DATA ANALYSIS WORKFLOW FOR GAS CHROMATOGRAPHY MASS SPECTROMETRY-BASED METABOLOMICS STUDIES

By

Yan Ni

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2014

Approved by:

_____
Dr. Xiuxia Du

_____
Dr. Anthony Fodor

_____
Dr. Brian Cooper

_____
Dr. Cynthia Gibas

_____
Dr. Pinku Mukherjee

ABSTRACT

YAN NI. Data analysis workflow for gas chromatography mass spectrometry-based metabolomics studies. (Under the direction of DR. XIUXIA DU)

Metabolomics has emerged as an integral part of systems biology research that attempts to comprehensively study low molecular weight organic and inorganic metabolites under certain conditions within a biological system. Technological advances in the past decade have made it possible to carry out metabolomics studies in a high-throughput fashion using gas chromatography coupled with mass spectrometry. As a result, large volumes of data are produced from these studies and there is a pressing need for algorithms that can efficiently process and analyze the data in a high-throughput fashion as well. To address this need, we have developed computational algorithms and the associated software tool named an Automated Data Analysis Pipeline (ADAP). ADAP allows data to flow seamlessly through the data processing steps that include de-nosing, peak detection, deconvolution, alignment, compound identification and quantitation. The development of ADAP started in 2009 and the past four years have witnessed continuous improvements in its performance from ADAP-GC 1.0, to ADAP-GC 2.0, and to the current ADAP-GC 3.0. As part of the performance assessment of ADAP-GC, we have compared it with three other software tools. In this dissertation, I will present the computational details about these three versions of ADAP-GC, the capabilities of the software tool, and the results from software comparison.

DEDICATION

To my family and teachers who introduced me into science.

ACKNOWLEDGEMENTS

My greatest appreciation goes to my advisor, Dr. Xiuxia Du for her invaluable guidance, patience, and support in my dissertation research over the past five years. She provided me with a fertile ground and free academic environment for my research, and helped me transition from a pharmaceutics background to the world of bioinformatics.

I would also like to thank members in our research group for their enormous help in many ways. Dr. Wenxin Jiang, who was the first lab member working on the ADAP project, brought me into the world of R programming and helped me learn the details about his work. Kyle Suttlemyre and Peter Pham, who were the primary force behind the ADAP software development, collaborated closely with me implementing the computational algorithms that Dr. Jiang and I developed.

I would also like to thank the research group of Dr. Wei Jia. Dr. Jia was the primary collaborator when we started developing ADAP and provided our research with raw mass spectrometry data from his metabolomics research and his in-house spectral library of standard compounds. My thanks also go to Dr. Yunping Qiu, a staff scientist in Dr. Jia's group, whose expertise in chemistry and his insight on metabolomics data helped us tremendously in our efforts to find the solution to a number of computational hurdles and to improve the overall user-friendliness of the ADAP software.

I most sincerely thank my dissertation committee members, Dr. Anthony Fodor, Dr. Brian Cooper, Dr. Cynthia Gibas, and Dr. Pinku Mukherjee for spending their precious time on meeting with me and providing invaluable suggestions and ideas for me to continuously make progress on my research.

Finally, I thank the Department of Bioinformatics and Genomics for providing me with the opportunity to pursue Ph.D. in Bioinformatics and the Graduate Assistant Support Plan (GASP) for providing me with the financial support in the past years. My appreciation also goes to all of the staff members in my department and the international students and scholars office for their friendly help.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background of Metabolomics

1.1.1 What is Metabolomics?

System biology is focused on the study of biological components and their complex interaction to define emergent properties of biological systems [1, 2]. Metabolomics (also known as metabonomics) has emerged as an integral part of systems biology research that attempts to comprehensively study low molecular weight organic and inorganic metabolites (typically <1,500 Da) under certain conditions within a biological system [3, 4]. In parallel to the terms 'transcriptome' and 'proteome', the set of metabolites synthesized by a biological system constitute its 'metabolome' [3]. Metabolites are regarded as the end products of cellular regulatory processes, and the changes in their levels in cells, blood, or tissues reflect the ultimate response of biological systems to diseases, genetic changes, or environmental perturbations.

1.1.2 Applications of Metabolomics

Metabolomics is a small-molecule-based science in the “omics” field, which enables the dynamic and holistic measurement of endogenous metabolites in the biological systems in response to genetic or environmental changes [5]. Compared to other ‘omics’, metabolome is closest to the phenotype of a biological system. Monitoring the metabolome using metabolomics technologies allows a quick assessment of the overall system status (normal or abnormal) [6] and facilitates disease diagnosis [7-10]. In

the meantime, environmental, developmental, or genetic perturbations can cause changes in the identity and quantity of metabolites along the metabolic pathways [11]. Metabolomics allows researchers to capture these metabolic changes and then study the biochemical mechanisms of diseases, develop effective drugs [12-14], and carry out toxicology research [15, 16]. Lastly, metabolomics makes it possible to comprehensively assess nutritional status, which is becoming increasingly essential as our society realizes the importance of nutrition to our health and disease prevention [17, 18].

1.1.3 Metabolomics Approaches and Workflow

A metabolomics study typically involves five steps: study design, sample collection and storage, sample preparation and analysis, data processing and analysis, and final biochemical pathway analysis and interpretation [19]. In general, metabolomics experiments can be subdivided into targeted analyses and untargeted analyses [3]. Targeted metabolomics involves accurate quantitation of a list of metabolites from related metabolic pathways of interest, whereas untargeted analyses use a more global approach to measure as many metabolites as possible without bias [20]. Sometimes, metabolic fingerprinting is used to consider the total metabolic profile as a unique pattern characterizing a snapshot of the metabolism in a particular cell line or tissue [21].

Regardless of the approach that is used, the metabolomics workflow generally consists of six steps as shown in Figure 1.1.

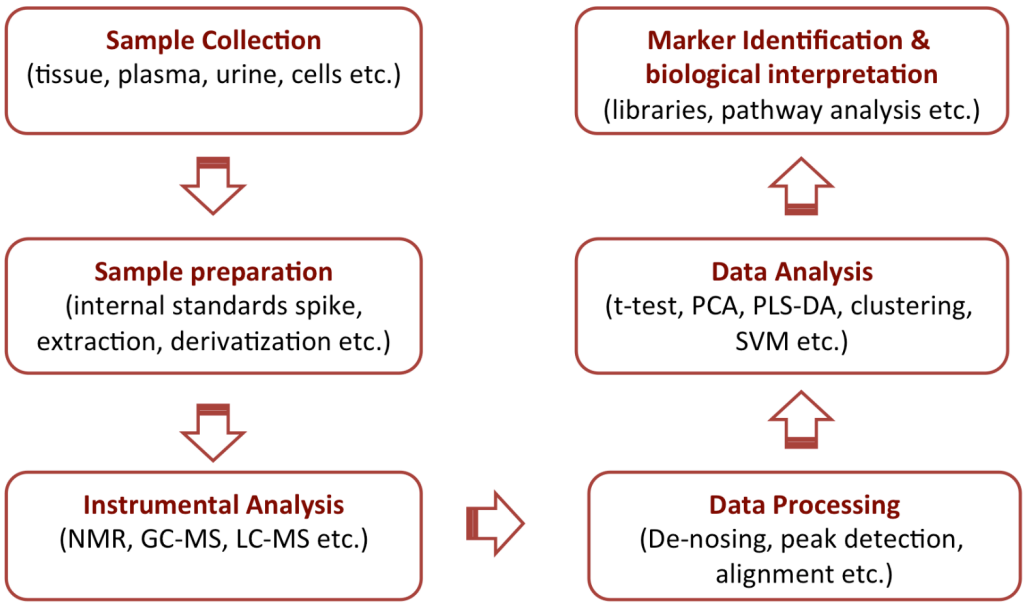

Figure 1.1. The general workflow of a metabolomics study

1.1.4 Analytical Platforms for Metabolomics

Many analytical platforms have been used for metabolomics studies [22], including nuclear magnetic resonance (NMR) spectroscopy, direct infusion mass spectrometry (MS), Fourier transform infrared (FT-IR) spectroscopy, gas chromatography coupled to mass spectrometry (GC-MS), two-dimensional GC coupled to MS (GCxGC-MS), liquid chromatography coupled to MS (LC-MS), and capillary electrophoresis coupled to MS (CE-MS) [19]. The advantages of NMR include high reproducibility, potential for high-throughput fingerprinting, minimal requirement for sample preparation, and non-destructive nature [23]. The disadvantages of NMR spectroscopy, however, are also obvious: first, the technique shows relatively low sensitivity; second, it consumes relatively large sample materials when compared to MS. In contrast, MS-based metabolomics method is highly sensitive, which makes it the method of choice for studies that involve the identification and quantitation of low-

concentration metabolites from complex samples.

For MS-based metabolomics studies, both GC-MS and LC-MS are usually required to obtain a good coverage of the metabolome. The GC-MS platform is well suited for metabolites that are volatile and thermally stable, provides high chromatographic resolution and permits separation of structurally similar compounds [24]. The LC-MS platform is not limited by sample volatility and thermal stability and can analyze groups of compounds that are not amenable to GC-MS. Figure 1.2 shows the complementary nature of these two platforms. Primary mass analyzers that are coupled to GC separation are quadrupole and TOF instruments. Compared to conventional GC-MS, GC-time-of-flight-MS has been one commonly used platform in metabolite profiling experiments, providing rapid metabolite detection with high mass accuracy, fast scan speed, and high mass resolution to increase laboratory throughput [25]. Primary mass analyzers that are coupled to LC separation include quadruple, TOF and Orbitrap.

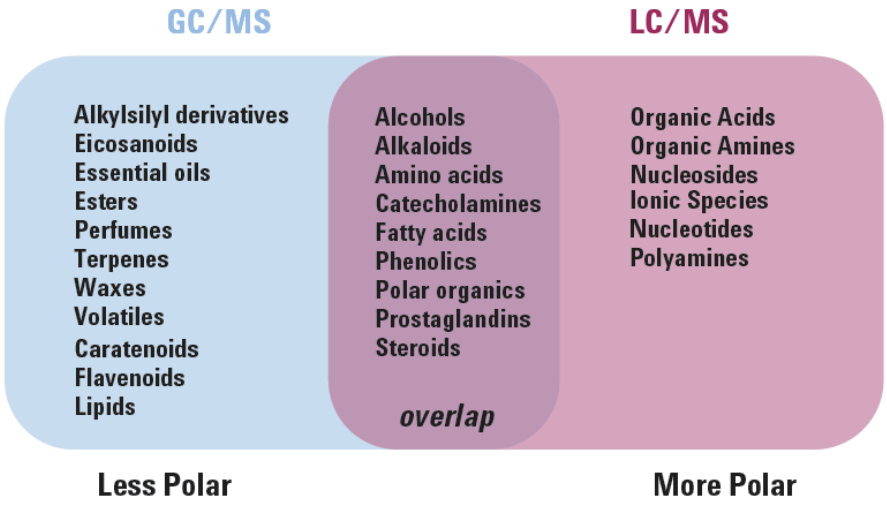

Figure 1.2. Coverage of identified metabolites using GC-MS and LC-MS platforms.

1.2 Metabolomics Bioinformatics

High-throughput mass spectrometry-based metabolomics studies usually generate very complex and large volume of data. In particular, running large-scale projects with hundreds to thousands of samples from clinical or epidemiology studies is on the verge of becoming routine. As a result, automated computational algorithms and software tools are necessary for extracting metabolite information from the raw mass spectrometry data and for making sense of the data. This need gave birth to the field of metabolomics bioinformatics. Specifically, the last three sequential steps in Figure 1.1 including data processing, data analysis, and data interpretation constitute the core of metabolomics bioinformatics research. Data processing extracts the qualitative and quantitative information of metabolites from the raw MS data. Data analysis determines statistically significant metabolites and identifies patterns of metabolite changes based on the quantitative metabolite information. Data interpretation places the metabolite information in the context of biological pathways using online databases such as KEGG (http://www.kegg.jp), human metabolome database (HMDB) (http://www.hmdb.ca), as well as available literature.

In the past decade, many free software tools have been developed that can handle one, two, or all of the three aforementioned bioinformatics tasks. These tools include AMDIS [26], XCMS [27, 28], MZmine [29, 30], MetAlign [31], MetaboAnalyst [32], MeltDB [33], MetaQuant [34], MathDAMP [1], MAVEN [35], MetabolomExpress [36], MetaboliteDectetor [37], MetIDEA [38], MetDAT [39], TargetSearch [40], and TagFinder [41, 42]. Among these, AMDIS, MetIDEA, MAVEN, spectconnect, TargetSearch and TagFinder are primarily for data processing; XCMS, MZmine and

MeltDAT provide some basic statistical analysis modules in addition to data processing; MetaboAnalyst and MeltDB primarily carry out data analysis based on the qualitative and quantitative metabolite information that is extracted by other software tools (e.g., XCMS and AMDIS). In addition to these free software tools, commercial tools have also been developed that include Mass Profiler Pro (Agilent), MarkerLynx (Waters), ChromaTOF (LECO), and AnalyzerPro (SpectralWorks). Despite the successful applications of these software tools, various limitations exist. These limitations include low throughput, inaccuracy in the extraction of qualitative and quantitative information, and incomplete workflow. In particular, any inaccuracy in the extracted qualitative and quantitative metabolite information will cause misleading results in the subsequent data analysis and interpretation. High throughput metabolomics studies call for a fully automated computational workflow that can handle all of the three data handling steps in an equally high-throughput fashion and overcome the existing limitations.

1.3 Algorithms Development, Implementation and Integration

My dissertation research focuses on developing novel algorithms for comprehensively processing and analyzing GC-MS data in metabolomics studies with the goal to improve the accuracy of metabolite identification and quantitation. This dissertation research is a part of the overarching ADAP bioinformatics system for mass spectrometry-based metabolomics studies. Specifically, an automated data analysis pipeline has been developed with full capabilities of de-noising, peak detection, deconvoluion, alignment, and compound identification and quantitation. Among them, peak detection and deconvolution are two critical steps that have witnessed continuously optimization and improvements in compound identification and quantitation from

ADAP-GC 1.0, to ADAP-GC 2.0, and to the current ADAP-GC 3.0. As part of the performance assessment of ADAP-GC 3.0, we have compared it with three existing software tools, i.e., ChromaTOF, AMDIS, and AnalyzerPro. Finally, the data processing algorithms have been integrated with visualization and statistical analysis package together to provide an automated and integrated software tool for users (Figure 1.4).

