	264.1474	498.2057	13.62
52	371.1407	15.22	13.99	HMDB01065	2-Hydroxyphenethlamine - Isomer	137.0841	371.1424	13.77
53	365.1527	15.24	14.01	HMDB01645	L-Norleucine	131.0946	365.1529	14.11
54	354.0703	15.34	14.12	HMDB00192	L-Cystine	240.0238	354.0702	14.11
55	416.1172	15.80	14.58	HMDB00755	Hydroxyphenyllactici acid	182.0579	416.1162	14.39
56	414.1244	16.63	15.42	HMDB01889	Theophylline	180.0647	414.1230	15.42
57	551.2324	17.03	15.82	HMDB28940	Leucyl-Tryptophan	317.1739	551.2323	15.77
58	512.2234	17.05	15.84		Phenyl-Leucine	278.1631	512.2214	15.90
59	385.1216	17.57	16.37	HMDB01859	Acetaminophen	151.0633	385.1216	16.35
60	300.1034	17.63	16.42	HMDB00214	Ornithine	132.0899	300.1033	16.58
61	512.2225	17.68	16.48	HMDB13243	Leucyl-phenylalanine	278.1630	512.2214	16.59
62	546.2066	17.76	16.56	HMDB13302	Phenylalanylphenylalanine	312.1474	546.2057	16.55
63	460.1655	17.81	16.61	HMDB28878	Histidinyl-Alanine	226.1066	460.1649	16.69
64	386.1057	18.29	17.09	HMDB00020	p-Hydroxyphenylacetic acid	152.0473	386.1057	16.91
65	307.1104	18.67	17.47	HMDB00182	L-Lysine	146.1055	307.1111	17.47
66	460.1654	18.79	17.61	HMDB28689	Alanyl-Histidine	226.1066	460.1649	17.62
67	400.1200	19.06	17.92	HMDB02199	Desaminotyrosine	166.0630	400.1213	18.04
68	353.1050	21.14	20.27	HMDB29105	Tyrosyl-Glycine	238.0954	353.1060	20.19
69	278.1082	22.19	21.46	HMDB01414	1-4-diaminobutane	88.1000	278.1083	21.27
70	356.0949	22.28	21.56	HMDB00750	3-Hydroxymandelic acid - COOH	168.0423	356.0951	21.64
71	353.1060	22.51	21.82	HMDB28853	Glycyl-Tyrosine	238.0954	353.1060	21.63
72	360.1138	22.58	21.91	HMDB28699	Alanyl-Tyrosine	252.1110	360.1138	21.85
73	501.2423	22.97	22.34	HMDB01932	Metoprolol	267.1834	501.2418	22.09
74	285.1159	23.10	22.49	HMDB02322	Cadaverine	102.1157	285.1162	22.39
75	324.5955	23.25	22.66	HMDB00158	L-Tyrosine	181.0739	324.5953	22.65
76	374.1294	23.39	22.80	HMDB29118	Tyrosyl-Valine	280.1423	374.1295	22.83
77	328.1002	23.78	23.19	HMDB00228	Phenol	94.0419	328.1002	23.16
78	417.6356	23.86	23.27	HMDB29095	Tryptophyl-Tyrosine	367.1532	417.6349	23.25

79	381.1384	24.37	23.78	HMDB28941	Leucyl-Tyrosine	294.1580	381.1373	23.98
	381.1384	24.37	23.78	HMDB29109	Tyrosyl-Leucine	294.1580	381.1373	23.77

6.3.3 Multivariate analysis

To relate the three health states to the metabolome profiles, multivariate analysis was performed on the training data set. Principle component analysis (PCA) was first used to obtain an overview of the NA, MCI and AD saliva data in an unsupervised approach. The PCA score plot is shown in Figure 6.3A. As we can see, due to large biological variations, there are some overlaps of the samples from different health states. However, we can still see some separation; saliva samples from AD patients are located on the top right corner marked in green. The separation between NA and MCI samples is not very obvious, indicating that the metabolomic differences between MCI and NA samples are not as large as those of AD and NA.

We applied OPLS-DA to study the metabolomic variations in NA, MCI and AD with the score plot shown in Figure 6.3B. To evaluate the quality of the OPLS-DA model, an internal validation method using a seven-fold cross-validation step was applied, from which the values of Q^2Y (predictive ability of the model) and R^2Y (goodness of fit parameter) were calculated. The score plot shows a very clear separation among the three groups with high validation parameters (R^2Y=0.93 and Q^2Y =0.87), indicating the robustness of the model. To view the progression of saliva metabolomic variations from NA to MCI, and then to AD, a 3D OPLS-DA plot is shown in Figure 3C. There is a clear trajectory of the metabolic changes through the three health states (i.e., the AD clusters are far away from the NA's, compared to MCI's).

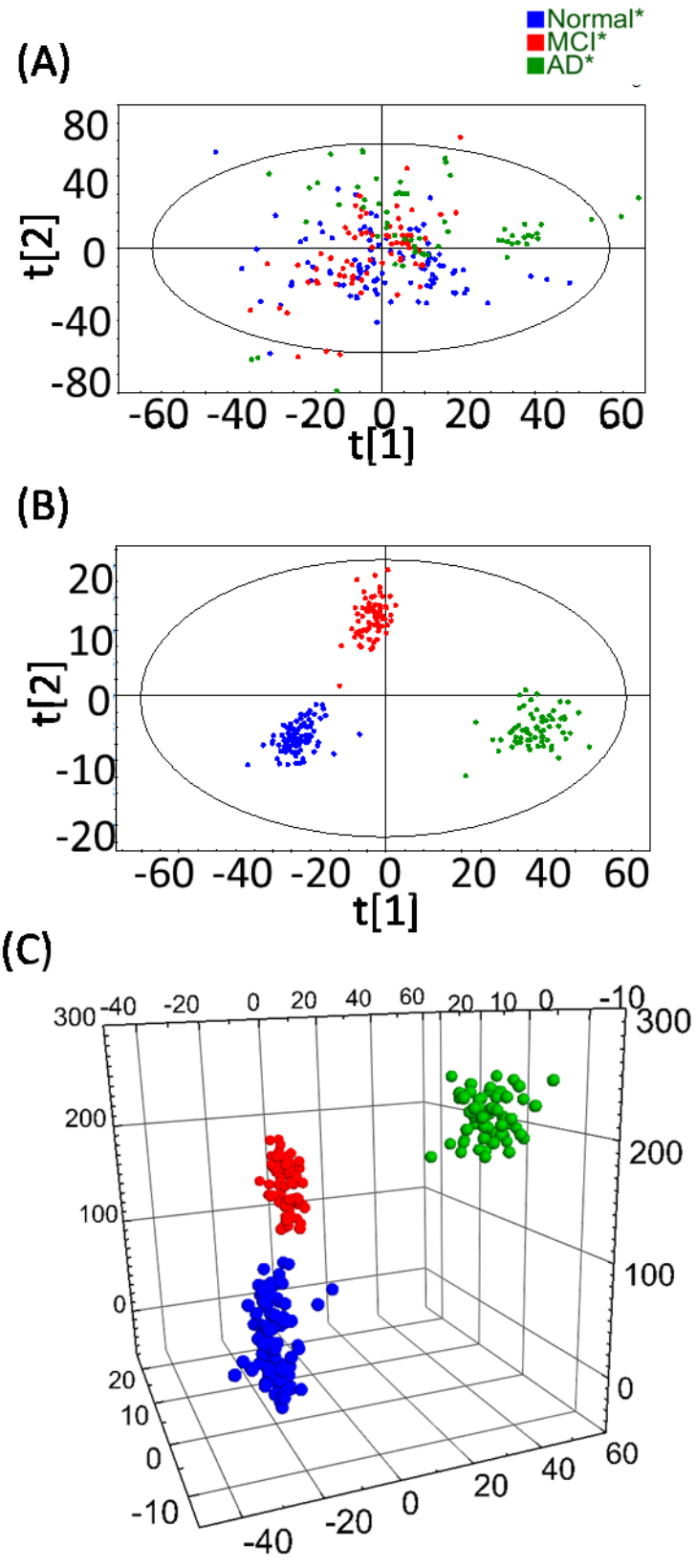

(A)
Normal*
MCI*
AD*
80
40
0
-40
-80
t[2]
-60 -40 -20 0 20 40 60
t[1]
(B)
20
10
0
-10
-20
t[2]
-60 -40 -20 0 20 40 60
t[1]
(C)
300
200
100
0
-40 -20 0 20 40 60
20 10 0 -10
300
200
100
0
60
40 20 0 -20 -40
20
10
0
-10

Figure 6.3 (A) PCA score plot, (B) 2D OPLS-DA score plot, and (C) 3D OPLS-DA score plot of AD, MCI, and NA

Table 6.2 shows a misclassification table applied to the internally cross-validated OPLS-DA model. Only one sample was misclassified and the overall P value (Fisher probability: 1.8×10^{-35}) is very small, which validates the group separation.

Table 6.3 Misclassification table of the OPLS-DA analysis for AD, MCI, and NA

	Members	**Correct**	**NA**	**MCI**	**AD**	**No class (YPred < 0)**
NA	35	100%	35	0	0	0
MCI	25	100%	0	25	0	0
AD	22	100%	0	0	22	0
No class	0		0	0	0	0
Total	82	100%	35	25	22	0
Fishers prob.	1.20E-41					

6.3.4 Discovery and validation of potential biomarkers

The ultimate goal of this study is to find potential metabolite biomarkers for diagnosis of MCI and AD. To determine the significant metabolites that differentiate paired groups (i.e., MCI vs. NA, AD vs. MCI and AD vs. NA) with relatively high confidence, both multivariate (OPLS-DA) and univariate (Volcano plot) statistical tools were applied to cross-select the important metabolites. Common metabolites found by both tools were extracted and then ROC analysis was applied to evaluate their diagnostic performance. Metabolites with high diagnostic performance were considered as the potential biomarkers and further externally evaluated using

another set of samples (i.e., validation set). We strived to use as few metabolites as possible (i.e., the top three ranked metabolites) to build a diagnostic model, as the use of a larger number of metabolites may not be practical or cost-effective in clinical applications.

Using the above approach, OPLS-DA analysis was first applied for pair-wise comparisons and the resultant score plots are shown in Figure 4. Notably, all OPLS-DA models demonstrate clear group separation with high validation metrics, confirming the goodness of fit and good predictive capabilities of the proposed models. From OPLS-DA analysis, metabolites with VIP score of larger than 1.5 in all three comparisons were retained (see Supplemental Table T6.3). Next, volcano plots analysis was performed to find metabolites with high fold-change and low *p* values and the results are shown in Figure 5. Thresholds of *p*-value of 0.01 and fold-change of 1.2 were used to discriminate between the significantly up and down-regulated metabolites (see Supplemental Table T6.4).

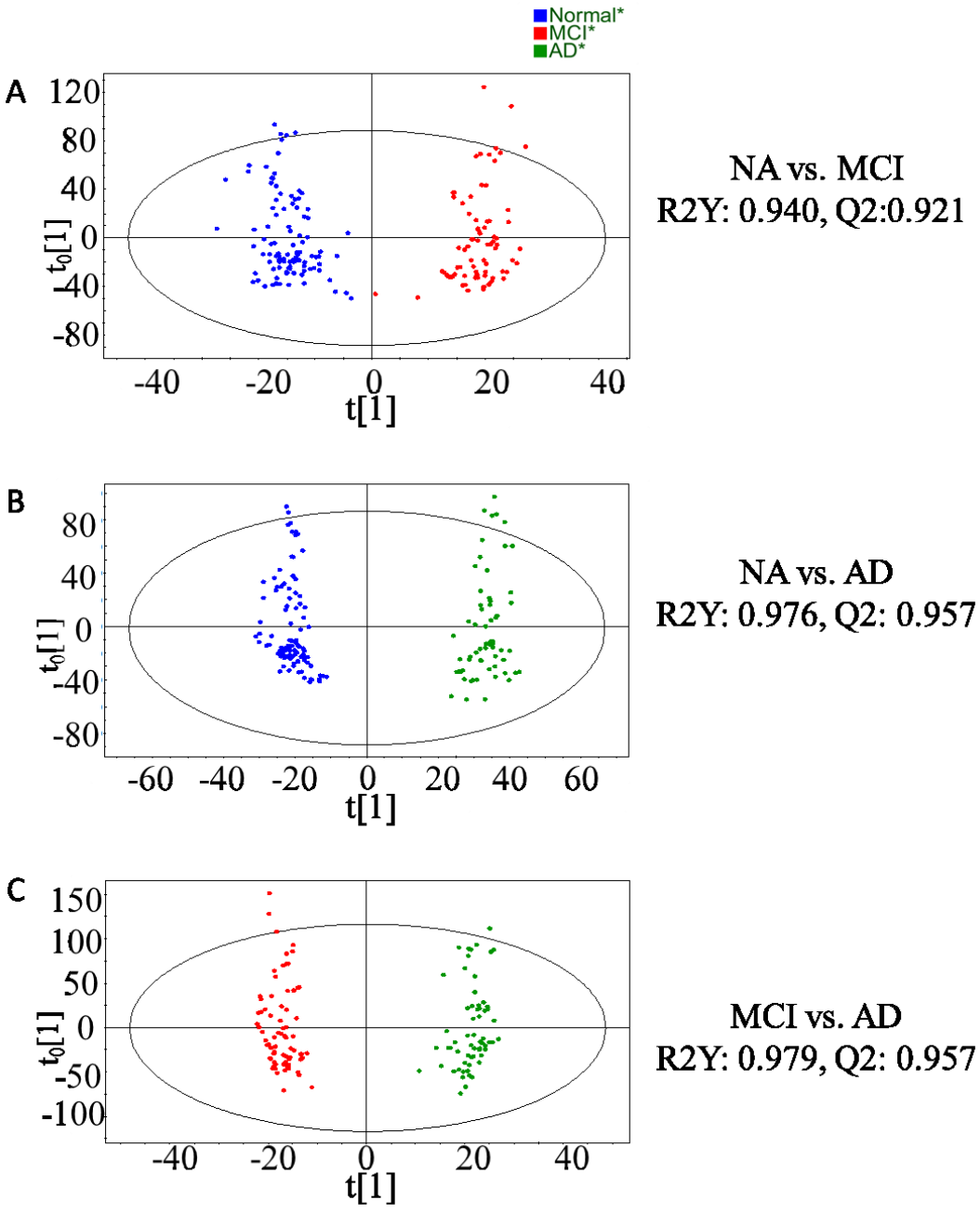


Figure 6.4 OPL-DA score plots for pair-wise comparisons of: (A) NA vs. MCI, (B) NA vs. AD, and (C) MCI vs. AD.

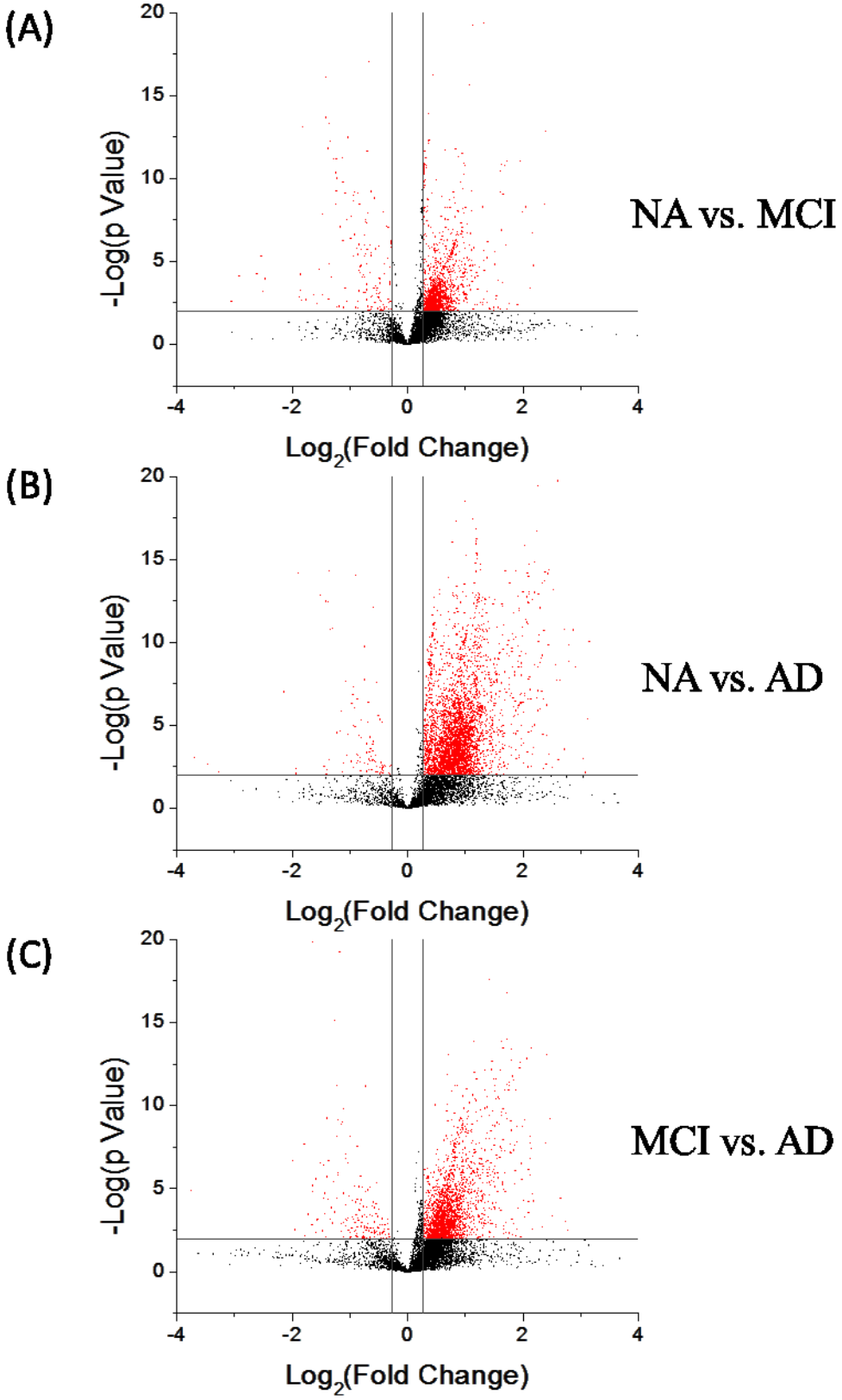

(A)
-Log(p Value)
Log2(Fold Change)
NA vs. MCI
(B)
-Log(p Value)
Log2(Fold Change)
NA vs. AD
(C)
-Log(p Value)
Log2(Fold Change)
MCI vs. AD

Figure 6.5 Volcano plots for pair-wise comparisons of: (A) NA vs. MCI, (B) NA vs. AD, and (C) MCI vs. AD.

The results of OPLS-DA and Volcano plot analyses were compared and only the metabolites shown up as significant in both analyses were further considered. ROC analysis was performed on these common metabolites and an AUC cut-off value of 0.75 was used to generate the final list of significant metabolites that were deemed to be potentially biomarkers. There were 175 significant metabolites found in AD vs. NA comparison, 142 metabolites in AD vs. MCI comparison and 59 metabolites in MCI vs. NA comparison (see Supplemental Table T6.5).

To validate some of the significant metabolites found in the training set, another set of 27 samples were collected separately from the training set. Following the same experimental protocol, each individual sample was labeled by ^{12}C-DnsCl and experimental triplicate was performed. Also, the same ^{13}C-labeled pooled sample was used and mixed with each ^{12}C-labeled individual sample at a 1:1 sample amount, followed by LC-MS analysis. An average of 2981 peak pairs or metabolites were detected per sample with a total of 4157 peak pairs detected in the 27 samples. Among them, 3184 peak pairs were commonly found in more than 50% of the samples, which is lower than the 3801 common peak pairs found in 50% of the samples in the training set. The lower numbers are due to the fact that the low abundance peak pairs are not recovered using the zero-fill program as efficiently in a smaller size of samples as that in a larger size of samples.

The purpose of this validation dataset is to validate the discovered biomarkers and diagnostic models and thus only those significant metabolites discovered in the training dataset were studied in the validation dataset. Based on the criteria of AUC > 0.75 in both the training set and validation set, we were able to identify 63, 48, and 2 common metabolites that have

consistently good ROC performance for AD vs. NA, AD vs. MCI and MCI vs. NA comparisons in both the training and validation dataset, respectively. These metabolites are listed in Supplemental Table T6.6. Among them, 4 metabolites were definitively identified for differentiating AD and NA, while 3 metabolites were definitively identified for differentiating AD and MCI. All these validated biomarkers have less than 20% missing values in both the training and validation datasets.

6.3.5 Development and validation of diagnostic model

A single metabolite alone may have a good prediction power for disease diagnosis. However, a diagnostic model using multiple biomarkers may give a better performance.[190] We used a linear support vector machine tool (linear SVM) in MetaboAnalyst[26] to develop a diagnostic model for each of the three pair-wise comparisons using the training data set. Its diagnostic performance was further evaluated using the validation data set. Table 3 shows the summary of the three linear SVM based diagnostic models.

As Table 3 shows, using the top 3 metabolites (#6112, #7628 and #4489), we could distinguish AD from NA with AUC=0.998 (0.992-1.000 at 95% CI) in the training set. This result was validated in the validation set with AUC=0.989. The diagnostic sensitivity was 94.1% and the specificity was 98.1%. Similarly, using three metabolites (#1429, #3731 and #943), we could separate AD from MCI with AUC=0.998 (0.991-1.000 at 95% CI) in TS and AUC=0.997 in VS. The sensitivity was 98.5% and the specificity was 98.6%. In the case of MCI vs. NA, using two metabolites (#3731 and #7500), MCI and NA could be differentiated with AUC=0.774 (0.672-0.852 at 95% CI) in TS and AUC=0.889 in VS. The sensitivity was 70.7% and the specificity was 79.0%. It is clear that the diagnostic performance on MCI vs. NA is not as good as the other two pair-wise comparisons. Adding more metabolites did not improve the diagnostic

performance. This indicates that it is more difficult to differentiate the metabolic changes from normal aging to mild cognition impairment. Nevertheless, using the two metabolites found, we could separate NA and MCI with good sensitivity and specificity. Future work of using different labeling chemistries targeting different groups of submetabolomes may discover other classes of metabolites that could improve the overall diagnostic performance.

