Computational Analysis of LC-MS/MS Data for Metabolite Identification

Bin Zhou

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

In

Electrical Engineering

Yue J. Wang (Chair)

Luiz A. DaSilva

Chang-Tien Lu

November 30th, 2011
Arlington, VA

Keywords: Liquid chromatography-mass spectrometry, Metabolomics, Spectral matching, Outlier screening, Support vector machine

Computational Analysis of LC-MS/MS Data for Metabolite Identification

Bin Zhou

ABSTRACT

Metabolomics aims at the detection and quantitation of metabolites within a biological system. As the most direct representation of phenotypic changes, metabolomics is an important component in system biology research. Recent development on high-resolution, high-accuracy mass spectrometers enables the simultaneous study of hundreds or even thousands of metabolites in one experiment. Liquid chromatography-mass spectrometry (LC-MS) is a commonly used instrument for metabolomic studies due to its high sensitivity and broad coverage of metabolome.

However, the identification of metabolites remains a bottle-neck for current metabolomic studies. This thesis focuses on utilizing computational approaches to improve the accuracy and efficiency for metabolite identification in LC-MS/MS-based metabolomic studies. First, an outlier screening approach is developed to identify those LC-MS runs with low analytical quality, so they will not adversely affect the identification of metabolites. The approach is computationally simple but effective, and does not depend on any preprocessing approach. Second, an integrated computational framework is proposed and implemented to improve the accuracy of metabolite identification and prioritize the multiple putative identifications of one peak in LC-MS data. Through the framework, peaks are likely to have the m/z values that can give appropriate putative identifications. And important guidance for the metabolite verification is provided by prioritizing the putative identifications. Third, an MS/MS

spectral matching algorithm is proposed based on support vector machine classification. The approach provides an improved retrieval performance in spectral matching, especially in the presence of data heterogeneity due to different instruments or experimental settings used during the MS/MS spectra acquisition.

Acknowledgements

First of all, I dedicate this thesis to my parents for their support and encouragement all these years through this long and sometimes difficult education journey.

Thanks for the academic guidance and financial support provided by my advisors Dr. Habtom W. Ressom and Dr. Yue Wang. Especially, thank Dr. Ressom's kindness, inspiration and patience which guide me toward the maturity and sophistication.

Thank Dr. Jianhua Xuan for his help and support during my graduate study.

Thank my colleges at Virginia Tech and Georgetown University with whom I have the privilege to work. Among them, Dr. Junfeng Xiao and Dr. Amrita Cheema from Georgetown University provide most of the data used in this thesis. Also, I want to thank Dr. Yuanjian Feng, Dr. Chen Wang, Dr. Li Chen, Tsung-Heng Tsai, and Mohammad R Nezami Ranjbar from Virginia Tech for many helpful discussions and advices.

Finally, I want to express my appreciation to all my friends for their help during the past years.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Since 2000, a significant trend in bioscience research is the systems biology paradigm, where a biological system is investigated with a holistic view rather than reductionist consideration of each component. Metabolomics plays an important role in systems biology research as one of the new "omics" approach. It involves the detection, identification and quantitation of small molecules involved in metabolism. With recently development of biotechnology, it is now feasible to acquire metabolomics data with high throughputs using technical platforms such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS).

Liquid chromatography-mass spectrometry (LC-MS), which couples MS detection with chromatography separation, has gained popularity for metabolomic analyses due to its sensitivity, accuracy and coverage of metabolome. By analyzing spectral features from LC-MS data, metabolites can be detected and their abundances in samples can be quantitated, relatively or absolutely. However, metabolite identification presents as a major gap between the LC-MS signals and metabolomic knowledge. While LC-MS signals provide information regarding the mass-to-charge ratio (m/z) and the retention time of the detected ions, they must be mapped to metabolites to understand the biochemical processes they describe. Lack of efficient and accurate metabolite identification is the current bottleneck in LC-MS based metabolomic studies.

In this thesis, a computational framework for LC-MS data analysis is presented, with focus on metabolite identification. Before metabolite identification, LC-MS data are preprocessed to remove outliers, detect peaks corresponding to metabolites and align peaks across multiple samples. The peak list obtained from preprocessing is subjected to statistical analysis to select the peaks of interest. Then, the tandem MS (MS/MS) data of these peaks are acquired. Integrating LC-MS data and MS/MS data, the

proposed framework uses multiple computational approaches to deduce and prioritize putative identifications for the selected metabolites, including ion annotation, mass-based search, isotopic pattern analysis, spectral interpretation and spectral matching. Finally a support vector machine (SVM)-based approach is proposed to improve the retrieval performance in spectral matching, especially when data are heterogeneous from difference instruments. It was demonstrated that the proposed framework and algorithms can improve the efficiency, accuracy and coverage for metabolite identification, which enable LC-MS users to obtain more accurate identification for a larger number peaks in LC-MS based metabolomics with less experiments and costs comparing to the traditional approach.

1.2 Background

Metabolism is the set of chemical reactions which happens in a biological organism. These chemical reactions are usually divided into two categories. Catabolism breaks down large molecules to harvest energy and provide building blocks to synthesize organic compounds. Anabolism uses the energy and building blocks from catabolism to synthesize protein, lipids, nucleotides and etc. The substrates, intermediates and products involved in metabolism are metabolites. Common types of metabolites include amino acids, carbohydrates, nucleic acids, and lipids. By organizing the chemical reactions within metabolism into metabolic pathways, one metabolite is sequentially transformed into another metabolite or its polymers through a chain of reactions, with the catalysis of enzymes. Metabolism plays an important role in supporting the normal function of a biological system by allowing the growth and reproduction of the organism, maintaining its structure and adapting to ever-changing environment.

Metabolic analysis, the study of metabolite composition and abundance in a biological organism, can be dated back to ancient China, where doctors used ants to detect high glucose level in human urine, and hence diagnose diabetes [1]. The concept that individuals might have a “metabolic profile” that reflected in the makeup of their biological fluids was introduced by Roger Williams in the late 1940s who used paper chromatography to identify characteristic metabolic patterns in urine and saliva, which were

associated with diseases such as schizophrenia [2]. However, these methods are largely qualitative and can detect only a few metabolites in one experiment.

As bioscience research is moving towards systems biology paradigm in the past decade, the study of metabolism receives increasing attention in recent years. When a biological system is perturbed by environment stimuli, its response can be reflected by alterations in multiple levels: gene sequence, the transcription of genes, the expression and post-translational modification of proteins, and the composition and abundance of metabolites. Focusing on the interactions within and between these levels, systems biology adopts a holistic approach mainly through the analysis of high-throughput "omics" data. As a result, metabolomics, which is the detection, quantitation and identification of complete set of metabolites (metabolome) in a biological system, joins the "omics" cascade and becomes an important component of systems biology research. The specific advantage of metabolomics is that it provides the closest and most direct description of the phenotype of the biological system, as shown in Figure 1. By analyzing hundreds or thousands of metabolites simultaneously, we can acquire a snapshot of what is exactly happening in the biological system. At present, metabolomic investigations have been applied in various research areas including environmental and biological stress studies, functional genomics, biomarker discovery, and integrative systems biology [3-5]. Those studies facilitate understandings of biochemical fluxes and discoveries of metabolites which are indicative of unusual biological or environmental perturbations.

A related term to "metabolomics", "metabonomics" was coined by Nicholson [6] to represent studies of changes in metabolic activities in response to patho-physiological stimuli or genetic modifications. Both terms overlap by a large degree in practice especially within the field of human disease research and they are often in effect synonymous [7].

In practice, metabolomics generally involves the analysis of all small molecules in a biological sample with molecular weights less than 1,800 Da. This type of analyses is usually performed using either nuclear magnetic resonance (NMR) or mass spectrometry (MS). Comparing to NMR, MS detection has a higher sensitivity, usually at pg level. As a result, it is widely used in measuring metabolites in complex biological samples.

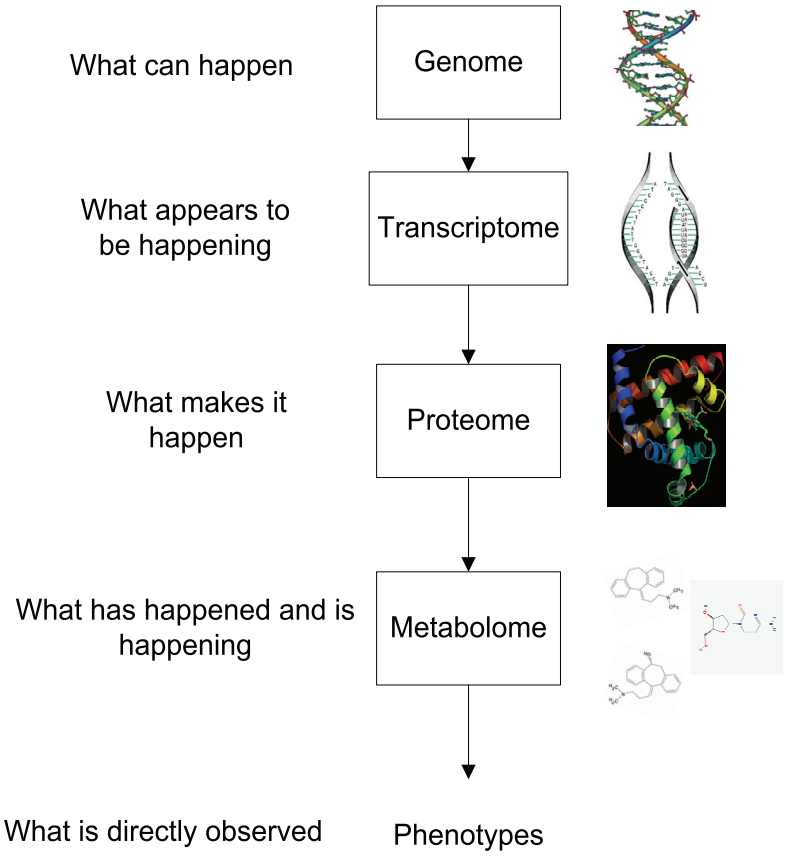


Figure 1. The role of metabolomics in the "omics cascade" [8]

1.3 LC-MS based metabolomics

Mass spectrometry is often hyphenated with chromatography separation in metabolomic studies to provide another dimension of sample separation and improve signal-to-noise ratio. LC-MS is one of the common configurations, which performs liquid chromatography separation followed by mass spectrometry detection. Comparing to gas chromatography-mass spectrometry (GC-MS), LC-MS can analyze non-volatile metabolites without derivatization, thus offers a good coverage of the metabolome.

A schematic diagram of LC-MS instruments is shown in Figure 2. After a sample is injected into the LC-MS instrument, it is first subjected to liquid chromatography separation. During the separation, the sample is dissolved into a liquid fluid called the "mobile phase". The sample-carrying solution is then forced by a high pressure to go through a column that is packed with the "stationary phase" composed of

small particles, a porous monolithic layer, or a porous membrane. Different compounds in the sample solution elute out of the column at different times. The specific time at which a compound elutes out of the column is called its retention time. The retention time of a compound is determined by its interaction strength with stationary phase, and is often subjected to great variation depending on the experimental conditions. However since only a small number of metabolites share the same or similar retention time in an experiment, liquid chromatography reduces the sample complexity and alleviates the background noise in mass spectrometry detection.

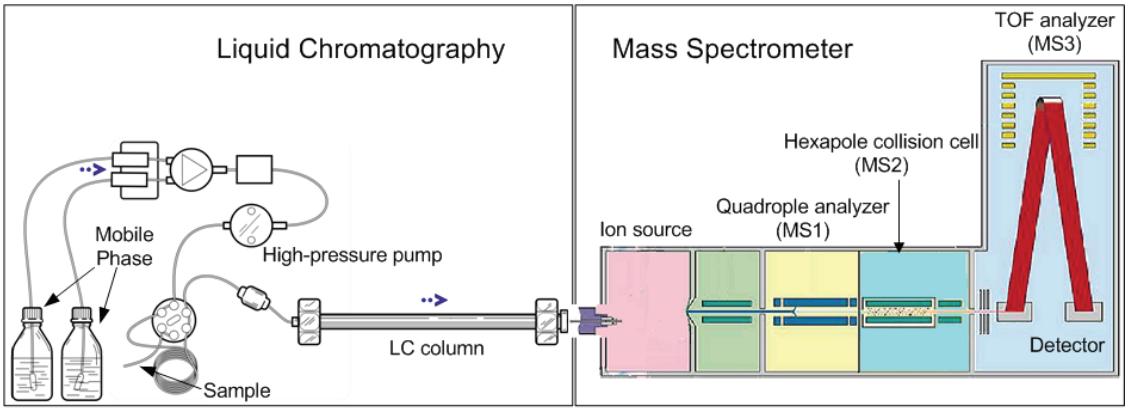


Figure 2. A Schematic diagram of LC-MS instrument

After eluting from the liquid chromatography, compounds are injected into the mass spectrometer. Depending on the specific technology used to implement a mass spectrometer, there are different types of mass spectrometers. However, they all consist of three basic modules: ion source, mass analyzer, and detector. The ion source converts electrically neural compounds in the sample into charged molecular ions. This conversion can be achieved through different techniques such as electrospray ionization (ESI), atmospheric pressure chemical ionization, atmospheric pressure photoionization, fast atom bombardment (FAB), and etc. Among them, ESI is by far the method of choice in most LC–MS-based metabolomic studies. It can ionize a wide range of metabolites with various molecular weight and compound polarity. Particular, ESI is a “soft ionization” approach which generally forms intact molecular ions, thus aids initial identification of metabolites [9]. Due to the diverse chemical properties of metabolites, it is often

required to analyze the biological sample in both +ve (positive) and –ve (negative) ionization modes to maximize metabolome coverage [10].

Mass analyzer separates ions according to their m/z values by applying electrical or magnetic fields on them. Commonly used mass analyzers include quadrupole, ion trap, time-of-flight (TOF), Orbitrap and Fourier transform ion cyclotron (FT-ICR). Although their resolution and accuracy can differ significantly in practice, their working principles are the same. Based on Lorentz force law and Newton's second law, the motion of charged a particle in an electro-magnetic filed in vacuum is governed by:

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1.1}$$

$$\mathbf{F} = m\mathbf{a} \tag{1.2}$$

where **F** is the force exerted on the ion in the electro-magnetic field, m and Q are the mass and charge of the ion respectively, **v** and **a** are its velocity and acceleration, **E** and **B** are the electric field and magnetic field. By combining the above two equations, we have

$$(m / z)\mathbf{a} = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1.3}$$

where e is the elementary charge. The Equation (1.3) suggests that the motion of an ion is determined by its mass-to-charge ratio *m/z*, given the electrical field, magnetic field and the initial condition of the ion. Through the manipulation of electric field and magnetic field, mass analyzers separate ions with different m/z values. Modern high resolution mass spectrometers can achieve the m/z separation with a resolution over 1,000,000 while keep measurement accuracy under 1 ppm (parts-per-million). In addition, more than one mass analyzer are often used in modern mass spectrometers to perform tandem mass spectrometry experiment, which is indispensable in metabolite identification as introduced in the next section. Typical examples of this type of hybrid instruments include quadrupole TOF (QTOF), triple-quadrupole (QqQ), quadrupole-ion trap, and etc.