In the next three chapters, chapter two focuses on the development of novel computational algorithms for GC-MS data processing and methodological advances in each release of ADAP-GC. ADAP-GC 1.0 and 2.0 has been published in 2010 [43] and 2012 [44], respectively. The current version ADAP-GC 3.0 optimized based on previous versions will be submitted for review. Chapter three focuses on comparative evaluation of ADAP-GC 3.0 with three existing software tools in terms of compound identification and quantitation. Finally, chapter four introduces the development of visualization software and statistical methods for GC-MS data analysis in metabolomics studies. These two manuscripts on comparative evaluation with existing algorithms (chapter three) and ADAP-GC software (chapter four) will be submitted for review soon.

Table 1.1. Comparison of representative software tools for GC-MS data analysis

Software	Decon	Other capabilities for GC/MS data analysis	Limitations
AMDIS [26]	Yes	Peak detection, identification and interactive visualization.	Without alignment and statistical analysis.
MZmine 2 [29]	No	Smoothing and filtering, peak detection, alignment, statistical analysis, interactive visualization and identification.	Mainly developed for LC/MS data.
MetAlign [31]	No	Baseline correction, peak picking, filtering, alignment, and univariate analysis.	Without interactive visualization and multivariate analysis methods.
MetaboAnalyst 2.0 [45]	No	Statistical analysis, functional enrichment analysis, pathway analysis, visualization and report generation.	Use XCMS package for data processing.
MeltDB [33]	No	Identification, statistical analysis, visualization and pathway analysis.	Use third party software for data processing.
MetaboliteDetector [37]	Yes	Data filtering, baseline and noise analysis, peak detection, deconvolution, compound identification, and relative quantitation, alignment and visualization.	Without statistical analysis
MET-IDEA [38]	No	Automatic peak alignment, annotation and integration.	Mainly developed for quantitative analysis and utilize the output from AMDIS
Spectconnect [46]	No	Conserved metabolite identification without reference library.	Use AMDIS for data processing, and without interactive visualization and statistical analysis.
TargetSearch [40]	No	Baseline correction, peak detection, retention time correction and identification.	Without statistical analysis and interactive visualization.
TagFinder [42]	No	Retention index calculation, alignment and identification.	Without peak picking, statistical analysis and interactive visualization.

Table 1.1 (Continued)

XCMS online [47]	No	Feature detection, retention time correction, alignment, annotation, statistical analysis, and data visualization.	Mainly developed for LC/MS data and without interactive visualization.
ChromaTOF	Yes	Baseline correction, peak finding, deconvolution, alignment, compound identification and quantitation, user-friendly interface	Commercial, without statistical analysis methods
AnalyzorPro	Yes	Baseline correction, smoothing, peak finding, deconvolution, compound identification and quantitation, user-friendly interfaces	Commercial, without alignment, statistical analysis methods

Note: Decon is short for deconvolution.

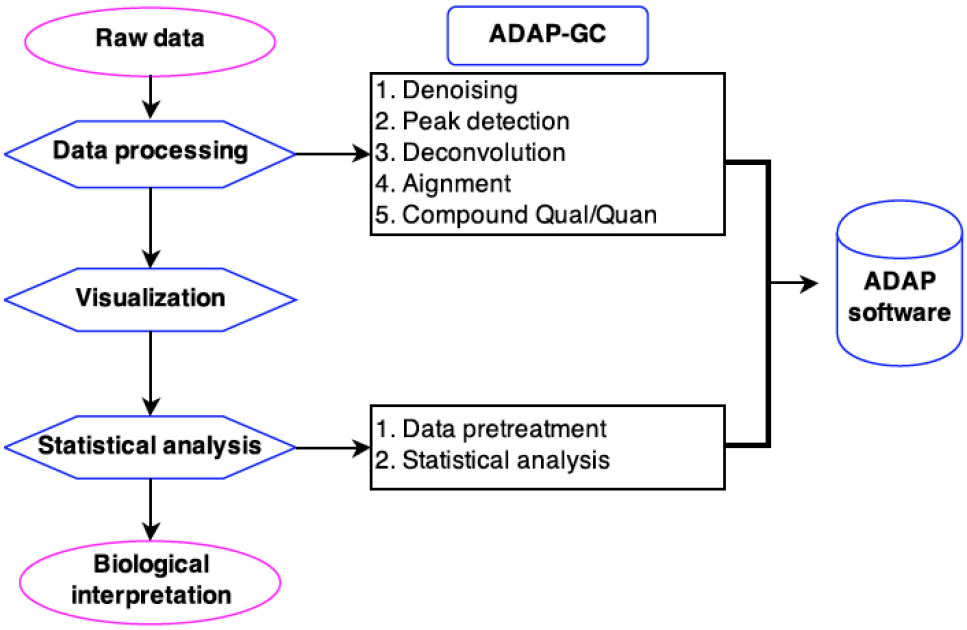

Figure 1.4. The stucture of integreating data processing algotithms, visualization, and statitstical analysis together for ADAP-GC pipeline.

CHAPTER 2: DEVELOPMENT OF COMPUTATIONAL ALGORITHMS FOR PROCESSING GC-MS DATA

2.1 Introduction

Processing of GC-MS-based metabolomics data involves five steps including de-nosing, peak detection, spectral deconvolution, chromatogram alignment, and compound identification and quantitation [48]. The MS signals resulting from GC-MS measurements can be contaminated by different sources of technical variations that de-noising makes it possible to remove the random noises from signals [49]. Peak detection aims to detect all the peaks with different peak widths and shapes in total ion current chromatogram (TIC) and each extraction ion current chromatogram (EIC). A peak is an observed, temporal, and bell-shaped signal intensity pattern in the chromatogram and is numerically represented by one peak apex, one left and one right boundary, and the signal intensity pattern between the two boundaries. Deconvolution is a critical process for extracting pure mass spectrum of a same compound for identification and quantitation, which is particularly challenging in GC-MS where compounds with similar biochemical properties frequently co-elute from GC column that produce a mass spectrum with a mixture of fragments from multiple compounds (Figure 2.1). In GC-MS-based metabolomics, retention time variations/drifts are common in chromatography due to the variations in column performance or column overloading with sample especially when a large number of samples are analyzed [50]. Thus, after deconvolution, alignment is

important to align peaks originating from a same metabolite to an identical retention time. Finally, the combination of compound identification and quantitation results enables further statistical analysis and biological interpretation.

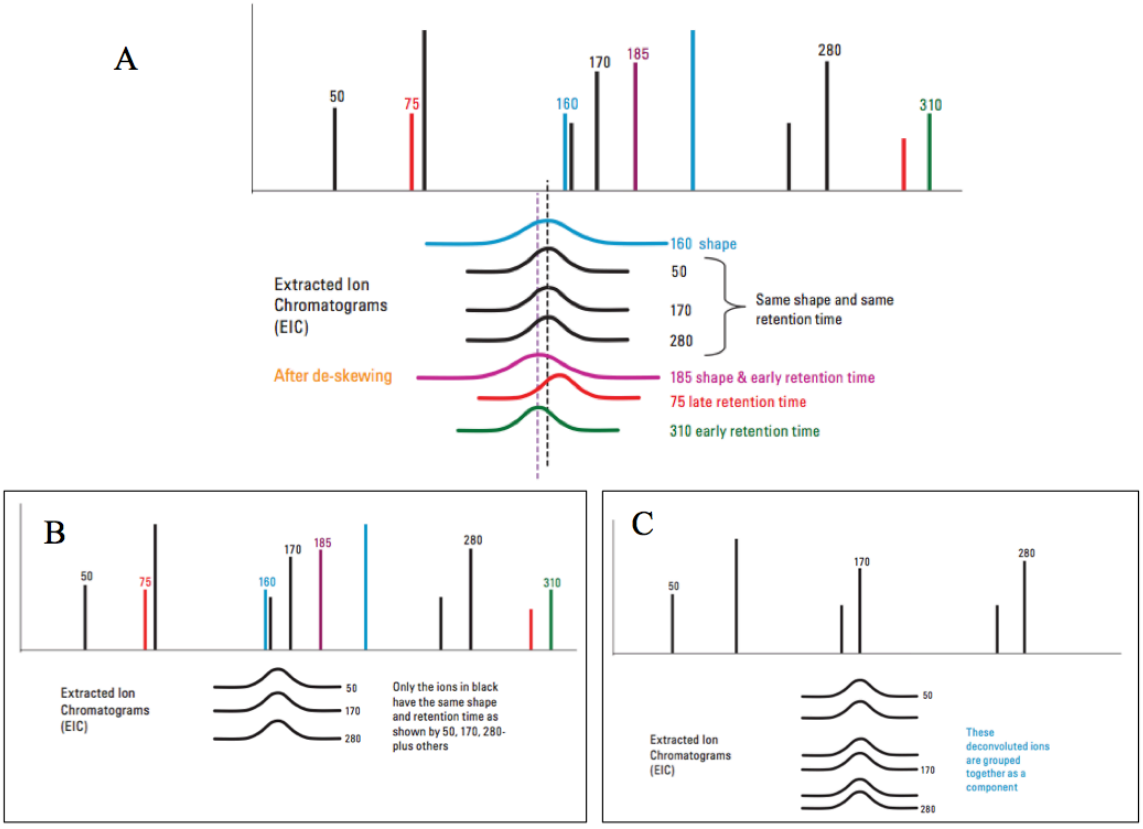

Figure 2.1. A simplified diagram of extracting ion chromatograms of a compound in the process of deconvolution. (**A**) A raw mass spectrum consists of fragments from mulitple compounds. (**B**) EICs from a same compound have similar chromatograms in terms of similar peak shapes and RTs. (**C**) Deconvolution is a process to extract fragment ions with similar peak shapes and RTs that correspond to a same compound. This figure is from the Agilent website http://www.chem.agilent.com/Library/applications/5990-5052EN.pdf.

The success to identify and quantify compounds from complex biological samples lies in the robustness and accuracy of data processing. Unfortunately, the development of data processing algorithms is slow due to the technical challenges and complexities, thus most of existing software tools that have been applied for GC-MS data analysis in

metabolomics do not have comprehensive capabilities of data processing. For example, MetabAnalyst, MetIDEA, and Spectconnect focus only on quantitation and statistical analysis so that they have to rely on third party software (e.g., AMDIS and XCMS) for data processing when dealing with raw MS data. Furthermore, only four individual software tools have their own capability of spectral deconvolution (AMDIS, AnalyzerPro, ChromaTOF, and MetaboliteDectctor) (Table 1.1). A comparative evaluation work of the deconvolution performance of AMDIS, ChromaTOF, and AnalyzerPro [51] has discussed that none of existing programs met the challenges and needs for metabolomics, and thus called for more efficient, automated, flexible and reliable data handling systems.

In this chapter, we present the novel computational algorithms that we have developed for extracting the qualitative and quantitative metabolite information from GC-MS metabolomics data. Among the five steps of data processing,, peak detection and deconvolution are the two most critical components and have witnessed continuous optimization and improvements in compound identification and quantitation from ADAP-GC 1.0, to ADAP-GC 2.0, and to the current ADAP-GC 3.0. Next, the computational algorithms of data processing in each version of ADAP-GC pipeline are introduced and compared in detail.

2.2 Experimental Procedures and Testing Datasets

2.2.1 Testing Datasets

(1) Calibration curve (CC) samples for testing ADAP-GC 1.0. Ten calibration curve samples were prepared at different dilutions from the original mixture of 20 fatty

acid standards (Table 2.3). These fatty acids vary in biochemical properties with different number of carbons and double bonds, or different double bond positions.

(2) Liver injury (LI) samples for testing ADAP-GC 1.0. Serum samples were collected from male Sprague-Dawley rats. Ten rats had acute liver injury and 10 served as healthy control.

(3) Mixture of standard compounds (Sample I): a total of 38 standard compounds were carefully selected and mixed together with known ratios (Table 2.1). Criteria for selecting those compounds are: (i) They should contain different classes of compounds that include amino acids, organic acids, fatty acids, polyamines, and ketones; (ii) They should be common in human urine or blood samples, and (iii) The retention times of compounds are spaced across the entire 30-minute time range. Both ADAP-GC 1.0 and ADAP-GC 2.0 applied the datasets of sample I to evaluate their performance in metabolite identification.

(4) Mixture of standard compounds (Sample II): seven calibration curve samples with each containing 27 standard compounds were prepared at different concentrations (0.2, 0.4, 0.6, 0.8, 1, 2 and 5 μg/ml of each compound). We designed this sample sets carefully, requiring standards from different compound classes spaced across 30 minutes of the entire elution time span and having 4 pairs of co-eluting compounds. This enabled us to evaluate the performance of ADAP-GC 2.0 and 3.0 in terms of identifying and quantifying different classes of compounds and co-eluting compounds.

(5) Urine samples with standard mixtures spiked in (Sample III): sample III was prepared by spiking each of the seven calibration curve samples of Sample II and an additional sample consisting of 0.1 μg/ml of each standard compound into a pooled urine

sample. Sample III was used for evaluating the performance of ADAP-GC 2.0 and 3.0 in terms of processing complex samples.

Table 2.1. List of 38 standard compounds in mixed standards samples.

No	Compounds	ET	Injection (ng)	NIST Score
1	L-α-alanine	6.2908	70	949
2	L-leucine	8.51	105	896
3	L-Proline	8.8917	94	762
4	Glycine	9.0025	141	957
5	Succinic acid	9.13	94	924
6	L-Serine	9.721	94	772
7	pipecolinic acid	9.815	141	886
8	beta-Alanine	10.645	188	953
9	4-Hydroxy-L-proline	11.8733	94	700
10	trans-Cinnamic acid	12.2384	188	866
11	L-cysteine	12.248	94	0
12	creatinine，anhydrous	12.248	141	937
13	α-Ketoglutaric acid	12.5142	188	933
14	L-asparate	12.605	188	954
15	L-Phenylalanine	13.1708	141	604
16	n-Dodecanoic acid	13.5467	141	864
17	L-(+)-Arabionse	13.5758	141	736
18	DL-Homocysteine	13.6375	281	822
19	L-Asparagine	14.2058	188	797
20*	L-(+)-Rhamnose monohydrate	14.3375 14.4433	141	910 876
21	L-(-)-arabitol	14.3833	234	952
22	1,4-diaminobutane	14.675	117	938
23	L－Ornithine monohydrochloride	15.8158	234	718
24	1,5-Diaminopentane dihydrochloride/Cadaverine	16.13	281	947
25	n-Tetradecanoic acid(myristic acid)	16.3825	141	548
26*	D-Fructose	16.6117 16.7575	141	940 943
27	L-Histidine	17.3875	281	766
28	Indol-3-acetic acid	18.1075	188	925
29	Palmitic acid/hexadecanoic acid	19.5183	234	938
30	Dopamine hydrochloride	19.9408	234	946
31	3-indolepropionic acid	20.0175	117	938
32	Oleic acid	22.2967	281	946
33	n-Octadecanoic acid	22.7617	469	956
34	uridine	25.36	375	803

35	n-Eicosanoic acid	25.3725	703	912

Table 2.1 (continued)

36	Sucrose	26.7083	234	950
37	estradiol	27.1392	375	955
38*	Testosterone	27.3892 27.4458	938	967 973
	Average			855

Note: Searched against in-house library

2.2.2 GC-TOF-MS Instrument Analysis

Five different types of samples were analyzed on a GC-TOF-MS platform. All the standard mixtures, serum and urine samples were prepared, derivatized, and analyzed following previously published protocols [52, 53]. Briefly, after TMS derivatization, each 1 μL aliquot of the derivatized solution was injected in splitless mode into an Agilent 6890N GC system (Santa Clara, CA, USA) that was coupled with a Pegasus HT TOF-MS (LECO Corporation, St. Joseph, MI, USA). Separation was achieved on a DB-5 ms capillary column (30 m × 250 μm I.D., 0.25-μm film thickness; Agilent J&W Scientific, Folsom, CA, USA), with helium as the carrier gas at a constant flow rate of 1.0 ml/min. The temperature of injection, transfer interface, and ion source was set to 260°C, 260°C, and 210°C, respectively. The GC temperature programming was set to 2 min isothermal heating at 80°C, followed by 10°C/min oven temperature ramps to 220 °C, 5 °C/min to 240°C, and 25°C/min to 290 °C, and a final 8 min maintenance at 290°C. Electron impact ionization (70 eV) at full scan mode (m/z 40-600) was used, with an acquisition rate of 20 spectra/second in the TOF-MS setting.