Table 6.4 Metabolic diagnostic models for pair-wise comparison using top-ranked biomarkers

	ID	RT (min)	Molecular weight	Putative ID	Training AUC (95% CI)	Validation AUC (95% CI)	Sensitivity	Specificity
AD vs NA	#6112	18.86	297.1087	Methylguanosine	0.998 (0.992 – 1.000)	0.989	94.10%	98.10%
	#7628	22.51	302.1379	Histidinyl-Phenylalanine				
	#4489	15.16	330.19					
AD vs MCI	#1429	7.8	125.0448		0.998 (0.991 – 1.000)	0.997	98.50%	98.60%
	#3731	13.93	468.1979	Glucosylgalactosyl hydroxylysine-H_2O				
	#943	5.91	105.0791	Aminobutyric acid + H_2				
MCI vs NA	#3731	13.93	468.1573	Glucosylgalactosyl hydroxylysine-H_2O	0.774 (0.672 – 0.852)	0.889	70.70%	79.00%
	#7500	22.29	289.1635	Glutamine + Carnitine				

The potential biomarkers listed in Table 3 have not been positively identified. We have also built the diagnostic models using the positively identified metabolites only and the results are shown in Table 4. For separating AD and NA, using three identified metabolites,

phenylalanyl-proline, phenylalanyl-phenylalanine and Urocanic acid, we could obtain AUC 0.832 (0.772-0.920 at 95% CI) in TS and 0.754 in DS with 71.2% sensitivity and 81.9% specificity. For separating AD and MCI, alanyl-phenylalanine and phenylalanyl-proline could be used with AUC 0.874 (0.787-0.948 at 95% CI) in TS and 0.792 in DS with 83.9% sensitivity and 82.3% specificity. Although these models did not achieve high AUC as the top 3 biomarkers, these positively identified metabolites could be more readily transferred to a real clinical application using targeted LC-MS/MS metabolite analysis. However, for future validation work using CIL LC-MS, it is worth monitoring the identified metabolites as well as the un-identified metabolites. If some of the high-performance biomarkers are validated in multi-center large scale validation studies, more efforts could be devoted to identify these biomarkers using fractionation methods and multiple characteristic tools such as tandem MS, NMR and synthesis of standards.

Table 6.5 Metabolic diagnostic model for pair-wise comparison using identified biomarkers

	Training AUC (95% CI)	Validation AUC (95% CI)	Sensitivity	Specificity
AD vs NA : phenylalanyl-L-Proline Phenylalanylphenylalanine Urocanic acid	0.832 (0.772 – 0.920)	0.754	71.4%	80.0%
AD vs MCI: Alanyl-phenylalanine phenylalanyl-L-Proline	0.874 (0.787 – 0.948)	0.792	85.7%	80.0%

6.4 Conclusions

In this work, an improved dansylation isotope labelling LC-FTICR-MS method has been developed for metabolite biomarker discovery using human salivary samples. Even though only a very small amount of starting material (5 μL of individual saliva) was used for the experiments,

a total of 6230 metabolites in the amine/phenol submetabolome could be detected from the 83 samples used as the training set. Using the top 3 metabolites commonly found by both OPLS-DA and volcano plot analyses, excellent sensitivity (~99%) and specificity (~99%) could be achieved for differentiating AD from NA or AD from MCI and good sensitivity (~71%) and specificity (~82%) could be obtained for separating MCI from NA. These results were validated using another set of 27 samples. This study has shown the promise of using salivary biomarkers for diagnosis of MCI and AD.

Chapter 7

Conclusion and Future Work

LC-MS technique has been widely used in the application of metabolomics due to its high sensitivity and high throughput. However, because of the great chemical diversity and wide dynamic range, it is difficult to detect and identify a entire metabolome using one technique. To improve the separation and detection in LC-MS based metabolomics, our group has applied a divide-and-conquer approach by using chemical isotope labeling reagents to target specific submetabolome. One approach, for example, uses dansylation chemistry to target amine-/phenol-containing submetabolome. This approach has been successfully applied to the study of urine, serum, CSF, saliva, and cell lysis solutions. My thesis research focuses on the development and application of CIL LC-MS platform. Based on the research objectives, my thesis work is composed of three parts. The first part aims at the development computer programs to improve LC-MS-base metabolic data processing. The second part describes the solutions to metabolite identification in LC-MS-base metabolomics. The third part addresses an application of CIL LC-MS for the clinical biomarker discovery. The major achievements of each research project are summarized below.

In Chapter 2, Metabolomics requires quantitative comparison of individual metabolites present in an entire sample set. Unfortunately missing intensity-values in one or more samples are very common. Because missing values can have a profound influence on metabolomic results, the extent of missing values found in a metabolomic dataset should be treated as an important parameter for measuring the analytical performance of a technique. In this work, we report a study of the scope of missing values and a robust method of filling the missing values in a chemical isotope labeling (CIL) LC-MS metabolomics platform. Unlike conventional LC-MS,

CIL LC-MS quantifies the concentration differences of individual metabolites in two comparative samples based on the mass spectral peak intensity ratio of a peak pair from a mixture of differentially labeled samples. We show that this peak-pair feature can be explored as a unique means of extracting metabolite intensity information from raw mass spectra. In our approach, a stringent peak-pair peaking algorithm, IsoMS, is initially used to process the LC-MS dataset to generate a CSV file or table that contains metabolite ID and peak ratio information (i.e., metabolite-intensity table). A zero-fill program is developed to automatically find a missing value in the CSV file and go back to the raw LC-MS data to find the peak pair, then calculate the ratio and enter the ratio value into the table. Most of the missing values are found to be low abundance peak pairs. We demonstrate the performance of this method in analyzing experimental and technical replicate dataset of human urine metabolome. Furthermore, we propose a standardized approach of counting missing values in replicate dataset as a way of gauging the extent of missing values in a metabolomics platform. Finally, we illustrate that applying the zero-fill program, in conjunction with dansylation CIL LC-MS, can lead to a marked improvement in finding significant metabolites that differentiate bladder cancer patients and their controls in a metabolomics study of 109 subjects.

In Chapter 3, generating precise and accurate quantitative information on metabolomic changes in comparative samples is important for metabolomics research where technical variations in the metabolomic data should be minimized in order to reveal biological changes. We report a method and software program, IsoMS-Quant, for extracting quantitative information from a metabolomic dataset generated by chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS). Unlike previous work of relying on mass spectral peak ratio of the highest intensity peak pair to measure relative quantity difference of a differentially labeled

metabolite, this new program reconstructs the chromatographic peaks of the light- and heavy-labeled metabolite pair and then calculates the ratio of their peak areas to represent the relative concentration difference in two comparative samples. Using chromatographic peaks to perform relative quantification is shown to be more precise and accurate. IsoMS-Quant is integrated with IsoMS for picking peak pairs and Zero-fill for retrieving missing peak pairs in the initial peak pairs table generated by IsoMS to form a complete tool for processing CIL LC-MS data. This program can be freely downloaded from the www.MyCompoundID.org website for non-commercial use.

In Chapter 4, high-performance chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS) is an enabling technology based on rational design of labeling reagents to target a class of metabolites sharing the same functional group (e.g., all the amine-containing metabolites or the amine submetabolome) to provide concomitant improvements in metabolite separation, detection and quantification. However, identification of labeled metabolites remains to be an analytical challenge. In this work, we describe a library of labeled standards and a search method for metabolite identification in CIL LC-MS. The current library consists of 273 unique metabolites, mainly amines and phenols, that are individually labeled by dansylation (Dns). Some of them produced more than one Dns-derivative (isomers or multiple labeled products), resulting in a total of 315 dansyl compounds in the library. These metabolites cover 42 metabolic pathways, allowing the possibility of probing their changes in metabolomics studies. Each labeled metabolite contains three searchable parameters: molecular ion mass, MS/MS spectrum and retention time (RT). To overcome RT variations caused by experimental conditions used, we have developed a calibration method to normalize RTs of labeled metabolites using a mixture of RT calibrants. A search program, DnsID*, has been developed in

www.MyCompoundID.org for automated identification of dansyl labeled metabolites in a sample based on matching one or more of the three parameters with those of the library standards. Using human urine as an example, we illustrate the workflow and analytical performance of this method for metabolite identification. This freely accessible resource is expandable by adding more amine and phenol standards in the future. In addition, the same strategy should be applicable for developing other labeled standards libraries to cover different classes of metabolites for comprehensive metabolomics using CIL LC-MS.

In Chapter 5, we report an analytical tool to facilitate metabolite identification based on MS/MS spectral match of an unknown to a library of predicted MS/MS spectra of possible human metabolites. To construct the spectral library, all the known endogenous human metabolites in the Human Metabolome Database (HMDB) (8,021 metabolites) and their predicted metabolic products via one metabolic reaction in the Evidence-based Metabolome Library (EML) (375,809 predicted metabolites) were subjected to *in silico* fragmentation to produce the predicted MS/MS spectra. This library is hosted at the public MCID website (www.MyCompoundID.org) and a spectral search program, MCID MS/MS*, has been developed to allow a user to search one or a batch of experimental MS/MS spectra against the library spectra for possible match(s). Using MS/MS spectra generated from standard metabolites and a human urine sample, we demonstrate that this tool is very useful for putative metabolite identification. It allows a user to narrow down many possible structures initially found by using accurate mass search of an unknown metabolite to only one or a few candidates, thereby saving time and efforts in selecting or synthesizing metabolite standard(s) for eventual positive metabolite identification.

In Chapter 6, we report an improved saliva metabolome profiling workflow to search for potential metabolite biomarkers for differentiating individuals with normal aging (NA), mild cognitive impairment (MCI) and Alzheimer's disease (AD). The workflow is based on a high-performance differential chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS) platform using dansylation derivatization for in-depth profiling of the amine/phenol submetabolome. A total of 82 saliva samples (35 NA, 22 MCI and 25 AD) were profiled as the training set (TS) and another 27 independent samples (10 NA, 10 MCI and 7 AD) were analyzed as the validation set (VS). In total, 6230 peak pairs or metabolites were detected in TS and 3590 of them (57.6%) could be mass-matched to the metabolites in metabolomic databases. In addition, 3801 metabolites could be consistently detected in more than 50% of the samples. They were subjected to analysis using multiple statistical tools in order to cross-select potential biomarkers with high statistical significance. Receiver operating characteristic (ROC) analysis was performed to determine the diagnostic power of each potential biomarker in TS. Metabolites with AUC>0.75 were further externally validated in VS. In total, 63, 47, and 2 metabolites were validated as biomarkers in AD vs. NA, AD vs. MCI, and MCI vs. NA comparison, respectively. Diagnostic model was developed for each pair-wise comparison using linear supportive vector machine. Our study showed the possibility of distinguishing AD from NA using three metabolites with AUC=0.998 (0.992-1.000 at 95% CI) in TS and AUC=0.989 in VS. We could separate AD from MCI using three metabolites with AUC=0.998 (0.991-1.000 at 95% CI) in TS and AUC=0.997 in VS. Using two metabolites, MCI and NA could be differentiated with AUC=0.774 (0.672-0.852 at 95% CI) in TS and AUC=0.889 in VS. While a large scale validation work is still required to confirm the specificity and sensitivity of these

biomarkers, this work demonstrates that saliva metabolites could potentially be used for diagnosis of MCI and AD.

Today, the metabolomics is far away from mature and there are still several challenges need to be resolved.

Firstly, the metabolic coverage of current analytical technique is still small comparing to the overall size of the metabolome. It is very critical to have good metabolome coverage as it is the first step towards the comprehensive metabolomics analysis. Even through the CIL LC-MS method developed in our group has been proven to be able to detect much more metabolites than regular LC-MS approach; we are still facing the issue of not being able to detect enough number of metabolites. To further improve the metabolic coverage, two dimensional LC-MS strategy has been proposed, in which case, metabolites were further separated in a second dimensional column to separate the co-eluting metabolites from the first dimension. Our preliminary result indicates that a high pH - low pH two dimensional LC-MS coupling with our CIL LC-MS method has a much better metabolome coverage.

Secondly, the integration of untargeted metabolomic profiling with biological pathways analysis can be a very popular research topic in the next few years. Biological pathway study, which usually performed using targeted metabolomic analysis, has the advantage of high sensitivity and specificity. However, focusing on one pathway may lead to the information loss of other related pathways. The benefit of integrating untargeted metabolomic profiling with biological pathway analysis allows the understanding of a much wider pathways and also the discovery of new pathways relate to different clinical stages.

Thirdly, metabolite identification is still a big challenge in the MS based metabolomics. Typically, the result of manual MS/MS interpretation varies greatly depending on the experience. Also, in the approach of matching experimental spectrum with standard spectra, it is very likely that the standard MS/MS spectra are not available. All these mentioned problems don't fit with requirement of metabolite identification in the large metabolomics data generated by high throughput MS based techniques. Automatic structural interpretation would be a great breakthrough in the field of metabolite identification if it can offer high accuracy and low false positive rate. Our lab has spent efforts to develop a predicted MS/MS library for the purpose of semi-automatic, batch mode based metabolite identification as described in Chapter 5. However, there are still lots of rooms to improve the prediction algorithm to generate better predicted MS/MS spectra.

References

1. Wishart, D.S., et al., *HMDB: the human metabolome database.* Nucleic acids research, 2007. **35**(suppl 1): p. D521-D526.
2. Lindon, J.C., J.K. Nicholson, and E. Holmes, *The handbook of metabonomics and metabolomics*. 2011: Elsevier.
3. Goodacre, R., et al., *Metabolomics by numbers: acquiring and understanding global metabolite data.* TRENDS in Biotechnology, 2004. **22**(5): p. 245-252.
4. Rhee, E.P. and R.E. Gerszten, *Metabolomics and cardiovascular biomarker discovery.* Clinical chemistry, 2012. **58**(1): p. 139-147.
5. Zhang, A., H. Sun, and X. Wang, *Power of metabolomics in biomarker discovery and mining mechanisms of obesity.* Obesity Reviews, 2013. **14**(4): p. 344-349.
6. Kim, Y.S., P. Maruvada, and J.A. Milner, *Metabolomics in biomarker discovery: future uses for cancer prevention.* 2008.
7. Wang, X., A. Zhang, and H. Sun, *Power of metabolomics in diagnosis and biomarker discovery of hepatocellular carcinoma.* Hepatology, 2013. **57**(5): p. 2072-2077.
8. Zhang, A., H. Sun, and X. Wang, *Serum metabolomics as a novel diagnostic approach for disease: a systematic review.* Analytical and bioanalytical chemistry, 2012. **404**(4): p. 1239-1245.
9. Quinones, M.P. and R. Kaddurah-Daouk, *Metabolomics tools for identifying biomarkers for neuropsychiatric diseases.* Neurobiology of disease, 2009. **35**(2): p. 165-176.
10. Kell, D.B., *Systems biology, metabolic modelling and metabolomics in drug discovery and development.* Drug discovery today, 2006. **11**(23): p. 1085-1092.
11. Wishart, D.S., *Applications of metabolomics in drug discovery and development.* Drugs in R & D, 2008. **9**(5): p. 307-322.
12. Davidov, E., et al., *Advancing drug discovery through systems biology.* Drug discovery today, 2003. **8**(4): p. 175-183.
13. Wang, X., et al., *Metabolomics coupled with proteomics advancing drug discovery toward more agile development of targeted combination therapies.* Molecular & Cellular Proteomics, 2013. **12**(5): p. 1226-1238.
14. Bundy, J.G., M.P. Davey, and M.R. Viant, *Environmental metabolomics: a critical review and future perspectives.* Metabolomics, 2009. **5**(1): p. 3-21.
15. Lin, C.Y., M.R. Viant, and R.S. Tjeerdema, *Metabolomics: methodologies and applications in the environmental sciences.* Journal of Pesticide Science, 2006. **31**(3): p. 245-251.
16. Taylor, J., et al., *Application of metabolomics to plant genotype discrimination using statistics and machine learning.* Bioinformatics, 2002. **18**(suppl 2): p. S241-S248.
17. Spratlin, J.L., N.J. Serkova, and S.G. Eckhardt, *Clinical applications of metabolomics in oncology: a review.* Clinical Cancer Research, 2009. **15**(2): p. 431-440.
18. Patti, G.J., O. Yanes, and G. Siuzdak, *Innovation: Metabolomics: the apogee of the omics trilogy.* Nature reviews Molecular cell biology, 2012. **13**(4): p. 263-269.
19. van der Greef, J. and A.K. Smilde, *Symbiosis of chemometrics and metabolomics: past, present, and future.* Journal of Chemometrics, 2005. **19**(5): p. 376-386.
20. Oliver, S.G., et al., *Systematic functional analysis of the yeast genome.* Trends in biotechnology, 1998. **16**(9): p. 373-378.

21. Dunn, W.B. and D.I. Ellis, *Metabolomics: current analytical platforms and methodologies.* TrAC Trends in Analytical Chemistry, 2005. **24**(4): p. 285-294.
22. Zhang, A., et al., *Modern analytical techniques in metabolomics analysis.* Analyst, 2012. **137**(2): p. 293-300.
23. Theodoridis, G., H.G. Gika, and I.D. Wilson, *LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics.* TrAC Trends in Analytical Chemistry, 2008. **27**(3): p. 251-260.
24. Smith, C.A., et al., *XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.* Analytical chemistry, 2006. **78**(3): p. 779-787.
25. Katajamaa, M., J. Miettinen, and M. Orešič, *MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data.* Bioinformatics, 2006. **22**(5): p. 634-636.
26. Xia, J., et al., *MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis.* Nucleic acids research, 2012. **40**(W1): p. W127-W133.
27. Zhou, R., et al., *IsoMS: automated processing of LC-MS data generated by a chemical isotope labeling metabolomics platform.* Analytical chemistry, 2014. **86**(10): p. 4675-4679.
28. Smith, C.A., et al., *METLIN: a metabolite mass spectral database.* Therapeutic drug monitoring, 2005. **27**(6): p. 747-751.
29. Kind, T., et al., *FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry.* Analytical chemistry, 2009. **81**(24): p. 10038-10048.
30. Johnson, C.H. and F.J. Gonzalez, *Challenges and opportunities of metabolomics.* Journal of cellular physiology, 2012. **227**(8): p. 2975-2981.
31. Gibney, M.J., et al., *Metabolomics in human nutrition: opportunities and challenges.* The American journal of clinical nutrition, 2005. **82**(3): p. 497-503.
32. Koal, T. and H.-P. Deigner, *Challenges in mass spectrometry based targeted metabolomics.* Current molecular medicine, 2010. **10**(2): p. 216-226.
33. Reo, N.V., *NMR-based metabolomics.* Drug and chemical toxicology, 2002. **25**(4): p. 375-382.
34. Verpoorte, R., Y. Choi, and H. Kim, *NMR-based metabolomics at work in phytochemistry.* Phytochemistry reviews, 2007. **6**(1): p. 3-14.
35. Zhang, S., et al., *Advances in NMR-based biofluid analysis and metabolite profiling.* Analyst, 2010. **135**(7): p. 1490-1498.
36. Lu, W., B.D. Bennett, and J.D. Rabinowitz, *Analytical strategies for LC–MS-based targeted metabolomics.* Journal of Chromatography B, 2008. **871**(2): p. 236-242.
37. Kanani, H., P.K. Chrysanthopoulos, and M.I. Klapa, *Standardizing GC–MS metabolomics.* Journal of Chromatography B, 2008. **871**(2): p. 191-201.
38. Ramautar, R., G.W. Somsen, and G.J. de Jong, *CE‐MS in metabolomics.* Electrophoresis, 2009. **30**(1): p. 276-291.
39. Spagou, K., et al., *Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies.* Journal of separation science, 2010. **33**(6‐7): p. 716-727.
40. Swartz, M.E., *UPLC™: an introduction and review.* Journal of Liquid Chromatography & Related Technologies, 2005. **28**(7-8): p. 1253-1263.

41. Want, E.J., et al., *Global metabolic profiling procedures for urine using UPLC–MS.* Nature protocols, 2010. **5**(6): p. 1005-1018.
42. Stoll, D.R., X. Wang, and P.W. Carr, *Comparison of the practical resolving power of one-and two-dimensional high-performance liquid chromatography analysis of metabolomic samples.* Analytical chemistry, 2008. **80**(1): p. 268-278.
43. Delahunty, C. and J.R. Yates Iii, *Protein identification using 2d-lc-ms/ms.* Methods, 2005. **35**(3): p. 248-255.
44. Guo, K. and L. Li, *Differential 12C-/13C-isotope dansylation labeling and fast liquid chromatography/mass spectrometry for absolute and relative quantification of the metabolome.* Analytical chemistry, 2009. **81**(10): p. 3919-3932.
45. Guo, K. and L. Li, *High-Performance Isotope Labeling for Profiling Carboxylic Acid-Containing Metabolites in Biofluids by Mass Spectrometry.* Analytical Chemistry, 2010. **82**(21): p. 8789-8793.
46. Li, L., et al., *MyCompoundID: using an evidence-based metabolome library for metabolite identification.* Analytical chemistry, 2013. **85**(6): p. 3401-3408.
47. Shortreed, M.R., et al., *Ionizable isotopic labeling reagent for relative quantification of amine metabolites by mass spectrometry.* Analytical Chemistry, 2006. **78**(18): p. 6398-403.
48. Tsukamoto, Y., et al., *Synthesis of the isotope-labeled derivatization reagent for carboxylic acids, 7-(N,N-dimethylaminosulfonyl)-4-(aminoethyl)piperazino-2,1,3-benzoxadiazole (d6) [DBD-PZ-NH2 (D)], and its application to the quantification and the determination of relative amount of fatty acids in rat plasma samples by high-performance liquid chromatography/mass spectrometry.* Biomedical Chromatography, 2006. **20**(4): p. 358-64.
49. Yang, W.C., et al., *Enhancement of amino acid detection and quantification by electrospray ionization mass spectrometry.* Analytical Chemistry, 2006. **78**(13): p. 4702-4708.
50. Abello, N., et al., *Poly(ethylene glycol)-Based Stable Isotope Labeling Reagents for the Quantitative Analysis of Low Molecular Weight Metabolites by LC-MS.* Analytical Chemistry, 2008. **80**(23): p. 9171-9180.
51. Yang, W.C., F.E. Regnier, and J. Adamec, *Comparative metabolite profiling of carboxylic acids in rat urine by CE-ESI MS/MS through positively pre-charged and (2)H-coded derivatization.* Electrophoresis, 2008. **29**(22): p. 4549-60.
52. Yang, W.C., et al., *Stable isotope-coded quaternization for comparative quantification of estrogen metabolites by high-performance liquid chromatography-electrospray ionization mass spectrometry.* Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 2008. **870**(2): p. 233-40.
53. Armenta, J.M., et al., *Sensitive and Rapid Method for Amino Acid Quantitation in Malaria Biological Samples Using AccQ.Tag Ultra Performance Liquid Chromatography-Electrospray Ionization-MS/MS with Multiple Reaction Monitoring.* Analytical Chemistry, 2010. **82**(2): p. 548-558.
54. Huang, Y.Q., et al., *Use of isotope mass probes for metabolic analysis of the jasmonate biosynthetic pathway.* Analyst, 2011. **136**(7): p. 1515-1522.
55. Wang, H., et al., *N-Alkylpyridinium isotope quaternization for matrix-assisted laser desorption/ionization Fourier transform mass spectrometric analysis of cholesterol and fatty alcohols in human hair.* Analytica Chimica Acta, 2011. **690**(1): p. 1-9.