Detector module converts the abundances of the ions from mass analyzer into electrical signals by recording the charge induced or current produced when an ion hits or passes through the detector. For example, microchannel plate (MCP) detector combined with time-to-digital (TDC) convertor is

commonly used for TOF analyzer. When an ion hits the channel wall of MCP, it produces an electrical signal which is amplified through a cascade of second emission. The signal is then registered to a small time bin by TDC, which then counts the number of arrival ions at a particular time. From Equation (1.3), the m/z value of the ion can be calculated from its arrival time because the motion of the ion in TOF analyzer is determined by its acceleration **a**, given the initial conditions of the ion. By summing a large number of ion arrival events, TDC generate a mass spectrum which is a function of ion abundance against m/z values.

Because samples are separated by both liquid chromatography and mass spectrometry, the raw data generated from LC-MS-based metabolomic experiment is a 3-dimension signal, as shown in Figure 3. One dimension of the signal is retention time, at which a metabolite elutes from the liquid chromatography. The second dimension records the m/z value of a metabolite. Since metabolite can rarely hold more than two charges, the m/z value is a direct reflection of its molecular weight. The third dimension provides a quantitation of the abundances of metabolites, which enables the quantative study of hundreds of metabolites.

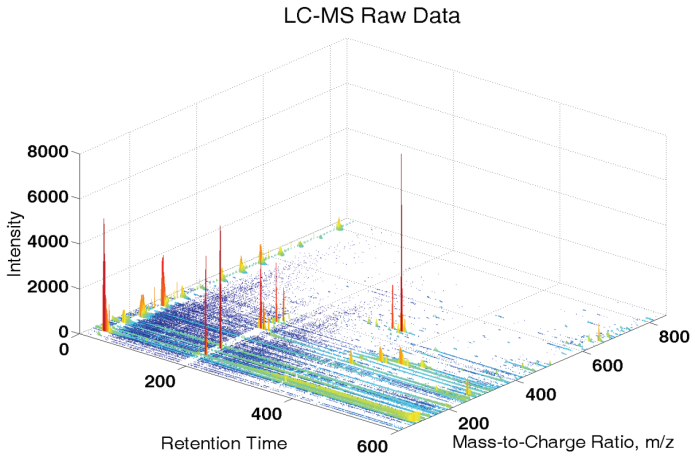


Figure 3 The raw data from a LC-MS-based metabolomic study

However, a gap remains between LC-MS signals and biochemical knowledge of the investigated organism. Although modern LC-MS instrument is able to accurately measure the m/z values, i.e. masses of the detected ions (with an accuracy ranges from 50 ppm to less than 1 ppm), it is still difficult to link

an observed ion to a specific metabolite. Partly it is because that metabolites with different elemental formulas may still share highly similar masses, as demonstrated in [11]. To make things even more difficult, there are also metabolites with exactly the same elemental formula but different structures. One example is provided in Figure 4. 1-Methylguanine, 6-O-Methylguanine and 7-Methylguanine are all of the same elemental formula C6H7N5O. As a result, all three metabolites have the identical mass at 165.07 Da. However, their structures are different from each other. Only LC-MS analysis is not capable of identifying these three metabolites even though it may separate them on the retention time scale. For metabolite identification with less ambiguity and higher confidence, current research often resorts to tandem mass spectrometry which will be discussed below.

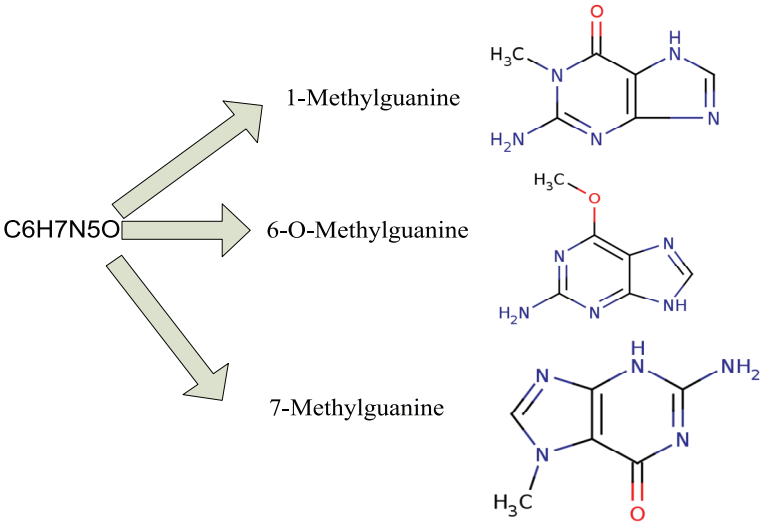


Figure 4 Metabolites of the same elemental formula but different structures

1.4 LC-MS/MS for metabolite identification

Tandem mass spectrometry (MS/MS) performs more than one stage of MS analysis, with molecular fragmentation occurring between stages. This can be achieved through multiple mass analyzers separated in space or a single mass analyzer operated differently in time. Several MS instruments support tandem MS capability, such as QTOF, QqQ and Orbitrap. In this section, the principles of tandem MS are demonstrated using a LC-QTOF instrument, as shown in Figure 2.

In Figure 2, there are two stages of mass analyzers: a quadrupole mass analyzer (MS1) and a TOF mass analyzer (MS3). Between the two stages, there is a hexapole collision cell (MS2). Contrary to LC-MS experiments, in which quadrupole analyzer operates as an ion guide to allow all the ions to pass through regardless of their masses, the quadruople works as a mass filter which transmits only the ions of a specific m/z value in LC-MS/MS experiments. In practice, the quadrupole analyzer usually performs ion filtering with a mass window of 1-3 Da wide. The selected parent ions are then accelerated to energies between 20 to 200 eV and transmitted through the hexapole collision cell. The parent ions collide with the neutral gas molecules (usually argon or nitrogen) which are filled in the collision cell. After the first few collisions, the parent ions are broken into smaller fragments. This process, called collision induced dissociation (CID), generates a series of product (daughter) ions by breaking certain chemical bonds in the molecular ions. Under the same experimental conditions, CID follows a specific fragmentation pathway, thus generates relatively stable species of product ions. The product ions, in addition to the remaining parent ions, are re-accelerated and enter into the TOF analyzer, in which they are separated by their m/z values. The mass spectrum recorded by the detector provides the information about both product ions and the remaining parent ion. It is called MS/MS spectrum to indicate that it is from a tandem MS experiment.

The MS/MS spectrum is of particular importance in metabolite identification as it provides a characteristic fingerprint for different metabolites even if these metabolites have the same elemental formula. An example is shown in Figure 5, where the MS/MS spectra of the three metabolites in Figure 4 are demonstrated. Although the three metabolites are of the same mass, their structure differences are captured by the MS/MS spectra.

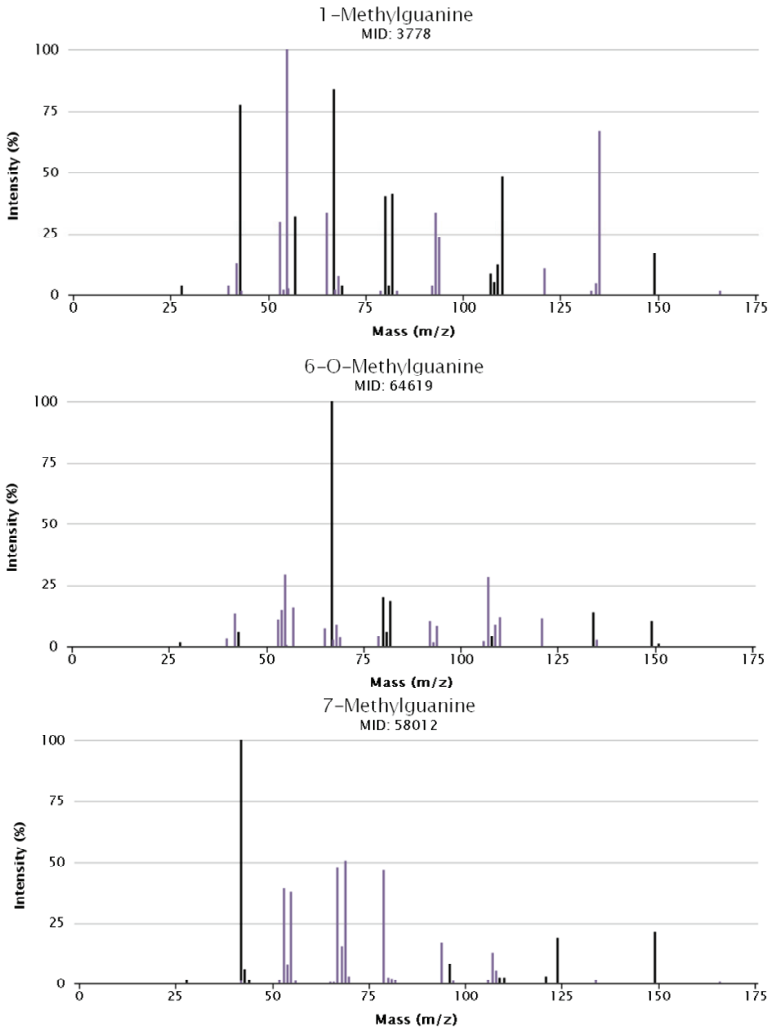


Figure 5. The MS/MS spectra of 1-methylguanine, 6-O-methylguanine and 7-methylguanine

1.5 Summary of contributions

The major contributions of this thesis are summarized below:

1. An integrated computational framework is proposed which increases the accuracy of metabolite identification and effectively prioritizes the multiple putative identifications for each observed peak from LC-MS data.
2. An improved MS/MS spectral matching algorithm is proposed which combines "peak similarity" with "profile similarity" through SVM classification. The proposed algorithm was demonstrated to

increase the F-measure of the metabolite identification system, especially when the spectral data are heterogeneous from different instruments and institutions.

3. We also develop an outlier screening approach based on Xrea quality measure of LC-MS data, which removes sample runs that substantially deviate from the others.

1.6 Organization of the thesis

In Chapter 2, we overview the current approaches for LC-MS/MS data analysis. In this chapter, computational techniques used in LC-MS data preprocessing, difference detection, metabolite identification and metabolite verification are overviewed. We emphasize on the peak detection and peak alignment in LC-MS data preprocessing, as the preprocessing of the raw data will directly affect the following steps in data analysis and determine the set of peaks which needs identification.

Chapter 3 summarizes the major contributions of this thesis. We first briefly discuss the outlier screening in LC-MS data before preprocessing. An outlier screening algorithm based on Xrea quality measure is proposed. Then we present an integrated computational framework for metabolite identification. The presented framework is applied to a set of peaks selected from data preprocessing and difference detection. It demonstrates to be successful in improving the accuracy of metabolite identification and prioritizing the putative identification of metabolites. Last, an improved spectral matching algorithm using SVM is presented. The algorithm combines peak similarity with profile similarity to construct a feature set. SVM is used to improve the accuracy of MS/MS spectral matching, especially when there is a large degree of heterogeneity in the data, e.g. the MS/MS spectra are acquired using different instruments. The improved performance of the algorithm is demonstrated using an in-house dataset in combination with a dataset downloaded from Human Metabolome DataBase (HMDB).

Chapter 4 concludes the thesis and discusses the future computational perspectives of LC-MS/MS-based metabolite identification.

Chapter 2 An Overview of LC-MS/MS Data Analysis[1]

The analysis of LC-MS/MS data generally involves preprocessing, statistical analysis, and metabolite identification. Preprocessing converts the raw LC-MS measurements into a list of detected peaks comparable across multiple samples. Statistical analysis identifies a subset of peaks which reflects phenotypic changes of biological samples. Metabolite identification finds the identity of the observed peaks through a side-by-side comparison of the biological sample and authentic compounds of putative identifications. In this section, the general workflow of a LC-MS/MS data analysis is reviewed with an emphasis on data preprocessing.

2.1 LC-MS data preprocessing

To convert the raw LC-MS data into a peak list which can be easily interpreted and compared across runs, multiple pre-processing steps need to be performed as outlined below:

Outlier Screening aims to eliminate LC-MS runs or peaks which exhibit an unacceptable deviation from the majority of their replicates (analytical or biological). While the LC-MS variability is unavoidable, the outlier runs/peaks with excessive amount of bias need to be removed from the subsequent analysis. In practice, principal component analysis (PCA) is often used to identify sample outliers by visually inspecting the 2-D (or 3-D) score plot of the data. Those runs that deviate significantly from the majority are considered as possible outliers. R package OutlierD uses quantile regression on MA plot to detect outlier peaks from LC-MS/MS data [12]. However, these methods generally depend on other preprocessing steps, such as peak detection and peak alignment. And some samples may be declared as outliers because of poor preprocessing rather than samples themselves. An outlier screening approach

[1] Part of this section has been published in "LC-MS-based metabolomics", Bin Zhou, Jun Feng Xiao, Leepika Tuli and Habtom W. Ressom, Molecular BioSystem, 01 Nov 2011, [Epub ahead of print]

is proposed in the next chapter, which is able to identify low-quality LC-MS runs without resorting to other preprocessing steps.

Filtering is used to remove the noise and contaminants from LC-MS data. One major requirement for filtering is to suppress the noise while preserving the peaks in the data. For example, Savitzky-Golay filter reduces the noise while keeping the peaks by preserving high-frequency components in LC-MS signals. By locally fitting a high-order polynomial function to the observed data, Savitzky-Golay filter is particularly successful in preserving the sharpness of peaks.

Baseline correction estimates the low-frequency baseline, and then subtracts the estimated baseline from the raw signal. Baseline shift is often observed as the baseline of the intensities is elevated with increasing retention time, and the elevated baseline results in an over-estimate of the intensities of those late eluting analytes. A low-order Savitzky-Golay filter can be used to remove the baseline from LC-MS signal [13]. PCA can also be used for baseline correction by first estimating the noise sub-space and then subtracting the projected signal in the noise sub-space from the raw data [14]. An example of baseline correction is shown in Figure 6.