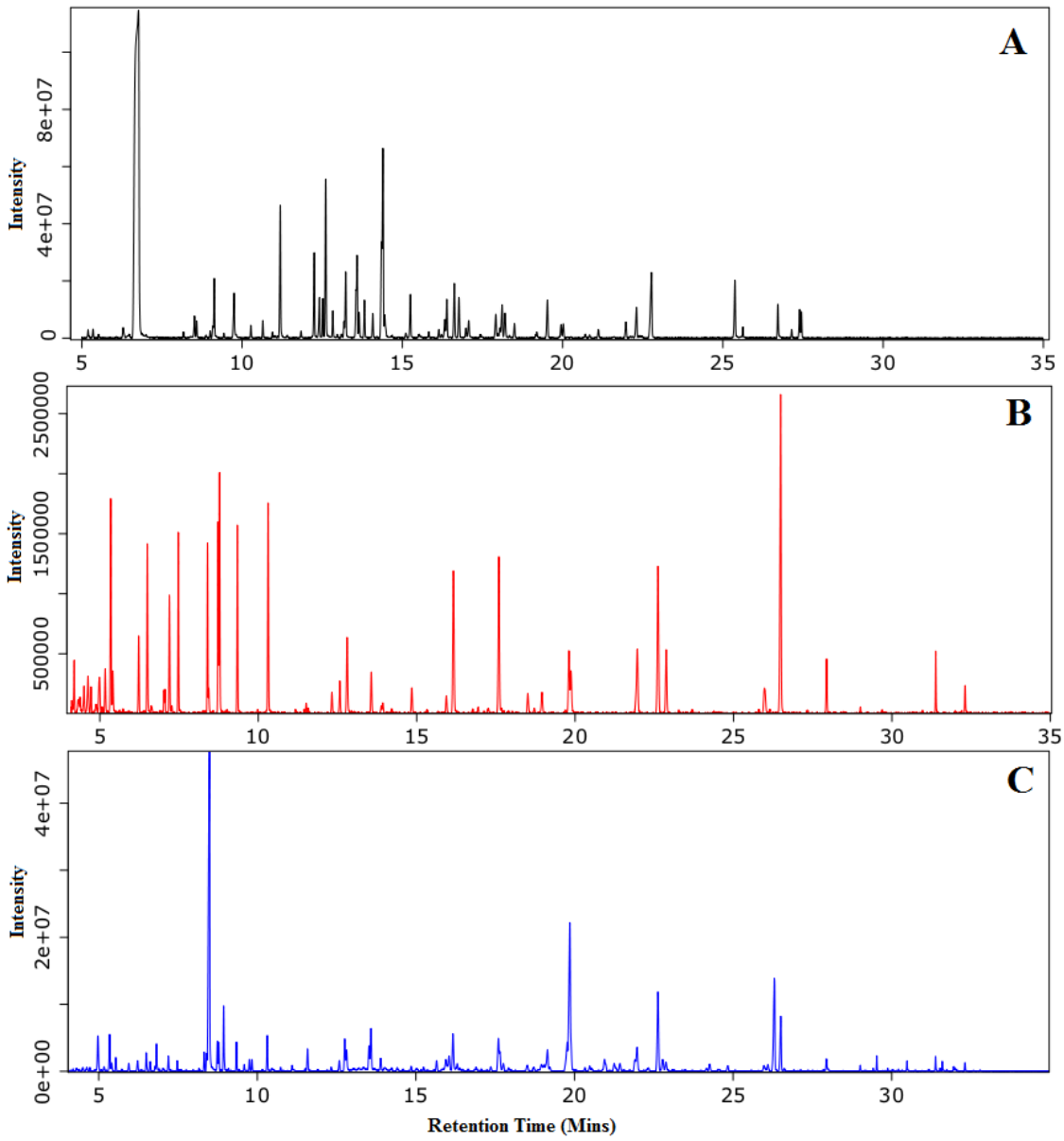

Figure 2.2. The TICs of selected three datasets from Sample I (A), II (B) and III (C), respectively.

2.3 Computational Algorithms for GC-MS Data Processing

The raw data consisting of original mass spectra and chromatogram information is exported as NetCDF format from GC-TOF-MS platform after sample analysis. The TICs of three representative samples are illustrated in Figure 2.2. In order to extract pure mass

spectra of compounds for their identification and quantitation, deconvolution is one of the most critical steps in data processing. De-noising and peak detection are two prerequisite steps in order to remove noises and reduce their interferences and to then detect peak apex and boundaries of all the peaks in both TIC and EIC. After deconvolution, alignment is performed to correct retention time shifts of a same compound among different samples. With the goal to develop a fully integrated and robust pipeline, the computation algorithms have been witnessed the continuous progress and improvement from ADAP-GC 1.0, ADAP-GC 2.0 to the current ADAP-GC 3.0 (Figure 2.3). Next, computational algorithms to extract pure mass spectra (de-noising, peak detection and deconvolution) in three versions of pipelines are introduced in detail.

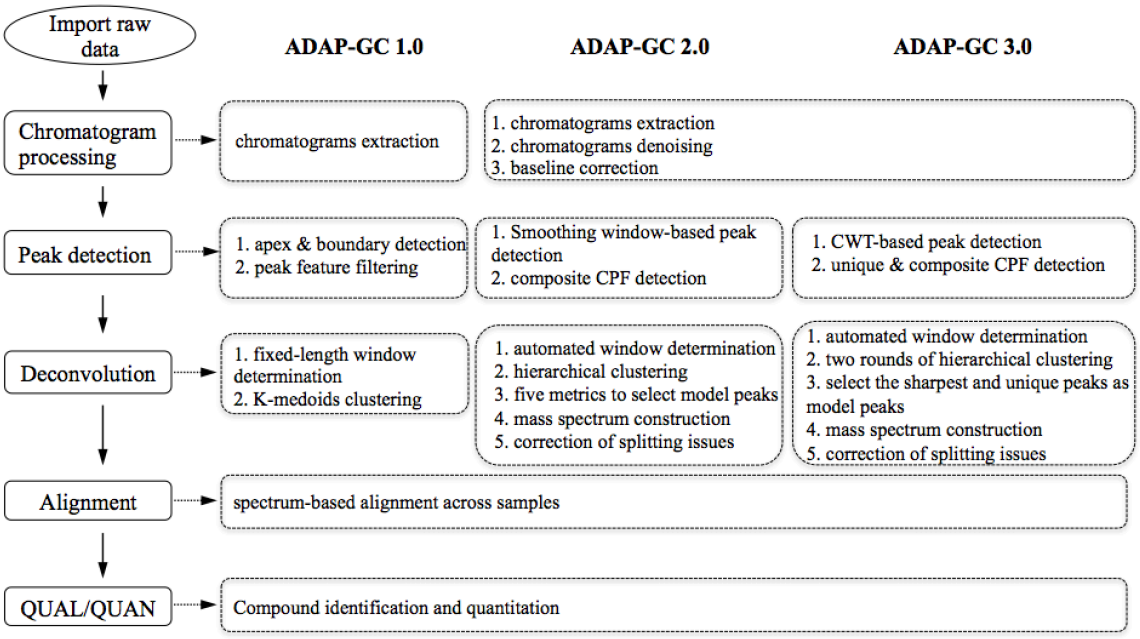

Figure 2.3. Improvement of compuational algorithms in three versions of ADAP-GC pipeline.

2.3.1 Peak Detection and Deconvolution in ADAP-GC 1.0

2.3.1.1 Smoothing Window-based Peak Detection

A number of algorithms have been developed and a thorough examination and comparison of these existing algorithms can be found in the review by Yang et al [54]. These algorithms usually perform a de-noising step that includes chromatogram smoothing and/or baseline correction before peak detection. Here, peak detection is performed before de-noising so that all of the EIC peaks can be extracted. This prevents the removal of true EIC peaks from happening that can be caused by imperfection in the de-noising algorithm, and ultimately benefits identification of compounds in terms of both confidence and total number of identifications. The rationale behind this is three-fold. Firstly, the observation of a larger number of fragments that belong to the same compound increases the likelihood that the identification is correct. Secondly, observing fragments of large mass has a positive impact on compound identifications because a larger mass is usually given a heavier weight in library search [26]. Since fragments of large masses tend to produce low intensity peaks in a spectrum, measures that are taken to preserve these peaks will facilitate identifications. Lastly, many compounds of interest such as biomarkers in biological studies are in the low concentration range. Due to these reasons, we chose to preserve as much information as possible at each stage of the data processing.

Here, peak detection consists of two sequential steps: peak picking and peak filtering. Peak picking is accomplished by first searching for the apex within a time window (Figure 2.4). Here, this window spans nine scans, which translates to 9/20 seconds. The window width is a parameter that is specified by users based on the

characteristics of their data. If the window were too narrow, the peak picking process would be very susceptible to noise. On the other hand, if the window is too wide, true apexes can be missed. Following the apex detection, the corresponding left and right boundaries of each EIC peak are determined, and peak height and shape are recorded. After all of the peaks in an EIC have been characterized, peak filtering is performed to remove peaks that most likely have resulted from noise. Specifically, the EIC is divided into equal-length time windows. Within each window, a window-specific threshold is calculated as the product of the lowest peak intensity and a preset factor (that is like a signal-to-noise ratio). Any other peaks with intensity below the threshold are filtered out.

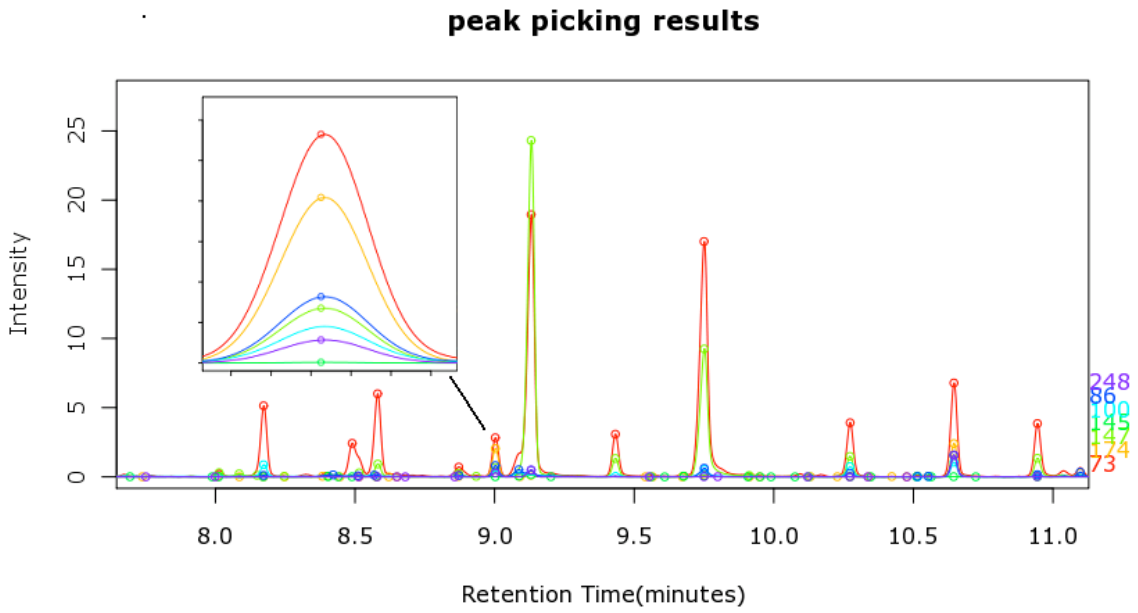

Figure 2.4. Peak detection in an EIC. Different extracted ion chromatograms are denoted by different colors with their peak apexes marked as small circles and their masses shown on the far right. The inset is a zoomed-in depiction of a small segment of the EIC and shows the EIC peaks that have been detected.

2.3.1.2 K-medoids-based Deconvolution

When compounds in a sample are fully resolved by GC, their mass spectra can be easily constructed by simply assigning peak intensity values obtained above to the corresponding fragment mass. However, when two or multiple compounds elute from the GC system in close proximity, peaks of fragments ions from different compounds will overlap and deconvolution has to be performed in order to construct their mass spectra. Traditional deconvolution was based on the assumption that fragment ions with similar apex elution time belong to the same component [55, 56]. However, this assumption will not hold when the apex elution time of different components are indistinguishable. This scenario is not uncommon in complex samples and Figure 2.5 A depicts one such scenario where EICs of fragment ions from two components have nearly the same apex elution time.

A closer examination of the EIC peaks reveals that a distinguishing feature of the coeluting components lies in the shape of the EIC peaks (or profiles). This shape difference can be captured by the normalized dot product. Specifically, let the abundance of the EIC profiles of two peaks be represented as two vectors. The similarity between two EIC profiles can be measured by the normalized dot product.

$$\vec{e}_x = \{a_1, a_2, \cdots, a_n\}$$
$$\vec{e}_y = \{b_1, b_2, \cdots, b_n\}$$

$$r = \frac{\vec{e}_x \bullet \vec{e}_y}{\left|\vec{e}_x\right| \cdot \left|\vec{e}_y\right|} \quad \text{(Equation 2.1)}$$

where a_i and b_i are the abundance values at retention time t_i, and $\bullet$ represents the dot product.