56. Yuan, W., et al., *Amine Metabolomics of Hyperglycemic Endothelial Cells using Capillary LC-MS with Isobaric Tagging.* Journal of Proteome Research, 2011. **10**(11): p. 5242-5250.
57. Dai, W.D., et al., *Comprehensive and Highly Sensitive Urinary Steroid Hormone Profiling Method Based on Stable Isotope-Labeling Liquid Chromatography Mass Spectrometry.* Analytical Chemistry, 2012. **84**(23): p. 10245-10251.
58. Mazzotti, F., et al., *Light and heavy dansyl reporter groups in food chemistry: amino acid assay in beverages.* Journal of Mass Spectrometry, 2012. **47**(7): p. 932-939.
59. Toyo'oka, T., *LC-MS determination of bioactive molecules based upon stable isotope-coded derivatization method.* Journal of Pharmaceutical and Biomedical Analysis, 2012. **69**: p. 174-84.
60. Leng, J.P., et al., *A highly sensitive isotope-coded derivatization method and its application for the mass spectrometric analysis of analytes containing the carboxyl group.* Analytica Chimica Acta, 2013. **758**: p. 114-121.
61. Tayyari, F., et al., *N-15-Cholamine-A Smart Isotope Tag for Combining NMR- and MS-Based Metabolite Profiling.* Analytical Chemistry, 2013. **85**(18): p. 8715-8721.
62. Zhang, S.J., et al., *Analysis of estrogenic compounds in environmental and biological samples by liquid chromatography-tandem mass spectrometry with stable isotope-coded ionization-enhancing reagent.* Journal of Chromatography A, 2013. **1280**: p. 84-91.
63. Mochizuki, T., et al., *Isotopic variants of light and heavy L-pyroglutamic acid succinimidyl esters as the derivatization reagents for DL-amino acid chiral metabolomics identification by liquid chromatography and electrospray ionization mass spectrometry.* Analytica Chimica Acta, 2014. **811**: p. 51-59.
64. Li, S.F., et al., *A novel method of liquid chromatography-tandem mass spectrometry combined with chemical derivatization for the determination of ribonucleosides in urine.* Analytica Chimica Acta, 2015. **864**: p. 30-38.
65. Huan, T. and L. Li, *Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform.* Analytical chemistry, 2014. **87**(2): p. 1306-1313.
66. Xia, J., et al., *MetaboAnalyst: a web server for metabolomic data analysis and interpretation.* Nucleic acids research, 2009. **37**(suppl 2): p. W652-W660.
67. Sysi-Aho, M., et al., *Normalization method for metabolomics data using optimal selection of multiple internal standards.* BMC bioinformatics, 2007. **8**(1): p. 93.
68. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley Interdisciplinary Reviews: Computational Statistics, 2010. **2**(4): p. 433-459.
69. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in metabonomics.* Journal of proteome research, 2007. **6**(2): p. 469-479.
70. Bylesjö, M., et al., *OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification.* Journal of Chemometrics, 2006. **20**(8-10): p. 341-351.
71. Wishart, D.S., *Current progress in computational metabolomics.* Briefings in Bioinformatics, 2007. **8**(5): p. 279-293.
72. Worley, B. and R. Powers, *Multivariate analysis in metabolomics.* Current Metabolomics, 2013. **1**(1): p. 92-107.
73. Bijlsma, S., et al., *Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation.* Analytical chemistry, 2006. **78**(2): p. 567-574.

74. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
75. Wang, Y., et al., *PubChem: a public information system for analyzing bioactivities of small molecules.* Nucleic acids research, 2009. **37**(suppl 2): p. W623-W633.
76. Pence, H.E. and A. Williams, *ChemSpider: an online chemical information resource.* Journal of Chemical Education, 2010. **87**(11): p. 1123-1124.
77. Kind, T. and O. Fiehn, *Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry.* BMC bioinformatics, 2007. **8**(1): p. 105.
78. Ausloos, P., et al., *The critical evaluation of a comprehensive mass spectral library.* Journal of the American Society for Mass Spectrometry, 1999. **10**(4): p. 287-299.
79. Horai, H., et al., *MassBank: a public repository for sharing mass spectral data for life sciences.* Journal of mass spectrometry, 2010. **45**(7): p. 703-714.
80. Hill, A.W. and R.J. Mortishire-Smith, *Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach.* Rapid Communications in Mass Spectrometry, 2005. **19**(21): p. 3111-3118.
81. Heinonen, M., et al., *FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data.* Rapid Communications in Mass Spectrometry, 2008. **22**(19): p. 3043-3052.
82. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra.* Bmc Bioinformatics, 2010. **11**: p. 12.
83. Rasche, F., et al., *Identifying the Unknowns by Aligning Fragmentation Trees.* Analytical Chemistry, 2012. **84**(7): p. 3417-3426.
84. Rojas-Cherto, M., et al., *Metabolite Identification Using Automated Comparison of High-Resolution Multistage Mass Spectral Trees.* Analytical Chemistry, 2012. **84**(13): p. 5524-5534.
85. Hufsky, F., K. Scheubert, and S. Bocker, *Computational mass spectrometry for small-molecule fragmentation.* Trac-Trends in Analytical Chemistry, 2014. **53**: p. 41-48.
86. Ma, Y., et al., *MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra.* Analytical Chemistry, 2014. **86**(21): p. 10724-10731.
87. Ridder, L., et al., *In Silico Prediction and Automatic LC-MSn Annotation of Green Tea Metabolites in Urine.* Analytical Chemistry, 2014. **86**(10): p. 4767-4774.
88. Wang, Y.F., et al., *MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics.* Analytical Chemistry, 2014. **86**(19): p. 9496-9503.
89. Zhou, J.R., et al., *HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries.* Bioinformatics, 2014. **30**(4): p. 581-583.
90. Allen, F., R. Greiner, and D. Wishart, *Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification.* Metabolomics, 2015. **11**(1): p. 98-110.
91. Vaniya, A. and O. Fiehn, *Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics.* Trac-Trends in Analytical Chemistry, 2015. **69**: p. 52-61.
92. Liew, A.W.C., N.F. Law, and H. Yan, *Missing value imputation for gene expression data: computational techniques to recover missing data from available information.* Briefings in Bioinformatics, 2011. **12**(5): p. 498-513.

93. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-525.
94. Albrecht, D., et al., *Missing values in gel-based proteomics.* Proteomics, 2010. **10**(6): p. 1202-1211.
95. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC Bioinformatics, 2012. **13**: p. 9.
96. Bijlsma, S., et al., *Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation.* Analytical Chemistry, 2006. **78**(2): p. 567-574.
97. Hrydziuszko, O. and M.R. Viant, *Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline.* Metabolomics, 2012. **8**(1): p. S161-S174.
98. Gromski, P.S., et al., *Influence of missing values substitutes on multivariate analysis of metabolomics data.* Metabolites, 2014. **4**(2): p. 433-52.
99. Little, R. and B. Rubin, *Statistical Analysis with Missing Data*. 2002, Hoboken, NJ: Wiley.
100. Anderle, M., et al., *Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum.* Bioinformatics, 2004. **20**(18): p. 3575-3582.
101. Torres-Garcia, W., et al., *Integrative analysis of transcriptomic and proteomic data of Shewanella oneidensis: missing value imputation using temporal datasets.* Molecular BioSystems, 2011. **7**(4): p. 1093-1104.
102. Valledor, L. and J. Jorrin, *Back to the basics: Maximizing the information obtained by quantitative two dimensional gel electrophoresis analyses by an appropriate experimental design and statistical analyses.* Journal of Proteomics, 2011. **74**(1): p. 1-18.
103. Schwammle, V., I.R. Leon, and O.N. Jensen, *Assessment and Improvement of Statistical Tools for Comparative Proteomics Analysis of Sparse Data Sets with Few Experimental Replicates.* Journal of Proteome Research, 2013. **12**(9): p. 3874-3883.
104. Jung, K., et al., *Adaption of the global test idea to proteomics data with missing values.* Bioinformatics, 2014. **30**(10): p. 1424-1430.
105. Koopmans, F., et al., *Empirical Bayesian Random Censoring Threshold Model Improves Detection of Differentially Abundant Proteins.* Journal of Proteome Research, 2014. **13**(9): p. 3871-3880.
106. Sangster, T.P., et al., *Investigation of analytical variation in metabonomic analysis using liquid chromatography/mass spectrometry.* Rapid Communications in Mass Spectrometry, 2007. **21**(18): p. 2965-2970.
107. Sysi-Aho, M., et al., *Normalization method for metabolomics data using optimal selection of multiple internal standards.* Bmc Bioinformatics, 2007. **8**.
108. Dunn, W.B., et al., *Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system.* Journal of Chromatography, B: Analytical Technologies in the Biomedical and Life Sciences, 2008. **871**(2): p. 288-298.
109. Begley, P., et al., *Development and Performance of a Gas Chromatography-Time-of-Flight Mass Spectrometry Analysis for Large-Scale Nontargeted Metabolomic Studies of Human Serum.* Analytical Chemistry, 2009. **81**(16): p. 7038-7046.

110. Veselkov, K.A., et al., *Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery.* Analytical Chemistry, 2011. **83**(15): p. 5864-5872.
111. Mattarucchi, E. and C. Guillou, *Comment on "Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery".* Analytical Chemistry, 2011. **83**(24): p. 9719-9720.
112. Veselkov, K.A., et al., *Response to Comment on "Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery".* Analytical Chemistry, 2011. **83**(24): p. 9721-9722.
113. Mattarucchi, E. and C. Guillou, *Critical aspects of urine profiling for the selection of potential biomarkers using UPLC-TOF-MS.* Biomedical Chromatography, 2012. **26**(4): p. 512-517.
114. Mak, T.D., et al., *MetaboLyzer: A Novel Statistical Workflow for Analyzing Postprocessed LC-MS Metabolomics Data.* Analytical Chemistry, 2014. **86**(1): p. 506-513.
115. Theodoridis, G.A., et al., *Liquid chromatography-mass spectrometry based global metabolite profiling: A review.* Analytica Chimica Acta, 2012. **711**: p. 7-16.
116. Katajamaa, M. and M. Oresic, *Processing methods for differential analysis of LC/MS profile data.* BMC Bioinformatics, 2005. **6**: p. 12.
117. Dunn, W.B., et al., *Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy.* Chemical Society Reviews, 2010. **40**(1): p. 387-426.
118. Scalbert, A., et al., *Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research.* Metabolomics, 2009. **5**(4): p. 435-458.
119. Goodacre, R., et al., *Proposed minimum reporting standards for data analysis in metabolomics.* Metabolomics, 2007. **3**(3): p. 231-241.
120. Guo, K. and L. Li, *Differential C-12/C-13-Isotope Dansylation Labeling and Fast Liquid Chromatography/Mass Spectrometry for Absolute and Relative Quantification of the Metabolome.* Analytical Chemistry, 2009. **81**(10): p. 3919-3932.
121. Zhou, R., K. Guo, and L. Li, *5-Diethylamino-naphthalene-1-sulfonyl chloride (DensCl): a Novel Triplex Isotope Labeling Reagent for Quantitative Metabolome Analysis by Liquid Chromatography Mass Spectrometry.* Analytical Chemistry, 2013.
122. Bruheim, P., H.F.N. Kvitvang, and S.G. Villas-Boas, *Stable isotope coded derivatizing reagents as internal standards in metabolite profiling.* Journal of Chromatography A, 2013. **1296**: p. 196-203.
123. Zhou, R., et al., *IsoMS: Automated Processing of LC-MS Data Generated by a Chemical Isotope Labeling Metabolomics Platform.* Analytical Chemistry, 2014. **86**(10): p. 4675-4679.
124. Phinney, K.W., et al., *Development of a Standard Reference Material for Metabolomics Research.* Analytical Chemistry, 2013. **85**(24): p. 11732-11738.
125. Peng, J., et al., *Development of a Universal Metabolome-Standard Method for Long-Term LC-MS Metabolome Profiling and Its Application for Bladder Cancer Urine-Metabolite-Biomarker Discovery.* Analytical Chemistry, 2014. **86**(13): p. 6540-6547.
126. Bueschl, C., et al., *Isotopic labeling-assisted metabolomics using LC-MS.* Analytical and Bioanalytical Chemistry, 2013. **405**(1): p. 27-33.

127. Tang, Z.M. and F.P. Guengerich, *Dansylation of Unactivated Alcohols for Improved Mass Spectral Sensitivity and Application to Analysis of Cytochrome P450 Oxidation Products in Tissue Extracts.* Analytical Chemistry, 2010. **82**(18): p. 7706-7712.
128. Song, P., et al., *In Vivo Neurochemical Monitoring Using Benzoyl Chloride Derivatization and Liquid Chromatography-Mass Spectrometry.* Analytical Chemistry, 2012. **84**(1): p. 412-419.
129. Toyo'oka, T., *LC-MS determination of bioactive molecules based upon stable isotope-coded derivatization method.* Journal of Pharmaceutical and Biomedical Analysis, 2012. **69**: p. 174-184.
130. Zhou, R.K., K. Guo, and L. Li, *5-Diethylamino-naphthalene-1-sulfonyl Chloride (DensCl): A Novel Triplex Isotope Labeling Reagent for Quantitative Metabolome Analysis by Liquid Chromatography Mass Spectrometry.* Analytical Chemistry, 2013. **85**(23): p. 11532-11539.
131. Wang, L., et al., *Qualitative and quantitative analysis of enantiomers by mass spectrometry: Application of a simple chiral chloride probe via rapid in-situ reaction.* Analytica Chimica Acta, 2014. **809**: p. 104-108.
132. Sun, X.H., et al., *An in-advance stable isotope labeling strategy for relative analysis of multiple acidic plant hormones in sub-milligram Arabidopsis thaliana seedling and a single seed.* Journal of Chromatography A, 2014. **1338**: p. 67-76.
133. Ulbrich, A., et al., *Organic Acid Quantitation by NeuCode Methylamidation.* Analytical Chemistry, 2014. **86**(9): p. 4402-4408.
134. Liu, P., et al., *Profiling of Thiol-Containing Compounds by Stable Isotope Labeling Double Precursor Ion Scan Mass Spectrometry.* Analytical Chemistry, 2014. **86**(19): p. 9765-9773.
135. Hao, L., et al., *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**(2): p. 467-475.
136. Chen, G.Y., et al., *Development and application of a comparative fatty acid analysis method to investigate voriconazole-induced hepatotoxicity.* Clinica Chimica Acta, 2015. **438**: p. 126-134.
137. Huan, T. and L. Li, *Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform.* Analytical Chemistry, 2015. **87**(2): p. 1306-1313.
138. Zheng, J.M., R.A. Dixon, and L. Li, *Development of Isotope Labeling LC-MS for Human Salivary Metabolomics and Application to Profiling Metabolome Changes Associated with Mild Cognitive Impairment.* Analytical Chemistry, 2012. **84**(24): p. 10802-10811.
139. Wu, Y. and L. Li, *Development of isotope labeling liquid chromatography-mass spectrometry for metabolic profiling of bacterial cells and its application for bacterial differentiation.* Analytical Chemistry, 2013. **85**(12): p. 5755-63.
140. Fu, F.F., et al., *Comparative Proteomic and Metabolomic Analysis of Staphylococcus warneri SG1 Cultured in the Presence and Absence of Butanol.* Journal of Proteome Research, 2013. **12**(10): p. 4478-4489.
141. Smith, C.A., et al., *METLIN - A metabolite mass spectral database.* Therapeutic Drug Monitoring, 2005. **27**(6): p. 747-751.
142. Wishart, D.S., et al., *HMDB: the human metabolome database.* Nucleic Acids Research, 2007. **35**: p. D521-D526.

143. Horai, H., et al., *MassBank: a public repository for sharing mass spectral data for life sciences.* Journal of Mass Spectrometry, 2010. **45**(7): p. 703-714.
144. Li, L., et al., *MyCompoundID: Using an Evidence-Based Metabolome Library for Metabolite Identification.* Analytical Chemistry, 2013. **85**(6): p. 3401-3408.
145. Weber, R.J.M. and M.R. Viant, *MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways.* Chemometrics and Intelligent Laboratory Systems, 2010. **104**(1): p. 75-82.
146. Benton, H.P., et al., *Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling.* Analytical Chemistry, 2015. **87**(2): p. 884-891.
147. Guo, K., F. Bamforth, and L. Li, *Qualitative Metabolome Analysis of Human Cerebrospinal Fluid by C-13-/C-12-Isotope Dansylation Labeling Combined with Liquid Chromatography Fourier Transform Ion Cyclotron Resonance Mass Spectrometry.* Journal of the American Society for Mass Spectrometry, 2011. **22**(2): p. 339-347.
148. Boswell, P.G., et al., *A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments.* Journal of Chromatography A, 2011. **1218**(38): p. 6732-6741.
149. Cao, M.S., et al., *Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics.* Metabolomics, 2015. **11**(3): p. 696-706.
150. Creek, D.J., et al., *Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction.* Analytical Chemistry, 2011. **83**(22): p. 8703-8710.
151. Hagiwara, T., et al., *HPLC Retention time prediction for metabolome analysi.* Bioinformation, 2010. **5**(6): p. 255-8.
152. Sumner, L.W., et al., *Proposed minimum reporting standards for chemical analysis.* Metabolomics, 2007. **3**(3): p. 211-221.
153. Wu, Y. and L. Li, *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**(24): p. 10723-31.
154. Wu, Y. and L. Li, *Development of isotope labeling liquid chromatography-mass spectrometry for metabolic profiling of bacterial cells and its application for bacterial differentiation.* Anal Chem, 2013. **85**(12): p. 5755-63.
155. Wu, M.H., et al., *Liquid chromatography/mass spectrometry methods for measuring dipeptide abundance in non-small-cell lung cancer.* Rapid Communications in Mass Spectrometry, 2013. **27**(18): p. 2091-2098.
156. Guo, K., C.J. Ji, and L. Li, *Stable-isotope dimethylation labeling combined with LC-ESI MS for quantification of amine-containing metabolites in biological samples.* Analytical Chemistry, 2007. **79**(22): p. 8631-8638.
157. Huan, T. and L. Li, *Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry.* Analytical Chemistry, 2015. **in press.**
158. Tang, Y., et al., *PEP Search in MyCompoundID: Detection and Identification of Dipeptides and Tripeptides Using Dimethyl Labeling and Hydrophilic Interaction Liquid Chromatography Tandem Mass Spectrometry.* Analytical Chemistry, 2014. **86**(7): p. 3568-3574.

159. Yuan, M., et al., *A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue.* Nature Protocols, 2012. **7**(5): p. 872-881.
160. Wei, R., G.D. Li, and A.B. Seymour, *High-Throughput and Multiplexed LC/MS/MRM Method for Targeted Metabolomics.* Analytical Chemistry, 2010. **82**(13): p. 5527-5533.
161. Tsugawa, H., et al., *MRMPROBS: A Data Assessment and Metabolite Identification Tool for Large-Scale Multiple Reaction Monitoring Based Widely Targeted Metabolomics.* Analytical Chemistry, 2013. **85**(10): p. 5191-5199.
162. Xu, W., et al., *Development of High-Performance Chemical Isotope Labeling LC-MS for Profiling the Human Fecal Metabolome.* Analytical Chemistry, 2015. **87**(2): p. 829-836.
163. Brown, M., et al., *Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics.* Analyst, 2009. **134**(7): p. 1322-1332.
164. Dunn, W.B., et al., *Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics.* Metabolomics, 2013. **9**(1): p. S44-S66.
165. Oberacher, H., G. Whitley, and B. Berger, *Evaluation of the sensitivity of the 'Wiley registry of tandem mass spectral data, MSforID' with MS/MS data of the 'NIST/NIH/EPA mass spectral library'.* Journal of Mass Spectrometry, 2013. **48**(4): p. 487-496.
166. Stein, S., *Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification.* Analytical Chemistry, 2012. **84**(17): p. 7274-7282.
167. Tautenhahn, R., et al., *An accelerated workflow for untargeted metabolomics using the METLIN database.* Nature Biotechnology, 2012. **30**(9): p. 826-828.
168. Wang, N. and L. Li, *Exploring the precursor ion exclusion feature of liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry for improving protein identification in shotgun proteome analysis.* Analytical Chemistry, 2008. **80**(12): p. 4696-4710.
169. Storandt, M., *Cognitive deficits in the early stages of Alzheimer's disease.* Current Directions in Psychological Science, 2008. **17**(3): p. 198-202.
170. Mueller, S.G., et al., *Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI).* Alzheimer's & Dementia, 2005. **1**(1): p. 55-66.
171. Blennow, K., et al., *Cerebrospinal fluid and plasma biomarkers in Alzheimer disease.* Nature Reviews Neurology, 2010. **6**(3): p. 131-144.
172. Shaw, L.M., et al., *Biomarkers of neurodegeneration for diagnosis and monitoring therapeutics.* Nature reviews Drug discovery, 2007. **6**(4): p. 295-303.
173. Blennow, K., *Cerebrospinal fluid protein biomarkers for Alzheimer's disease.* NeuroRx, 2004. **1**(2): p. 213-225.
174. Blennow, K., E. Vanmechelen, and H. Hampel, *CSF total tau, Aβ42 and phosphorylated tau protein as biomarkers for Alzheimer's disease.* Molecular neurobiology, 2001. **24**(1-3): p. 87-97.
175. Ray, S., et al., *Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins.* Nature medicine, 2007. **13**(11): p. 1359-1362.
176. Song, F., et al., *Plasma biomarkers for mild cognitive impairment and Alzheimer's disease.* Brain research reviews, 2009. **61**(2): p. 69-80.
177. Twamley, E.W., S.A.L. Ropacki, and M.W. Bondi, *Neuropsychological and neuroimaging changes in preclinical Alzheimer's disease.* Journal of the International Neuropsychological Society, 2006. **12**(05): p. 707-735.