Peak detection is a transformation which converts the raw continuous LC-MS data into centroided discrete data so each ion is represented as a peak. This transformation offers two advantages: (1) part of the noise in the continuous data is removed; (2) data dimension is reduced without much information loss. Peak detection is generally carried out in two steps by first calculating the centroids of peaks over m/z range and then searching across retention time range for chromatographic peaks. Currently, main efforts of peak detection algorithms focus on centroiding over retention time, as many instruments are capable of providing centroided m/z measurements. In this thesis, we used open source software XCMS [15] for peak detection over retention time, in which a matched filtering approach with the second-order derivative of Gaussian function is used for peak detection. An example of peak detection using XCMS is shown in Figure 6.

Peak alignment enables the comparison of LC-MS-based metabolomic data across samples. The retention time of an ion may drift across different samples, even if those samples are replicates. The drift

is generally non-uniform over the retention time range and cannot be completely controlled during experiments. For large-scale studies involving multiple samples, peak alignment is needed to ensure that the same ion is compared across samples. In XCMS, peak alignment is performed in two consecutive steps. First, the detected peaks from different samples are matched based on their similarity in m/z and retention time. The peaks within a small m/z interval across different samples are grouped using a kernel density estimator. After grouping, “well-behaved” peak groups to which very few samples have no peak assigned or have more than one peaks assigned are used as landmarks for alignment. The deviations of the retention times of these landmarks from the median values of the corresponding peak groups are regressed against the retention time. Those regions on chromatogram without “well-behaved” peaks can then be interpolated and used for retention time correction. The aligned peaks are grouped again to match peaks with corrected retention times. This procedure is usually carried out iteratively for two or three times to make sure the retention time drift is sufficiently corrected for. A critical assessment of several popular alignment methods concludes that XCMS gives the best performance for alignment of LC-MS metabolomics data [16]. An example of a set of EIC peaks before and after alignment is shown in Figure 6.

Normalization of peak intensities helps to reduce the systematic variation of LC-MS data. One way to normalize LC-MS data is to add same amount of internal standards into all the samples. Relative abundance is calculated by adjusting other ions' intensities based on their ratios to internal standards. For normalization techniques without using internal standard, normalization to osmolality and normalization to "total useful MS signal" are recommended. [17].

Transformation of LC-MS data is sometimes needed to modify the data distribution so it is more suitable for subsequent statistical analysis. For example, transformations which lead to a more normal-distributed dataset or compress the dynamic range of data are often used. These transformations are usually heuristic. Z-score, log-transformation, and square-root transformation of peak intensities are common choices of transformations.

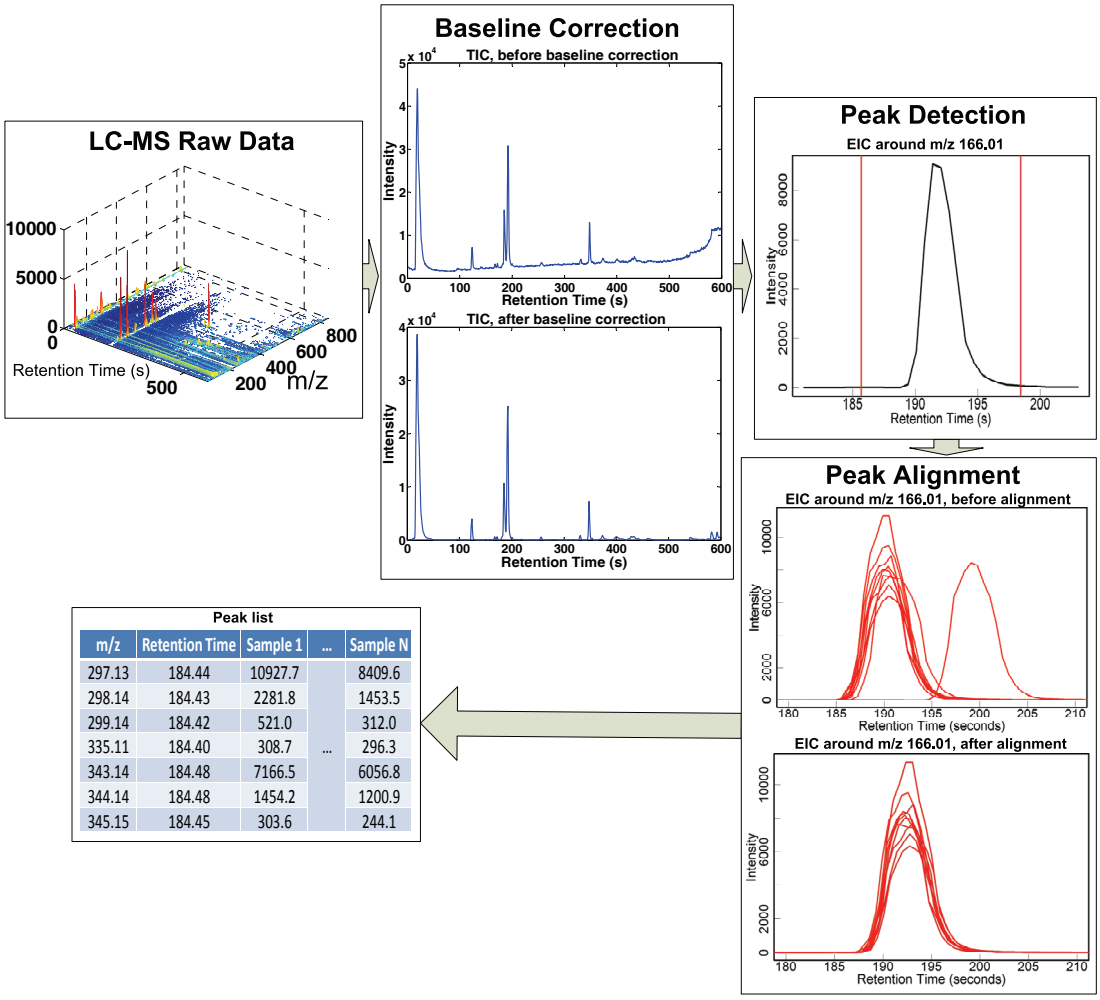


Figure 6. Major steps in the preprocessing of LC-MS data

2.2 Statistical analysis

After pre-processing, the LC-MS raw data are summarized by a peak list. The statistical analysis aims to detect those peaks whose intensity levels are significantly altered between distinct biological groups.

The statistical analysis methods can be categorized as univariate and multivariate analysis. The univariate approach assesses the statistical significance of each peak separately. Commonly used univariate techniques include t-test, fold-change analysis, Wilcoxon rank-sum test, analysis of variance (ANOVA), etc. P-values are usually assessed in univariate methods, either through parametric approaches or permutation tests. Because thousands of metabolites can be simultaneously measured in an untargeted study, the multiple hypothesis testing problem will result in a high chance of false discovery even with a

small p-value threshold. False discovery rate (FDR) is used to estimate the chance of false discovery at a given test statistics threshold or to control the total number of false discoveries.

Multivariate analysis considers the combinatorial effect of multiple variables. It can be further categorized as unsupervised and supervised techniques. Unsupervised learning refers to methods that identify hidden structure in the data without knowing the class labels. One of the most popular unsupervised techniques in LC-MS-based metabolomic study is PCA, which finds a series of orthogonal projection directions that maximize the variance of the projected data. Other unsupervised techniques such as self-organizing map (SOM) [18] or two-mode clustering [19] have also been used in LC-MS-based metabolomic studies.

In contrary to unsupervised techniques, supervised learning uses the class label information to construct a model to interpret the LC-MS data. Partial least square-discriminant analysis (PLS-DA) is a supervised technique widely used in LC-MS-based metabolomics data analysis. PLS-DA finds the projection direction which gives largest covariance between the data and the labels. It is successful in identifying the projection that separates the pre-defined groups and finding the discriminant metabolites for the separation [20]. In addition to PLS-DA, other supervised learning methods such as random forest [21] and SVM [22] have also been used in assessing the discriminative power of metabolites.

In the thesis, we focus on univariate analysis approaches for statistical analysis of pre-processed data. Specifically, fold change, Wilcoxon rank-sum test, and FDR were used to select or rank the peaks which are most indicative of the two distinct biological groups.

2.3 Metabolite identification

One of the key challenges in metabolomics studies is the identification of metabolites. Compared to peptide identification in LC-MS/MS based proteomics, it is more difficult to identify metabolite. Proteins consist of 20 amino acids repeatedly arranged in different linear orders and their biological functions are determined by the monomer order. Because of this linear structure, the fragmentation patterns and MS/MS spectra of proteins can be predicted to a certain degree of confidence. With the completion of

human genome project, a large fraction of the human protein sequences have become readily available. Based on the sequence database and the understanding of fragmentation mechanism, the hypothetical fragmentation spectra of peptides can be generated and used to compare with experimental MS/MS spectra for peptide identification.

Metabolite, on the other hand, is a general term for low-weight molecules. The structure of metabolites is not a repeated pattern of limited "alphabets" but a combination of elements such as C, H, O, S, N, and P with diverse structures. Their chemical and physical diversity make it difficult to derive general rules to predict their fragmentation patterns. Many types of metabolism reaction further complicate the identification task. While it is estimated that there are 2,000 major metabolites in human body [23], the total number of possible metabolites can reach up to 1,000,000 [24].

At present, metabolite identification in untargeted metabolic analysis is mainly achieved through mass-based search followed by manual verification. First, the m/z value of a molecular ion of interest is searched against database(s) [25-28]. The molecules having molecular weights within a specified tolerance range to the query molecular weight are retrieved from databases as putative identifications. The mass-based search can seldom provide unique identifications for the ions of interest due to three reasons. First, it has been shown that even with an accuracy of less than 1 ppm, which is a remarkably better accuracy than most analytical platforms can achieve, it is still not sufficient for unambiguous metabolite identification due to the presence of compounds with extremely similar molecular weights [11]. Second, mass-based metabolite identification cannot discriminate isomers which have the same elemental composition but different structures. Third, all the metabolite databases are of limited coverage. Generally less than 30% of the detected ions in a typical LC-MS-based metabolomic experiment can be uniquely identified through mass-based search, leaving most of the ions either unidentified or with multiple putative identifications. Improved approaches, such as those involving isotope labeling, can be used to reduce the ambiguities from the mass-based search. But they cannot guarantee unique identification either [29].

To verify the mass-based search results, authentic compounds of those putative identifications are subjected to MS or tandem MS experiments together with the sample. By comparing the retention times or tandem MS spectra of the authentic compounds with the ions of interest in the sample, the identities of the metabolites can be confirmed. It may be necessary to extend the MS^2 to MS^3 or MS^4 level for more confident identifications of some metabolites. It was suggested that at least two independent and orthogonal data (retention time and mass spectrum, accurate mass and tandem mass spectrum, etc.) relative to an authentic compound analyzed under identical experimental conditions are necessary to verify a putative metabolite identification [30]. Several examples of metabolite verification using authentic compounds are demonstrated in Section 3.2.7. The limiting factor of verification is that it is often costly and time consuming. The authentic compounds of putative identifications need to be acquired. More experiments need to be performed. Sometimes, a molecular ion can have more than 100 putative identifications which make manual verifications extremely laborious.

Chapter 3 Major Contributions of the Thesis

The major contributions of the thesis are presented in this section. They consist of three parts: (1) a computationally simple outlier screening approach is developed using raw LC-MS data to exclude low quality sample runs from analysis; (2) an integrated computational framework for improved metabolite identification is proposed and implemented by either developing in-house software or adapting publicly available tools; (3) an improved approach is developed to increase the retrieval performance in MS/MS spectral matching when heterogeneous data are used.

3.1 Outlier screening

Outlier screening aims to eliminate LC-MS runs which exhibit a significant deviation from the majority of their replicates. While the LC-MS variability is unavoidable, the outliers with excessive amount of variations need to be removed. Especially, low-quality runs due to analytical or instrumental reasons will affect following peak detection and peak alignment steps, which in turn will have adversary effects on metabolite identification as preprocessing may result in false detection and quantitation of metabolites. On the other hand, analytical runs with normal variation should be kept so the sample size of the experiment is not unnecessarily reduced simply because of variations such as retention time drifts, which can be corrected during the preprocessing. One example of low quality runs is shown in Figure 7 in comparison with a normal LC-MS run. The low quality run is distinct from the normal run with a relatively flat total ion chromatogram (TIC). Exploiting this characteristic, we proposed an outlier screening approach to detect and remove the low quality runs from a LC-MS dataset.

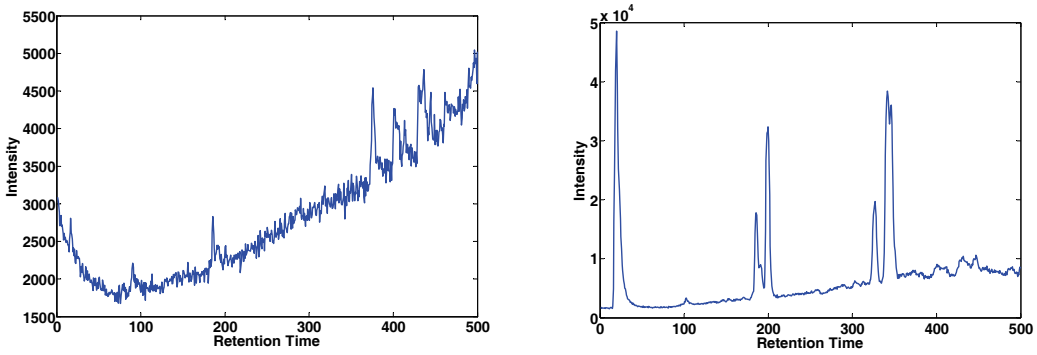


Figure 7. The comparison of TICs of a low quality LC-MS run (left) and a normal LC-MS run (right)

A quality assessment measure, Xrea, was previously developed to evaluate the quality of MS spectra [31] and MS/MS spectra [32]. We will show that the measure is equally applicable to the TIC of LC-MS data. Based on the cumulative intensity normalization, Xrea measures the "peakness" of the TIC. First, the intensity of TIC at each retention time is normalized by the total intensity. The cumulative normalized intensity (CNI) at each retention time t is the sum of the normalized intensities over all the time points with intensities smaller than or equal to the normalized intensity at t, i.e.

$$CNI(t) = \frac{\sum_{x} (I(x) \mid Rank(x) \geq Rank(t))}{\sum_{x} I(x)} \tag{3.1}$$

where *CNI(t)* is the cumulative normalized intensity at t, *I(t)* is the original intensity at t, and *Rank(x)* is the rank order of the intensity at retention time x by sorting intensities in descending order. By plotting the CNIs in ascending order, it is demonstrated that CNIs are indicative of the "peakness" of TIC. As in Figure 8, CNIs are closer to the diagonal line when the corresponding TIC is flatter. To quantify the "peakness", the area between the diagonal line and the cumulative normalized intensities is used as an indicator of the qualities of LC-MS data. The Xrea quality descriptor is defined as

$$Xrea = \frac{\frac{N+1}{2} - \sum_{t} CNI(t)}{\frac{N+1}{2} + \alpha} \tag{3.2}$$

where α is the correction term to account for cases in which the highest intensity is significantly larger than other data points and is set to the highest normalized intensity following [33]. Generally, a lower *Xrea* indicates a flatter TIC.