To separate co-eluting components, the r values of all pairs of EIC peaks that are in a narrow deconvolution time window (10 scans in this study) are calculated and a similarity matrix is formed for this window. Compounds that elute within this window are considered co-eluting and thus indistinguishable based on their apex elution time only. Subsequently, *k*-medoids clustering is applied on this matrix to cluster the fragment ions. Figure 2.5 B depicts the clustering results for one time window within which two compounds co-elute. The k-mediods clustering requires an initial assignment of *k*, the number of clusters. However, *k* is unknown prior to deconvolution. To resolve this issue, the silhouette score [57] is used in this study to assess the clustering quality and determine the *k* value.

$$S = \frac{d_{\text{inter}} - d_{\text{intra}}}{\max(d_{\text{inter}}, d_{\text{intra}})} \quad \text{(Equation 2.2)}$$

where *S* is the silhouette score of a cluster, d_{intra} is the intra-cluster distance and is calculated as the average pairwise distance between objects within the cluster, d_{inter} is the inter-cluster distance and is the minimum average distance from all of the other clusters to this cluster. The clustering is performed for different values of *k* and the *k* that results in the largest silhouette score is ultimately selected. To avoid falsely splitting fragments from the same component into two or more groups, an intra-cluster distance threshold is specified. When d_{intra} is smaller than this threshold, the corresponding *k* is accepted and the search stops. After deconvolution, each component that is detected in one sample is associated with its sample-specific mass spectrum and apex elution time. Due to

differences in experimental conditions such as temperature and column conditions, the apex elution time that is observed for the same compound is usually shifted differently across samples and, as a result, alignment is needed to correct this shift.

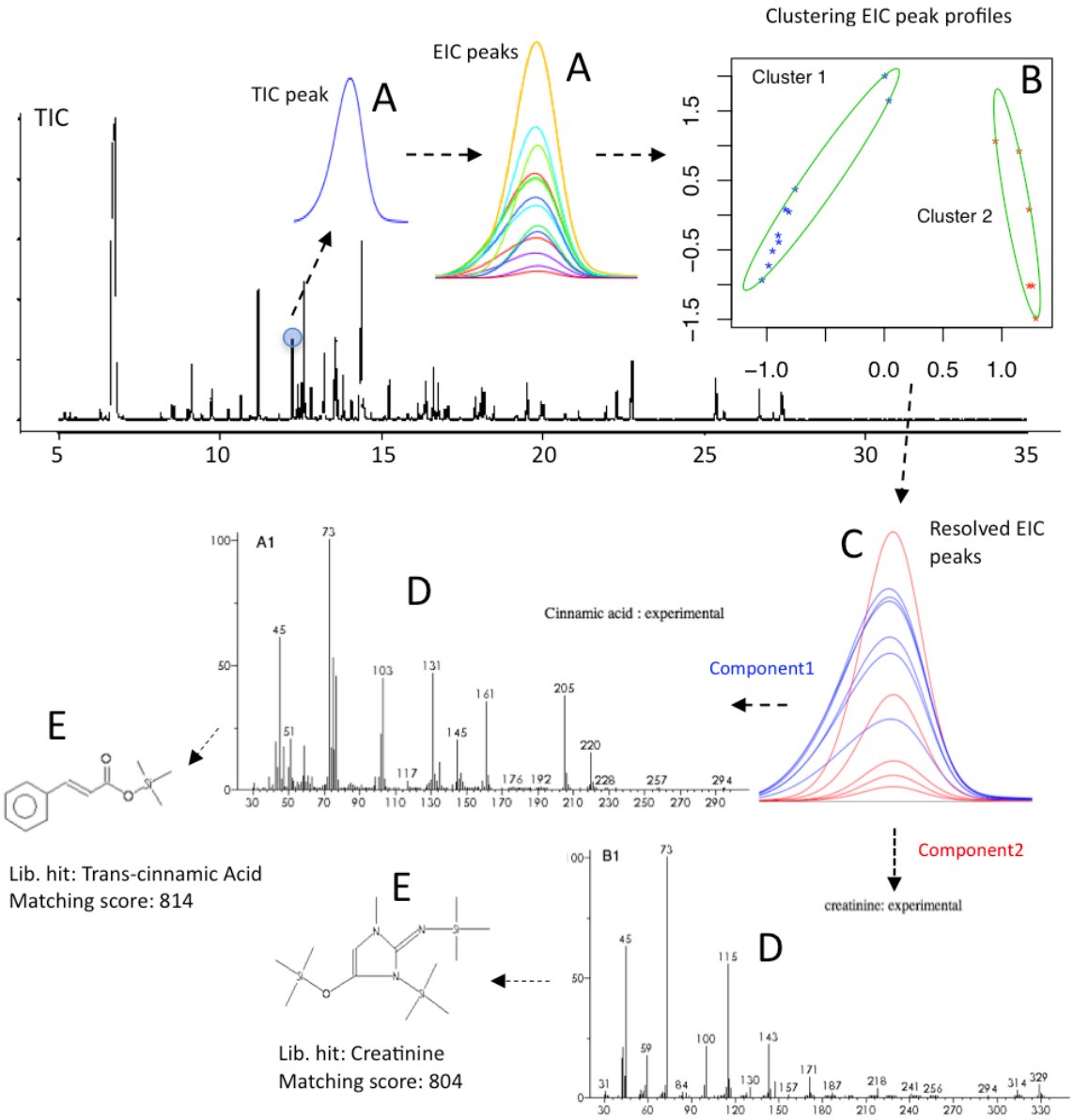

Figure 2.5. Illustration of deconvolution and identification of components. (A) Unresolved TIC peak and EIC peaks. The apex elution time of these EIC peaks is hardly distinguishable. (B) Optimal clustering of the EIC peaks into two groups that are shown in red and blue, respectively. (C) Separation of EIC peaks based on shapes of elution profiles. (D) Construction of mass spectra from the separated EIC peaks. (E) Identification of compounds by searching the spectra obtained in (D) against the NIST

reference library. The two coelutin compounds are found to be trans-cinnamic acid and creatinine.

2.3.2 Peak Detection and Deconvolution in ADAP-GC 2.0

Through testing, k-medoids-based deconvolution method has limitations in identifying and quantifying co-eluting compounds that two or more compounds elute from the chromatography column and their TIC and common EIC peaks overlap partially or completely in retention time. Specifically, for a fragment ion that is produced by only one of the co-eluting components, the grouping is usually quite successful in terms of correctly assigning it to its originating component. For a fragment ion that is produced by more than one co-eluting components, its chromatographic peak actually results from summation of signals that are produced by these components. However, this observed peak could be assigned to only one component using k-medoids-based deconvolution method. Consequently, the intensity pattern of the fragmentation spectrum constructed for this component deviates from the true pattern due to the abnormally high intensity of this fragment ion; the fragmentation spectra constructed for the other co-eluting components are incomplete since this fragment ion is missing. Ultimately, both the identification and quantification of these co-eluting metabolites are affected (Figure 2.6). To resolve this issue, an observed, shared peak should be decomposed back into individual peaks each of which corresponds to its originating co-eluting compound.

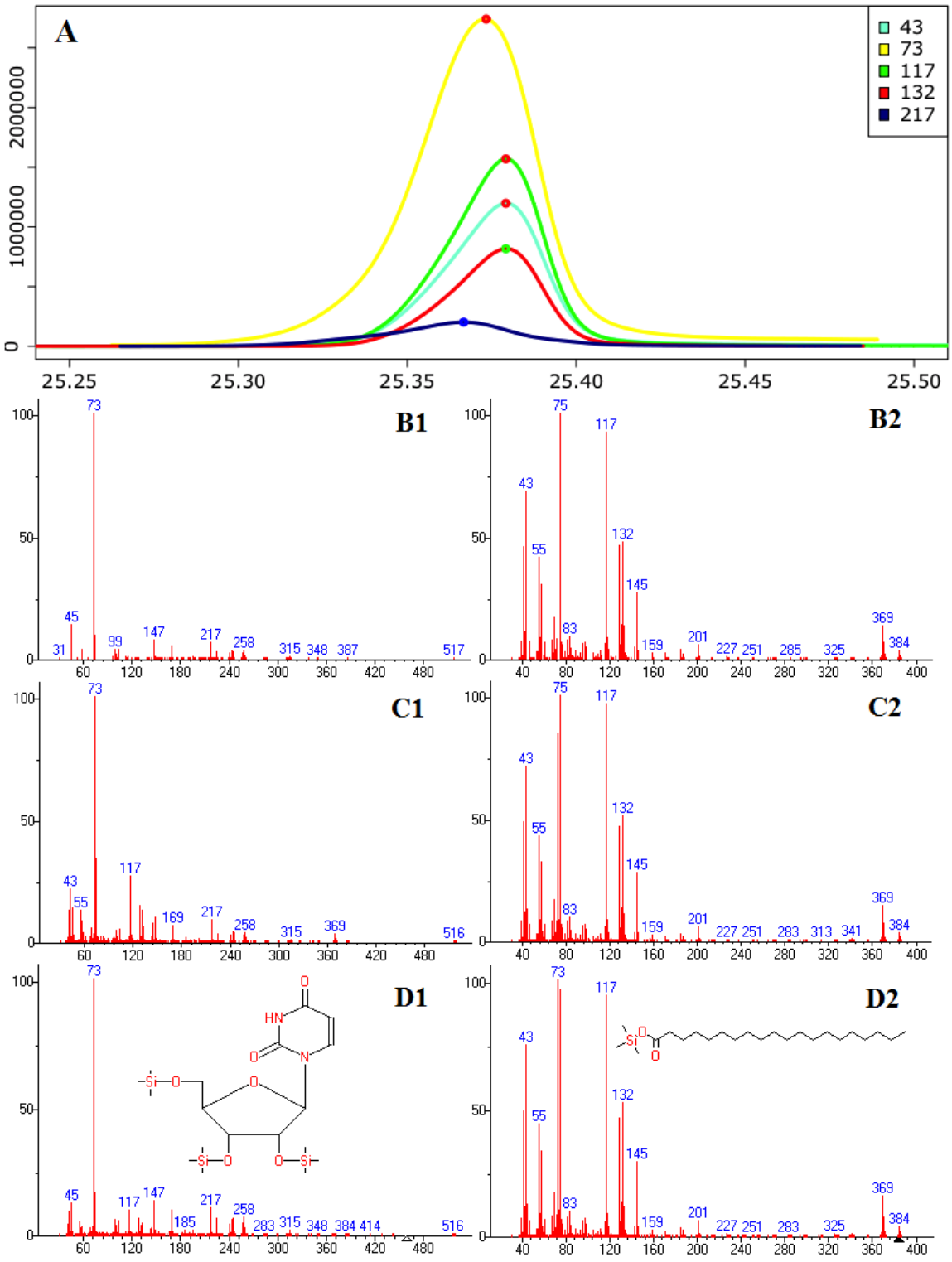

Figure 2.6. Comparison of constructed mass spectra and subsequent metabolite identification results with and without accurate deconvolution of shared peaks from two co-eluting compounds, uridine (Left) and n-Eicosanoic acid (Right). (A) Raw EICs of

selected masses. Mass 43, 73, and 117 marked with red circles are shared by both compounds. Mass 217 is unique to uridine while mass 132 is unique to n-Eicosanoic acid. (B1-2) Constructed mass spectra of uridine (B1) and n-Eicosanoic acid (B2) after deconvolution using ADAP-GC 1.0. The shared masses 43, 73, and 117 are only included either in the spectrum for n-Eicosanoic acid or in uridine. Their matching scores are 810 and 881, respectively. (C1-2) Constructed mass spectra after deconvolution that decomposes shared peaks. Each of the shared masses, 43, 73, and 117, is included in the spectra for both n-Eicosanoic acid and uridine. Their matching scores are 909 and 948, respectively. (D1-2) Reference spectra from an in-house library.

2.3.2.1 Simple and Composite Peak Detection

To our knowledge, a peak could result from the elution of either a single or multiple co-eluting components. In the latter case, the peak overlaps with its neighboring peaks and they must participate in the subsequent deconvoluton process together as a whole. In order to discriminate these two cases, we define chromatographic peak features (CPFs). A CPF is the elution profile of a minimum number of components that makes the elution profile complete, with “complete” meaning that the elution profile lasts from the beginning to the end of the elution of the component(s). A CPF that results from a single component is defined as a simple CPF and a CPF that results from summing signals of two or more components is defined as a composite CPF. A simple CPF has only one local maximum and a composite CPF could have one, two, or more local maxima. Figure 2.7 shows an example of each.

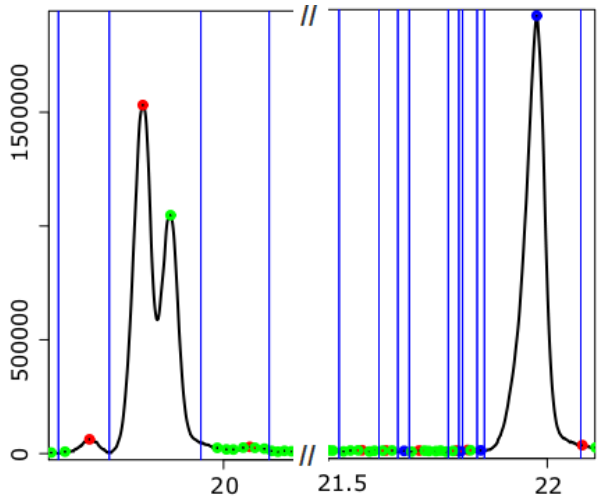

Figure 2.7. Examples of a simple and composite chromatographic peak feature

The MS signals resulting from GC-MS measurements can be contaminated by different sources of technical variations that can be removed by prior processing steps. In particular, de-noising makes it possible to remove random noises from signals [49]. The de-noising process consists in removing noise while preserving the useful information in the signal. In ADAP-GC 2.0, the de-noising and following peak detection are performed on each extracted ion chromatogram (EIC), thus the EIC for every observed mass is first extracted from the raw data. After extraction, each EIC undergoes smoothing and baseline correction. A moving average is used for smoothing while the baseline is identified for every EIC using the LOWESS (locally weighted scatterplot smoothing) regression algorithm and is subsequently subtracted from the EIC.

After de-noising, in order to detect peak apexes, a window of certain width moves along the entire EIC in one direction one unit of time at a time. The width of the moving window needs to be such that the window can cover about half of the peak width for most of the relatively low-intensity peaks. If the window width is 2n+1 units of time, then a

peak apex is found when the same point on the EIC is the maximum for n+1 consecutive moving windows. Peak boundaries are determined using a similar approach, except that the minimum is used. Each resulting quadruple that includes the left and right boundary time, apex time, and the relative intensity pattern between the two boundaries form a peak. To determine if a peak is a simple CPF or part of a composite CPF, we calculate the ratio of the intensity values at the boundaries to the intensity value at the peak apex. If one of the ratios is higher than an empirical threshold (e.g. 0.3), this peak is part of a composite CPF. All of the neighboring incomplete peaks are then merged to form a composite CPF.