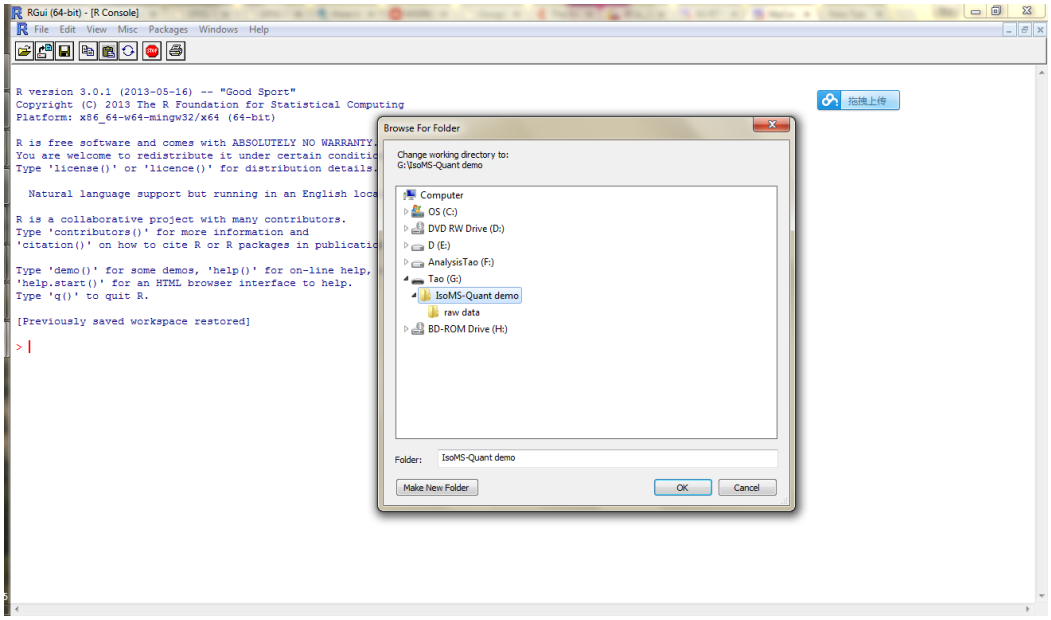
178. Trushina, E., et al., *Identification of altered metabolic pathways in plasma and CSF in mild cognitive impairment and Alzheimer's disease using metabolomics.* 2013.
179. Ibáñez, C., et al., *Toward a predictive model of Alzheimer's disease progression using capillary electrophoresis–mass spectrometry metabolomics.* Analytical chemistry, 2012. **84**(20): p. 8532-8540.
180. Sato, Y., et al., *Identification of a new plasma biomarker of Alzheimer's disease using metabolomics technology.* Journal of lipid research, 2012. **53**(3): p. 567-576.
181. Graham, S.F., et al., *Untargeted Metabolomic Analysis of Human Plasma Indicates Differentially Affected Polyamine and L-Arginine Metabolism in Mild Cognitive Impairment Subjects Converting to Alzheimer's Disease.* PloS one, 2015. **10**(3): p. e0119452.
182. Inoue, K., et al., *Blood-based diagnosis of Alzheimer's disease using fingerprinting metabolomics based on hydrophilic interaction liquid chromatography with mass spectrometry and multivariate statistical analysis.* Journal of Chromatography B, 2015. **974**: p. 24-34.
183. González-Domínguez, R., et al., *Metabolomic screening of regional brain alterations in the APP/PS1 transgenic model of Alzheimer's disease by direct infusion mass spectrometry.* Journal of pharmaceutical and biomedical analysis, 2015. **102**: p. 425-435.
184. Dallmann, R., et al., *The human circadian metabolome.* Proceedings of the National Academy of Sciences, 2012. **109**(7): p. 2625-2629.
185. Zhang, A., H. Sun, and X. Wang, *Saliva metabolomics opens door to biomarker discovery, disease diagnosis, and treatment.* Applied biochemistry and biotechnology, 2012. **168**(6): p. 1718-1727.
186. Sugimoto, M., et al., *Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles.* Metabolomics, 2010. **6**(1): p. 78-95.
187. Zheng, J., R.A. Dixon, and L. Li, *Development of isotope labeling LC–MS for human salivary metabolomics and application to profiling metabolome changes associated with mild cognitive impairment.* Analytical chemistry, 2012. **84**(24): p. 10802-10811.
188. Wu, Y. and L. Li, *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**(24): p. 10723-10731.
189. Wishart, D.S., et al., *HMDB 3.0—the human metabolome database in 2013.* Nucleic acids research, 2012: p. gks1065.
190. Xia, J., et al., *Translational biomarker discovery in clinical metabolomics: an introductory tutorial.* Metabolomics, 2013. **9**(2): p. 280-299.

Appendix

IsoMS-Quant User Manual

- IsoMS-Quant is a program written in R for recalculating the peak intensity ratio using the chromatographic peak area information. This program is part of the data processing software used for the chemical isotope labeling (CIL) LC-MS metabolomics platform.
- The IsoMS-Quant script is freely available for non-commercial use from www.mycompoundid.org.
- The instruction for using the IsoMS-Quant program is given below.

1) Download the IsoMS-Quant script from MyCompoundID.org.
2) Assign the folder of IsoMS-Quant as the working folder of RGui by clicking: File → Change dir… (see below).



3) Open the IsoMS-Quant script (see below) and change the parameters therein (see Table 1 for the explanation of these parameters).

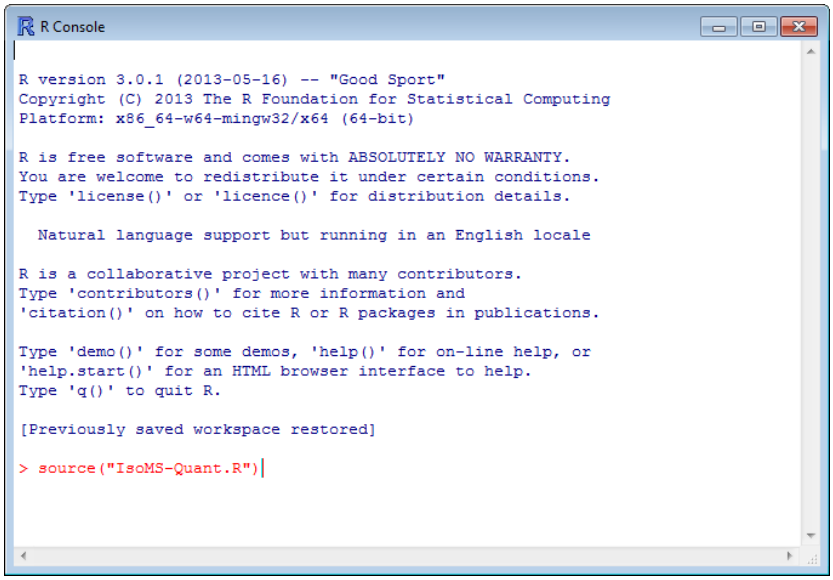
```
########################################
# This is a script to do peak reconstruction after the zero-fill process
# Tao Huan, APril, 09, 2015
# Copyright @ University of Alberta


########################################
# This is the setting part
file.path <- "G:/IsoMS-Quant demo/"
raw.file.path <- "G:/IsoMS-Quant demo/raw data/"
mz.tol = 5 # default 5 ppm
rt.tol = 30 # default 30 seconds
int.uplimit <- 1e7 # Saturation intensity threshold
########################################
```

Table 1. IsoMS-Quant parameters that need to be changed according to the user's LC-MS instrumental conditions.

Parameter	Function
file.path	Set the data path to the folder that contains the zero-filled metabolite-intensity matrix
raw.file.path	Set the data path to the folder that contains all the raw LC-MS data containing all the peak information
mz.tol	Set the mz tolerance for the IsoMS-Quant processing
rt.tol	Set the retention time tolerance for the IsoMS-Quant processing
int.uplimit	Set the mass intensity saturation threshold

4) Save the parameter changes to the script. Type in the command code in RGui as shown in red in the following screen shot and press enter to run the script.



5) After running the script, a new csv file named "After_reconstruction_ratio.csv" will be created. This csv file contains the IsoMS-Quant result.

Supplemental Note for Dns-lib Work Instrumental Settings

LC-MS for Constructing the Dansyl Library. To construct the dansyl library, an individual ^{12}C-dansyl labeled standard was injected into a Bruker HD Impact QTOF mass spectrometer (Billerica, MA, USA) with electrospray ionization (ESI) linked to an Agilent 1100 HPLC system (Palo Alto, CA, USA). Reversed-phase Zorbax Eclipse C18 column (2.0 mm × 100 mm, 1.7 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=0.0 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 sample injection volume was 2 µL. All the spectra were collected using the positive ion mode. For MS/MS, multiple reaction monitoring (MRM) using the Bruker QTOF-MS was used to generate an averaged

collision-induced dissociation (CID) spectrum at the collision energies of 20 and 50 eV for each dansyl standard; half of the acquisition time was spent at 20 eV and another half was spent at 50 eV. In this way, both low and high mass fragment ions were detected, providing better coverage for spectral comparison.

LC and LC-MS Settings. Various instrumental settings were used in this work for different purposes and are given below.

(1) LC-QTOF-MS

Instrument: Bruker HD Impact QTOF system (Billerica, MA, USA) equipped with an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA).

Column: An Agilent reversed phase Eclipse Plus C18 column (2.1 mm × 10 cm, 1.8 μm particle size, 95 Å pore size) for separation.

LC gradient: Solvent A was 0.1% (v/v) formic acid in 5% (v/v) acetonitrile, and solvent B was 0.1% (v/v) formic acid in acetonitrile. The chromatographic conditions were: t = 0 min, 20% B; t = 3.5 min, 35% B; t = 18 min, 65% B; t = 21 min, 95% B; t = 21.5 min, 95% B; t = 23 min, 98% B; t = 24.5 min, 98% B; t = 26.5 min, 99% B; t = 28.5 min, 99% B; t = 29.5 min, 20% B. The flow rate was 180 μL/min and the injection volume was 2 μL.

QTOF instrument parameters

1. MS instrument parameter: m/z scan range: 150 to 1000
2. Ion mode: positive ion

3. Source parameters

   End plate offset: 500 V, Capillary: 4500 V, Nebulizer: 1.8 Bar, Dry gas: 8.0 L/min, Dry temperature: 230 ºC.

4. Tune parameters

   Funnel 1 RF 250.0 Vpp, Funnel 2 RF: 150.0 Vpp, Hexapole RF: 110.0 Vpp, Quadrupole ion energy: 3.0 eV, Low mass: 100.00, Collision RF 1500.0 Vpp, Transfer time: 80.0 µs. Pre pulse storage 10.0 µs.

(2) LC-QTOF-MRM for MS/MS Spectral Library Construction

Instrument: Bruker HD Impact QTOF system (Billerica, MA, USA) equipped with an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA).

Column: A Phenomenex Kinetex C18 column (2.1 mm × 5 cm, 1.7 µm particle size, 100 Å pore size) was used for chromatographic separation.

LC gradient: Solvent A was 0.1% (v/v) formic acid in 5% (v/v) acetonitrile, and solvent B was 0.1% (v/v) formic acid in acetonitrile. The chromatographic conditions were: t=0.0, 20%B; t = 1.0, 20%B; t = 1.01, 99% B, t = 10.0, 99% B, t = 10.01, 20%B; t = 18.0, 20%B. The flow rate was 180 µL/min and the injection volume was 5 µL.

QTOF instrument parameters

1. MS instrument parameter: m/z scan range: 20 to 1000
2. Ion mode: positive ion
3. Source parameters

End plate offset: 500 V, Capillary: 4500 V, Nebulizer: 1.0 Bar, Dry gas: 6.0 L/min, Dry temperature: 230 ºC.

4. Tune parameters

   Funnel 1 RF 200.0 Vpp, Funnel 2 RF: 200.0 Vpp, Hexapole RF: 50.0 Vpp, Quadrupole ion energy: 5.0 eV, Low mass: 50.00, Pre pulse storage 5.0 μs. Basic stepping: Collision RF 200.0 - 700.0 Vpp (50% - 50%), Collision Energy 20 – 50 ev (timing 50% - 50%).

5. MRM

   Precursor ion m/z: specified to the mz_light, width: 6.0, isCID: 0.0 eV, collision energy: 40 eV, x Acq: 2.0

(3) LC-Qtrap-MS/MS

Instrument: QTRAP 4000 system (Applied Biosystems, Foster City, CA) equipped with an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA).

Column: A Phenomenex Kinetex C18 column (2.1 mm × 5 cm, 1.7 μm particle size, 100 Å pore size) was used for chromatographic separation.

LC gradient: Solvent A was 0.1% (v/v) formic acid in 5% (v/v) acetonitrile, and solvent B was 0.1% (v/v) formic acid in acetonitrile. The chromatographic conditions were: t = 0 min, 20% B; t = 1 min, 20% B; t = 1.01 min, 99% B; t = 2 min, 99% B; t = 2.01 min, 20% B; t = 8 min, 20% B. The flow rate was 180 μL/min and the injection volume was 5 μL.

Qtrap instrument parameters:

All MS/MS spectra were obtained in the positive ion mode with enhanced product ion scan. Mass range was *m/z* 50-700 for HMDB03134 and *m/z* 50-500 for HMDB28878, with a scan rate of 1000 Da/s. Dynamic fill time was selected. Curtain gas was set to 10 psi, CAD gas was set to high, IS was 4800, TEM was 200, GS1 and GS2 were set to 12 and 0, respectively and ihe was on.

(4) LC-QTOF-MS/MS for Running Labeled Samples

Instrument: Bruker HD Impact QTOF system (Billerica, MA, USA) equipped with an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA).

Column: An Agilent reversed phase Eclipse Plus C18 column (2.1 mm × 10 cm, 1.8 μm particle size, 95 Å pore size) for separation.

LC gradient: Solvent A was 0.1% (v/v) formic acid in 5% (v/v) acetonitrile, and solvent B was 0.1% (v/v) formic acid in acetonitrile. The chromatographic conditions were: t = 0 min, 20% B; t = 3.5 min, 35% B; t = 18 min, 65% B; t = 21 min, 95% B; t = 21.5 min, 95% B; t = 23 min, 98% B; t = 24.5 min, 98% B; t = 26.5 min, 99% B; t = 28.5 min, 99% B; t = 29.5 min, 20% B. The flow rate was 180 μL/min and the injection volume was 2 μL.

QTOF instrument parameters:

1. m/z scan range: 150 to 1000
2. Ion mode: positive ion
3. Source parameters

End plate offset: 500 V, Capillary: 4500 V, Nebulizer: 1.8 Bar, Dry gas: 8.0 L/min, Dry temperature: 230 ºC.

4. Tune parameters

   Funnel 1 RF 250.0 Vpp, Funnel 2 RF: 150.0 Vpp, Hexapole RF: 110.0 Vpp, Quadrupole ion energy: 3.0 eV, Low mass: 100.00, Pre pulse storage 10.0 µs. Basic stepping: Collision RF 200.0 - 1500.0 Vpp (50% - 50%), Collision energy: 20 - 50 ev (timing: 50% - 50%).

5. Auto MS/MS

   Precursor exclusion mode. Smart exclusion 5X, Active exclusion: exclude after 3 spectra, release after 1.00 min.

(5) LC-FTICR-MS

Instrument: Bruker 9.4 Tesla Apex-Qe Fourier transform ion-cyclotron resonance (FTICR) mass spectrometer (Bruker, Billerica, MA) linked to an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA).

Column: The samples were injected onto an Agilent reversed phase Eclipse Plus C18 column (2.1 mm × 10 cm, 1.8 µm particle size, 95 Å pore size) for separation.

LC gradient: Solvent A was 0.1% (v/v) formic acid in 5% (v/v) acetonitrile, and solvent B was 0.1% (v/v) formic acid in acetonitrile. The chromatographic conditions were: t = 0 min, 20% B; t = 3.5 min, 35% B; t = 18 min, 65% B; t = 21 min, 95% B; t = 21.5 min, 95% B; t = 23 min, 98% B; t = 24.5 min, 98% B; t = 26.5 min, 99% B; t = 28.5 min, 99% B; t = 29.5 min, 20% B. The flow rate was 180 µL/min. A splitter of 2:1 ratio was used and 120 µL/min of the flow injected into the MS instrument. The sample injection volume was 2 µL.

FTICR instrument parameters:

1. m/z scan range: 200 to 1000
2. Ion mode: positive ion
3. Source parameters

   Capillary: 4200 V, Nebulizer: 2.3 Bar, Dry gas: 7.0 L/min, Dry temperature: 190 ºC.
4. Tune parameters

   Acquisition 256k, Average spectra 2, source accumulation 0.1 seconds, collision cell accumulation 1.0 seconds, TOF 0.0007 seconds.

Dns-lib Tutorial

1. Workflow. The workflow for metabolite identification using Dns-library is shown below (Figure 1).

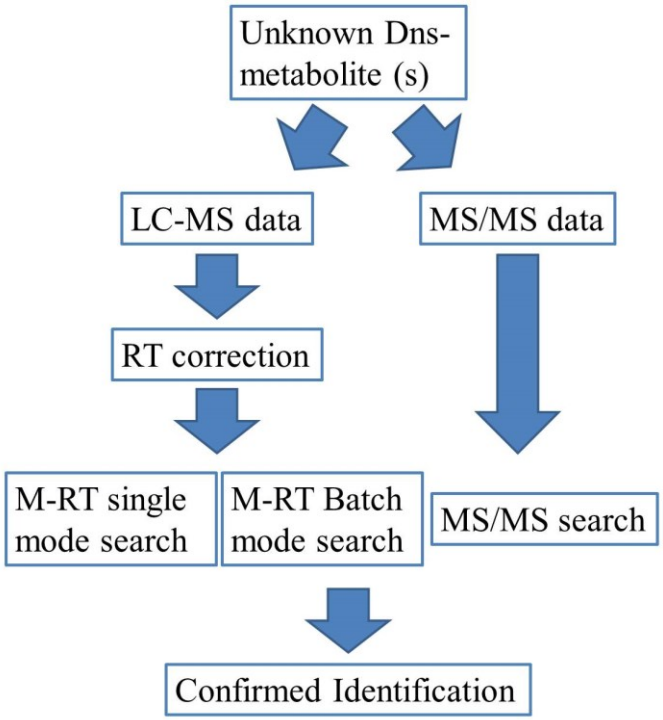


Figure 1. Workflow for M-RT search and MS/MS search.

2. Dns-library database. The current Dns-library consists of 273 unique metabolites with a total of 315 entries. The Dns-library view on the sidebar lists all these Dns-metabolites with their mz and normalized RT information. Figure 2 shows a screenshot of the Dns-library database. The user can view the HMDB number, monoisotopic molecular mass, mz_light, normalized or corrected RT for each of the Dns-metabolite standards from the table. In addition, the hyperlinks for each Dns-metabolite to HMDB and KEGG databases are provided. These databases provide detailed biological information about the metabolite.

#	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Corrected RT	HMDB link	KEGG link	Show detail
1	HMDB00001	1_Methylhistidine	169.0851	403.1434	2.17	Link	Link	Detail
2	HMDB00002	1-3_Diaminopropane	74.0844	308.1427	2.63	Link	Link	Detail
3	HMDB00002	1-3_Diaminopropane - multi-tags	74.0844	271.0583	20.49	Link	Link	Detail
4	HMDB00020	p-Hydroxyphenylacetic acid	152.0473	386.1057	16.91	Link	Link	Detail
5	HMDB00021	Iodotyrosine	306.9705	387.5436	23.88	Link	Link	Detail
6	HMDB00022	3_Methoxytyramine	167.0946	317.6056	25.49	Link	Link	Detail
7	HMDB00045	Adenosine monophosphate	347.0631	581.1214	1.75	Link	Link	Detail
8	HMDB00050	Adenosine	267.0968	501.1551	3.94	Link	Link	Detail
9	HMDB00051	Ammonia	17.0266	251.0849	5.82	Link	Link	Detail
10	HMDB00056	Beta-Alanine	89.0477	323.1060	7.24	Link	Link	Detail
11	HMDB00064	Creatine	131.0695	365.1278	3.02	Link	Link	Detail
12	HMDB00070	D-Pipecolic acid	129.0790	363.1373	13.23	Link	Link	Detail
13	HMDB00085	Deoxyguanosine	267.0968	501.1551	8.49	Link	Link	Detail
14	HMDB00087	Dimethylamine	45.0578	279.1162	15.07	Link	Link	Detail
15	HMDB00089	Cytidine	243.0855	477.1438	5.87	Link	Link	Detail
16	HMDB00089	Cytidine - H2O	243.0855	459.1333	7.38	Link	Link	Detail
17	HMDB00095	Cytidine monophosphate	323.0519	557.1102	1.88	Link	Link	Detail
18	HMDB00095	Cytidine monophosphate - Isomer	323.0519	557.1102	2.87	Link	Link	Detail
19	HMDB00099	L-Cystathionine	222.0674	345.0920	13.34	Link	Link	Detail
20	HMDB00099	L-Cystathionine - Isomer	222.0674	345.0920	13.69	Link	Link	Detail
21	HMDB00101	Deoxyadenosine	251.1018	485.1602	8.72	Link	Link	Detail
22	HMDB00112	Gamma-Aminobutyric acid	103.0633	337.1216	7.79	Link	Link	Detail
23	HMDB00112	Gamma-Aminobutyric acid - H2O	103.0633	319.1144	13.57	Link	Link	Detail

Figure 2. Screenshot of a partial Dns-library table.