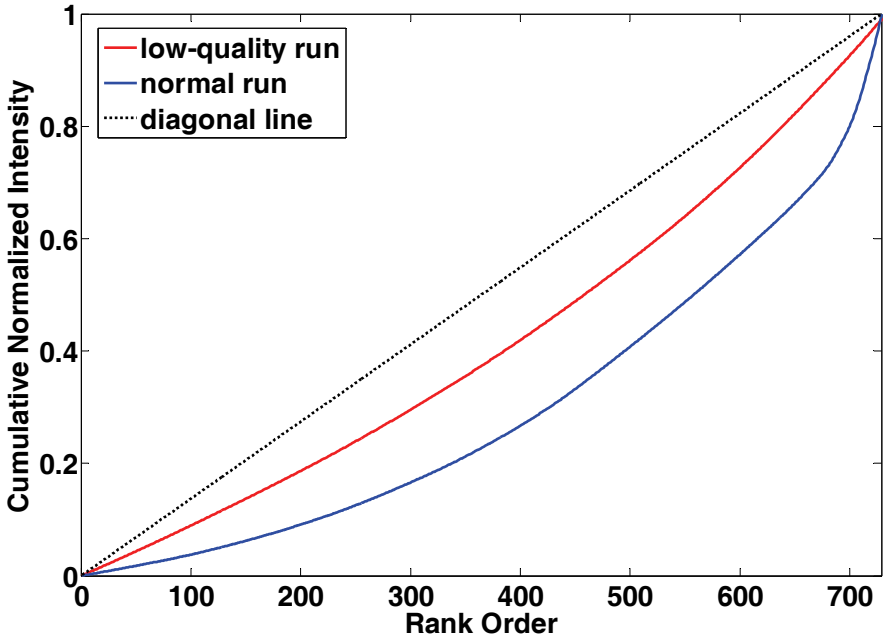


Figure 8. CNIs for the low-quality LC-MS run and the normal LC-MS run in Figure 7

After Xrea values are calculated for samples in a dataset, they are visualized using a box-plot. 25-percent quantile $Q1$ and 75-percent quantile $Q3$ are calculated. As a common criteria in statistics, an LC-MS run is considered as an outlier if

$$Xrea < Q1 - 1.5IQR \tag{3.3}$$

where $IQR=Q3-Q1$. Only the LC-MS runs with abnormally small Xrea values are considered as outliers because only low Xrea values indicate low-quality LC-MS runs.

The proposed outlier screening approach is applied to an LC-MS dataset with 337 runs. The box-plot of the Xrea values is shown in Figure 9. Out of 337 LC-MS runs, 23 runs are detected as outliers. Among 23 outliers, 9 LC-MS runs are instrument calibration runs from water and standard chemical solution rather than biological samples; 12 LC-MS runs are confirmed as abnormal runs due to insufficient amount

of samples at the experiment; the other 2 are normal biological samples which are false positive. In this dataset, the proposed approach achieves 99.4% specificity and 100% sensitivity.

Conclusion: The proposed approach is computationally simple but effective to exclude the low-quality runs as outliers from LC-MS data. It does not depend on any particular preprocessing method and does not exclude samples with reasonable variations (such as retention time drift) which can be corrected during preprocessing.

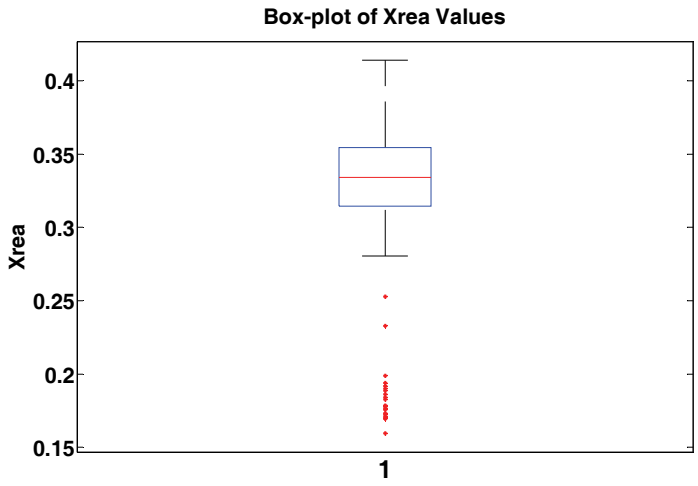


Figure 9. Box-plot of Xrea values from a dataset of 337 LC-MS runs

3.2 An integrated computational framework for improved metabolite identification[2]

As mentioned in Section 2.3, metabolite identification remains one of the major bottlenecks in LC-MS-based metabolomic studies. Although metabolite verification is generally considered as the "gold standard" and thus indispensable in metabolite identification [30], the verification using authentic compounds is often laborious and costly, as one LC-MS peak may have multiple putative identifications. Computational approaches can assist the identification process by increasing the coverage of

[2] This part of the thesis has been published in "A Computational Pipeline for LC-MS/MS Based Metabolite Identification", Bin Zhou, Jun Feng Xiao, and Habtom Ressom, the proceedings of International Conference on Bioinformatics and Biomedicine, 2011

identification and providing important guidance for metabolite verification through prioritization of putative identifications. Several computational algorithms have been developed to facilitate the identification of metabolites [34-37]. However, these algorithms are largely developed in isolated manners.

In this thesis, we propose an integrated computational framework, which adapts and incorporates multiple software tools and databases to assist metabolite identification before metabolite verification is performed. The proposed framework is shown in Figure 10.

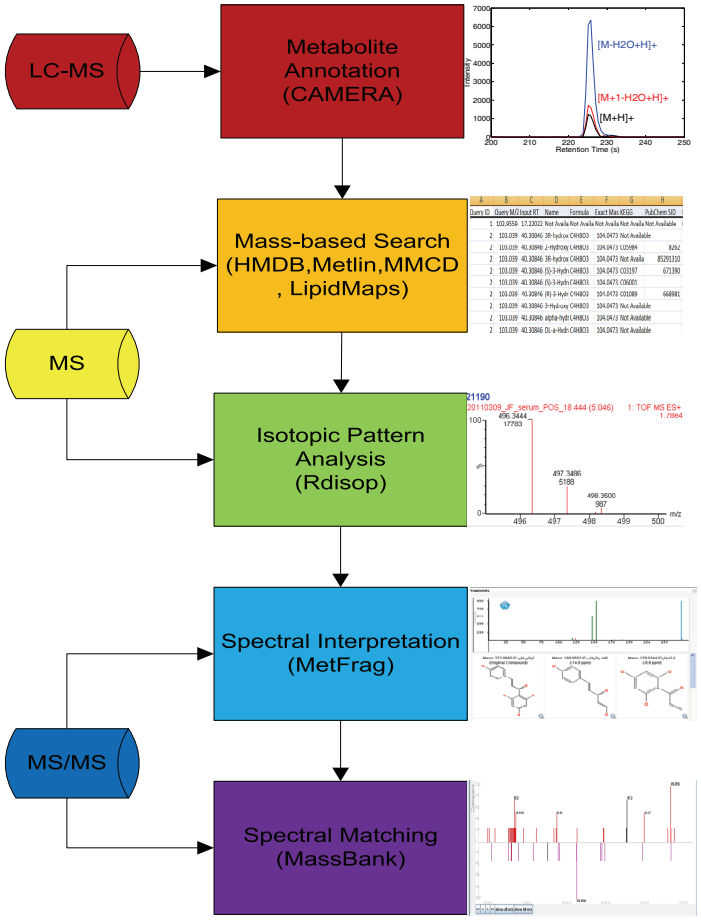


Figure 10. Computational framework for metabolite identification

The proposed framework integrates three types of information one can generally acquired from an LC-MS/MS experiment: (1) retention times and elution profiles, (2) exact mass and mass spectra, (3) and MS/MS spectra. Each type of information can be utilized to facilitate the metabolite identification. As

shown in the figure, the framework involves five steps: ion annotation, mass-based search, isotopic pattern analysis, spectral interpretation, and spectral matching. It should be noted that not all ions of interest can go through the entire framework due to the availability of the data or the coverage of the database(s). However, each step in the framework refines the identification results from previous steps. Details of each step are introduced in the following sections.

3.2.1 Ion annotation

Ion annotation is a procedure to recognize group of ions which are likely to originate from the same compound. In LC-MS based metabolomics, one metabolite is often represented by multiple peaks with distinct m/z values but at similar retention times. Recognition of those peaks from the same metabolite can facilitate the metabolite identification.

Generally, one metabolite can generate three types of ions in LC-MS data: adduct, isotope, and in-source fragment. An adduct ion is "an ion formed by interaction of two species, usually an ion and a molecule, and often within the ion source, to form an ion containing all the constituent atoms of one species as well as an additional atom or atoms" [38]. The most common adduct ions in LC-MS are protonated ion $[M+H]^+$ or deprotonated ion $[M-H]^-$ (although deprotonated ion is the loss of a proton rather than addition, it is generally considered as an adduct). In addition, there could be other types of adducts, such as sodium adduct, potassium adduct, and etc. Some most common forms of adducts are listed in Table 1, while more complete information concerning adduct in mass spectrometry can be found in [39, 40].

Isotopes are variants of atoms of the same chemical elements, which have the same number of protons but different number of neutrons. As a result, the atoms of the same element may have different masses depending on the number of neutrons they have. Common metabolites are composed of elements carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphor (P), and sulfur (S). Most of them have at least one naturally-existing, stable isotope. So metabolites are usually a mixture of several isotopic species. During mass spectrometry analysis, different isotopic species are separated, which will generate a series of peaks separated on m/z by around 1 Da difference. Among them, the peak with the lowest m/z

(so all the constituent elements of its corresponding molecular ions are of the smallest mass) is defined as the monoisotopic peak.

Ionization	Formation	Ion Mass
Positive	$[M+H]^+$	m+1.0073
	$[M+2H]^{2+}$	m/2+1.0073
	$[M+Na]^+$	m+22.9892
	$[M+K]^+$	m+38.9632
	$[M+NH4]^+$	m+18.03382
Negative	$[M-H]^-$	m-1.0073
	$[M-2H]^{2-}$	m/2-1.0073
	$[M-2H+Na]^-$	m+20.9747
	$[M-2H+K]^-$	m+36.9486

Table 1. Common types of adducts in LC-MS. M is the molecule with molecular weight m

The third type of ions is in-source fragments. Although ESI is generally considered as a soft-ionization approach which mainly generates intact molecular ion, fragmentation may still happen during ionization. One commonly seen in-source fragment is water-loss fragment $[M+H-H2O]^+$ or $[M-H-H2O]^-$, where a water molecule is lost during the ionization process.

Different adducts/isotopes/water-loss products of the same compound theoretically share the same retention time in chromatograms. As long as the scan rate is properly adjusted and enough scanning points are acquired to define the chromatographic peaks, the ions from the same compound share similar-shaped elution profiles which can be represented by their extracted ion chromatograms (EICs). Thus ion annotation can be accomplished by clustering similar elution profiles together.

Different ion formations of the same metabolite will differ in their m/z values. The observed m/z of an ion derived from a metabolite with a monoisotopic molecular weight M is

$$s = \frac{nM + \alpha + \beta M_{neutron}}{z} \tag{3.4}$$

where n is the number of molecules in the ion, α is the mass of the adducts (or fragments), $M_{nuetron}$ is the mass of the neutron, β is the extra number of neutrons in isotopes, and z is the charge of the ion. In LC-MS many types of adducts and fragments are known, such as $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, and $[M+H-H2O]^+$ etc. As a result, the m/z relationships between these known ion formations are often known-a-priori.

An R-package CAMERA (Collection of Algorithms for MEtabolite pRofile Annotation), was previously developed in [34]. It performs ion annotation in two steps. In the first step, the detected peaks with similar retention times are roughly grouped together using a sliding retention time window. Within each group, the EICs of the peaks are extracted and the peaks are clustered into smaller groups based on the Pearson correlation between their EICs. The m/z difference between each peak pair within in a group is calculated and compared to known m/z relationships between different ion formations. The two ions are considered to come from the same compound if their m/z difference can be explained by one of the known m/z relationships.

3.2.2 Mass-based search

After grouping peaks together by ion annotation, the exact monoisotopic masses of these compounds can be calculated. The calculated masses are then used to search against metabolite databases like Human Metabolome Database (HMDB) [41], Metlin [27] and Madison Metabolomics Consortium Database (MMCD) [36] or more general chemical databases like PubChem or ChemSpider. Metabolites having masses within pre-specified tolerance range of the query mass are retrieved from these databases. An in-house software tool is developed to search four metabolite databases: HMDB, Metlin, MMCD and LipidMaps [42]. Because the same metabolite may appear in more than one database, the results from different databases are merged together based on the InChI Key of the retrieved metabolites. The InChI Key is the hashed version of International Chemical Identifier (InChI) and contains information about

molecular formula, atom connection, and stereochemistry information of a compound. Because the stereochemistry is generally less a concern in the untargeted metabolomic study, the metabolites which share the same first 14 characters in InChI key are considered as the same compound and merged together. The merged results will be used as the putative identifications for the ions of interest.

The developed software was tested using a peak list from an untargeted metabolomic study. There were 229 m/z values in the peak list acquired under negative ionization mode. The peak list was subjected to a mass-based search with 20 ppm tolerance. The merged results are compared with retrieval results from each individual database in Table 2.

	HMDB	MMCD	Metlin	LipidMaps	MetaboSearch
A	204	462	719	429	1010
B	60	90	81	54	102

Table 2. Comparison of the merged results with retrieval results from individual databases regarding the number of putative identifications (A) and the number of m/z values that have at least one putative identification (B).

3.2.3 Isotopic pattern analysis

The putative identifications from the mass-based search for an ion may include both isomers (compounds with different structures but the same elemental formula) and compounds with similar molecular weights but different elemental formulas. For those putative identifications with different elemental formulas, they can be prioritized through analysis of their isotopic patterns.

The isotopes of each element found in nature are of constant relative abundances or "natural abundances", as shown in Table 3. When these elements are combined into metabolite molecules, the m/z values and relative abundances of the isotopic species of the metabolites can be calculated theoretically, which is the isotopic pattern for the metabolite. Even for two metabolites with very similar molecular weights, their isotopic patterns will differ as long as they are of different elemental formulas. By

comparing the observed isotopic pattern with the theoretical isotopic patterns of putative identifications, the putative identifications are prioritized.