2.3.2.2 Model Peak-based Deconvolution

The new deconvolution is based on a chromatographic model peak approach. For a component that does not co-elute with other components, it is no more than collecting all the corresponding simple CPFs and forming a fragmentation spectrum using all of the mass and apex intensity pairs. But for a component that does co-elute with other components, deconvolution decomposes composite CPFs into simple features and then groups the resultant simple features based on their CPF similarities. The general process of deconvolution consists of four sequential steps: (a) determination of deconvolution windows, (b) selection of model CPFs for decomposing composite CPFs, (c) construction of the mass spectrum for each observed component, and (d) correction of splitting issues. The concept of model CPF is the same as that of "model peak" used by Dromey et al [58] and Stein in AMDIS [26]. In comparison with AMDIS, ADAP-GC 2.0 employs multiple factors including "sharpness" (to be described) for evaluating CPFs whereas AMDIS uses the sharpness value only.

(1) Determination of deconvolution windows: A deconvolution window delimits the temporal span wherein deconvolution is carried out. These windows are determined based on TIC CPFs detected in the previous step of peak detection. Basically, the left and right boundaries of each TIC CPF define a deconvolution window. Any EIC CPF whose peak apex falls in this window will participate in the window-specific deconvolution. It is worthwhile to point out that part of the EIC features could be outside of the deconvolution window. This TIC-based determination of deconvolution windows is fully automated and produces windows that are data-dependent, which avoids possible problematic issues associated with fixed windows. After deconvolution windows are determined, deconvolution proceeds sequentially for all the windows.

(2) Selection of model CPFs: A model CPF is defined as the elution profile of a compound when it elutes from a chromatographic system alone and its concentration is within the linear dynamic range of the mass analyzer of the mass spectrometer. As such, the elution profile is produced from this compound only with less interference from neighboring compounds and has a relatively high signal-to-noise ratio (SNR). ADAP-GC 2.0 constructs/selects the best model CPF for each observed component based on all of the EIC CPFs that correspond to this component. The construction/selection process is deconvolution window-specific and consists of two sequential steps: selecting good candidate CPFs, and determining the number of components and the model CPF for each component.

(a) Select good candidates of model CPFs. For each EIC CPF within a deconvolution window, five metrics are used to measure how well it can be used as a

candidate model peak. These include sharpness, SNR, apex intensity, Gaussian similarity, and mass. Details on each metric are as follows.

(i) The sharpness indicates how quickly the abundance values of the corresponding mass change with time. The higher the sharpness value, the more likely the EIC feature is generated by a single component. The sharpness value is calculated as

$$\text{sharpness}=\sum_{i=2}^{p}\frac{I_i - I_{i-1}}{I_{i-1}} + \sum_{i=p}^{n-1}\frac{I_i - I_{i+1}}{I_{i+1}} \quad \text{(Equation 2.3)}$$

where n is the total number of time points for a CPF, p is the time index of the apex, and I_i is the abundance value at time index i.

(ii) The SNR is estimated based on the high and low frequency signal components of the CPF, which is calculated using the continuous wavelet transform that has been described in the section of peak detection [59, 60].

(iii) The apex intensity is used to gauge how well the measured peak profile represents the true concentration of the component in the sample. The higher the intensity, the more robust the intensity measurement by the mass analyzer, in that the intensity measurement is less likely affected by background noise. Clearly, the apex intensity compensates SNR.

(iv) The Gaussian similarity measures how well a CPF can be modeled by a Gaussian curve. The reason why we use it to select good candidate CPFs is that, based on our observation of GC-TOF-MS data, the elution profile of a large portion of compounds exhibits a symmetric bell shape when they elute alone. The similarity score between an EIC CPF and the Gaussian curve that best fits it is calculated as the normalized dot product between them [61].

In the meantime, we have observed that some compounds exhibit either fronting peaks (the left side of the CPF spans a longer time range than the right side) or tailing peaks (the right side of the CPF spans a longer time range than the left). Tailing can occur due to various reasons, including column contamination, poor column installation, or co-eluting compounds. In the latter case, the shorter side of a tailing CPF is usually produced by a single component and therefore is still valuable for selecting model CPFs. To make use of it, ADAP-GC 2.0 constructs a complete CPF by appending a mirror image to the shorter side and calculates its Gaussian similarity.

(v) The mass value of an EIC CPF is indicative of the likelihood that it is unique to a component. The higher the mass value, the more likely it is unique.

With all the five aforementioned metrics calculated for measuring the qualities of EIC CPFs, two-step screening method to select good candidates of model CPFs is applied. Firstly, three separate filters are used to remove those CPFs with very low SNR, sharpness, or Gaussian similarity scores. Secondly, for each CPF that passes all the three filters, a composite score is calculated as:

$$\text{Score} = C_1 \times \text{Mass} + C_2 \times \text{Gaussian Similarity} + C_3 \times \text{Intensity} + C_4 \times \text{SNR} \quad \text{(Equation 2.4)}$$

To be noticed, the sharpness value is not considered in the composite score because we have found from many rounds of testing that the sharpness value is a very reliable measure in itself and including it in the total score does not have significant influence on the final performance. The four weights, c_1, c_2, c_3, c_4 , have been systematically tested and adjusted, and ultimately set as 0.1, 0.3, 0.2, and 0.2 for optimal performance for the analytical platform we used. Based on the total score of all of the CPFs that have passed the first filtering step, a threshold is calculated as. Those CPFs

whose composite score is higher than the threshold are considered good candidates of model CPFs (Figure 2.8).

Threshold = min (total scores) + 0.25 × range (total score) (Equation 2.5)

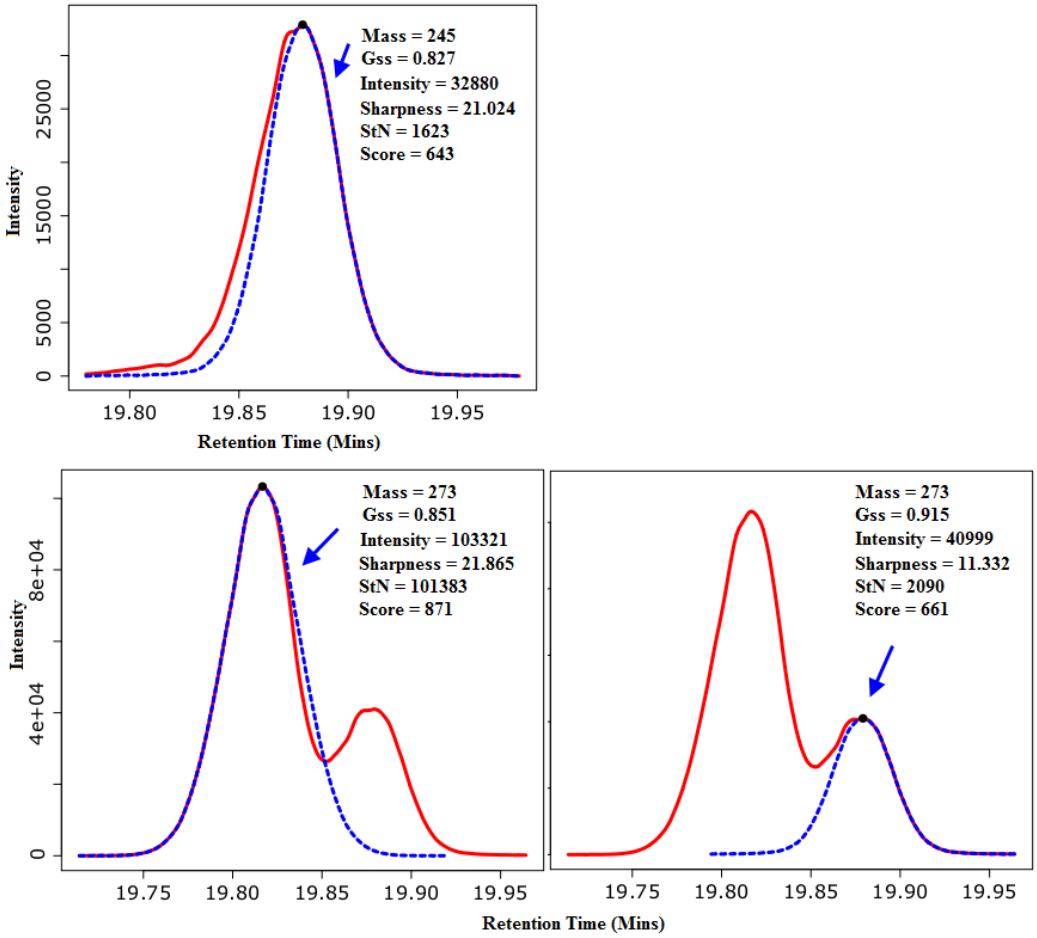

Figure 2.8. Formation of candidate model CPFs and calculation of their total scores based on Equation (2). (Top) A candidate model CPF in blue produced from a simple CPF in red. The mirror image of the shorter side (i.e., the right side in this case) of the CPF is appended to the shorter side to form a candidate model CPF. (Bottom) Two candidates of model CPFs in blue produced from a composite CPF in red. The far left side of the composite CPF has most likely been produced by a single component and therefore can be used to construct a candidate model CPF. The same is true for the far right side of the composite CPF. One candidate model CPF is produced from the far left side of the composite CPF (Bottom Left) and the other candidate is produced from the far right side of the composite CPF (Bottom Right).

(b) Determine the total number of components and the model CPF for each component. The good candidates of model CPFs then participate in a hierarchical clustering for determining the most likely number of components in the current deconvolution window. The pair-wise peak feature dissimilarity is used as the distance measure in the clustering and the threshold for obtaining the clusters is determined in an empirical fashion. Each resulting cluster corresponds to a specific component and the CPF with the highest total score within this cluster is selected as the model peak. With model peaks determined, we are ready to decompose each EIC composite CPF into simple features and construct a fragmentation mass spectrum for each component.

(3) Construction of the mass spectrum for each observed component. Each composite CPF results from a linear summation of simple CPFs. In order to determine the constituent simple features for each composite feature, we apply constrained optimization [62] by minimizing the residual between the composite CPFs and a linear combination of the model peaks. The residue is calculated as:

$$E = \sum_{i=1}^{n} \left(X[i] - \sum_{k=1}^{K} a_k M_k[i] \right)^2 \quad \text{(Equation 2.6)}$$

where X represents the composite CPF, n is the total number of time points, K is the total number of model peaks within this deconvolution window, and M_k, a_k, $k = 1,2,\cdots,K$ represent model peaks and corresponding weighting coefficients, respectively. The optimization gives rise to the weights a_1, a_2, ...a_k. For all of the CPFs within the current deconvolution window, the resulting weights that correspond to the same model peak yield the mass spectrum of a component. Clearly, the intensity pattern of different masses of the spectrum is reflected in the relative magnitudes of the weights.

(4) Correction of splitting issues. It could happen that two or more model CPFs are constructed/selected for the same component within a deconvolution window. As a result, all of the EIC features that contain this component are split into two or more groups of simple CPFs, which give rise to more than one mass spectrum. Since one single spectrum is split into more than one and different masses could be split differently, the accuracy of the resulting mass spectra and the estimated concentration of the component will be reduced. To resolve this issue, a post-deconvolution checking step is performed by computing pair-wise mass spectrum similarity within each deconvolution window. When highly similar spectra are found, the model peak with the highest total score calculated in Eqn. (2) is selected to represent this component and the other similar model CPFs are discarded. A second deconvolution is then carried out to produce a more accurate mass spectrum. In computing the mass spectrum similarity, both the signal similarity and the time shift between two spectra are considered. Figure 2.9 depicts one example where the splitting issue was corrected.

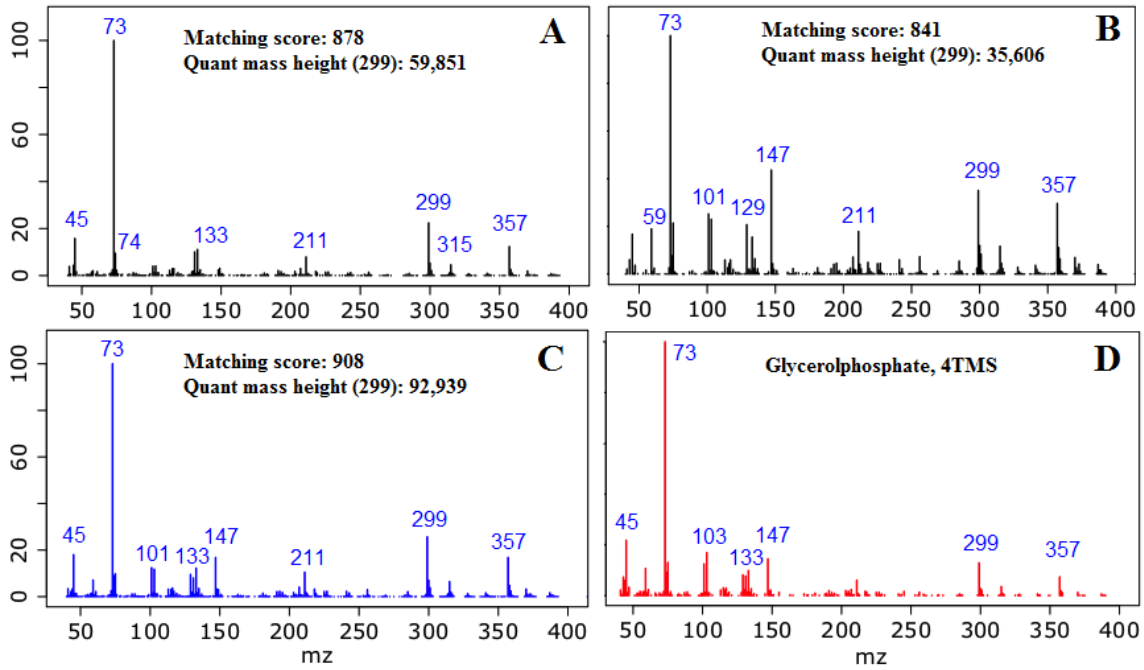

Figure 2.9. Resolving compound splitting issues and improving ADAP-GC 2.0 performance in terms of both identification and quantitation. This is an example from one of the datasets in Sample III. ADAP-GC 2.0 originally selected two model CPFs (corresponding to masses 357 and 286) with a three-scan time shift between them. (A) Resolved spectrum based on model peak mass 357. (B) Resolved spectrum based on model peak mass 286. Both spectra were matched with glycerolphosphate and mass 299 was selected as the quantitation mass. The mass spectra shown in (A) and (B) are highly similar with a matching score between them > 850, so a second deconvolution was performed using the model peak 357 because it has a higher total score as calculated in Eqn. (2). (C) Resolved spectrum after the second deconvolution. Based on this spectrum, both of the matching score and the peak height of the quantitation mass increased considerably, with the latter reflecting the true concentration of the compound in the sample. (D) Reference spectrum of glycerolphosphate in the user library.