The user can click the “Show Detail” button, which guides the user to a page with more detailed information about the dansyl labeled metabolite (Figure 3). An LC-MS chromatogram and MS/MS spectrum are provided on this page. These data were collected using pure standard

compound and can be used to compare with the user's experimental data. Details on the preparation of the Dns-standards can be found in the materials and methods part of the paper.

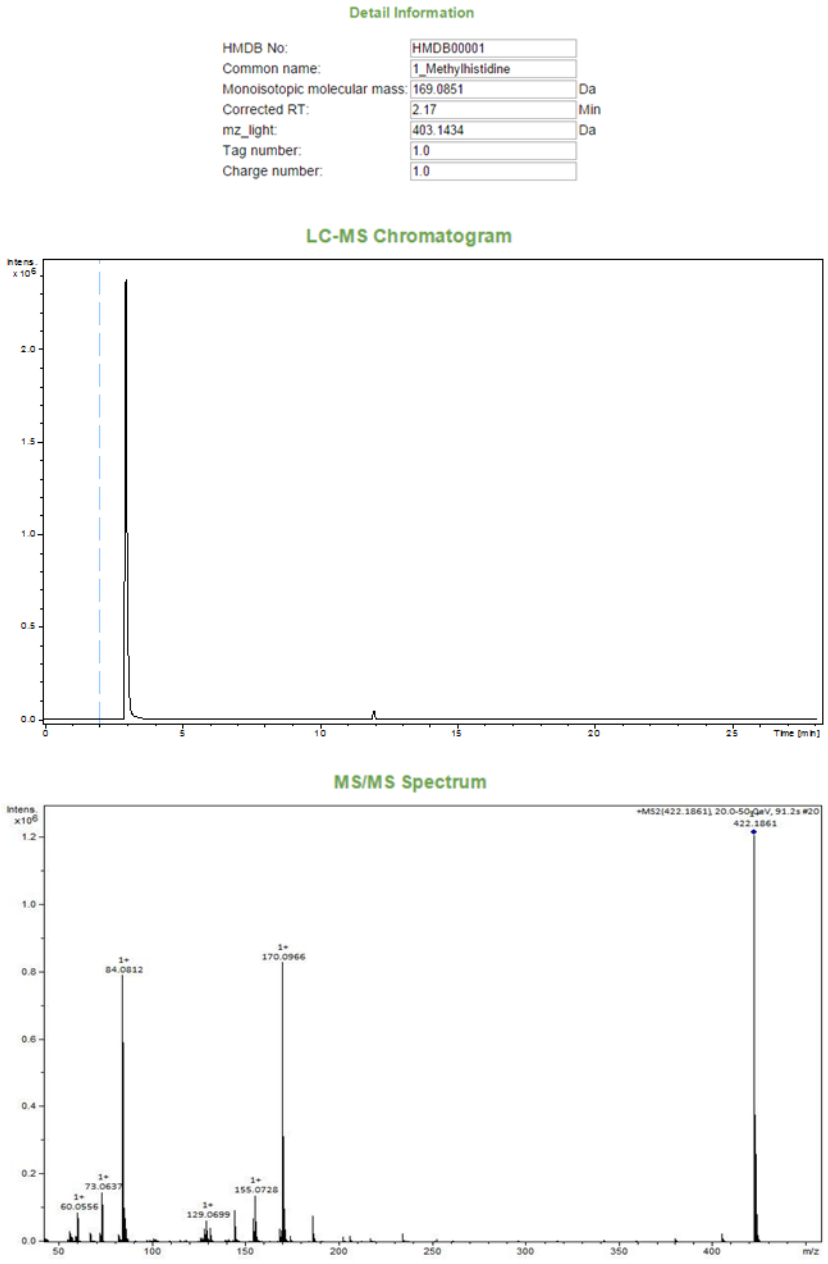


Figure 3. Screenshot of the "Show detail" page.

3. **M-RT single mode search.** M-RT single mode search allows a user to search the Dns-library by submitting a single metabolite feature with its RT and mass (M+H). Also, a calibration file

needs to be submitted to correct the retention time of the single metabolite feature. Figure 4 shows the screenshot of the single mode search.

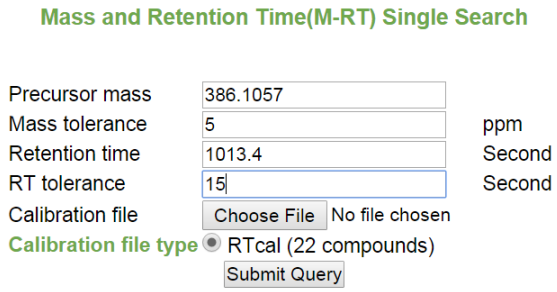


Figure 4. M-RT single mode search parameters.

There are six search parameters.

1) **Precursor Mass.** The user needs to input the precursor mass of the metabolite feature.
2) **Mass tolerance.** The user needs to define a mass tolerance for the precursor mass search. 5 ppm is normally used for data collected using high resolution MS such as TOF and FT (10 ppm or higher may be used for very low abundance peaks). If the experiment is performed using a low resolution MS instrument, a larger mass tolerance should be considered.
3) **Retention time.** The user needs to input the retention time of the metabolite feature.
4) **RT tolerance.** The user needs to define a retention time tolerance for the M-RT search;15 seconds is normally used. If no close matches are found, a wider retention time window should be considered with caution. For LC with lower retention time precision, a larger RT tolerance may be used.
5) **Calibration file.** A calibration file needs to be uploaded to adjust the retention time of the metabolite feature to match the retention time of metabolites in the Dns-library. The template of the calibration file is shown in the "**user example**". The user needs to

download the template and fill in the retention time information for each of the calibration standards used in the calibration file. The retention time has a unit of second.

6) **Calibration file type.** In the current Dns-library RT correction method, a 22 Dns-standards file is used. We will include other types of the calibration files for different applications in the future.

7) **Submit query.** Once all the parameters have been set, the user can click on the "submit query" to start the M-RT single mode search.

4. M-RT batch mode search. M-RT batch search mode allows a user to search the Dns-library using the entire dansyl-labeled LC-MS file. Figure 5 shows the screenshot of the batch mode search.

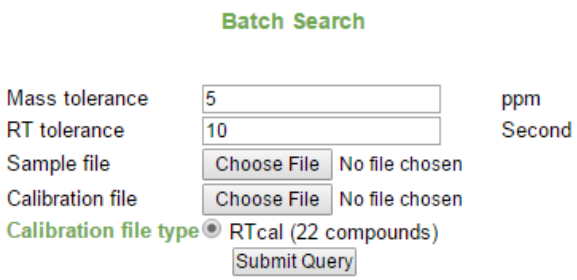


Figure 5. M-RT batch mode search parameters.

The parameters include:

1) **Mass tolerance.** The user needs to define a mass tolerance for the precursor mass search. 5 ppm is normally used for data collected using high resolution MS such as TOF and FT (10 ppm or higher may be used for very low abundance peaks). If the experiment is performed using a low resolution MS instrument, a larger mass tolerance should be considered.

2) **RT tolerance.** The user needs to define a retention time tolerance for the M-RT search;15 seconds is normally used. If no close matches are found, a wider retention time window should be considered with caution. For LC with lower retention time precision, a larger RT tolerance may be used.
3) **Sample file.** A sample file needs to be uploaded onto the website for batch mode search. The sample file is the metabolite-intensity matrix after processing the raw LC-MS data in IsoMS, Iso-Align, and Zero-fill.
4) **Calibration file.** A calibration file needs to be uploaded for adjusting the retention time of the metabolite feature to match with the retention time of the metabolites in the Dns-library. The template of the calibration file is shown in the "**user example**". The user needs to download this template and fill in the retention time information for each of the calibration standards used in the calibration file. The retention time has a unit of second.
5) **Calibration file type.** In the current Dns-library RT correction method, a 22-Dns-standards file is used. We will include other types of the calibration files for different applications in the future.
6) **Submit query.** Once all the parameters have been set, the user can click on the "submit query" to start the M-RT batch mode search.

5. MS/MS search. The MS/MS search function allows a user to identify a dansyl labeled metabolite using MS/MS information. Figure 6 shows the screenshot of the MS/MS search function.

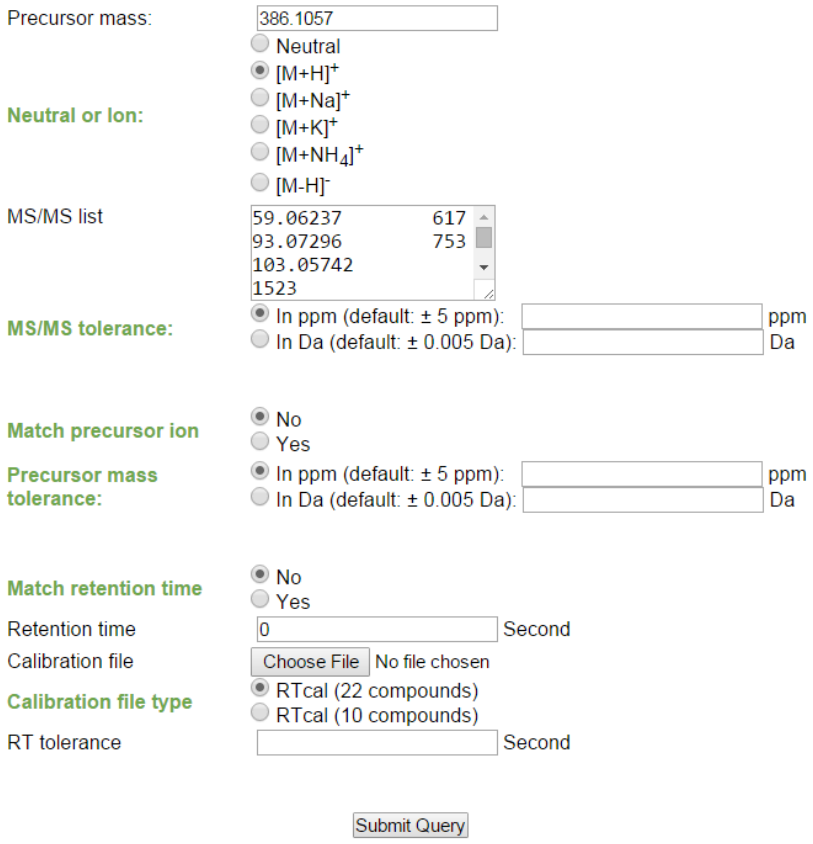


Figure 6. MS/MS search parameters.

The parameters include:

1) **Precursor mass.** The user needs to input the precursor mass of the metabolite feature.
2) **Neutral or ion.** The user can define the type of the precursor mass. It can be either an M+H ion or a neutral mass.
3) **MS/MS list.** The user needs to input a list of MS/MS fragment ion masses with their associated intensities.
4) **MS/MS tolerance.** The user needs to set a mass tolerance for the MS/MS fragment ions to perform the matching with the MS/MS information in the Dns-library.
5) **Match precursor ion.** The user has the option of defining the precursor ion mass for MS/MS search. If this option is enabled, only the Dns-metabolites that match with the

precursor mass will be further used to compare the MS/MS fragment ions. If this option is disabled, the MS/MS match is performed on all 273 Dns-metabolites.

6) **Precursor mass tolerance.** The user needs to define a mass tolerance for the precursor mass search. 5 ppm is normally used for data collected using high resolution MS such as TOF and FT (10 ppm or higher may be used for very low abundance peaks). If the experiment is performed using a low resolution MS instrument, a larger mass tolerance should be considered.
7) **Match retention time.** The user has the option of including RT for MS/MS search. If this option is on, only the Dns-metabolites that match with the retention time will be further used to compare the MS/MS fragment ions. If this option is off, the MS/MS match is performed on all 273 Dns-metabolites.
8) **RT tolerance.** The user needs to define a retention time tolerance for the M-RT search;15 seconds is normally used. If no close matches are found, a wider retention time window should be considered with caution. For LC with lower retention time precision, a larger RT tolerance may be used.
9) **Calibration file.** A calibration file needs to be uploaded for adjusting the retention time of the metabolite feature to be consistent with the retention time of metabolites in the Dns-library. The template of the calibration file is shown in the "**user example**". The user needs to download that template and fill in the retention time information for each of the calibration standard used in the calibration file. The retention time has a unit of second.

10) **Calibration file type.** In the current Dns-library RT correction method, a 22-Dns-standards file is used. We will include other types of the calibration files for different applications in the future.

6. M-RT search result display. Figure 8 shows the screenshot of the M-RT search result. The search result table is similar to the Dns-library table with several extra columns.

Search Result

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	Show detail
1	375.0785	2.05	1.94	HMDB00224	O-Phosphoethanolamine	141.0191	375.0774	2.02	0.0011	0.08	Link	Link	Detail
2	359.0743	2.31	2.18	HMDB00251	Taurine	125.0147	359.0730	2.24	0.0013	0.06	Link	Link	Detail
3	403.1443	2.32	2.19	HMDB00001	1_Methylhistidine	169.0851	403.1434	2.17	0.0009	0.02	Link	Link	Detail
4	403.1443	2.32	2.19	HMDB00479	3_methyl-histidine	169.0851	403.1434	2.01	0.0009	0.18	Link	Link	Detail
5	408.1708	2.61	2.47	HMDB00517	L-Arginine	174.1117	408.1700	2.44	0.0008	0.03	Link	Link	Detail
6	343.0781	2.64	2.50	HMDB00965	Hypotaurine	109.0197	343.0781	2.47	0.0000	0.03	Link	Link	Detail
7	351.1124	2.81	2.67	HMDB00128	Guanidoacetic acid	117.0538	351.1121	2.74	0.0003	0.07	Link	Link	Detail
8	366.1132	3.09	2.94	HMDB00168	L-Asparagine	132.0535	366.1118	3.0	0.0014	0.06	Link	Link	Detail
9	422.1861	3.21	3.06	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.0	0.0005	0.06	Link	Link	Detail
10	359.1547	3.28	3.14	HMDB01861	3_Methylhistamine	125.0953	359.1536	3.27	0.0011	0.13	Link	Link	Detail

Figure 8. Screenshot of M-RT search result.

7. MS/MS search result display. Figure 9 shows the screenshot of the M-RT search result. The search result table is similar to the Dns-library table with several extra columns.

Search Result

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	MS/MS score ▲	Show detail
7	581.1214	NA	NA	HMDB00045	Adenosine monophosphate	347.0631	581.1214	1.75	0.0000	NA	Link	Link	1.00	Detail
155	581.1214	NA	NA	HMDB01341	ADP	427.0294	661.0877	1.49	79.9663	NA	Link	Link	0.74	Detail
142	581.1214	NA	NA	HMDB01044	2'-Deoxyguanosine 5'-monophosphate	347.0631	581.1214	5.57	0.0000	NA	Link	Link	0.52	Detail
151	581.1214	NA	NA	HMDB01173	5'-Methylthioadenosine	297.0896	531.1479	6.97	49.9735	NA	Link	Link	0.19	Detail
8	581.1214	NA	NA	HMDB00050	Adenosine	267.0968	501.1551	3.94	79.9663	NA	Link	Link	0.18	Detail
293	581.1214	NA	NA	HMDB60003	Isovanillic acid	168.0423	402.1006	15.69	179.0208	NA	Link	Link	0.16	Detail
21	581.1214	NA	NA	HMDB00101	Deoxyadenosine	251.1018	485.1602	8.72	95.9612	NA	Link	Link	0.14	Detail
146	581.1214	NA	NA	HMDB01069	2_Phenylaminoadenosine	358.1390	592.1973	8.73	11.0759	NA	Link	Link	0.04	Detail
18	581.1214	NA	NA	HMDB00095	Cytidine monophosphate - Isomer	323.0519	557.1102	2.87	24.0112	NA	Link	Link	0.03	Detail

Figure 9. Screenshot of MS/MS search.

Examples of M-RT and MS/MS Search

1. An example of using M-RT to do single mode search.

1). For the M-RT single mode search, the user enters a precursor mass (359.0730) and retention time (425.28 seconds), together with their mass tolerance (5 ppm) and RT tolerance (15 seconds) (see Figure 1). A calibration file with 22 calibration standards also needs to be uploaded. The template of the calibration file can be found below. The user needs to download it and change the retention time according to the calibration file performed with the metabolite feature. After filling out the retention time, click the "Submit Query" to start the M-RT single mode search.

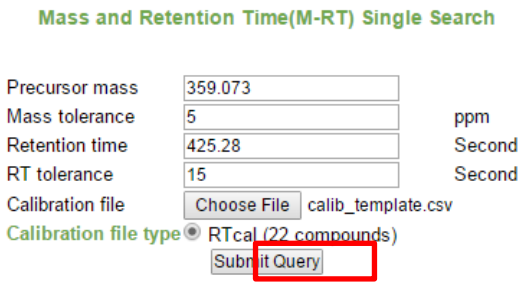


Figure 1. Single mode search parameter

2). The search result is shown in Figure 2.

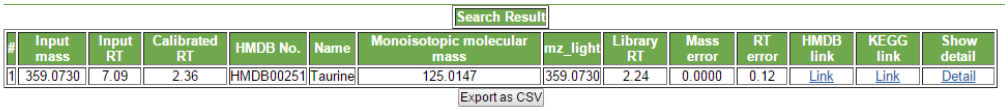


Figure 2. Single search result.

2. An example of using M-RT to do batch mode search.

1). For the M-RT batch mode search, the user enters a mass tolerance (5 ppm) and RT tolerance (15 seconds) (see Figure 1). The user also needs to upload a sample file and a calibration file. The template of the sample file and calibration file are attached. For the calibration file, the user needs to download it and change the retention time according to the calibration file. After it's all done, click the "Submit Query" to start the M-RT single mode search.

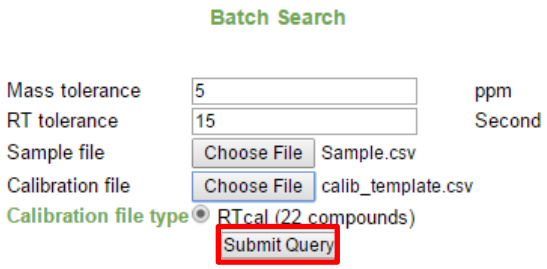


Figure 3. Batch mode search parameters.

2). The search result is shown in Figure 4.

Search Result

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	Show detail
1	375.0785	2.05	1.94	HMDB00224	O-Phosphoethanolamine	141.0191	375.0774	2.02	0.0011	0.08	Link	Link	Detail
2	359.0743	2.31	2.18	HMDB00251	Taurine	125.0147	359.0730	2.24	0.0013	0.06	Link	Link	Detail
3	403.1443	2.32	2.19	HMDB00001	1_Methylhistidine	169.0851	403.1434	2.17	0.0009	0.02	Link	Link	Detail
4	403.1443	2.32	2.19	HMDB00479	3_methyl-histidine	169.0851	403.1434	2.01	0.0009	0.18	Link	Link	Detail
5	408.1708	2.61	2.47	HMDB00517	L-Arginine	174.1117	408.1700	2.44	0.0008	0.03	Link	Link	Detail
6	343.0781	2.64	2.50	HMDB00965	Hypotaurine	109.0197	343.0781	2.47	0.0000	0.03	Link	Link	Detail
7	351.1124	2.81	2.67	HMDB00128	Guanidoacetic acid	117.0538	351.1121	2.74	0.0003	0.07	Link	Link	Detail
8	366.1132	3.09	2.94	HMDB00168	L-Asparagine	132.0535	366.1118	3.0	0.0014	0.06	Link	Link	Detail
9	422.1861	3.21	3.06	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.0	0.0005	0.06	Link	Link	Detail
10	359.1547	3.28	3.14	HMDB01861	3_Methylhistamine	125.0953	359.1536	3.27	0.0011	0.13	Link	Link	Detail
11	436.2014	3.44	3.29	HMDB03334	Symmetric dimethylarginine	202.1430	436.2013	3.05	0.0001	0.24	Link	Link	Detail

Figure 4. Batch mode search result.

3). At the end of the search result table, there is an "Export as CSV" button (Figure 5). By clicking this button, the user can export the search results into a CSV table shown in Figure 6

113	311.0703	24.64	24.44	HMDB00152	Gentisic acid - multi-tags	154.0266	311.0716	24.69	0.0013	0.25	Link	Link	Detail
114	311.0703	24.64	24.44	HMDB01856	Protocatechuic acid	154.0266	311.0716	24.51	0.0013	0.07	Link	Link	Detail
115	322.1045	24.88	24.67	HMDB00259	Serotonin	176.0950	322.1058	24.65	0.0013	0.02	Link	Link	Detail
116	318.0783	24.94	24.74	HMDB00130	Homogentisic acid	168.0423	318.0794	24.84	0.0011	0.10	Link	Link	Detail
117	317.6047	25.56	25.46	HMDB00022	3_Methoxytyramine	167.0946	317.6056	25.49	0.0009	0.03	Link	Link	Detail
118	317.6047	25.56	25.46	HMDB02182	Phenylephrine	167.0946	317.6056	25.39	0.0009	0.07	Link	Link	Detail
119	302.5998	25.88	25.80	HMDB00306	Tyramine	137.0841	302.6004	25.83	0.0006	0.03	Link	Link	Detail

Export as CSV

Figure 5. "Export as CSV" button.