Element	Isotopes	Relative Abundance (%)	Mass (Da)
Hydrogen	^{1}H	99.985	1.008
	^{2}H	0.015	2.0141
Carbon	^{12}C	98.890	12.0
	^{13}C	1.110	13.0034
Nitrogen	^{14}N	99.634	14.0031
	^{15}N	0.366	15.0011
Oxygen	^{16}O	99.762	15.9949
	^{17}O	0.038	16.9991
	^{18}O	0.200	17.9992
Phosphor	^{31}P	100	30.9738
Sulfur	^{32}S	95.020	31.9721
	^{33}S	0.750	32.9715
	^{34}S	4.210	33.9679
	^{36}S	0.020	35.9671

Table 3. The relative abundance and masses of isotopes

R-package Rdisop [43] is used in the framework to score different elemental formulas. The score is computed as a posterior probability $P(M_i|D, B)$ where M_i is the elemental formula of the ith putative identification; D is the extracted isotopic pattern of the ion from its chromatographic apex and B stands for any prior knowledge about M_i. The posterior probability is calculated as

$$P(M_i \mid D,B) = \frac{P(M_i \mid B)P(D \mid M_i,B)}{\sum_j P(M_j \mid B)P(D \mid M_j,B)} \tag{3.5}$$

$$P(D \mid M_i,B) = \prod_j P(M_{i,j} \mid m_j)\prod_j P(f_{i,j} \mid p_j) \tag{3.6}$$

and $P(M_{i,j}|m_j)$ is the conditional probability to observe jth isotopic peak at $M_{i,j}$ when the theoretical m/z of the jth isotopic peak of the elemental formula M_i is at m_j; $P(f_{i,j}|p_j)$ is the conditional probability to observe jth isotopic peak with relative intensity $f_{i,j}$ when the theoretical relative intensity of the jth isotopic peak is P_j. In the algorithm, both conditional probabilities are assumed to be normal distribution.

3.2.4 Spectral interpretation

Spectral interpretation deduces the possible structure or sub-structure of an unknown molecular ion by comparing its MS/MS spectrum with hypothetic spectra predicted through in-silico fragmentation approaches.

MetFrag is used for spectral interpretation in the framework [35]. The in-silico fragmentation spectra are first generated through combinatorial disconnection of chemical bonds in the compound structure. The structures of putative identifications are expanded to consider possible rearrangements by applying known neutral loss rules. Then each structure (includes the expanded ones) is considered as the root node of a fragmentation tree and fragments are generated by splitting a molecular bond (or two bonds for ring structures). The splitting procedure is carried out using an iterative, breadth-first approach to generate all possible fragments. For those bond cleavages which produce fragments matching to the m/z values of fragment ions in the observed MS/MS spectra, the bonds and fragments are recorded for later scoring. This combinatorial disconnection procedure is performed until all the observed fragments in the MS/MS spectra are matched with in-silico fragments or a certain depth in the fragmentation tree is reached.

After the observed fragment ions in MS/MS data are matched against in-silico fragments, the candidate structure is scored according to the number of matches and the total bond dissociation energy needed to generate these fragments. The score of the ith particular putative identification is

$$Score_i = \frac{1}{\max(w)} w_i - \frac{1}{2\max(e)} e_i \tag{3.7}$$

$$w_i = \sum_{f \in F_i} (\text{int}_f)^{0.6} (\text{mass}_f)^3 \tag{3.8}$$

$$e_i = \frac{1}{|F_i|} \sum_{f \in F_i} \sum_{b \in B_f} BDE_b \tag{3.9}$$

where F_i is the set of fragment peaks from the MS/MS spectrum which match the in-silico fragments. int_f and $mass_f$ are the intensity and m/z values of the particular matched fragment ion *f*. BDE_b is the bond dissociation energy of a particular bond *b*. In Equation (3.7), w_i measures the weighted total number of explained fragments for the *i*th putative identification, and e_i is the penalty term considering bond dissociation energy. The compound with a structure which explains most of the fragment ions in the experimental MS/MS spectrum with small bond dissociation energy is favored as the identification for the ion of interest.

3.2.5 Spectral matching

Spectral matching mimics the manual verification of metabolites using MS/MS spectra. Instead of acquiring the MS/MS spectrum of the authentic compound each time, previously acquired MS/MS spectra are assembled into a spectral library and used to compare with the spectra acquired from biological samples. Several spectral libraries have already been constructed and accessible for public [25, 27, 37, 41]. Among them, MassBank is used in the proposed framework for spectral matching. 9,276 ESI-MS^n spectra of 2,337 metabolites are acquired using authentic compounds. For many metabolites, more than one spectrum is acquired with different collision energies or instruments, which increases the transferability of the spectral library among different experiments. One important aspect in spectral matching is the design of spectral similarity scores. MassBank uses a dot-product based approach where similarity score *S* is defined as

$$S = \frac{(\sum W_L W_Q)^2}{\sum W_L^{\,2} \sum W_Q^{\,2}} \tag{3.10}$$

where W_L and W_Q are scaled and mass-weighted intensities of the library spectrum and query spectrum respectively. The spectra is intensity-scaled and mass-weighted as

$$W_i = [f_i]^m [mz_i]^n \tag{3.11}$$

where f_i and mz_i are the intensity and the m/z value of each peak in the spectrum. *m* and *n* are empirically determined to be 0.5 and 2 for ESI-MS/MS spectra.

One problem of spectral matching is relatively low reproducibility across different instruments and experiments. It can be partially attributed to the instability of ESI source. In addition, there is a lack of standard protocol on how MS/MS should be acquired due to the diversity of metabolites. As a result, this data heterogeneity issue decreases the performance of spectral matching, especially when the dot-product is used as the similarity measure. A spectral matching algorithm with improved performance under data heterogeneity is discussed in Section 3.3.

3.2.6 Experiment

We used the proposed framework to assist the metabolite identification task in an LC-MS/MS-based metabolomic study. An untargeted metabolic analysis was performed using Waters UPLC-QToF Premier instrument on human serum samples. Both positive and negative ionization modes were used. The acquired data were preprocessed. Statistical analysis was performed to obtain a list of potential biomarkers that are significantly altered between case and control samples. Seven peaks from the list are presented in Table 4 to illustrate our framework. Prior to metabolite verification, we used our framework on the seven peaks to find the most probable identifications. The results obtained from each stage of the framework are presented in the following sections.

m/z	Retention time
432.31	227.14
433.31	227.14
432.31	224.09
433.32	224.06
450.32	224.68
450.32	227.17
378.24	261.80

Table 4. m/z values and retention times of seven selected peaks from an untargeted analysis

The results of ion annotation performed using CAMERA are shown in Table 5. It is evident that the seven peaks are derived from three unknown metabolites. We denoted these metabolites as M1-M3. We now focus on the last three peaks which represent the protonated and deprotonated ions of the target metabolites ($[M1+H]^+$, $[M2+H]^+$and $[M3-H]^-$). The remaining four peaks represent the water-loss fragments and the isotopes of the water-loss fragments of M1 and M2.

m/z	Retention time	Annotation
432.31	227.14	$[M1-H2O+H]^+$
433.31	227.14	$[M1+1-H2O+H]^+$
432.31	224.09	$[M2-H2O+H]^+$
433.32	224.06	$[M2+1-H2O+H]^+$
450.32	224.68	$[M2+H]^+$
450.32	227.17	$[M1+H]^+$
378.24	261.80	$[M3-H]^-$

Table 5. Ion annotation for the seven selected peaks

The exact masses of the three metabolites M1-M3 were calculated from their annotation and were subjected to mass-based search. A list of putative identifications was retrieved from HMDB, Metlin, MMCD, and LipidMaps databases. The results from the four databases were integrated using our in-house software. We used a 10 ppm tolerance for the positive mode data and a 20 ppm tolerance for the negative mode data. The results are shown in Table 6. For each of the metabolite, there are two putative identifications. GDCA (glycodeoxycholic acid) and GCDCA (glycochenodeoxycholic acid) are isomers with the same elemental formula but different structures. S-1-P (Sphingosine-1-phosphate) and Celiprolol are of different elemental formulas but with very similar masses of only 4 ppm difference. At this step, we do not know yet the right identification or the relative priorities of putative identifications for any of these metabolites.

Since the two putative identifications for M3 have different formulas, isotopic pattern analysis was performed using Rdisop to score the two candidate compounds. The isotopic pattern of deprotonated ion of M3 was extracted from the apex of chromatographic peak. Rdisop gave a higher matching score for S-1-P (0.80) compared with Celiprolol (0.12). As a result, S-1-P was given a higher priority than Celiprolol.

Metabolite	Query m/z	Name	Formula	Exact Mass
M1	450.32	GDCA	C26H43NO5	449.31
		GCDCA	C26H43NO5	449.31
M2	450.32	GDCA	C26H43NO5	449.31
		GCDCA	C26H43NO5	449.31
M3	378.24	Celiprolol	C20H33N3O4	379.24
		S-1-P	C18H38NO5P	379.24

Table 6 Mass-based search for metabolites M1-M3

After isotopic pattern analysis, MS/MS experiments were conducted on the protonated (or deprotonated) ions of M1-M3. For the protonated ions of M1 and M2, LC-MS/MS experiments were performed using Waters UPLC-QToF Premier instrument with precursor ion scan at m/z 450.3. For the deprotonated ion of M3, MS/MS experiments were performed using AB QSTAR instrument with precursor ion scan at m/z 378.2. However, since the intensity of M1 is too low in the sample, we only acquired valid MS/MS spectra for M2 and M3. The two spectra were subjected to spectral interpretation. The spectral interpretation results for M2 and M3 are shown in Table **7**, in which the priorities of putative identifications are denoted with an asterisk (*). For M2, GCDCA has a higher score than GDCA. However, both compounds have high matching scores and many peaks that can be interpreted by the in-silico fragmentation. For M3, S-1-P has a clear advantage over Celiprolol.

Metabolite	Score	# of interpreted peaks	Name
M2	1.0	20	GCDCA (*)
	0.901	32	GDCA
M3	1.0	20	S-1-P (*)
	0.0	4	Celiprolol

Table 7 Spectral interpretation for M2 & M3

By looking up the putative identifications of M2 and M3 in MassBank, we found both putative identifications for M2 in the database, but no Celiprolol existing in MassBank for M3. So the priorities of the putative identifications of M2 can be further refined with spectral matching. However no spectral matching was performed for M3. The spectral matching results for M2 are shown in Table 8 and Figure 11. The library spectrum of GDCA shows a slightly higher matching score compared with that of GCDCA. Because we generally have a higher confidence for experimentally acquired MS/MS spectra than those generated by in-silico fragmentation, we concluded that GDCA to have a higher priority than GCDCA as a possible identification for M2.

Metabolite	Score	Name
M2	0.537891	GDCA (*)
	0.530468	GCDCA

Table 8 Spectral matching results for M2

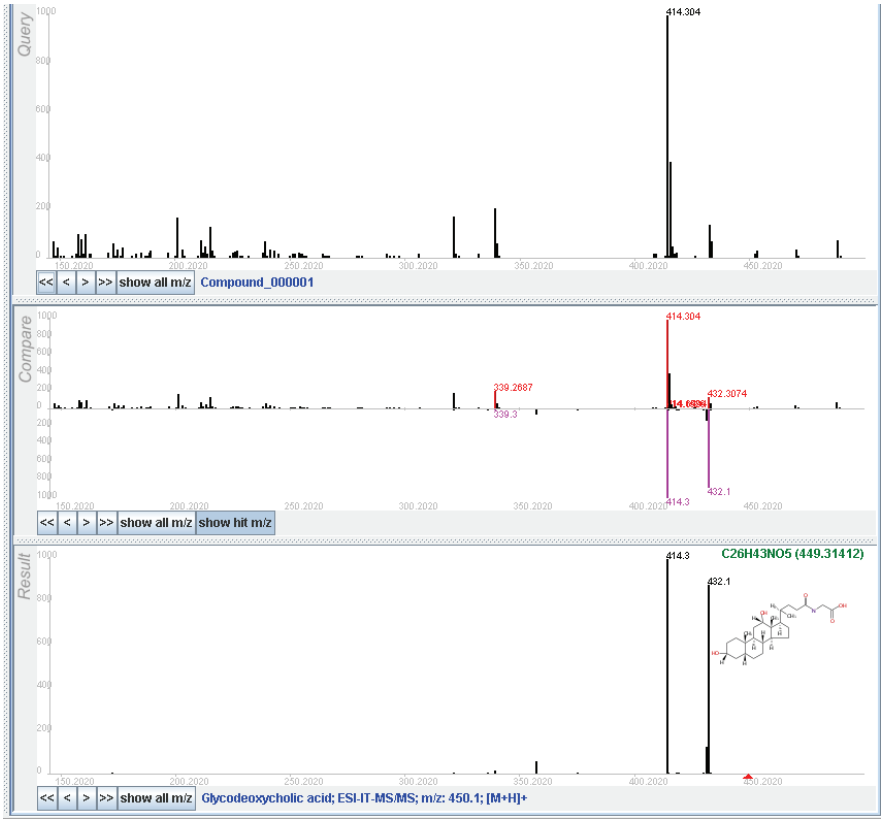


Figure 11. MS/MS spectrum of GDCA compared with M3. A: Experimental MS/MS spectrum of M3. C: Library MS/MS spectrum of GDCA; B: Comparison of A and B

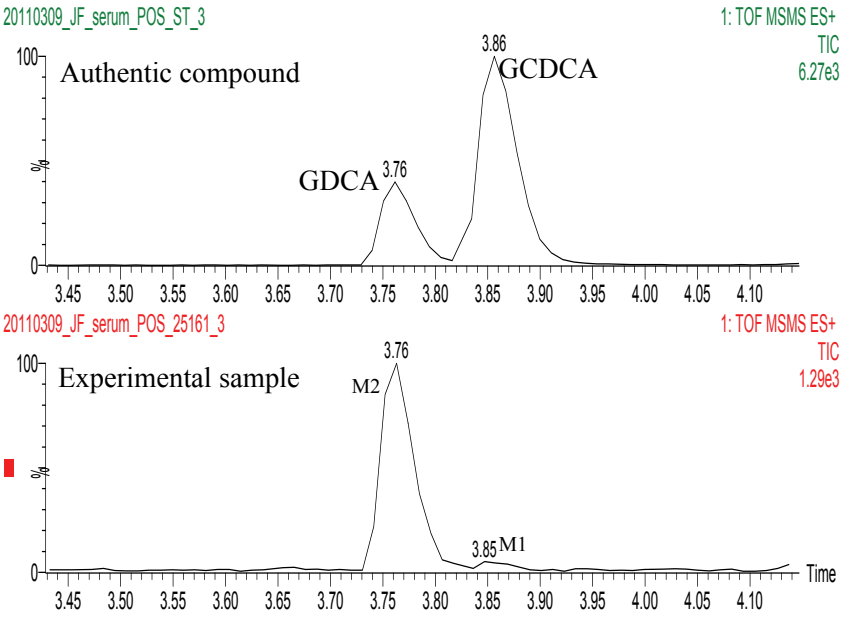
3.2.7 Discussion

Table 9 summarizes the prioritization results obtained by our framework for the seven peaks. Because two metabolites M1 and M2 share the same set of putative identifications of GDCA and GCDCA, and GDCA had been assigned a higher priority for M2, GCDCA was given a higher preference for M1. As a comparison, the table also presents putative identifications from a direct mass-based search with the same tolerance range as used in our experiment. Metabolite verification was then performed using the authentic standards of GDCA, GCDCA and S-1-P. The chromatograms and MS/MS spectra of protonated (or deprotonated) ions of the authentic compounds are compared with those of the experimental samples as illustrated in Figure 12 and Figure 13. The results verified that M1 is GCDCA, M2 is GDCA, and M3 is S-1-P. Similar experiments were also performed for adducts and water-loss fragment ions. The results