2.3.3 Peak Detection and Deconvolution in ADAP-GC 3.0

2.3.3.1 Continuous Wavelet Transform (CWT)-based Peak Detection

The previous peak detection method relies on finding local extrema within a window of specified width. Since different chromatographic peak features in a same EIC can have different widths, it is almost impossible to find one width parameter that fits all real peak features while ignoring noise signals. So the challenge with peak detection is to

develop a method that is robust to varying peak width. To our knowledge, wavelet transform represents a one-dimensional signal in a two-dimensional space with the second dimension representing scale, as a result, each peak apex that wavelet transform detects has a peak width (i.e., scale) associated with it (Figure 2.10). Therefore, it is not necessary to specify the window width parameter that the local maximum method requires since wavelet transform automatically finds both the peak apex and peak width. Here, we applied a package of wavelet methods for time series analysis called "wmtsa" in R and modified parameters accordingly to fit the characteristics of GC-TOF-MS data.

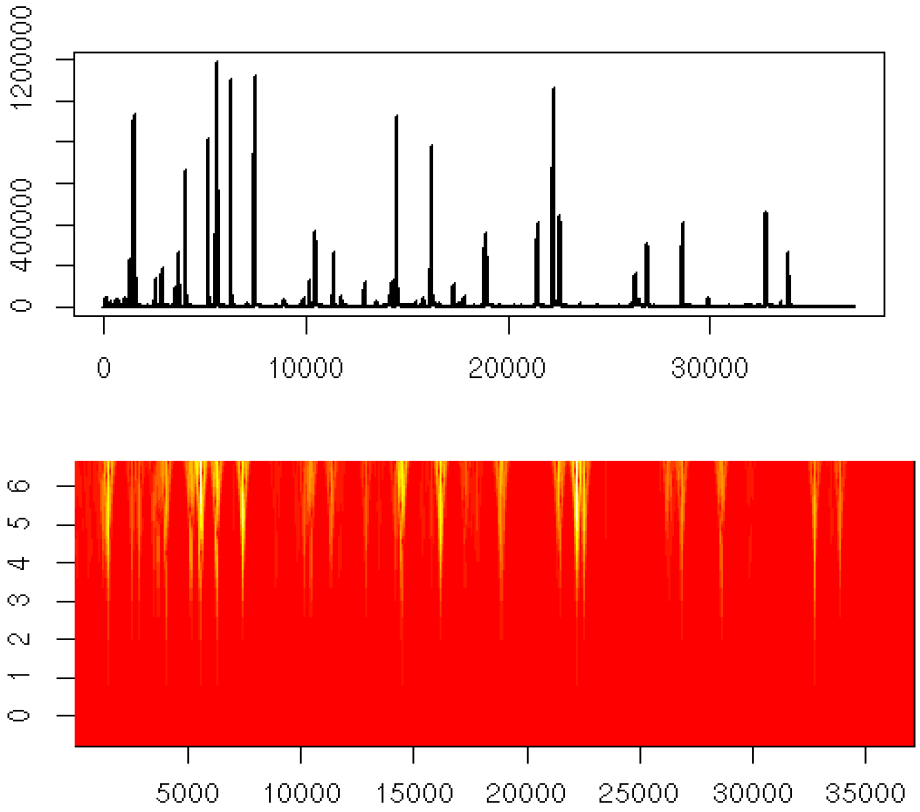

Figure 2.10. Wavelet transform-based peak detection on an EIC.

Each individual peak feature detected using wavelet method will be further examined to determine whether it is a simple/unique peak feature or belongs to a

composite peak feature. Here, two criteria are used for the determination: (1) boundary to peak apex: the ratio of intensity values at the left and right boundary, respectively, over the intensity value at the peak apex; (2) boundary difference to peak apex: the ratio of the intensity difference between the left and right boundary over the intensity value at the peak apex. Any peak feature that either of three ratios is higher than an empirical threshold (e.g. 0.3) is considered as a part of composite peak feature, which is going to be combined with neighboring peaks together for deconvolution. And those features that all of three ratios are smaller than the threshold are considered as unique peaks as the candidates of model peaks. Meanwhile, the continuous wavelet transform calculates the signal to noise ratio (SNR) based on the high and low frequency signal components of peak features. With empirical SNR cutoff, noisy peaks with low SNR values (e.g. < 10) are not considered. In order to examine the purity of unique peak features, local maximal method is simultaneously applied to check if there exists small peaks, which may not be detected by wavelet transform (Figure 2.11). For those initially assigned as unique peak features do exist small peaks between the left and right boundary, they are corrected as composite peak features. The combination of wavelet transforms and local maximum method makes peak detection more flexible and robust to determine unique and composite peak features, and select high-quality unique peak features for model peaks.

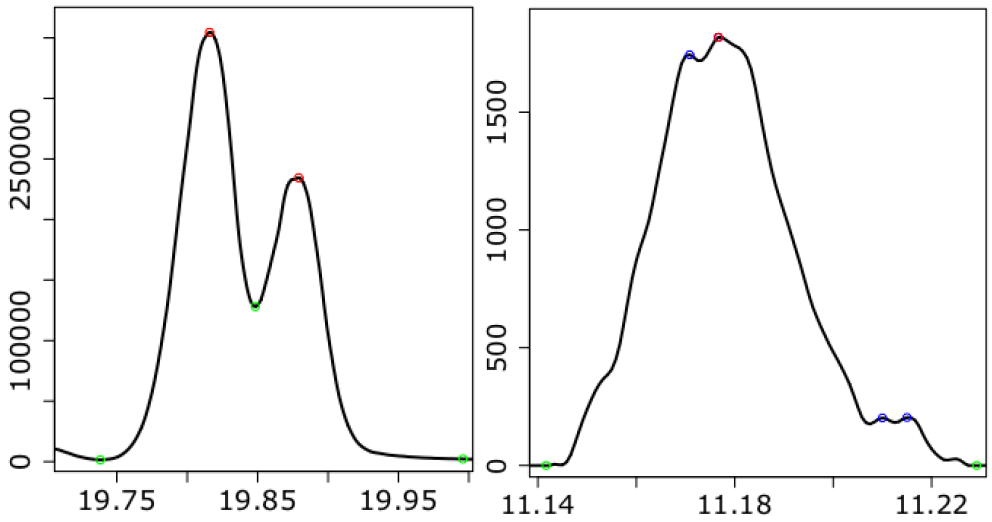

Figure 2.11. (A) An example of composite peak feature successfully detected by wavelet transform with two significant peak apexes (red circles) and their corresponding boundaries (green circles). (B) An example of a peak feature initially defined as a simple peak feature by wavelet transform with one peak apex and then corrected by the local maximum by finding more minor peaks (blue circles) around indicating it as a composite peak feature.

2.3.3.2 Model Peak-based Deconvolution

The challenge of deconvolution is to find a model peak that could represent the real elution profile for each co-eluting compound, which determines the purity of mass spectrum for compound identification and quantitation. The second version of deconvolution favors model peaks that are symmetric and resemble a Gaussian curve by using five metrics of peak qualities: signal to noise ratio, sharpness value, Gaussian curve fitting, absolute peak apex intensity and mass value. But later, we observed that some compounds especially at low concentrations or multiple compounds co-eluting together with large concentration variation, might not exhibit significant, smooth and Gaussian curve-like peak features. Thus, the criteria are so strict that their model peak features can be filtered out, leading to those compounds undetected. In order to solve this problem, the

third version of deconvolution does not require Gaussian peak shape; instead, it focuses on detecting the unique peak features for each component and selects the one with the highest sharpness value as the model peak. In addition, two sequential hierarchical clustering are applied in order to improve the determination of co-eluting compounds within a deconvolution window.

(1) The first hierarchical clustering calculates the eluting time distance among all unique and composite peak features within a deconvolution window in order to determine the minimal number of co-eluting compounds. For compounds eluting with more than 60 scans away (which equals to 3 seconds in our study), they can be easily distinguished while it could be relatively difficult for those compounds eluting within 60 scans with overlapping peak features. Thus, the distance cutoff has been set as 60 scans to meet three goals: (1) peaks with greater than 60 scans distance can be analyzed separately so that those peak features within 60 scans distance can be grouped together for the second hierarchical clustering without the bias of distant nodes; (2) an individual cluster with less than a total of two peak features could be considered as random noise or background signals and will be removed to minimize their interference in the following analysis; (3) the number of clusters decides the minimal number of components within this deconvolution window. If the total number of components is less than this number after the 2nd hierarchical clustering, which means there exists at least one cluster that has not any component detected, thus the mass spectrum at the median position will be extracted directly to reduce the possibility of missing any true positive compound.

(2) Determine the total number of components and the model peak for each component. At first, a simple filtering using an empirical threshold of signal to noise ratio

(e.g. 50) is used to select unique peak features as candidates for model peaks, because they are able to represent the corresponding compound elution profiles with minimal interference from noise and/or co-eluting compounds. After that, the 2nd hierarchical clustering is performed on these candidates within each RT cluster produced in the first step. As a result, the individual cluster may produce one or more components and the total number of co-eluting components can be determined within each deconvolution window.

The next step is to select the best candidate as the model peak for each component. Here, the new deconvolution approach applies a characteristic of sharpness to evaluate the degree of purity of model peak candidates. Our previous method to calculate sharpness does not consider peak width factor, so that some wide peaks could have high sharpness values due to cumulative summary of point-to-point change. This makes our previous method insensitive to describe the true sharpness characteristics of peaks. Borrowing the idea of how AMDIS calculates the sharpness of peak shape, we provide a simple but effective measure. Sharpness values between the maximum abundance A_{max} and an abundance value located n scans from the maximum A_n are defined as:

$$\text{Sharpness} = \frac{\sum_{n=1}^{N}(A_{\max} - A_n)/n}{N} \quad \text{(Equation 2.7)}$$

The median sharpness values on each side are found and then averaged, and then the averaged sharpness value is used to describe the individual unique peak feature. Three reasons are behind for this sharpness calculation as compared to AMDIS: (1) GC-TOF-MS data always has more collection/sampling points, so it's not necessary to perform fitting and time shifting as AMDIS indicated; (2) ADAP performs baseline correction

before peak detection, so that the noise factor could not reflect the estimated noise any more in our case; (3) More abundant peaks usually present sharper and smoother peak shape with fewer effects of noise. Thus, the absolute abundance difference between the A_{max} and A_n is chosen so that abundant peak features with larger sharpness values have more chances being selected as model peaks.

Next, as the same approach in the deconvolution of ADAP-GC 2.0, the constrained optimization method is applied by minimizing the residual between the detected peak features (both unique and composite peak features) and a linear combination of the model peaks. As a result, the mass spectrum of each compound is constructed based on the resulting weights corresponding to the representative model peak. It might be possible that two or more model peaks are selected for a same compound which further affects the accuracy of compound identification and quantitation, thus the pairwise mass spectra similarity within a same deconvolution window is calculated in order to correct such splitting issue.

2.3.4 Alignment

A number of algorithms have been developed for chromatogram alignment [63]. Most of them were designed primarily for aligning TIC. However, TIC-based alignment can be inaccurate when different compounds within the same TIC peak shift differently along the retention time axis. A better approach is to do component-based alignment. Specifically, the same components across samples are identified based on their spectrum similarity, component-specific time shifts are determined, and ultimately, components are aligned accordingly.

To search for the same components across samples, a measure of confidence needs to be defined that takes into account both the spectra similarity and retention time similarity between two spectra. In ADAP, this measure is:

$$\text{score}_{\text{total}}(s_i, s_j) = 0.9\text{score}_{\text{spec}}(s_i, s_j) + 0.1\text{score}_{\text{RT}}(s_i, s_j) \quad \text{(Equation 2.8)}$$

where s_i and s_j denote two spectra. $\text{score}_{\text{spec}}$ is adopted from the spectra similarity measure used in AMDIS and is a linear, weighted combination of the pure and impure score [56]. score_{RT} is calculated by:

$$\text{score}_{\text{RT}} = 1 - |\Delta\text{RT}|/w \quad \text{(Equation 2.9)}$$

with ΔRT being the difference in their apex retention time and w being the maximum retention time shift that is acceptable for components in a particular experiment. Components whose retention time difference exceeds w should be considered as different components.

The incorporation of score_{RT} into $\text{score}_{\text{total}}$ facilitates distinguishing components that have similar spectra and whose apex elution time difference is less than w, particularly in the case of isomers. This approach that uses a combination of mass measurement and elution time information has been used in the proteomics field to identify and quantify peptides [64, 65]. The final $\text{score}_{\text{total}}$ is scaled so that it is between 0 and 999. This is the same numerical scale that the NIST library search uses.

The alignment process starts with the earliest component and sequentially aligns every component for which a spectrum has been constructed. What lies at the core of aligning a component in ADAP is the accomplishment of two tasks: 1) identification of mass spectra that correspond to the same component across samples, and 2) selection of

the best representative spectrum for the component. These two tasks are accomplished via a two-phase searching algorithm as follows:

Phase 1. Among all the samples, identify the earliest component that has not been aligned and define a global alignment window that starts with this component and is of width w. Within this alignment window, use the spectrum of the earliest component as a reference and search for components in other samples that produce $\text{score}_{\text{total}}$ greater than a certain threshold (750 in this study). From these high-scoring components, select the one that produces the highest $\text{score}_{\text{total}}$ for each sample.

Phase 2. Use each sample-specific best-matching component as a reference and repeat the searching process in Phase 1 to find the best matching components in other samples. As a result, a group of spectra are identified for each reference.

For each component, the best representative spectrum across all the samples is determined by selecting the component that produces the highest average $\text{score}_{\text{total}}$ when it is used as a reference in Phase 2. This component will be used as the final reference to align spectra across samples. The rationale behind this is as follows. Each mass spectrum that is constructed for a component consists of two parts: the pure part that corresponds to the component itself and the impure part due to experimental noise and/or interference from coeluting components. Since the spectrum that primarily consists of the pure part should be the most reproducible and, consequently, give rise to the highest average $\text{score}_{\text{total}}$, it is apparently the best representation.