8	search result:										
9	#	Input mas	Input RT	Calibratec	HMDB No.	Name	Monoisot	mz_light	Library RT	Mass erro	RT error
10	1	375.0785	2.05	1.94	HMDB002	O-Phosphoethanolamine	141.0191	375.0774	2.02	0.0011	0.08
11	2	359.0743	2.31	2.18	HMDB002	Taurine	125.0147	359.073	2.24	0.0013	0.06
12	3	403.1443	2.32	2.19	HMDB000	1_Methylhistidine	169.0851	403.1434	2.17	0.0009	0.02
13	4	403.1443	2.32	2.19	HMDB004	3_methyl-histidine	169.0851	403.1434	2.01	0.0009	0.18
14	5	408.1708	2.61	2.47	HMDB005	L-Arginine	174.1117	408.17	2.44	0.0008	0.03
15	6	343.0781	2.64	2.5	HMDB009	Hypotaurine	109.0197	343.0781	2.47	0	0.03
16	7	351.1124	2.81	2.67	HMDB001	Guanidoacetic acid	117.0538	351.1121	2.74	0.0003	0.07
17	8	366.1132	3.09	2.94	HMDB001	L-Asparagine	132.0535	366.1118	3	0.0014	0.06
18	9	422.1861	3.21	3.06	HMDB006	Homo-L-arginine	188.1273	422.1856	3	0.0005	0.06
19	10	359.1547	3.28	3.14	HMDB018	3_Methylhistamine	125.0953	359.1536	3.27	0.0011	0.13
20	11	436.2014	3.44	3.29	HMDB033	Symmetric dimethylarginine	202.143	436.2013	3.05	0.0001	0.24
21	12	380.1288	3.44	3.3	HMDB006	L-Glutamine	146.0691	380.1275	3.32	0.0013	0.02
22	13	380.1288	3.44	3.3	HMDB034	D-Glutamine	146.0691	380.1275	3.32	0.0013	0.02
23	14	359.1538	3.49	3.34	HMDB018	3_Methylhistamine	125.0953	359.1536	3.27	0.0002	0.07
24	15	409.1551	3.65	3.5	HMDB009	Citrulline	175.0957	409.154	3.74	0.0011	0.24
25	16	399.1049	3.83	3.68	HMDB020	Methionine Sulfoxide	165.046	399.1043	3.72	0.0006	0.04
26	17	307.1223	3.93	3.78	HMDB015	Methylguanidine	73.064	307.1223	3.84	0	0.06

Figure 6. Exported CSV search result.

3. An example of performing M-RT and MS/MS search.

1). For the MS/MS search, the user inputs a precursor mass (581.1214) and select the ion type as [M+H]. Also, a MS/MS list needs to be uploaded. The MS/MS tolerance is defined at a default of 0.005 Da. The match precursor ion and match retention time functions are all turned off, which means the MS/MS search is based on the match of MS/MS fragments with the MS/MS standards. After all the parameters are set, click "Submit Query" to start the MS/MS search.

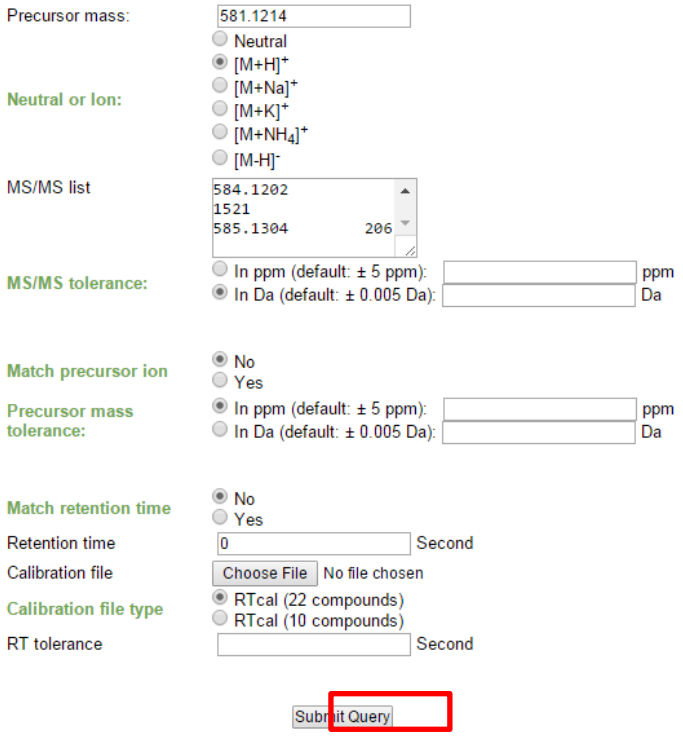


Figure 7. MS/MS search parameters.

2). The MS/MS search result is shown in Figure 8.

Search Result

#	Input mass	Input RT	Calibrated RT	HMDB No.	Name	Monoisotopic molecular mass	mz_light	Library RT	Mass error	RT error	HMDB link	KEGG link	MS/MS score ▲	Show detail
7	581.1214	NA	NA	HMDB00045	Adenosine monophosphate	347.0631	581.1214	1.75	0.0000	NA	Link	Link	1.00	Detail
160	581.1214	NA	NA	HMDB01341	ADP	427.0294	661.0877	1.49	79.9663	NA	Link	Link	0.97	Detail
146	581.1214	NA	NA	HMDB01044	2'-Deoxyguanosine 5'-monophosphate	347.0631	581.1214	5.57	0.0000	NA	Link	Link	0.53	Detail
155	581.1214	NA	NA	HMDB01173	5'-Methylthioadenosine	297.0896	531.1479	6.97	49.9735	NA	Link	Link	0.23	Detail
138	581.1214	NA	NA	HMDB00939	S-Adenosylhomocysteine	384.1216	426.1191	10.52	155.0023	NA	Link	Link	0.22	Detail
300	581.1214	NA	NA	HMDB60003	Isovanillic acid	168.0423	402.1006	15.69	179.0208	NA	Link	Link	0.21	Detail
299	581.1214	NA	NA	HMDB59966	3-5-Dimethoxyphenol	154.0630	388.1213	23.75	193.0001	NA	Link	Link	0.21	Detail
8	581.1214	NA	NA	HMDB00050	Adenosine	267.0968	501.1551	3.94	79.9663	NA	Link	Link	0.20	Detail
280	581.1214	NA	NA	HMDB28937	Leucyl-Proline	228.1474	462.2057	12.99	118.9157	NA	Link	Link	0.18	Detail
269	581.1214	NA	NA	HMDB28691	Alanyl-Leucine	202.1317	436.1901	11.36	144.9313	NA	Link	Link	0.18	Detail
234	581.1214	NA	NA	HMDB02991	Cysteamine	77.0299	311.0882	15.08	270.0332	NA	Link	Link	0.17	Detail

Figure 8. MS/MS search result.

Tutorial for MCID MS/MS Search

Part I. Introduction to MCID MS/MS Search

Part II. Examples of MCID MS/MS Search

- **Part II includes the instructions for file splitting and file merging in batch mode search using a large file of > 100 spectra:**
    - **2.1. Use "MCID-split.R" to split a big MS/MS data file**
    - **2.3. Use "MCID-merge.R" to combine all the search results**

Tutorial Part I. Introduction to MCID MS/MS Search

1. **Workflow.** The workflow for metabolite identification using MCID MS/MS search is shown in Figure 1. The precursor ion mass and fragment ion masses in an experimental MS/MS spectrum are entered into the program for comparison with the library metabolites and their predicted fragment ions. A match score (fit score) is generated in the search result which can be used to judge the quality of a match.

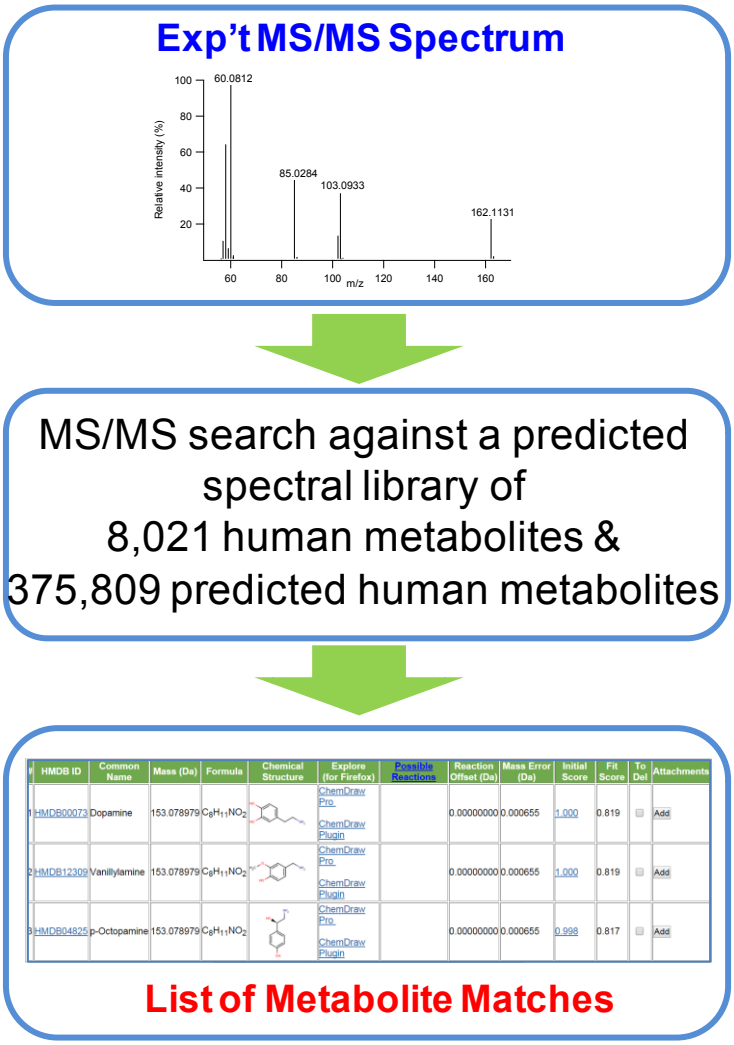


Figure 1. MCID MS/MS search workflow.

2. **MCID spectral library for MS/MS search.** The MCID database is composed of all the known endogenous human metabolites in the Human Metabolome Database (HMDB) (8,021 metabolites) and their predicted metabolic products via one metabolic reaction in the Evidence-based Metabolome Library (EML) (375,809 predicted metabolites). All the predicted MS/MS spectra are generated using *in silico* fragmentation algorithms. This spectral library is hosted at the public MCID website (www.MyCompoundID.org) and allows user to submit single experimental MS/MS spectrum or a batch of MS/MS spectra to search against the library spectra for possible match(s).

3. **MCID single-mode MS/MS search.** The MCID single-mode MS/MS search allows a user to search one experimental MS/MS data against the library spectra. Figure 2 shows the screenshot of MCID single mode MS/MS search interface.

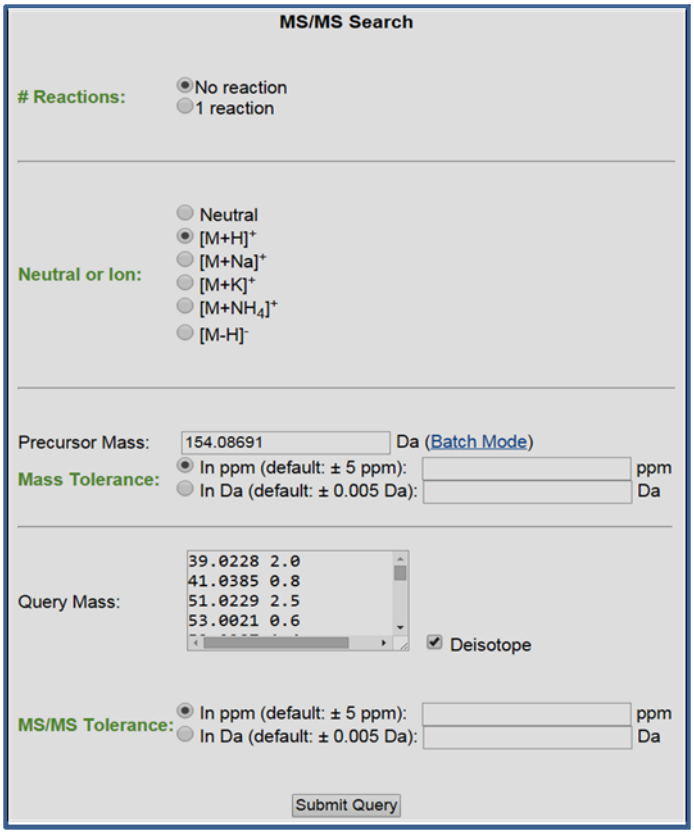


Figure 2. MCID single-mode MS/MS search interface.

a. **# Reaction.** The user needs to choose the type of library, either zero-reaction metabolite library (no reaction) or one-reaction metabolite library (one reaction).

b. **Neutral or Ion.** The user needs to define the type of precursor ion.

c. **Precursor Mass.** The user needs to input a precursor mass.

d. **Mass Tolerance.** The user needs to define a mass tolerance for the precursor mass. 0.005 Da is normally used for MS/MS data collected using high resolution MS such as TOF and FT. If the experiment is performed using a low resolution or low mass-accuracy MS instrument, a larger mass tolerance should be considered.

e. **Query Mass.** The user needs to input the list of MS/MS peaks with their intensities in this box. Once the "Deisotope" checkbox is checked, natural isotopic peaks will be excluded from the matching with the library MS/MS spectra to avoid false matching.

f. **MS/MS Tolerance.** The user needs to define a mass tolerance for the fragment ion peaks. 0.005 Da is normally used for data collected using high resolution MS such as TOF and FT. If the experiment is performed using a low resolution or low mass-accuracy MS instrument such as a triple quadrupole MS, a larger mass tolerance should be considered.

4. **MCID batch-mode MS/MS search.** The MCID batch-mode MS/MS search allows a user to search an entire experimental LC-MS/MS dataset for all the possible matches. Figure 3 shows the screenshot of the MCID batch-mode MS/MS search interface.

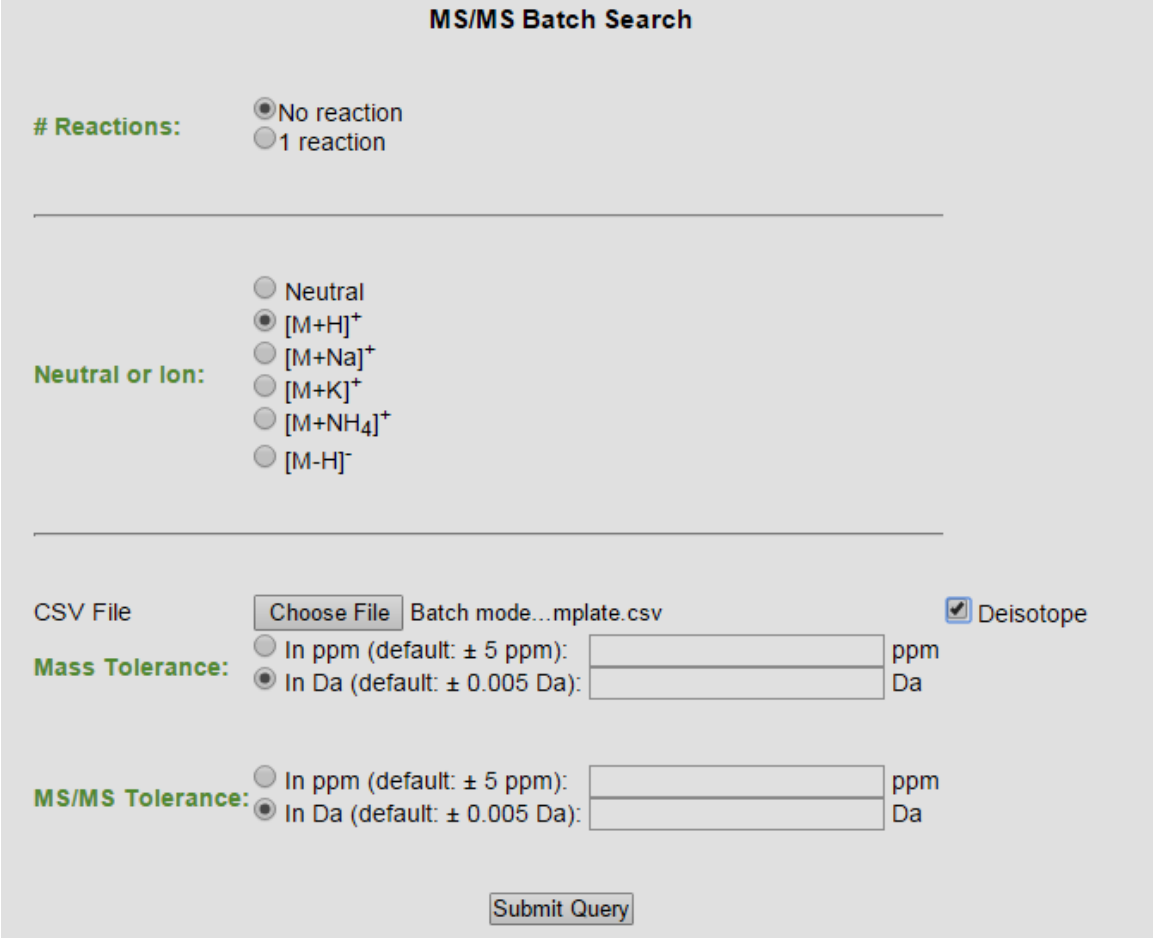


Figure 3. MCID batch-mode MS/MS search interface.

a. # Reactions. The user needs to choose the type of library, either zero-reaction metabolite library (no reaction) or one-reaction metabolite library (one reaction).

b. Neutral or Ion. The user needs to define the type of precursor ion. Usually $[M+H]^+$ is selected in a typical LC-MS/MS analysis.

c. CSV File. The user needs to upload a CSV file generated from LC-MS/MS analysis of a sample for batch-mode search. An example of the file format used (e.g., MSMS file example) can be downloaded from the website. The file size is limited to 100

spectra. If a large file is used, a file split program can be used to split the large file into several small files for uploading (see Instruction given in Part II, section 2).

d. **Deisotope.** Once the "Deisotope" checkbox is checked, natural isotopic peaks will be excluded from the matching with the library MS/MS spectra to avoid false matching.

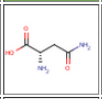
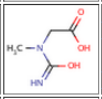
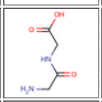
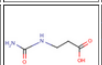
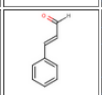
e. **Mass Tolerance.** The user needs to define a mass tolerance for the precursor mass. 0.005 Da is normally used for MS/MS data collected using high resolution MS such as TOF and FT. If the experiment is performed using a low resolution or low mass-accuracy MS instrument, a larger mass tolerance should be considered.

f. **MS/MS Tolerance.** The user needs to define a mass tolerance for the fragment MS peaks. 0.005 Da is normally used for data collected using high resolution MS such as TOF and FT. If the experiment is performed using a low resolution or low mass-accuracy MS instrument, a larger mass tolerance should be considered.

5. **Single-mode search result display.** Figure 4 shows the screenshots of the MCID MS/MS single-mode search results using L-Asparagine as an example. After MS/MS search, all the mass-matched candidates are listed in the result page shown in Figure 4A. The correct structure, L-Asparagine, has the highest fit score (0.984). To further interpret the match result, the user can click the web link in the "Initial Score" column to display another layer of the match result. For example, by clicking "1.000" in Initial Score from L-Asparagine, a new page is displayed as shown in Figure 4B. This page shows the matching quality of the predicted MS/MS spectrum against the experimental MS/MS spectrum. All the matched peaks are labeled in red and unmatched peaks are in grey. On the same page, all the experimental MS/MS peaks are listed in a table (see Figure 4C). By clicking in the "Detail"

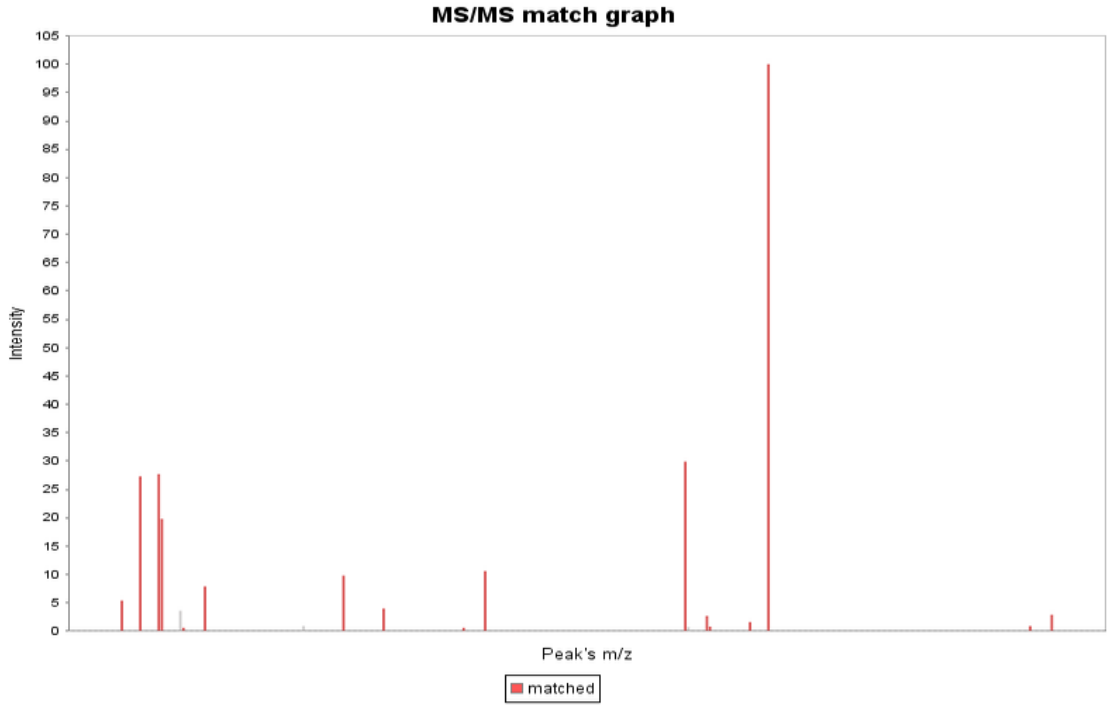
column, another page will be displayed as shown in Figure 4D. On this page, a specific experimental MS/MS peak is matched with a predicted MS/MS peak and the matched structure is displayed. The user can judge whether this matched structure is reasonable or not against the entire metabolite structure.