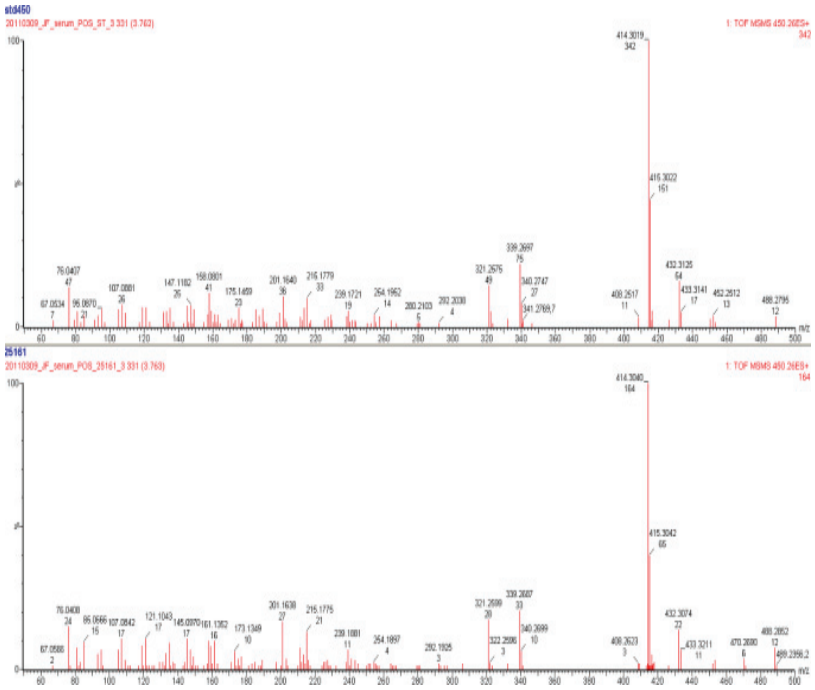
verify the accuracy of the annotation. Also, we found that putative identifications which received high priority by our framework led to correct identifications.

m/z	Retention time	Mass-based search	Proposed framework
432.31	227.14	Sphingofungin A	$[GCDCA-H2O+H]^+$ (*) $[GDCA-H2O]^+$
433.31	227.14	4,4'-Diaponeurosporenic acid; (20R,24R)-20-fluoro-1alpha,24-dihydroxy-26,27-cyclovitamin D3	$[GCDCA+1-H2O+H]^+$ (*) $[GDCA+1-H2O+H]^+$
432.31	224.09	Sphingofungin A	$[GDCA-H2O]^+$ (*) $[GCDCA-H2O+H]^+$
433.32	224.06	4,4'-Diaponeurosporenic acid; (20R,24R)-20-fluoro-1alpha,24-dihydroxy-26,27-cyclovitamin D3	$[GDCA+1-H2O+H]^+$ (*) $[GCDCA+1-H2O+H]^+$
450.32	224.68	GDCA; GCDCA	GDCA (*) GCDCA
450.32	227.17	GDCA; GCDCA	GCDCA (*) GDCA
378.24	261.80	Celiprolol; S-1-P	S-1-P (*) Celiprolol

Table 9 The putative identifications prioritized by our framework compared with direct mass-based search results for the seven peaks (* indicates higher priority).

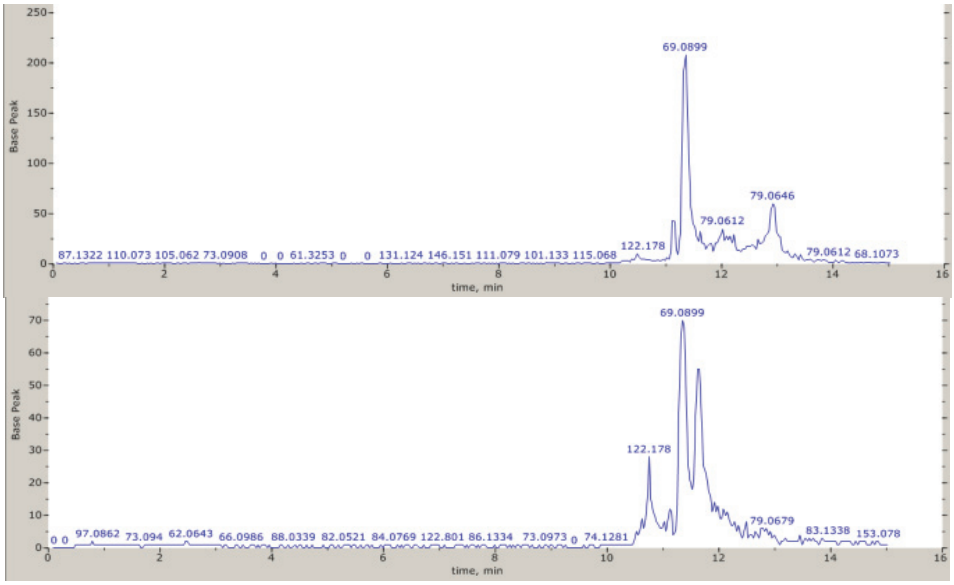


(A)

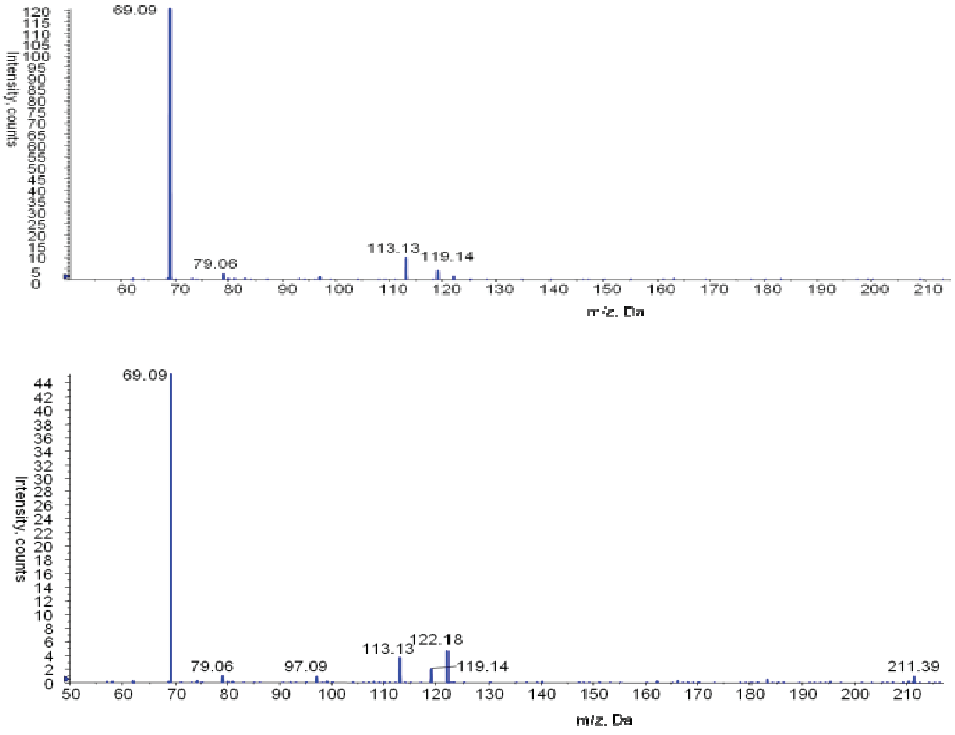


(B)

Figure 12. Metabolite verification for M1 & M2. A: Comparison of chromatograms of authentic compounds of GDCA & GCDCA (top) with M1 & M2 (bottom). B: Comparison of the MS/MS spectra of GDCA (top) with M2 (bottom).



(A)



(B)

Figure 13 Metabolite verification for M3. A: Comparison of the chromatogram of the authentic compound of S-1-P (top) with M3 (bottom). B: Comparison of the MS/MS spectra of S-1-P (top) with M3 (bottom)

Compared with direct mass-based search results in Table 9, our framework has several advantages including (1) with adducts, isotopes and water-loss fragments of metabolites correctly annotated, the framework increase the chance to find the right putative identifications for the peaks of interest; (2) It provides guidance for subsequent metabolite verification by prioritizing the putative identifications, thereby saving time and cost by minimizing the number of verification experiments involving authentic compounds.

3.3 SVM-based spectral matching for metabolite identification[3]

Spectral matching is an important component in the framework presented in Section 3.2. With sufficient database support and a proper scoring algorithm, spectral matching can provide highly confident identification results. During the past decade, several spectral matching algorithms have been developed for various applications and platforms [44-47]. While their exact forms differ from each other, these scoring algorithms are primarily some variants of dot product. The dot product of a query spectrum and a library spectrum intrinsically measures the correlation between the two spectra. The library spectrum which has the highest correlation with the query spectrum is considered to be the most probable identification.

While correlation generally performs well when the spectra are from highly controlled experiments, it degrades remarkably in real situations where the spectra are from different sources. This is because of the complexity and variation of process happening during the CID and also the instability of ESI source. Also experimental parameters such as collision energy have a significant impact on the intensity profile of a MS/MS spectrum. An example is shown in Figure 14, where two spectra of the same metabolite exhibit different spectral profiles when acquired under different conditions. As a result, the two spectra have a low correlation.

[3] This part of work has been published in "SVM-based spectral matching for metabolite identification", Bin Zhou, Amrita K Cheema, Habtom W Ressom, Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society 2010, 756-759 (2010)

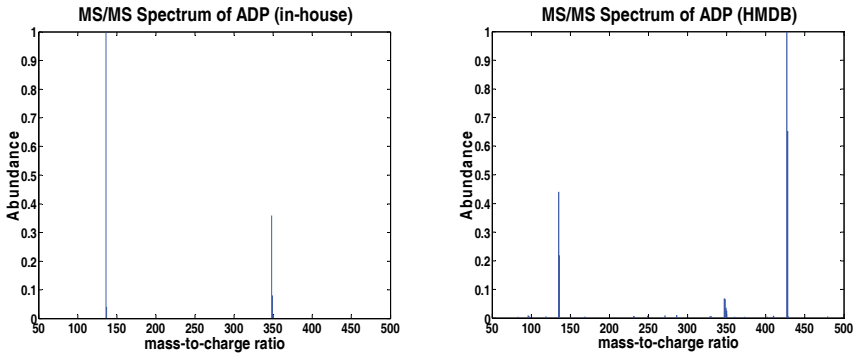


Figure 14 The MS/MS spectra for metabolite ADP from the in-house data (left panel) and from HMDB database (right panel)

When there is a high degree of data heterogeneity, correlation or dot product alone is not sufficient to measure the similarity between two spectra. Other similarity measures are needed. In this section, a SVM-based approach is presented to improve the retrieval performance for spectral matching.

3.3.1 SVM-based spectral matching

3.3.1.1 Feature construction

From the example given in Figure 14, we observe that while the intensity profiles of the two spectra are not similar, they share peaks appearing at the same positions over the m/z range. To take advantage of this observation, we propose to utilize "peak similarity", which measures the impact of the common peaks on the spectral comparison. Specifically, we define the following two measures of peak similarity:

$$N_{match} = N_{Q\&L} / \min(N_Q, N_L) \tag{3.12}$$

$$E_{match} = \sum_{i=1}^{N_{Q\&L}} (A_{Q,i} A_{L,i}) \tag{3.13}$$

where N_{match} is the normalized number of common peaks between the two spectra, and E_{match} is the total energy of common peaks. A_L and A_Q are the relative intensity of peaks in the library and query spectrum. N_L and N_Q are the total numbers of peaks in the library and query spectrum. $N_{Q\&L}$ is the number of common peaks appearing in both spectra.

We propose to combine the above peak similarity measures with a measure of profile similarity. To measure the profile similarity of two spectra, we use the Pearson correlation coefficient between the spectra. The correlation coefficient must be calculated between the vectors of the same length. Since the numbers of peaks in the two spectra are rarely equal, the spectra must be re-sampled before the calculation. Peak preserving re-sampling is used to re-sample the spectra [48]. The signal is re-constructed using a Gaussian kernel and the intensity at an m/z value is the maximum intensity of any contributing peaks. The Pearson correlation is then calculated using the re-sampled spectra.

The final identification is based on the overall consideration of both profile and peak similarity measures. We formulate the identification problem as a classification problem. Each comparison between two spectra is a sample for classification while the similarity measures are the features of the sample. And the label is binary: the sample is positive for comparison between the same metabolites while negative for comparison between different metabolites. The proposed metabolite identification algorithm is illustrated in Figure 15. Pair-wise comparisons are performed between query spectrum and the library spectrum. The normalized number of common peaks, the total energy of common peaks, and the Pearson correlation are utilized to train a SVM classifier to decide if the two spectra represent the same metabolite.

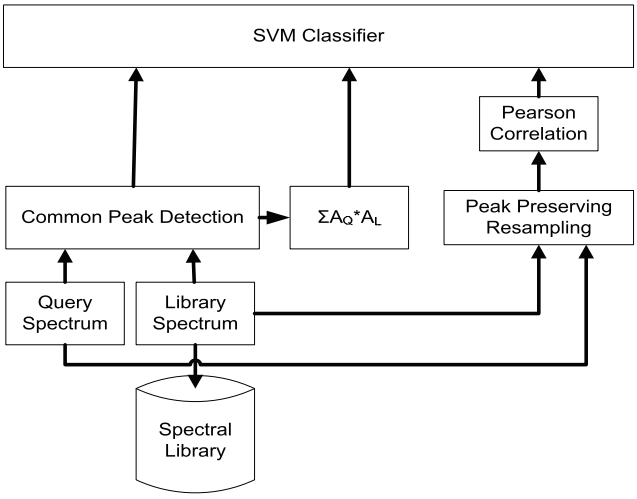


Figure 15 The algorithm diagram for the SVM-based approach

3.3.1.2 SVM classification

By examining the data, we find the separation hyperplane for positive and negative samples should be non-linear. The reason is that different similarity measures have different dynamic ranges and there is no explicit way to normalize them to make the data linearly separable. SVM with a Gaussian kernel function is a convenient and popular way to solve this kind of non-linear separation problems [49].