Ultimately, alignment is carried out with the best overall spectrum serving as the reference. ADAP requires that only these components that are observed in a sufficient number of samples be aligned. Phase 2 is a refinement of Phase 1 in that spectra

identified in Phase 1 may not be the best representation of the component in the corresponding samples. This can happen when the earliest spectrum that is used in Phase 1 is not the best representation. Therefore, the two-phase approach should improve the alignment performance compared to a one-phase approach where only Phase 1 is used. Specifically, more samples can be aligned and a better representative spectrum can be found for compound identification. Table 2.2 demonstrates these improvements. Figure 2.12 illustrates the necessity of alignment and compares the EICs before and after alignment. Deviations of retention time for the 19 standard compounds are depicted in Figure 2.12 K. This deviation profile is consistent with the temperature change in the experimental process, i.e. the maximum deviation occurs at the time when the temperature reaches its peak (approximately 20 min in this example).

Table 2.2. Performance Comparison of the One-Phase versus Two-Phase Alignment with the CC Samples

RT		One-phase alignment		Two-phase alignment	
(min)	Compound Name	SmpNr[1]	Score	SmpNr	Score
11.065	Decanoic acid	11	887	11	893
13.443	Laurate	11	855	11	863
14.559	Tridecanoic acid	11	883	11	890
16.626	Pentadecanoic acid	11	690	11	836
17.481	Palmitoleic acid	11	NA	11	880
17.74	Palmitic acid	3	605	11	819
18.852	Heptadecanoic acid	2	744	11	876
19.691	Linolic acid	10	479	11	798
19.768	Linolenic acid	6	NA	7	676
19.799	Oleic acid	NA[2]	NA	11	494
19.848	Elaidic acid	3	769	3	795
20.076	Stearic acid	11	749	11	914
20.464	8.11.14-Eicosatrienoic acid	NA	NA	11	840
21.254	Arachidonic acid	9	750	11	929
21.298	Eicosapentaenoic acid	2	651	11	812
21.567	11.14-Eicosadienoic acid	11	855	11	882
21.601	11-Eicosenoic acid	10	717	11	880
21.773	Arachidic acid	6	NA	11	NA

Table 2.2 (continued)

22.504	Decosahexaenoic acid	11	813	11	837
Average		8	746	10	829

Note: 1.SmpNr refers to the total number of samples in which a compound was observed. 2. NA means that the corresponding metabolite was not detected with high confidence because the matching score is below the cutoff value that was set at 750.

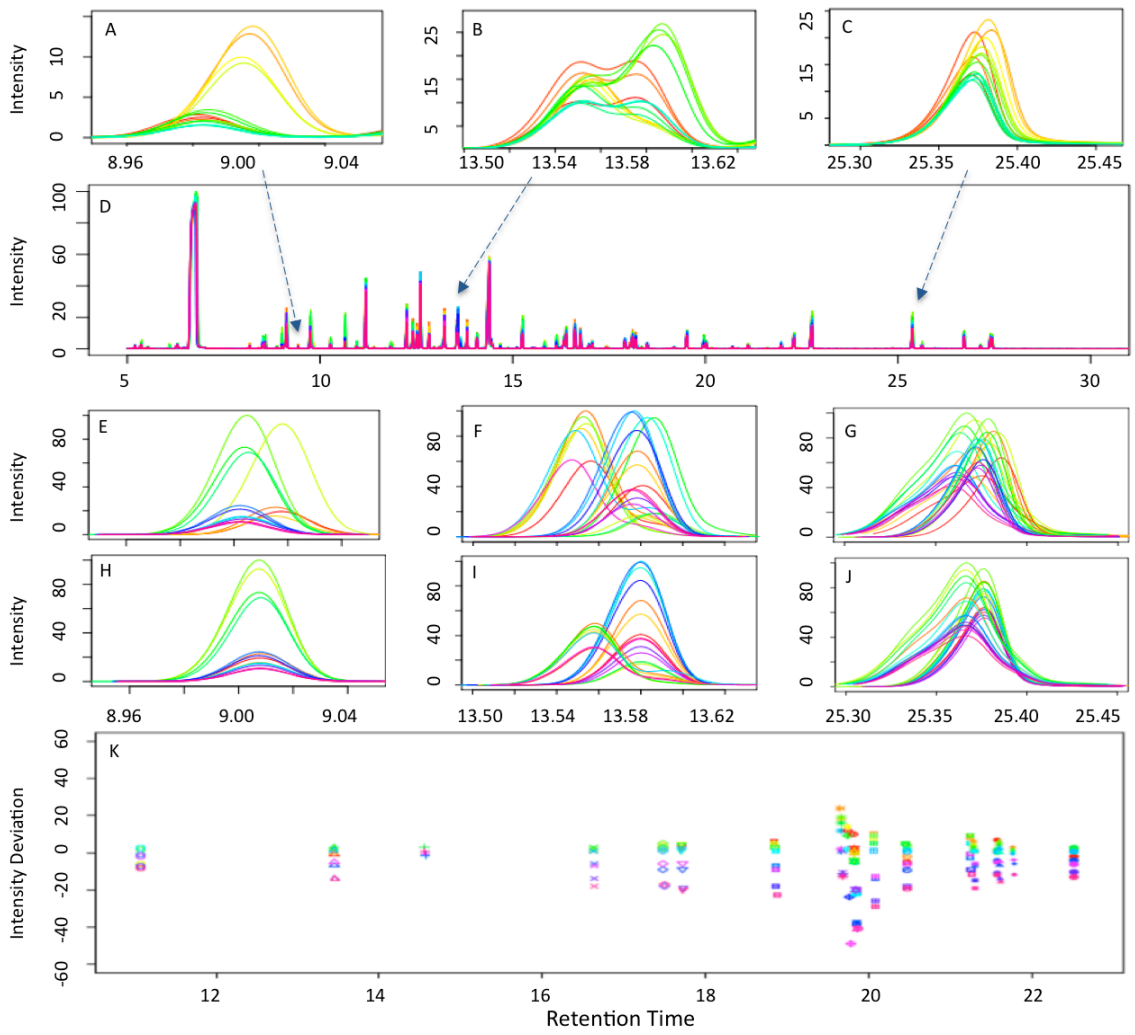

Figure 2.12. Necessity of alignment and comparison of EICs before and after alignment. The 15 MS samples are considered and each sample is represented by one unique color. (A-C) TICs within three different time intervals. (A) One component elutes with two distinct TIC peaks; (B) two components elute with two distinct TIC peaks; (C) two components elute with two peaks that are barely distinguishable; (D) TICs of the 15 MS samples. (E-G) EICs before alignment; (H-J) EICs after alignment. EIC pairs E-H, F-I, and G-J correspond to TIC segments (A), (B), and (C), respectively. For the two EIC pairs (F and I) and (G and J), two coeluting components became distinguishable only

after alignment. (K) Deviation of the elution time of 19 compounds in the MS samples with respect to the elution time of the alignment reference.

2.3.5 Compound Identification and Quantitation

2.3.5.1 Compound Identification or Qualification (QUAL))

Compound identity is determined by searching the corresponding mass spectrum against a library of spectra by measuring their similarities. ADAP-GC pipeline is equipped with the capabilities to perform library searching against user library or commercial library such as from NIST or to export extracted mass spectra in standard MSP format that can be read by NIST MS searching software [9] or other third-party software to perform library searching. The similarity of extracted mass spectrum against the standard mass spectrum in a library is measured by calculating their dot product of two mass spectra vectors:

$$\frac{(\sum(\vec{X} * \vec{Y})^{1/2})^2}{\sum \vec{X} * \sum \vec{Y}} \text{(Equation 2.10)}$$

After calculation, the similarity score is then normalized to 999 to keep consistent with AMDIS, the higher matching score indicates the more accurate mass spectrum extracted from raw data.

2.3.5.2 Compound Quantitation (QUAN)

Quantitation (QUAN) is achieved by selecting a quantitation mass of each compound and then calculates its peak height or area to represent the concentration of this compound in a sample. Each compound has all the fragment ions identified, and ADAP-GC provides three options for the compound quantitation: using the peak area or intensity values of the model CPF, most intense mass, or most intense unique mass. The

most intense mass is the mass with the highest overall intensity in the mass spectrum of each compound. “Unique” here means that a mass is not shared with neighboring compounds. In many cases, these three masses are the same for a compound. Since peak area and peak height are directly related to the concentration of compounds in a sample, either of them can be used for quantification.

At the end of the QUAL/QUAN, a table is generated that contains the identities and relative quantities of all compounds within each sample and is ultimately exported for statistical analysis and biological exploration (Figure 2.13).

ID	Compound 1	Compound 2	Compound 3	Compound 4	Compound 5	Compound 6
Compound name	Glucose	Cysteine	Glycine	Isoleucine	Leucine	Valine
Retention time	4.18583	4.18833	4.30416	4.33333	4.37	4.3775
Quantiation mass	105	86	148	105	222	207
Sample 1	46766	53010	4839	49072	13608	6720
Sample 2	35455	34508	3452	31291	4315	4117
Sample 3	38516	61141	4017	35320	9532	4928
Sample 4	42754	50202	4009	39198	7358	5555
Sample 5	32061	41937	0	31335	7582	4388
Sample 6	30139	45304	0	31129	5070	4533
Sample 7	23282	38414	2398	23606	2341	3630

Figure 2.13. An example of an QUAL/QUAN table obtained after data processing

2.4 Results and Discussion

2.4.1 ADAP-GC 1.0

ADAP-GC 1.0 is the first version of pipeline, which consists of four sequential steps: peak detection, deconvolution, alignment, and library search. ADAP-GC 1.0 was able to identify 37 out of 38 standards in Sample I with the average matching score 855 (Table 2.1) and 19 out of 20 fatty acids in the calibration samples with the average matching score 829 (Table 2.2). The r-squared values are calculated for the abundance values of 19 fatty acids vs. the true concentration of the compounds in the 11 CC

samples. All of their the r-squared values are close to 1 (data not shown), which indicates the accuracy of ADAP-GC 1.0 in extracting quantitative information of compounds.

ADAP-GC 1.0 allows data to flow seamlessly through these processing steps as a high-throughput pipeline: (1) it is fully automated and no human intervention is needed in the entire process; (2) the computationally intensive deconvolution and alignment are written in C++ and applies parallel computing using MPI, and (3) special care has been taken to accelerate computations by optimizing memory usage and data structure. Table 2.3 lists the number of samples and the corresponding processing time for the three sets of data. ADAP used less than 3 min to analyze 15 datasets from Sample I or 20 datasets from liver injury samples. To our knowledge, ADAP is much faster compared to other existing software tools including ChromaTOF.

Table 2.3. Measures of ADAP-GC 1.0 performance

	Sample I	CC	LI
Total number of Samples	15	11	20
Average sample data size (MB)	123	67	118
Average number of peaks detected per sample	125686	562018	281913
Average number of components detected per sample	405	596	388
Ttal number of components after alignment	478	304	277
Processing time (s)			
Peak picking + deconvoluion	75	62	103
Alignment	10	25	118
Total	85	97	132

From 20 rat serum samples of a liver injury experiment, a total of 277 components were produced after alignment. The resultant quantitation data was imported into the SIMCA-P 12.0 software package (Umetrics, Umeå, Sweden) for multivariate statistical analysis [66]. Specifically, mean-centering and auto-scaling were used for data

pretreatment [67]. Subsequently, PCA (Principal Component Analysis) was applied and a clear separation between the diseased and control groups (Figure 2.13 A) was observed with the first two components explaining 29.8% of the total variance. Lastly, a supervised PLS-DA model (Partial Least Squares Discriminant Analysis) was constructed (Figure 2.13 B) to identify the deferential metabolites that contribute to the separation between two groups. A total of 55 significant components were selected using VIP statistics (VIP ≥ 1, variable importance in the projection) and Pearson correlation coefficients (|Corr(t, X)| ≥ 0.45) of the cross-validated PLS-DA model [68]. The cutoff value of correlation coefficients was used to select the variables that were most correlated with the PLS-DA discriminant scores (PC1) at a significant univariate level of 0.05. Ten compounds have been identified via a NIST library search and they are Alanine, Lysine and Phenylalanine (amino acids), Citrate and 2-Oxoglutarate in TCA cycle (energy metabolism), ornithine and urea in urea cycle, Linoleate (unsaturated fatty acid), Creatinine and Cholesterol. Among them, Alanine, Urea and Phenylalanine were also identified in the tissue samples in a previous study [69]. These analysis results provide valuable pointers for further biological investigations about liver injury-induced metabolic disorder.

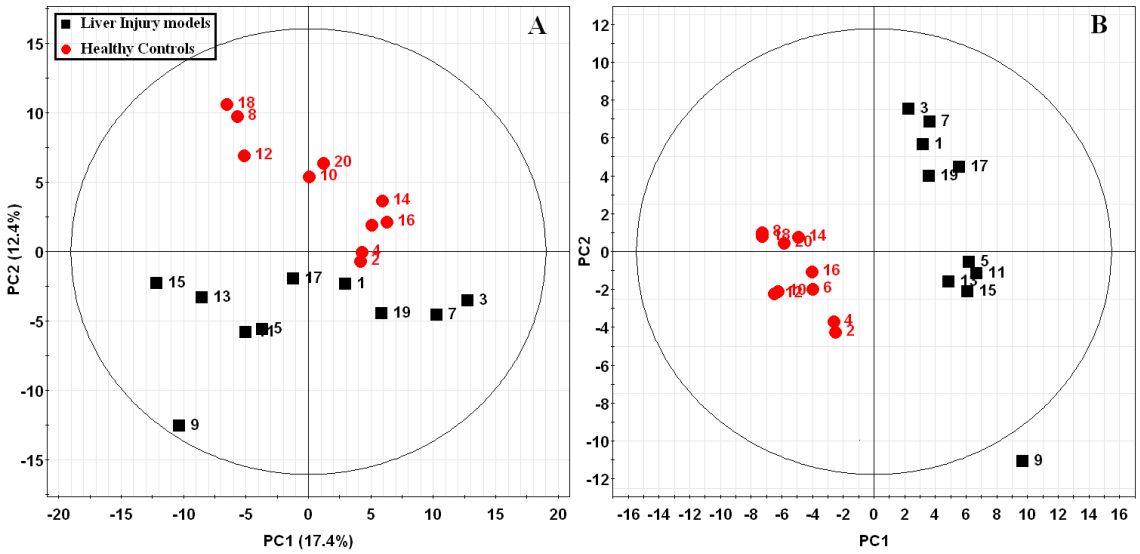

Figure 2.13. Multivariate statistical analysis of the quantitative metabolites data extracted from the LI samples by ADAP. Black and red markers correspond to liver injury (n=10) and healthy controls (n=10) samples, respectively. (A) PCA score plot. The first four principal components account for 45.6% of the total variance. (B) PLS-DA score plot. R2Y = 0.996 and Q2Y = 0.641 using two principal components in total.