(A)

#	HMDB ID	Common Name	Mass (Da)	Formula	Chemical Structure	Explore (for Firefox)	Possible Reactions	Reaction Offset (Da)	Mass Error (Da)	Initial Score	Fit Score	To Del	Attachments
1	HMDB00168	L-Asparagine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	1.000	0.984		Add
2	HMDB12265	N-Carbamoylsarcosine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.943	0.928		Add
3	HMDB11733	Glycyl-glycine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.861	0.847		Add
4	HMDB00026	Ureidopropionic acid	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.483	0.475		Add
5	HMDB03441	Cinnamaldehyde	132.057515	C_9H_8O		ChemDraw Pro ChemDraw Plugin		0.00000000	-0.004015	0.062	0.061		Add

(B)

NC(=O)CC(N)C(=O)O
Initial Score= 1.000
Fit Score= 0.984



(C)

Experimental peak	Intensity	Matched simulated peaks	Detail information	Experimental peak	Intensity	Matched simulated peaks	Detail information
42.0337	5.4	1	Detail	43.0177	27.3	2	Detail
44.013	27.7	1	Detail	44.0494	19.8	1	Detail
45.0448	3.6	0		45.0523	0.6	1	Detail
46.0287	7.9	1	Detail	51.0228	0.9	0	
53.0023	9.8	1	Detail	55.0179	4.0	1	Detail
59.037	0.6	2	Detail	60.0446	10.6	1	Detail
70.0291	29.9	1	Detail	70.0656	0.7	0	
71.013	2.7	1	Detail	71.0326	0.8	1	Detail
73.029	1.6	1	Detail	74.0243	100.0	1	Detail
87.0555	0.9	1	Detail	88.0394	2.9	1	Detail

(D)

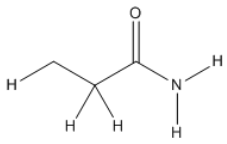
Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
71.0371	-1	70.0287	70.0291	0.0004	

Figure 4. Screenshots of single-mode MS/MS search results.

6. **Batch-mode search result display.** Figure 5 shows a screenshot of the MCID batch-mode MS/MS search result. As displayed at the top of the table, the user can further filter the search results table using precursor mass, intensity, number of fragments, number of hits, and the fit score. Also, the entire search results table can be exported as a CSV file by clicking the "Download Table Result". Figure 6 shows the screenshot of the exported search results. The web link provided at the end of each row allows the user to manually check the matching result from the MCID website. The user merely needs to cut and paste the link name to the internet and the search result displayed for a given match will be the same as a single-spectrum search result. The user can follow the instruction given in Section 5 to interpret the search results.

Min Precursor Mass: Max Precursor Mass:
Min Intensity: Max Intensity:
Min Fragments: Max Fragments:
Min Hits: Max Hits:
Min Fit Score: Max Fit Score:

Show 10 entries Search:

#	Retention Time	Precursor Mass	Precursor Intensity	No. of Fragments	No. of Hits	Max Fit Score	Show Details	Save Result
1	5.24	360.14059	19914	62	1	0.44	Show detail	CSV
2	5.26	346.12504	123448	64	2	0.31	Show detail	CSV
3	5.31	190.01729	29686	52	1	0.22	Show detail	CSV
4	5.34	110.06063	114212	20	1	0.89	Show detail	CSV
5	5.42	150.07856	42852	20	4	0.36	Show detail	CSV
6	5.44	282.12013	478864	84	4	0.98	Show detail	CSV
7	5.50	86.09722	498752	2	0	0.00	Show detail	CSV
8	5.56	197.00612	24810	30	0	0.00	Show detail	CSV
9	5.65	132.10302	99708	22	6	0.99	Show detail	CSV
10	5.66	223.02053	52658	137	0	0.00	Show detail	CSV

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

Download Table Result

Figure 5. Screenshot of batch-mode search results.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Index	Retention	Precursor	Precursor	No of Frag	Max Fit Sc	No of Can	HMDB No	Common	Formula	Mass	Reaction I	Reaction I	Initial Sco	Fit Score	Link
2	1	5.2375	360.1406	19914	62	0.440387	1	HMDB050	Rabeprazc	C18H21N3	359.1304	Zero Reac	0	1	0.440387	http://mci
3	2	5.263317	346.125	123448	64	0.309317	2	HMDB019	Omeprazc	C17H19N3	345.1147	Zero Reac	0	1	0.309317	http://mci
4	2	5.263317	346.125	123448	64	0.309317	2	HMDB050	(S)-Esome	C17H19N3	345.1147	Zero Reac	0	1	0.309317	http://mci
5	3	5.314833	190.0173	29686	52	0.221146	1	HMDB048	Lanthioni	C6H7NO4S	189.0096	Zero Reac	0	1	0.221146	http://mci
6	4	5.34065	110.0606	114212	20	0.890741	1	HMDB011	4-Aminop	C6H7NO	109.0528	Zero Reac	0	1	0.890741	http://mci
7	5	5.418	150.0786	42852	20	0.358484	4	HMDB115	1-Methyla	C6H7N5	149.0701	Zero Reac	0	1	0.358484	http://mci
8	5	5.418	150.0786	42852	20	0.358484	4	HMDB116	7-Methyla	C6H7N5	149.0701	Zero Reac	0	1	0.358484	http://mci
9	5	5.418	150.0786	42852	20	0.358484	4	HMDB116	3-Methyla	C6H7N5	149.0701	Zero Reac	0	0.898588	0.32213	http://mc
10	5	5.418	150.0786	42852	20	0.358484	4	HMDB020	6-Methyla	C6H7N5	149.0701	Zero Reac	0	0.59099	0.211861	http://mci
11	6	5.435267	282.1201	478864	84	0.980416	4	HMDB060	3'-O-Meth	C11H15N5	281.1124	Zero Reac	0	1	0.980416	http://mci
12	6	5.435267	282.1201	478864	84	0.980416	4	HMDB043	2'-O-Meth	C11H15N5	281.1124	Zero Reac	0	0.999502	0.979928	http://mci
13	6	5.435267	282.1201	478864	84	0.980416	4	HMDB033	1-Methyla	C11H15N5	281.1124	Zero Reac	0	0.999349	0.979778	http://mci
14	6	5.435267	282.1201	478864	84	0.980416	4	HMDB040	N6-Methy	C11H15N5	281.1124	Zero Reac	0	0.999101	0.979534	http://mci
15	7	5.504067	86.09722	498752	2	0	0									http://mci
16	8	5.5556	197.0061	24810	30	0	0									http://mci
17	9	5.6502	132.103	99708	22	0.993629	6	HMDB016	L-Norleuc	C6H13NO2	131.0946	Zero Reac	0	1	0.993629	http://mci
18	9	5.6502	132.103	99708	22	0.993629	6	HMDB001	L-Isoleuci	C6H13NO2	131.0946	Zero Reac	0	0.996482	0.990133	http://mci
19	9	5.6502	132.103	99708	22	0.993629	6	HMDB005	L-Alloisol	C6H13NO2	131.0946	Zero Reac	0	0.996482	0.990133	http://mci
20	9	5.6502	132.103	99708	22	0.993629	6	HMDB006	L-Leucine	C6H13NO2	131.0946	Zero Reac	0	0.992514	0.986191	http://mci
21	9	5.6502	132.103	99708	22	0.993629	6	HMDB036	Beta-Leuc	C6H13NO2	131.0946	Zero Reac	0	0.988603	0.982305	http://mci
22	9	5.6502	132.103	99708	22	0.993629	6	HMDB019	Aminocap	C6H13NO2	131.0946	Zero Reac	0	0.985773	0.979493	http://mci
23	10	5.658817	223.0205	52658	137	0	0									http://mci

Figure 6. Screenshot of the exported batch-mode search results.

Tutorial Part II. Examples of MCID MS/MS Search

1. An example of using MCID single-mode MS/MS search

Using L-Asparagine as an example, the MS/MS data are shown below.

Precursor ion(neutral): 132.0535

MS/MS list:

m/z	I %
42.0337	5.4
43.0177	27.3
44.0130	27.7
44.0494	19.8
45.0448	3.6
45.0523	0.6
46.0287	7.9
51.0228	0.9
53.0023	9.8
55.0179	4.0
59.0370	0.6
60.0446	10.6
70.0291	29.9
70.0656	0.7
71.0130	2.7
71.0326	0.8
73.0290	1.6
74.0243	100.0
75.0275	2.5
87.0555	0.9
88.0394	2.9

Referring to Figure 1, the user selects the spectral library as the zero-reaction library (i.e., No reaction), selects the type of precursor mass as Neutral, and enters the precursor mass (132.0535) along with the mass tolerance. In this case, the mass tolerance for the precursor mass is selected as the default (i.e., 0.005 Da). The user then enters the fragment ion masses and their corresponding intensities from the experimental MS/MS spectrum in the Query Mass box. Deisotope is selected as default to remove the ^{13}C-natural abundance peaks accompanied with the fragment ion peaks. The user enters the mass tolerance for the fragment ion masses or selects the default (0.005 Da). The user clicks the "Submit Query" to start the single-mode MS/MS search.

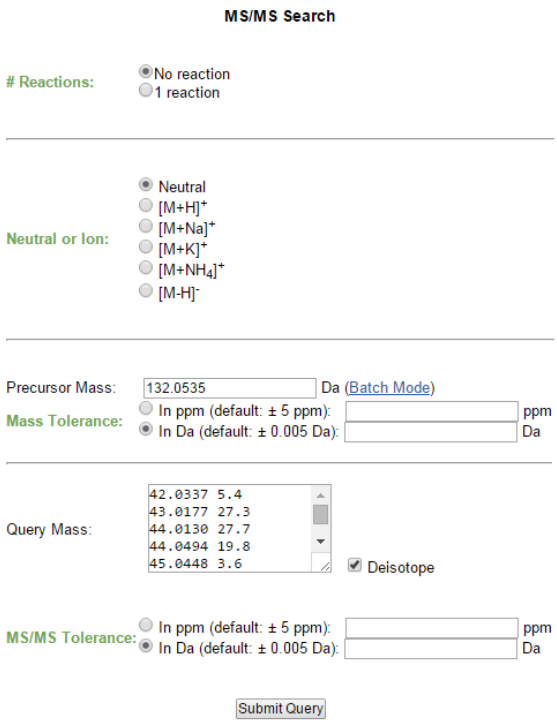


Figure 1. Screenshot of MCID single-mode MS/MS search settings.

The search result is shown in Figure 2A. To help interpret the match, the user can click the web link in the “Initial Score” column to display another layer of the match result. For example, by clicking "1.000" in Initial Score from L-Asparagine, a new page is displayed as shown in Figure 2B. This page shows the match quality of the predicted MS/MS spectrum against the experimental MS/MS spectrum. All the matched peaks are labeled in red and unmatched peaks are in grey. On the same page, all the experimental MS/MS peaks are listed in a table (see Figure 2C). By clicking in the “Detail” column, another page will be displayed as shown in Figure 2D. On this page, a specific experimental MS/MS peak is matched with a predicted MS/MS peak and the matched structure is displayed. The user can judge whether this matched structure is reasonable or not against the entire metabolite structure. The user can also follow the instruction given in Part I for more information on how to interpret the search results.

(A)

Search Result

Input Parameter Name	Parameter Value(s)
# Reactions	*0*
Ion Type	*Neutral*
Query Mass	*132.053500 Da*
Neutral Mass	*132.053500 Da*
Mass Tolerance	*0.005 Da*

Export as CSV | Check All | Uncheck All | Delete Selected Entries | Save Attachments

#	HMDB ID	Common Name	Mass (Da)	Formula	Chemical Structure	Explore (for Firefox)	Possible Reactions	Reaction Offset (Da)	Mass Error (Da)	Initial Score	Fit Score	To Del	Attachments
1	HMDB00168	L-Asparagine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	1.000	0.984	☐	Add
2	HMDB12265	N-Carbamoylsarcosine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.943	0.928	☐	Add
3	HMDB11733	Glycyl-glycine	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.861	0.847	☐	Add
4	HMDB00026	Ureidopropionic acid	132.053493	$C_4H_8N_2O_3$		ChemDraw Pro ChemDraw Plugin		0.00000000	0.000007	0.483	0.475	☐	Add
5	HMDB03441	Cinnamaldehyde	132.057515	C_9H_8O		ChemDraw Pro ChemDraw Plugin		0.00000000	-0.004015	0.062	0.061	☐	Add

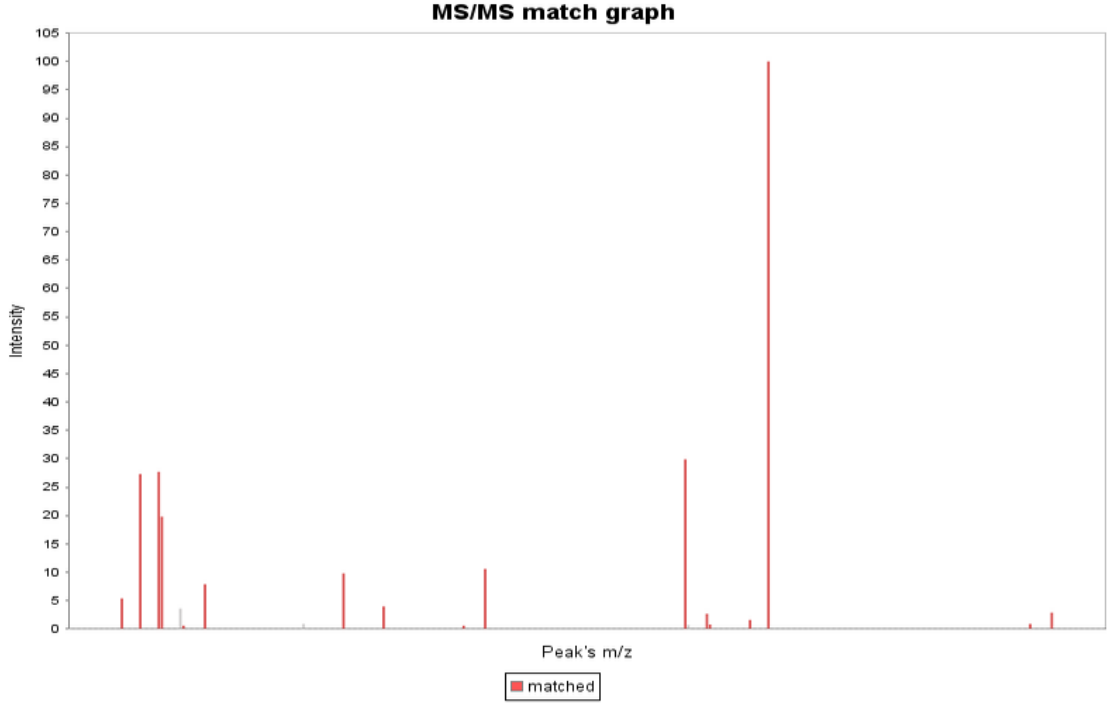
Export as CSV | Check All | Uncheck All | Delete Selected Entries | Save Attachments

(B)

NC(=O)CC(N)C(=O)O

Initial Score= 1.000

Fit Score= 0.984



(C)

Experimental peak	Intensity	Matched simulated peaks	Detail information	Experimental peak	Intensity	Matched simulated peaks	Detail information
42.0337	5.4	1	Detail	43.0177	27.3	2	Detail
44.013	27.7	1	Detail	44.0494	19.8	1	Detail
45.0448	3.6	0		45.0523	0.6	1	Detail
46.0287	7.9	1	Detail	51.0228	0.9	0	
53.0023	9.8	1	Detail	55.0179	4.0	1	Detail
59.037	0.6	2	Detail	60.0446	10.6	1	Detail
70.0291	29.9	1	Detail	70.0656	0.7	0	
71.013	2.7	1	Detail	71.0326	0.8	1	Detail
73.029	1.6	1	Detail	74.0243	100.0	1	Detail
87.0555	0.9	1	Detail	88.0394	2.9	1	Detail

(D)

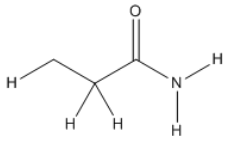
Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
71.0371	-1	70.0287	70.0291	0.0004	

Figure 2. Screenshots of single-mode MS/MS search results.

2. An example of using MCID batch-mode MS/MS search

2.1. Use "MCID-split.R" to split a big MS/MS data file

For the MCID batch-mode MS/MS search, we limit the size of the uploaded batch-mode file to 100 MS/MS spectra so that the server is not occupied for too long by a search work using a very big file. We provide an R based program, "MCID-split.R", for the user to split a big file into smaller files of up to 100 MS/MS spectra in each file. The user can download this program from

the MCID website and the latest R program from https://www.r-project.org/. To run the "MCID-split.R", the user needs to open the R program and assign the fold of "MCID-split.R" as the working folder of RGui by clicking: File → Change dir… (see Figure 3).

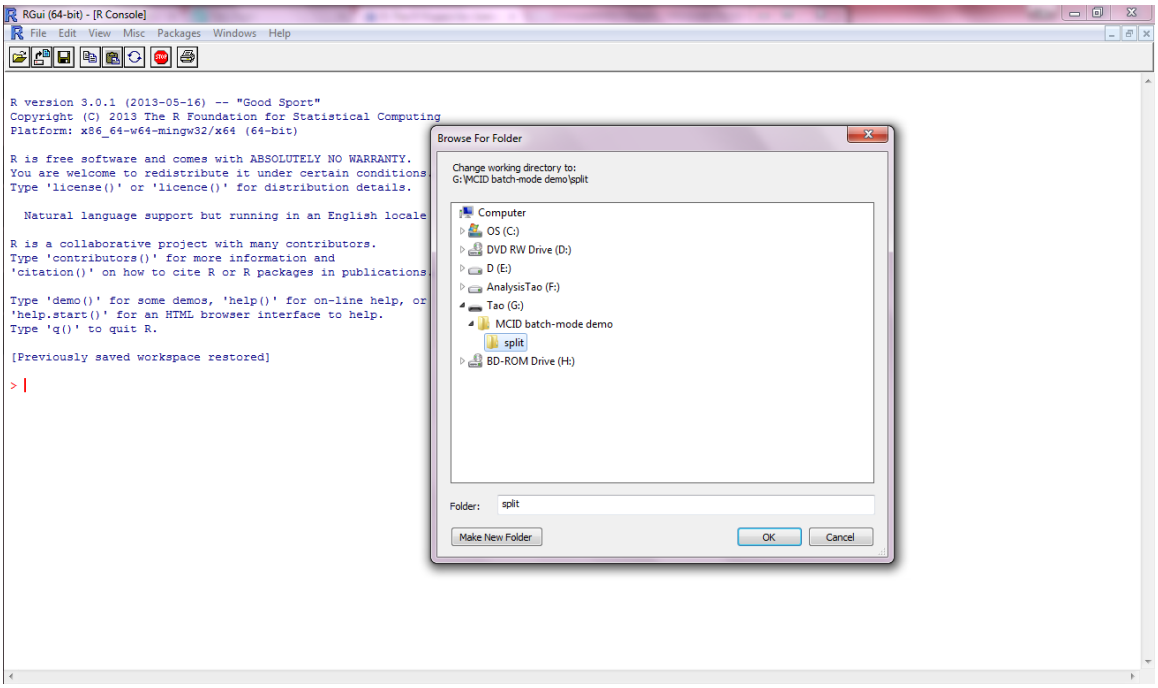


Figure 3. Screenshot of changing the working directory.

Then, the user opens the MCID-split.R script and changes the data path (data.path) (Figure 4) to the folder that contains the big file.

```
#########################################################################
# This is the setting part
data.path <- "G:/MCID batch-mode demo/split/"
file.name <- "run1.csv"
#########################################################################
```

Figure 4. Screenshot of setting the data path.

Next, the user needs to type in "source("MCID-split.R")" into the RGui and press enter to start the splitting process (see Figure 5).

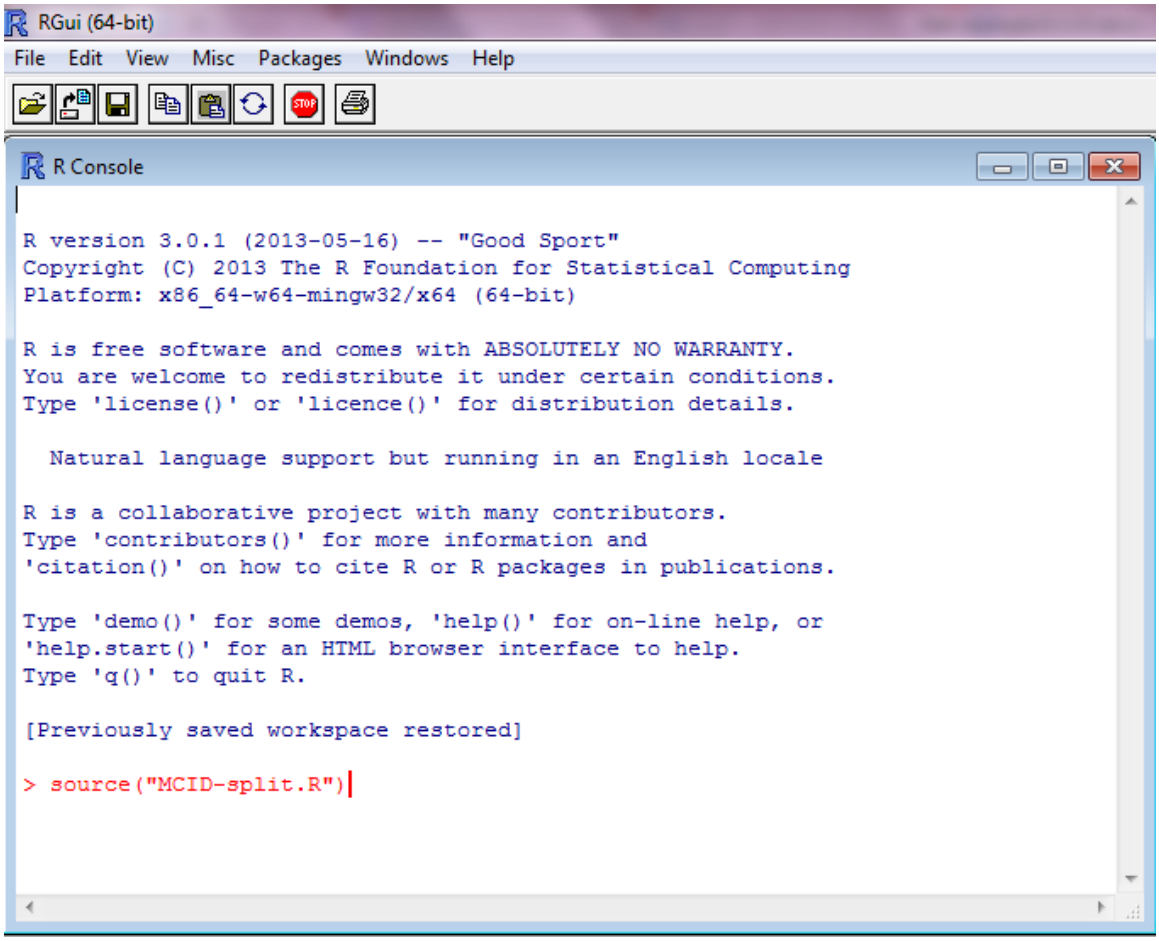


Figure 5. How to run the MCID-split.R.