Given a set of training data $(\mathbf{x}_i, y_i)$ i=1, ···, N, where $\mathbf{x}_i$ is the input vector of the ith data point and y_i is the corresponding label, the construction of a SVM model involves two basic operations: (1) nonlinear mapping of an input vector into a higher dimensional "feature space"; (2) construction of an optimal hyperplane to separate the data points through minimization of structural risk.

For the ease of the description, we started from the construction of the optimal separation hyperplane for binary classification. In this case $y_i \in \{-1,1\}$, and the hyperplane is defined as

$$\{x : f(x) = \boldsymbol{\beta}^T \mathbf{x} + \beta_0 = 0\} \tag{3.14}$$

The classification rule is

$$\hat{y}(\mathbf{x}) = sign(\boldsymbol{\beta}^T \mathbf{x} + \beta_0) \tag{3.15}$$

The minimization of structural risk is achieved through the maximization of the margin of the separation, which is the distance between the hyperplane and the closest data point to it. For the set of data points which are separable by the linear hyperplane, all data points satisfy the following constraints (through the proper scaling of $\boldsymbol{\beta}$ and β_0)

$$\begin{aligned} \boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 &\geq 1, \quad \text{for } y_i = 1 \\ \boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 &\leq -1, \text{ for } y_i = -1 \end{aligned} \tag{3.16}$$

The maximization of the margin of the separation amounts to

$$\min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\| \tag{3.17}$$

as the margin on either side of the hyperplane is $1/\|\boldsymbol{\beta}\|$. For a set of data points with overlapping between classes, a soft margin is introduced by defining the slack variables $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_N)$, which allow some data points to be on the wrong side of the margin

$$\begin{aligned} & y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \geq 1 - \varepsilon_i \\ & \varepsilon_i \geq 0 \end{aligned} \tag{3.18}$$

And the margin is still maximized while considers the slack variables at the same time.

$$\min_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}} (\frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\beta} + C \sum_{i=1}^{N} \varepsilon_i) \tag{3.19}$$

The Equations (3.18) and (3.19) defines a primal problem for optimization. Using the Lagrange multipliers α_i, i=1,2, ···,N, the corresponding dual problem is

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \tag{3.20}$$

with the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{3.21}$$

$$0 \leq \alpha_i \leq C \tag{3.22}$$

In addition, the Karush-Kuhn-Tucker conditions provide the following constraints on the optimum solution for (3.20)

$$\alpha_i [y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - (1 - \varepsilon_i)] = 0 \tag{3.23}$$

$$y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - (1 - \varepsilon_i) \geq 0 \tag{3.24}$$

According to Equations (3.23) and (3.24), when $y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) = (1 - \varepsilon_i)$, i.e. the data point lies between the hyperplanes $\boldsymbol{\beta}^T \mathbf{x} + \beta_0 = 1$ and $\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 = -1$, we have $\alpha_i \neq 0$. Since the solution for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \sum_{i=1}^{N_s} \alpha_i y_i \mathbf{x}_i^{(S)} \tag{3.25}$$

$\boldsymbol{\beta}$ is solely decided by those data points which exactly meet the equality constraint in (3.24). These data points are called the support vectors. N_s in the above equation is the number of support vectors in the training data. With $\boldsymbol{\beta}$ determined, the classification result for a data point $\mathbf{x}$ is

$$\hat{y} = sign(\sum_{i=1}^{N_s} \alpha_i y_i \mathbf{x}^T \mathbf{x}_i^{(S)} + \beta_0) \tag{3.26}$$

where $\beta_0 = 1 - \boldsymbol{\beta}^T \mathbf{x}_i^{(S)}$ can be calculated from any support vector with $\varepsilon_i = 0$.

The SVM discussed above can only provide a linear separation hyperplane. For nonlinear classification problem, a nonlinear mapping function $\phi(\bullet)$ is used to transform the input vector $\mathbf{x}_i$ into another Euclidean space, often of a higher dimension. So the Equations (3.20) and (3.26) become

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \tag{3.27}$$

$$\hat{y} = sign(\sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + \beta_0) \tag{3.28}$$

Using an inner-product kernel $K(\mathbf{x},\mathbf{x}_i)$, which satisfies Mercer's condition [50], there is

$$K(\mathbf{x}, \mathbf{x}_i) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \tag{3.29}$$

And Equations (3.27) and (3.28) become

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{3.30}$$

$$\hat{y} = sign(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + \beta_0) \tag{3.31}$$

In this case, we do not need to know the exact form of the mapping function $\phi(\bullet)$ to be able to optimize and use SVM. A common choice of the kernel function is Gaussian function $K(\mathbf{x}, \mathbf{x}_i) = \exp(-\|x - x_i\|^2 / \sigma^2)$ where σ^2 is a user-defined parameter for the Gaussian kernel.

The metabolite identification system is intrinsically an information retrieval system. One common characteristics of such system is the imbalance of the samples. In the study, more negative samples are present than positive samples. It is well-known that such imbalance will have an adversary impact on the

performance of SVM classifier [51]. Several ways have been devised to compensate for imbalance samples, including re-sampling (either down-sampling or over-sampling), cost-sensitive learning or ensemble learning [51-53]. Here, we use a random down-sampling approach on the majority group to acquire a balanced dataset to train the SVM classifier.

3.3.2 Experiment Results

3.3.2.1 Data acquisition

The in-house data consisted of 21 metabolites with molecular weights ranging from 107 to 428 Da. For each metabolite, authentic compound was used to acquire the MS/MS spectra on an ultra performance liquid chromatography quadrupole time of flight (UPLC-QTOF, Waters) instrument. The collision energy was tuned for each individual compound to acquire MS/MS spectra with a reasonable number of fragments. For some of the metabolites, more than one spectrum was acquired. Totally, our in-house data comprised of 45 MS/MS spectra representing 21 metabolites. The MS/MS spectra for the same 21 metabolites were also retrieved from the HMDB database. The spectra were acquired using a high performance liquid chromatography triple quadrupole (HPLC-QqQ, Waters) instrument with collision energy of 10eV. For each metabolite, only one MS/MS spectrum was available in HMDB.

3.3.2.2 Experiment design

To evaluate the performance of different algorithms, a 3-fold cross validation was performed. The 21 metabolites were randomly divided into two groups. In the training set, we had spectra for 14 metabolites from both the in-house data and the HMDB database. In the testing set, we had spectra for 7 metabolites from the in-house data and the HMDB database. The purpose of this stratification is to make sure that we have approximately the same degree of data heterogeneity in both training and testing sets. The training set was used to train a model to discriminate positive samples and negative samples. For a typical scoring method that uses a single variable to measure spectral similarity, the model is a score threshold, which maximizes the F-measure in the training set. For the proposed method, the model is a SVM classifier trained on the training set. The trained model was applied to the testing set to evaluate the identification

performance. This procedure was repeated 100 times to measure the performance of algorithms. In addition to the four algorithms previously introduced, we used the Pearson correlation coefficient alone as a similarity score for performance comparison.

Because we were particularly interested in evaluating the performance of various spectral matching algorithms on heterogeneous datasets, we carried out two experiments. In Experiment I, we match the spectra from one source against the spectra from the other source. For example, the query spectra from the in-house data were matched against a library composed of spectra from HMDB, or vice versa. In Experiment II, spectra from different sources were mixed together to form a new dataset and a spectrum in the dataset was matched against the rest of the dataset.

Because the data are highly imbalanced with significantly more negative samples, accuracy alone is not enough to evaluate the identification performance. Thus we utilized F-measure notion from information retrieval context to measure the performances of the algorithms, which is defined as

$$F = \frac{2 * \text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}, \text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN} \tag{3.32}$$

where TP, FP, FN are the number of true positive, false positive, and false negative, respectively. In addition, the precision-recall curve and its area under the curve (AUC) are also used to evaluate the performance of the algorithms.

3.3.2.3 Experiment results

The Pearson correlation coefficients for positive and negative samples in Experiments I and II are shown in Figure 16 and Figure 17, respectively. These figures illustrate the necessity to induce more similarity measures in addition to correlation. The comparisons between different metabolites (negative samples) generally show very small correlation coefficients as expected. However, the correlation coefficients of the comparisons between the same metabolites (positive samples) span a large range. For some of them, the spectral profiles from different platforms and experiments are similar, while for others there is a large variation between the spectra.

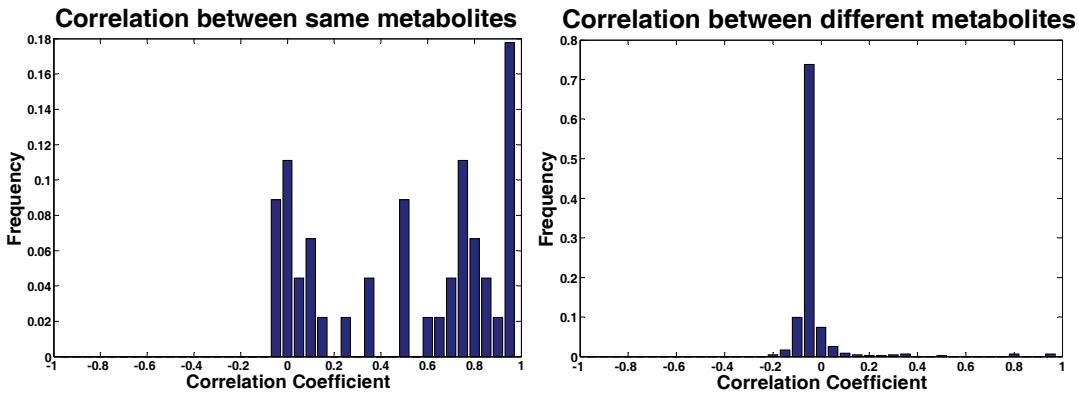


Figure 16 Correlation coefficients for comparison between same metabolites (left panel) and different metabolites (right panel) in Experiment I.

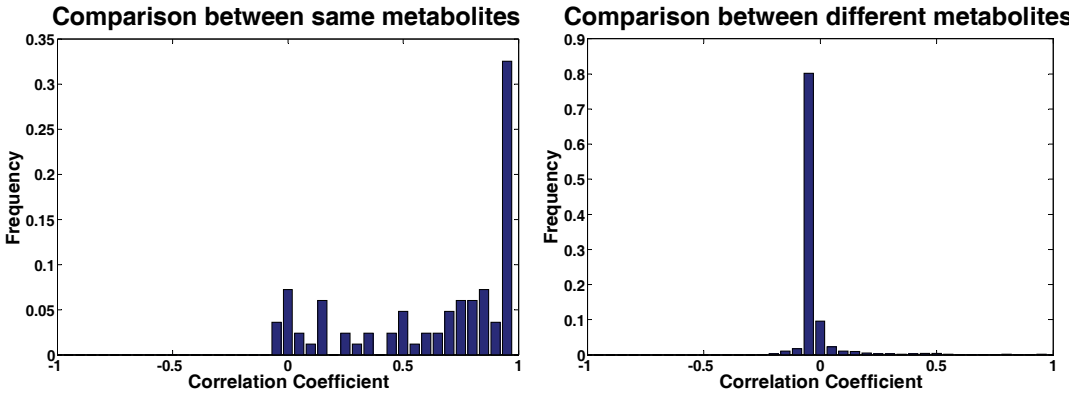


Figure 17 Correlation coefficients for comparison between same metabolites (left panel) and different metabolites (right panel) in Experiment II.

Table 10 presents the accuracy, F-measure and AUC of five spectral matching algorithms. Among the five algorithms, SVM gives the best performance on all three measures. While the accuracies of other algorithms are comparable, SVM achieved about 7% to 10% increase on F-measure. SVM also outperforms the other algorithms in terms of the AUC and precision-recall curve as shown in Table 10, Figure 18, and Figure 19.

Experiment	Method	F-measure (%)	Accuracy (%)	AUC (%)
I	NIST	74.0	93.8	79.7
	MassBank	72.1	93.3	83.5
	SpectraST	69.3	92.1	72.0
	Correlation	73.2	93.2	75.1
	SVM	**80.7**	**94.6**	**86.9**
II	NIST	77.7	95.3	84.3
	MassBank	77.3	95.2	86.1
	SpectraST	75.7	94.6	79.7
	Correlation	76.7	95.1	87.1
	SVM	**85.1**	**96.3**	**90.1**

Table 10 The performance of spectral matching algorithms

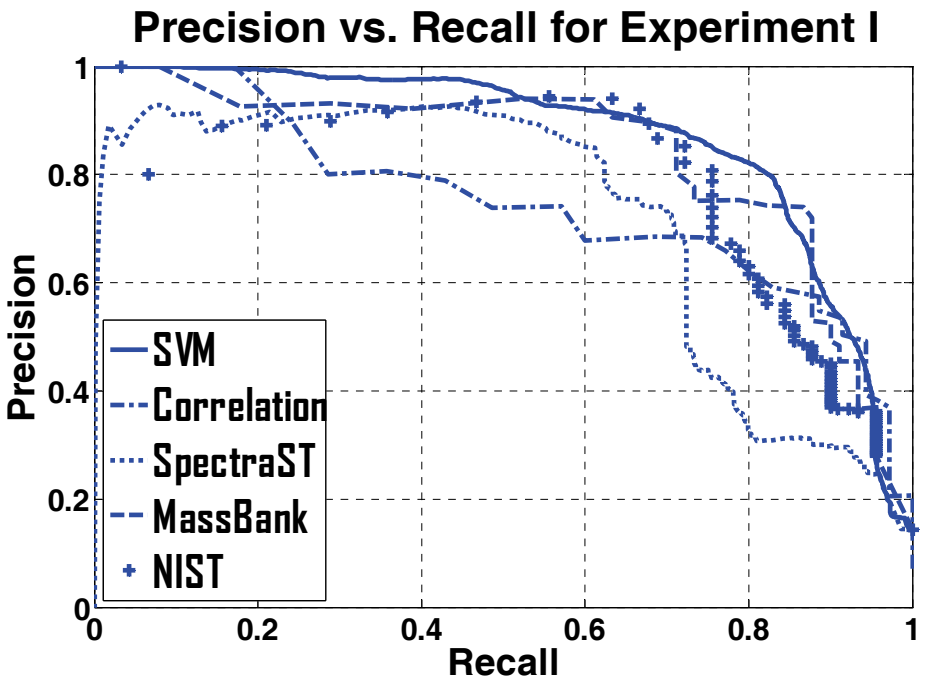


Figure 18 Precision-recall graph for the spectral matching algorithms in Experiment I