After running the program, the user can find a list of small files with each containing a maximum of 100 MS/MS spectra (see Figure 6). These files are ready to be used to do batch-mode search online.

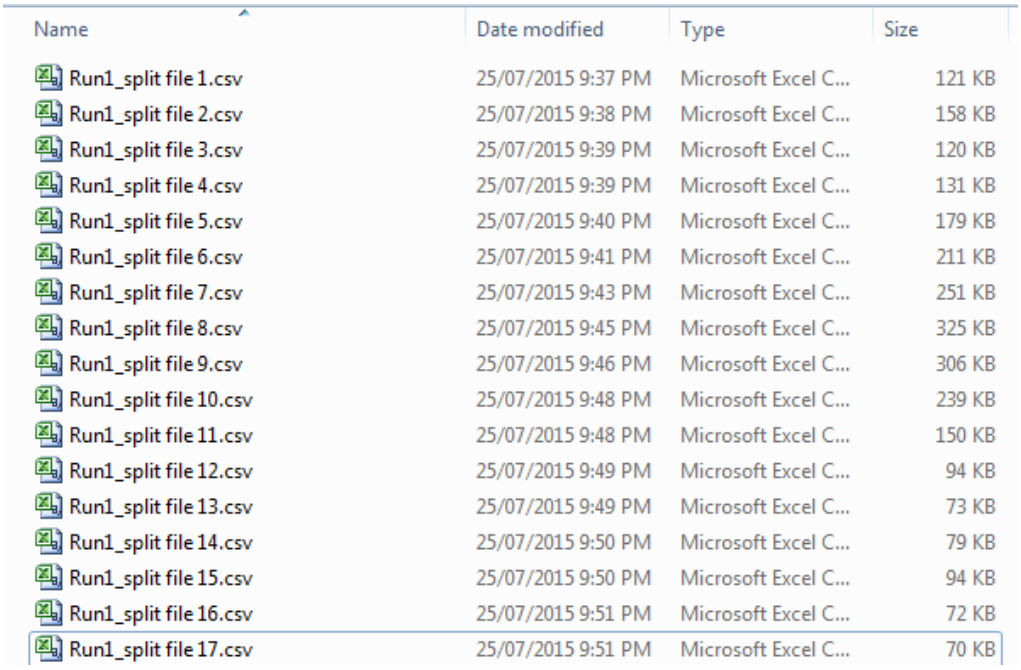


Figure 6. Screenshot of the file splitting results.

2.2. Batch-mode search parameters and results

To perform the batch-mode search, the user needs to define the reaction type (i.e., select the zero-reaction or one-reaction library), precursor ion type, precursor MS tolerance as well as MS/MS tolerance. Then, click the "Submit Query" to start the batch-mode search (see Figure 7). It takes about 2 min to complete a batch mode search with 100 MS/MS spectra using a precursor ion mass tolerance 0.005 Da. However, this search time may be longer if the server is busy to process many queries from multiple users.

Figure 8 shows a screenshot of the MCID batch-mode MS/MS search result. The user can follow the instructions in Part I to interpret the search results. As displayed at the top of the table, the user can further filter the search results table using precursor mass, intensity, number of fragments, number of hits (i.e., mass-matched candidates), and the fit score. Also, the entire search results table can be exported as a CSV file by clicking the "Download Table Result".

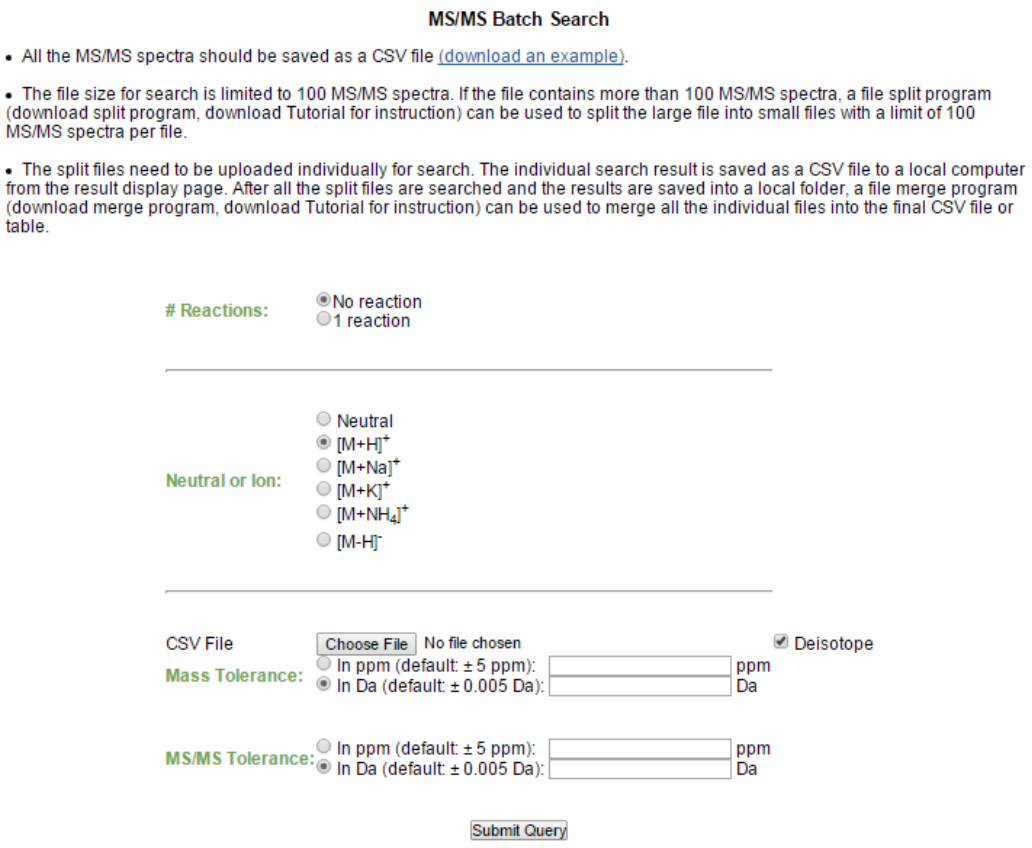


Figure 7. MCID batch-mode MS/MS search settings.

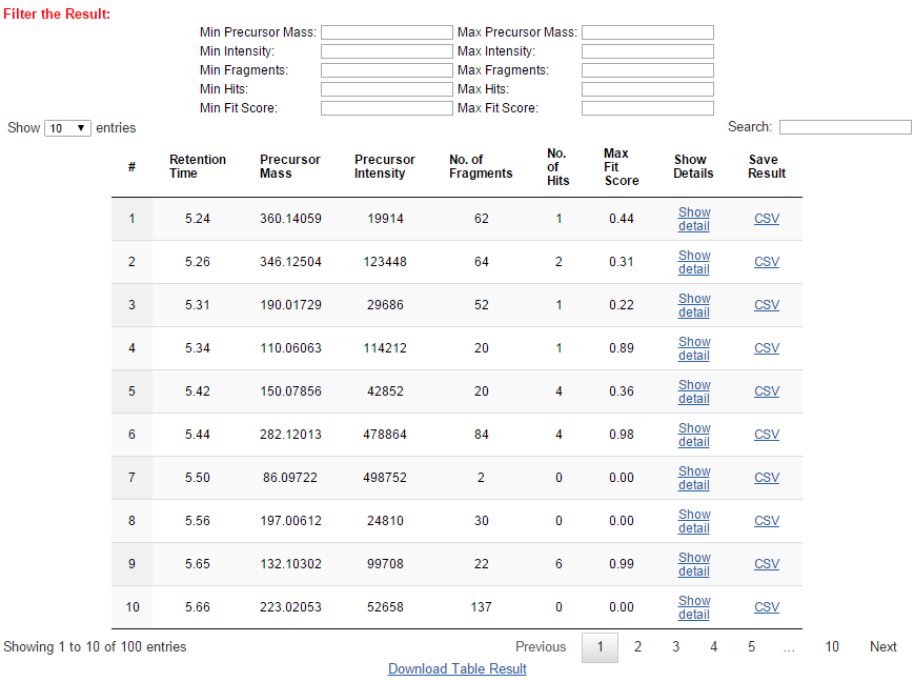


Figure 8. Screenshot of batch-mode MS/MS search results.

2.3. Use "MCID-merge.R" to combine all the search results

After all the search results in CSV are downloaded, another R program "MCID-merge.R" is used to combine all the individual search results files into one complete final results CSV table. To do so, similar to the use of "MCI-split.R", the user needs to open the RGui and assign the fold of "MCID-merge.R" as the working folder of RGui by clicking: File → Change dir… (see Figure 9).

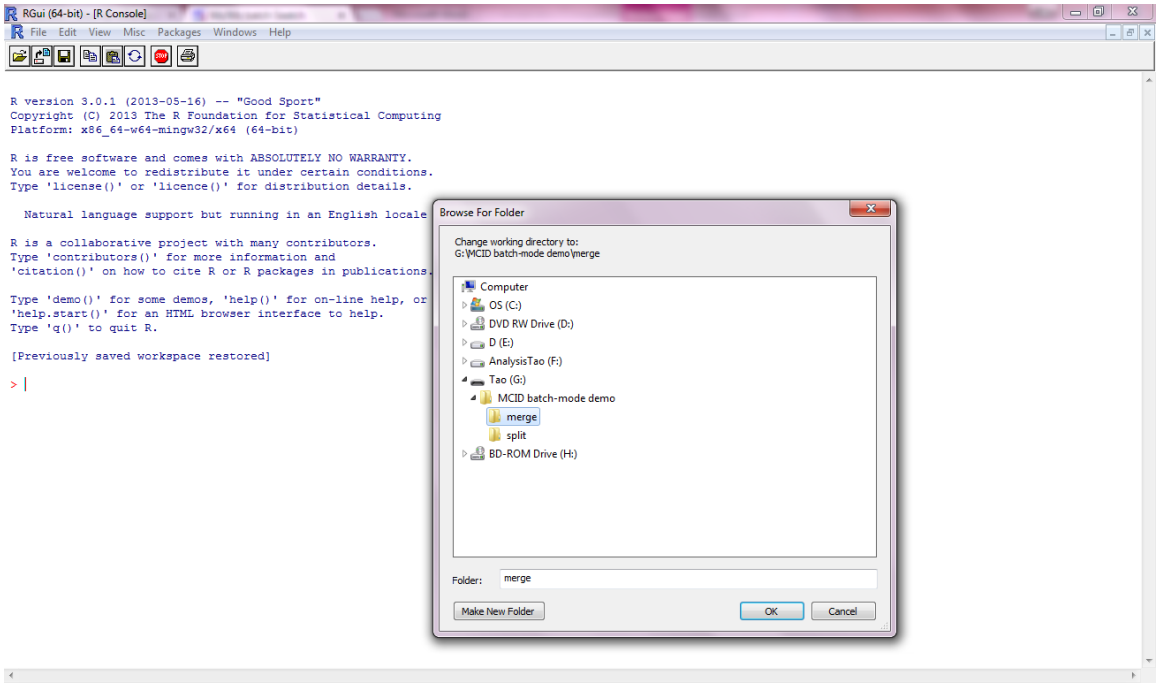


Figure 9. Screenshot of changing work directory.

Then, the user opens the MCID-merge.R script and changes the data path (data.path) (Figure 10) to the folder that contains all the search results files.

```
#########################################################################
# This is the setting part
data.path <- "G:/MCID batch-mode demo/merge/"
#########################################################################
```

Figure 10. Screenshot of data.path setting.

Next, the user needs to type in "source("MCID-merge.R")" into the RGui and press enter to start merging all the results files together (see Figure 11).

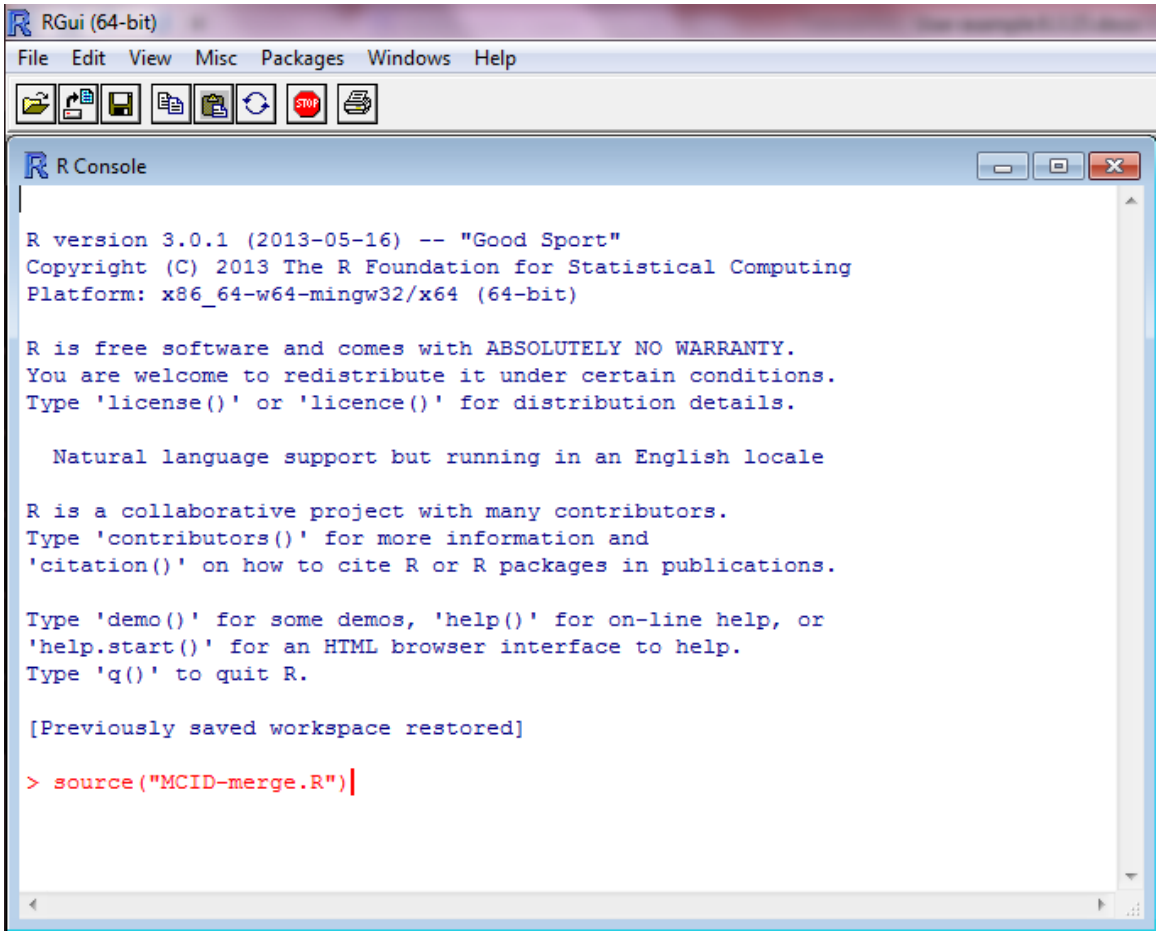
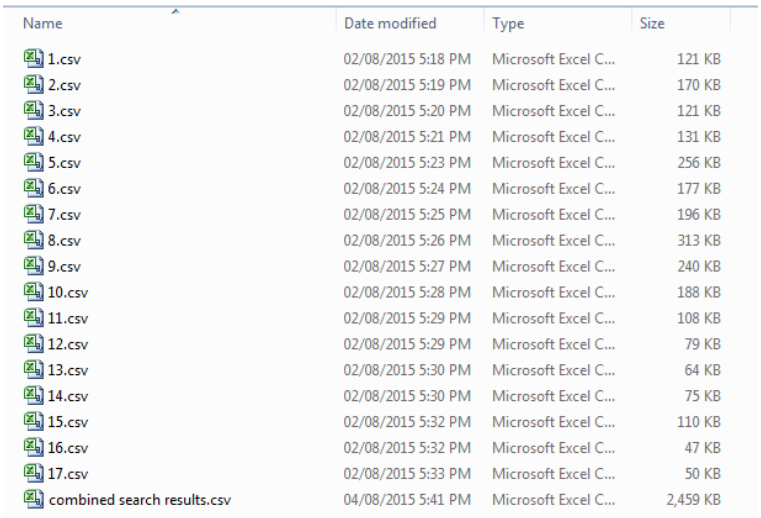


Figure 11. How to run the MCID-merge.R.

After the process is completed, a "combined search results.csv" file will be created (see Figure 12) and this file contains all the information from all the individual searches.



Name	Date modified	Type	Size
1.csv	02/08/2015 5:18 PM	Microsoft Excel C...	121 KB
2.csv	02/08/2015 5:19 PM	Microsoft Excel C...	170 KB
3.csv	02/08/2015 5:20 PM	Microsoft Excel C...	121 KB
4.csv	02/08/2015 5:21 PM	Microsoft Excel C...	131 KB
5.csv	02/08/2015 5:23 PM	Microsoft Excel C...	256 KB
6.csv	02/08/2015 5:24 PM	Microsoft Excel C...	177 KB
7.csv	02/08/2015 5:25 PM	Microsoft Excel C...	196 KB
8.csv	02/08/2015 5:26 PM	Microsoft Excel C...	313 KB
9.csv	02/08/2015 5:27 PM	Microsoft Excel C...	240 KB
10.csv	02/08/2015 5:28 PM	Microsoft Excel C...	188 KB
11.csv	02/08/2015 5:29 PM	Microsoft Excel C...	108 KB
12.csv	02/08/2015 5:29 PM	Microsoft Excel C...	79 KB
13.csv	02/08/2015 5:30 PM	Microsoft Excel C...	64 KB
14.csv	02/08/2015 5:30 PM	Microsoft Excel C...	75 KB
15.csv	02/08/2015 5:32 PM	Microsoft Excel C...	110 KB
16.csv	02/08/2015 5:32 PM	Microsoft Excel C...	47 KB
17.csv	02/08/2015 5:33 PM	Microsoft Excel C...	50 KB
combined search results.csv	04/08/2015 5:41 PM	Microsoft Excel C...	2,459 KB

Figure 12. Screenshot of the merged result.

When the "combined search results.csv" file is opened, all the information about the search results are shown (see Figure 13). The web link provided at the end of each row allows the user to manually check an individual match result from the MCID website. The user merely needs to cut and paste the link name to the internet and the search result displayed for a given match will be the same as a single-spectrum search result. The user can then follow the instruction given in Part I to interpret the search results.

Index	Retention	Precursor.	Precursor.	No.of.Fra	Max.Fit.Sc	No.of.Can	HMDB.No	Common.	Formula	Mass	Reaction.I	Reaction.I	Initial.Sco	Fit.Score	Link
1	0.035942	141.9587	19684	17	0	0				NA		NA	NA	NA	http://mcid
2	0.0445	158.003	9094	25	0	0				NA		NA	NA	NA	http://mcid
3	0.053058	159.9697	8954	29	0	0				NA		NA	NA	NA	http://mcid
4	0.311	106.9923	3560	1	0	0				NA		NA	NA	NA	http://mcid
5	0.457117	122.097	4338	23	0.645035	3	HMDB010	N-N-Dime	C8H11N	121.0891	Zero Reac	0	1	0.645035	http://mcid
5	0.457117	122.097	4338	23	0.645035	3	HMDB020	1-Phenyle	C8H11N	121.0891	Zero Reac	0	1	0.645035	http://mcid
5	0.457117	122.097	4338	23	0.645035	3	HMDB122	Phenyleth	C8H11N	121.0891	Zero Reac	0	0.957412	0.617564	http://mcid
6	0.525883	113.9648	3556	5	0	0				NA		NA	NA	NA	http://mcid
7	0.534442	158.0033	9094	16	0	0				NA		NA	NA	NA	http://mcid
8	0.534442	253.0928	2624	29	0.278611	1	HMDB000	Deoxyino	C10H12N4	252.0859	Zero Reac	0	1	0.278611	http://mcid
9	0.637575	97.96898	4714	1	0	0				NA		NA	NA	NA	http://mcid
10	0.671958	106.9928	3122	2	0	0				NA		NA	NA	NA	http://mcid
11	0.809483	338.3417	1018	4	0	0				NA		NA	NA	NA	http://mcid
12	0.8955	158.0029	9312	20	0	0				NA		NA	NA	NA	http://mcid
13	0.904058	141.9589	11380	11	0	0				NA		NA	NA	NA	http://mcid
14	0.947008	176.0147	2304	35	0	0				NA		NA	NA	NA	http://mcid
15	1.067383	97.96921	4714	1	0	0				NA		NA	NA	NA	http://mcid
16	1.075975	130.0089	2724	9	0	0				NA		NA	NA	NA	http://mcid
17	1.084533	141.9593	1610	8	0	0				NA		NA	NA	NA	http://mcid
18	1.2394	158.0032	9312	13	0	0				NA		NA	NA	NA	http://mcid
19	1.247958	90.94944	2490	2	0	0				NA		NA	NA	NA	http://mcid
20	1.256525	122.0978	4392	20	0.514689	3	HMDB010	N-N-Dime	C8H11N	121.0891	Zero Reac	0	1	0.514689	http://mcid
20	1.256525	122.0978	4392	20	0.514689	3	HMDB020	1-Phenyle	C8H11N	121.0891	Zero Reac	0	1	0.514689	http://mcid
20	1.256525	122.0978	4392	20	0.514689	3	HMDB122	Phenyleth	C8H11N	121.0891	Zero Reac	0	1	0.514689	http://mcid

Figure 13. Screenshot of the exported batch-mode search results.