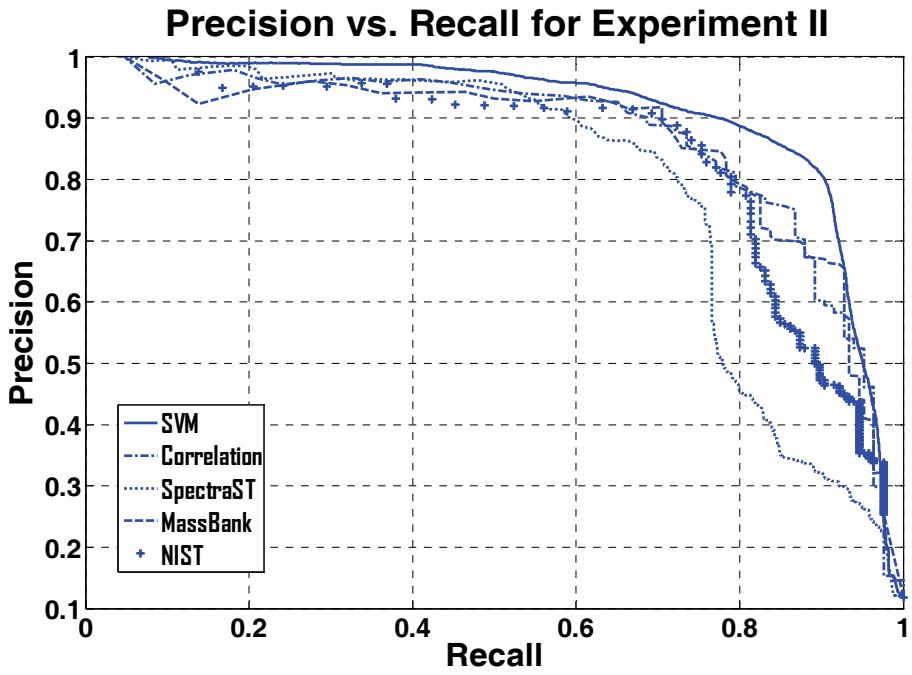


Figure 19 Precision-recall graph for the spectral matching algorithms in Experiment II

3.3.3 Discussion

We propose two metrics that measure the similarity of peaks present in two MS/MS spectra. The two metrics are combined with Pearson correlation coefficient through SVM. The proposed approach achieves more accurate spectral matching comparing with several existing algorithms, when the data are from difference sources and heterogeneous.

Chapter 4 Conclusion and Future Work

4.1 Conclusion

The identification of metabolites is the current bottle-neck in metabolomics studies. The most reliable way to unambiguously and confidently identify a metabolite is to compare its mass, retention time and fragmentation spectrum with those of authentic standards. As outlined in [30], at least two independent and orthogonal measurements analyzed under identical experiment conditions are needed to confirm the identity of a metabolite. However, various computational tools can often be utilized to reduce and prioritize the search space of metabolite identification, thereby improving the efficiency and reducing the cost. Although novel metabolites continue to be discovered, most of the metabolites we face in practice have already been found and identified in some other studies. The collection and utilization of information on these "known unknowns" poses major challenge for computational and informatics tools. This is in part because the information is often scattered in different sources in which spectra are acquired under different conditions.

In this thesis, several computational approaches are developed to improve the accuracy and efficiency of metabolite identification. First, a simple but effective outlier screening approach is proposed to identify low-quality LC-MS runs from a dataset. The approach does not depend on any preprocessing method, and can tolerate normal LC-MS data variations, while it is able to identify low-quality LC-MS runs with significantly less peaks in their TICs. Second, an integrated computational framework is proposed to exploit the mass, retention time and MS/MS spectrum information to acquire and prioritize the putative metabolite identifications for a selected subset of LC-MS peaks. The experimental results show that the proposed framework effectively improves the accuracy of metabolite identification and prioritizes the multiple putative identifications of peaks, which in turn saves the time and cost in the metabolite identification process. Third, an improved metabolite identification approach using MS/MS spectral matching is proposed. Both "peak similarity" and "profile similarity" between two MS/MS spectra are

measured. These features are combined through the non-linear classification capability of SVM. We demonstrate the ability of this approach to give more accurate identification of metabolites by comparing it with several existing spectral matching algorithms. We observe that the dot-product approach alone is not sufficient for identification in heterogeneous data. Our approach outperforms the existing spectral matching algorithms for metabolite identification, especially when there is a large degree of data heterogeneity when MS/MS spectra from different instruments are used at the same time.

4.2 Future work

Challenges exist in almost all the steps of metabolite identification. The quality of acquired MS/MS spectra is of great importance for successful identification. However, it is often affected by different experimental settings such as instrument type, collision energy, etc. A good quality MS/MS spectrum is often acquired through iterative adjustment of experimental parameters. Automated acquisition of high-quality MS/MS spectra is needed to increase the throughput of metabolite identification. Spectral libraries for GC-MS demonstrate great success in small molecule identification. LC-MS/MS spectral libraries for metabolomics studies such as HMDB, Metlin and MassBank continue to evolve and expand for increased metabolome coverage. In addition, the heterogeneity of spectral data poses a major challenge against the effective usage of spectral libraries. In some recent studies, promising results have been demonstrated to achieve certain degree of reproducibility of MS/MS spectra using different instruments from different laboratories [54, 55]. Through carefully designed experiments and appropriate spectral matching algorithms, compound identification with much higher accuracy can be achieved. Improved in-silico fragmentation models are needed to recognize complex ion-molecular interactions encountered in metabolites fragmentation. With improved specificity, such models will assist the identification of unknown metabolites with no spectral library coverage. In addition, since metabolites do not exist alone but within certain biological context such as metabolic networks and pathways, integration of contextual information into identification can potentially reduce the ambiguity in metabolite identification as illustrated in [56].

Reference

1. van der Greef, J. and A.K. Smilde, *Symbiosis of chemometrics and metabolomics: past, present, and future.* Journal of Chemometrics, 2005. **19**(5-7): p. 376-386.
2. Gates, S. and C. Sweeley, *Quantitative metabolic profiling based on gas chromatography.* Clin Chem, 1978. **24**(10): p. 1663-1673.
3. van der Greef, J., P. Stroobant, and R. van der Heijden, *The role of analytical sciences in medical systems biology.* Curr Opin Chem Biol, 2004. **8**(5): p. 559-65.
4. Khoo, S.H. and M. Al-Rubeai, *Metabolomics as a complementary tool in cell culture.* Biotechnol Appl Biochem, 2007. **47**(Pt 2): p. 71-84.
5. Goodacre, R., et al., *Metabolomics by numbers: acquiring and understanding global metabolite data.* Trends Biotechnol, 2004. **22**(5): p. 245-52.
6. Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.* Xenobiotica, 1999. **29**(11): p. 1181-9.
7. Robertson, D.G., *Metabonomics in toxicology: a review.* Toxicol Sci, 2005. **85**(2): p. 809-22.
8. Dettmer, K., P.A. Aronov, and B.D. Hammock, *Mass spectrometry-based metabolomics.* Mass Spectrom Rev, 2007. **26**(1): p. 51-78.
9. Sana, T.R., K. Waddell, and S.M. Fischer, *A sample extraction and chromatographic strategy for increasing LC/MS detection coverage of the erythrocyte metabolome.* Journal of Chromatography B, 2008. **871**(2): p. 314-321.
10. Theodoridis, G., H.G. Gika, and I.D. Wilson, *Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies.* Mass Spectrometry Reviews, 2011.
11. Kind, T. and O. Fiehn, *Metabolomics database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm.* BMC Bioinformatics, 2006. **7**: p. 234.
12. Cho, H., et al., *OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data.* Bioinformatics, 2008. **24**(6): p. 882-884.
13. Wang, W., et al., *Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards.* Analytical Chemistry, 2003. **75**(18): p. 4818-4826.
14. Bylund, D., *Chemometric tools for enhanced performance in liquid chromatography-mass spectrometry*, in *Faculty of Science, Technology and Media, Department of Natural Sciences*. 2001, Mid Sweden University: Uppsala. p. 40.
15. Smith, C.A., et al., *XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification.* Analytical Chem, 2006. **78**(3): p. 779 - 787.
16. Lange, E., et al., *Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements.* BMC Bioinformatics, 2008. **9**: p. 375.
17. Warrack, B.M., et al., *Normalization strategies for metabonomic analysis of urine samples.* Journal of Chromatography B, 2009. **877**(5-6): p. 547-552.
18. Kim, J.K., et al., *Time-course metabolic profiling in Arabidopsis thaliana cell cultures after salt stress treatment.* Journal of Experimental Botany, 2007. **58**(3): p. 415-424.

19. Hageman, J., M. Malosetti, and F. van Eeuwijk, *Two-mode clustering of genotype by trait and genotype by environment data.* Euphytica, 2010: p. 1-11.
20. Chen, C., et al., *Serum Metabolomics Reveals Irreversible Inhibition of Fatty Acid β-Oxidation through the Suppression of PPARα Activation as a Contributing Mechanism of Acetaminophen-Induced Hepatotoxicity.* Chemical Research in Toxicology, 2009. **22**(4): p. 699-707.
21. Enot, D., M. Beckmann, and J. Draper, *On the Interpretation of High Throughput MS Based Metabolomics Fingerprints with Random Forest*, in *Computational Life Sciences II*. 2006. p. 226-235.
22. Guan, W., et al., *Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines.* BMC Bioinformatics, 2009. **10**(1): p. 259.
23. Beecher, C.W.W., *THE HUMAN METABOLOME*, in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, R.G. George G. Harrigan, Editor. 2003, Springer. p. 352.
24. Wink, M., *Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores.* TAG Theoretical and Applied Genetics, 1988. **75**(2): p. 225-233.
25. Wishart, D.S., et al., *HMDB: the Human Metabolome Database.* Nucleic acids research, 2007. **35**(Database issue): p. D521 - 6.
26. Cui, Q., et al., *Metabolite identification via the Madison Metabolomics Consortium Database.* Nature Biotechnology, 2008. **26**(2): p. 162-164.
27. Smith, C.A., et al., *METLIN: A Metabolite Mass Spectral Database.* Therapeutic Drug Monitoring, 2005. **27**(6): p. 747-751.
28. Go, E., *Database Resources in Metabolomics: An Overview.* Journal of Neuroimmune Pharmacology, 2010. **5**(1): p. 18-30.
29. Giavalisco, P., et al., *13C Isotope-Labeled Metabolomes Allowing for Improved Compound Annotation and Relative Quantification in Liquid Chromatography-Mass Spectrometry-based Metabolomic Research.* Analytical Chemistry, 2009. **81**(15): p. 6546-6551.
30. Sumner, L., et al., *Proposed minimum reporting standards for chemical analysis.* Metabolomics, 2007. **3**(3): p. 211-221.
31. Schulz-Trieglaff, O., et al., *Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments.* BioData Mining, 2009. **2**(1): p. 4.
32. Na, S. and E. Paek, *Quality Assessment of Tandem Mass Spectra Based on Cumulative Intensity Normalization.* Journal of Proteome Research, 2006. **5**(12): p. 3241-3248.
33. Flikka, K., et al., *Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering.* Proteomics, 2006. **6**(7): p. 2086-2094.
34. Tautenhahn, R., C. Böttcher, and S. Neumann, *Annotation of LC/ESI-MS Mass Signals*, in *Bioinformatics Research and Development*, S. Hochreiter and R. Wagner, Editors. 2007, Springer Berlin / Heidelberg. p. 371-380.
35. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra.* BMC Bioinformatics, 2010. **11**(1): p. 148.
36. Cui, Q., et al., *Metabolite identification via the Madison Metabolomics Consortium Database.* Nat Biotech, 2008. **26**(2): p. 162-164.

37. 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.
38. McNaught, A.D. and A. Wilkinson, *IUPAC. Compendium of Chemical Terminology*. 2nd ed. 1997, Oxford: Blackwell Science.
39. Huang, N., et al., *Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data.* Journal of the American Society for Mass Spectrometry, 1999. **10**(11): p. 1166-1173.
40. Keller, B.O., et al., *Interferences and contaminants encountered in modern mass spectrometry.* Analytica Chimica Acta, 2008. **627**(1): p. 71-81.
41. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome.* Nucl. Acids Res., 2009. **37**(suppl_1): p. D603-610.
42. Fahy, E., et al., *LIPID MAPS online tools for lipid research.* Nucleic acids research, 2007. **35**(suppl 2): p. W606-W612.
43. Pervukhin, A., *Molecular formula identification using high resolution mass spectrometry: algorithms and applications in metabolomics and proteomics*. 2009, Friedrich-Schiller-Universität Jena.
44. Hisayuki, H. *Comparison of ESI-MS Spectra in MassBank Database*. 2008.
45. Henry, L., et al., *Development and validation of a spectral library searching method for peptide identification from MS/MS.* PROTEOMICS, 2007. **7**(5): p. 655-667.
46. Stein, S.E., *Estimating probabilities of correct identification from results of mass spectral library searches.* Journal of the American Society for Mass Spectrometry, 1994. **5**(4): p. 316-323.
47. Stein, S.E. and D.R. Scott, *Optimization and testing of mass spectral library search algorithms for compound identification.* Journal of the American Society for Mass Spectrometry, 1994. **5**(9): p. 859-866.
48. Lars, L., *Visual Analysis of Gel-Free Proteome Data.* IEEE Transactions on Visualization and Computer Graphics, 2006. **12**: p. 497-508.
49. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer series in statistics. 2001: Springer. 533.
50. Burges, C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121-167.
51. Musicant, D.R., V. Kumar, and A. Ozgur. *Optimizing F-Measure with Support Vector Machines*. in *PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE*. 2003: Haller AAAI Press.
52. Nguyen, G.H., A. Bouzerdoum, and S.L. Phung, *Learning Pattern Classification Tasks with Imbalanced Data Sets, Pattern Recognition*, in *Pattern Recognition*, P.-Y. Yin, Editor. 2009, INTECH. p. 193-208.
53. Tang, Y., et al., *SVMs modeling for highly imbalanced classification.* Trans. Sys. Man Cyber. Part B, 2009. **39**(1): p. 281-288.
54. Hopley, C., et al., *Towards a universal product ion mass spectral library – reproducibility of product ion spectra across eleven different mass spectrometers.* Rapid Communications in Mass Spectrometry, 2008. **22**(12): p. 1779-1786.

55. Oberacher, H., et al., *On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study.* Journal of Mass Spectrometry, 2009. **44**(4): p. 485-493.
56. Rogers, S., et al., *Probabilistic assignment of formulas to mass peaks in metabolomics experiments.* Bioinformatics, 2009. **25**(4): p. 512-518.