2.4.2 ADAP-GC 2.0

Figure 2.14 illustrates the sequential data analysis workflow of ADAP-GC 2.0 using a pair of co-eluting compounds in Sample II. The width of the moving window for detecting TIC and EIC peak apexes is 9 scans for analyzing Sample I, II, and III. Figure 2.14 A depicts TIC peak apexes marked by red, green, and blue circles. Red and green indicate peak apexes of co-eluting components with red for the locally most intense one. The combined peaks marked by red and green form composite CPFs indicating the existence of co-eluting components. CPFs marked in blue are simple CPFs. Based on TIC CPFs, deconvolution windows were automatically determined (blue vertical lines) with co-eluting components in the same window. For the deconvolution window ranging from 19.75 to 20.0 minutes, all of the EIC CPFs whose peak apexes were within the window were determined. Subsequently, the aforementioned two-step filtering process filtered out

EIC CPFs that did not meet the criteria, and finally selected 46 good candidates for model CPFs. Figure 2.14 B and 2.14 C depicts the raw chromatogram and the constructed mirror images of all candidates. These candidates then participated in a hierarchical clustering for determining the number of components (Figure 2.14 D). Based on the empirical distance cutoff indicated by the red dashed line, two clusters were identified, which indicated that two components existed in this deconvolution window. Within each cluster, the CPF with the highest total score was designated as a model CPF. The two model CPFs corresponding to masses 273 and 245 are displayed. Subsequently, all of the EIC CPFs were decomposed into a linear combination of the model CPFs using constrained optimization (Figure 2.14 E). The resulting weights gave rise to the mass spectra depicted in Figure 2.14 F. By searching the spectra against an in-house library, they were matched to citric and *iso*-citric acid with the matching score being 975 and 935, respectively. The peak elution time of the two compounds were found to be at 19.82 and 19.88 min, respectively.

2.4.2.1 QUAL/QUAN Analysis

To evaluate the overall performance of ADAP-GC 2.0 in terms of compound identification, we have analyzed Sample I that contains a mixture of 38 standard compounds and compared the results with that from ADAP-GC 1.0 (Table 2.4). Clearly, ADAP-GC 2.0 was able to identify all of the compounds. Furthermore, the matching score calculated by AMDIS [26] for most of the compounds is higher by using the mass spectra constructed from ADAP-GC 2.0. The average matching score sees a 40-point increase, which demonstrates the significant improvement of the overall identification performance compared to ADAP-GC 1.0 (pairwise student t-test, p=0.017). In particular,

ADAP-GC 2.0 was able to identify cysteine that was sandwiched between cinnamic acid and creatinine around 12.25 mins. Because they shared multiple intense fragments and ADAP-GC 1.0 did not have the capability to decompose EIC CPFs of shared masses, ADAP-GC 1.0 was not able to identify cysteine at all.

We used Sample II and III wherein mixtures of 27 standard compounds were prepared at different concentrations to evaluate the quantitation performance of ADAP-GC 2.0. In particular, we used Sample III to test the capability of ADAP-GC 2.0 to identify and quantify compounds from complex samples. Table 2.5 lists all of the 27 standard compounds, their matching scores, and the coefficients of determination R^2 (estimated quantity vs. true quantity) for Sample II and III, respectively. The magnitude of the R^2 is an indicator of the performance of ADAP-GC 2.0 in terms of accurately extracting the quantitative information of compounds. In this work, the masses of model peaks are selected as the quantitation mass. In table 2.5, 2-chlorophenylalanine is an internal standard with a constant concentration across samples. It was used to evaluate the stability of the analytical platform and facilitate normalization of the estimated quantity of other compounds. After normalization, the R^2 value of each compound across samples was calculated. The higher (approaching to 1) the value, the more accurate the quantitation that is based on the deconvolution results. As we can see, all of the 26 compounds (excluding the internal standard chlorophenylalanine) in Sample II show very good linearity with high R^2 values (>= 0.99) as well as good identification results with high matching scores (average score = 890).

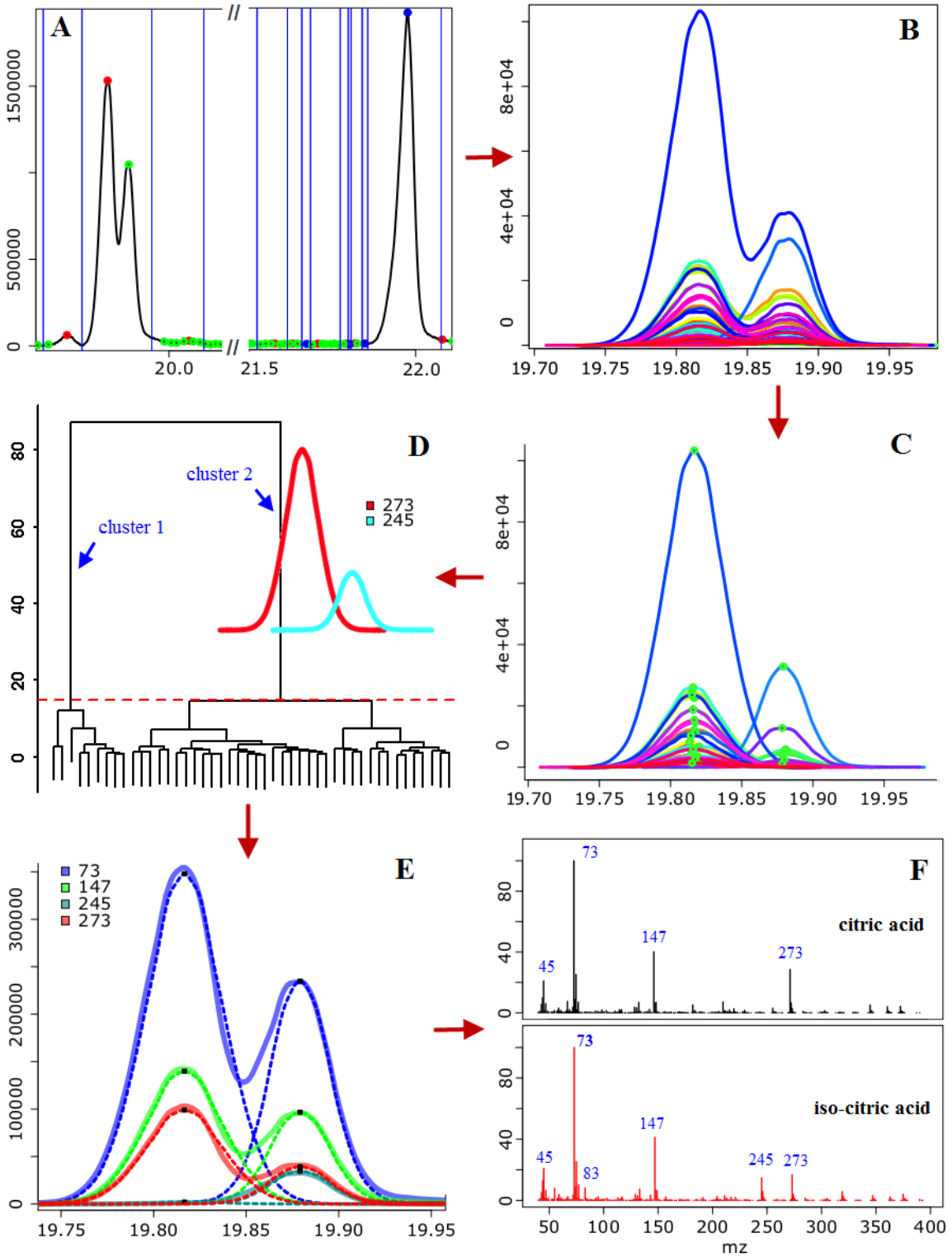

Figure 2.14. Illustration of the sequential data analysis workflow of ADAP-GC 2.0 using a pair of co-eluting compounds in Sample II. (A) Detection of CPFs from TIC and determination of deconvolution windows. Boundaries of deconvolution windows are marked by blue vertical lines. Two representative CPFs are displayed: one simple CPF

marked by a blue solid circle at the apex and one composite CPF marked by red and green solid circles at the apexes. Deconvolution of the EIC CPFs that have given rise to this composite TIC CPF is depicted in the subsequent sub-figures from (B) to (F). (B) Raw EICs of 46 good candidates. (C) The constructed mirror images of the 46 good candidates. (D) Determination of the number of components and corresponding model CPFs for each component using hierarchical clustering. The red dashed line indicates the empirical cutoff for determining the number of clusters. (E) The composite CPFs of mass 73, 147, 245 and 273 (solid line) were decomposed into simple CPFs (dashed line). (F) Two mass spectra were constructed with the maximum intensity normalized to 100. The two co-eluting components were identified as citric acid and *iso*-citric acid.

Table 2.4. Identification results of the 38 standard compounds from analyzing Sample I using ADAP-GC 2.0, in comparison with the results obtained using ADAP-GC 1.0.

No.	Compounds	ET (Min)	NIST (II)	Score	NIST (I)	Score
1	L-α-alanine	6.29	969		949	
2	L-leucine	8.51	910		896	
3	L-Proline	8.89	892		762	
5	Succinic acid	9.01	962		924	
4	Glycine	9.14	820		957	
6	L-Serine	9.72	908		772	
7	Pipecolinic acid	9.82	821		886	
8	β-Alanine	10.65	971		953	
9	4-Hydroxy-L-proline	11.88	922		700	
10	Trans-Cinnamic acid	12.25	929		866	
11	L-cysteine	12.25	658			
12	Creatinine	12.25	945		937	
13	α-Ketoglutaric acid	12.52	969		933	
14	L-asparate	12.61	883		954	
15	L-Phenylalanine	13.11	757		604	
16	n-Dodecanoic acid	13.55	961		864	
17	L-(+)-Arabionse	13.58	789		736	
18	DL-Homocysteine	13.65	856		822	
19	L-Asparagine	14.21	927		797	
20*	L-(+)-Rhamnose monohydrate	14.34	893		910	
	L-(+)-Rhamnose monohydrate	14.45	927		876	
21	L-(-)-arabitol	14.39	846		952	
22	1,4-diaminobutane	14.68	927		938	
23	L-Ornithine monohydrochloride	15.82	783		718	
24	1,5-Diaminopentane	16.14	978		947	

	dihydrochloride/Cadaverine			
25	n-Tetradecanoic acid(myristic acid)	16.39	963	548
26*	D-Fructose	16.62	913	940
	D-Fructose	16.76	980	943
27	L-Histidine	17.40	764	766
28	Indol-3-acetic acid	18.11	936	925
29	Palmitic acid/hexadecanoic acid	19.52	969	938
30	Dopamine hydrochloride	19.95	973	946
31	3-indolepropionic acid	20.03	957	938
32	Oleic acid	22.30	964	946
33	n-Octadecanoic acid	22.77	915	956
34	uridine	25.37	854	803
35	n-Eicosanoic acid	25.38	932	912
36	Sucrose	26.71	955	950
37	Estradiol	27.14	980	955
38*	Testosterone	27.40	982	967
	Testosterone	27.45	926	973
	Average		906	876

* The same compound was identified twice at two different elution times.

Table 2.5. Identification and quantification results of the 27 standard compounds from analyzing Sample II and III using ADAP-GC 2.0

		Sample II				Sample III			
No.	Compound Name	RT (Min)	Score	R^2	N^1	RT (Min)	Score	R^2	N
1	Pyruvic acid	5.17	925	1.000	7	5.17	949	0.978	8
2	Propanoic acid	5.34	974	0.991	7	5.34	970	0.998	8
3	β-Amino isobutyric acid	7.47	737	0.999	7	7.47	743	0.854	8
4	L-Leucine	8.40	915	0.999	7	8.40	863	1.000	8
5	Iso-leucine	8.73	750	0.998	7	8.74	838	0.998	8
6	Proline	8.78	960	0.999	7	8.78	894	0.996	8
7	Glyceric acid	9.34	973	0.999	7	9.34	969	0.999	8
8	Threonine	10.31	979	0.997	7	10.31	972	0.994	8
9	5-oxoproline	12.80	775	0.998	7	12.81	779	0.986	8
10	L-Cysteine	13.57	707	N.A.[2]	1	13.53	823	0.333	5
11	Creatinine	13.57	888	0.862	6	13.59	965	0.371	8
12	Citrulline	14.85	977	0.999	7	14.85	714	0.994	8
13	d-Xylose	15.94	955	1.000	7	15.94	781	0.978	8

14	Asparagine	16.15	798	0.987	7	16.16	760	0.993	7
13(2)	d-Xylose	16.16	955	0.997	7	16.17	968	0.992	8
15	1,4-Butanediamine	17.59	708	0.999	7	17.60	704	0.999	8
16	Glycerolphosphate	18.51	917	0.990	7	18.52	915	0.994	8
17	I.S.	18.95	951	/[3]	7	18.96	932	/[3]	8
18	Citric acid	19.81	964	0.999	7	19.85	980	0.000	8
19	Iso-citric acid	19.87	929	0.999	4	19.88	923	0.984	8
20	L-Histidine	21.92	925	0.999	5	21.95	933	0.980	8
21	L-Lysine	21.96	958	0.996	7	21.97	952	0.998	8
22	Mannitol	22.61	946	0.981	7	22.63	939	0.953	8
23	Galic acid	22.87	987	0.999	7	22.88	775	0.975	8
24	N-Acetyl glucosamine methoxime	25.97	792	0.999	6	25.97	703	0.999	4
25	L-Tryptophan	27.94	968	0.995	7	27.94	965	0.997	8
26	Adenosine	31.38	812	0.996	7	31.38	793	0.998	8
27	Guanosine	32.30	782	0.985	7	32.31	834	0.995	8
	Average value		890	0.991			869	0.901	

Note: 1. "N" is the number of samples from which a compound is identified. 2. N.A. means "not available." The R^2 value for L-Cysteine is not available because it was identified from only one sample. 3. "/" means "not calculated." The R^2 value was not calculated for 2-Chlorophenlyalanine because it served as an internal standard and its concentration was constant across all samples. I.S. is an internal standard: 2-Chlorophenylalanine.

It is worth noting that there are four pairs of co-eluting standard compounds with different degrees of overlapping in Sample II and III: (1) Asparagine and d-xylose: Higher and wider EIC peaks of d-xylose cover asparagine's smaller peaks under the left tails (Figure 2.15 A1), which caused shared EIC peaks to have more extended left side than the right (i.e., fronting); (2) Histidine and lysine: The degree of overlap between them is similar to that between asparagine and d-xylose, but the EIC of the early-eluting L-Lysine is not completely covered by that of the more intense L-Histidine, so the left tails of EIC peaks of shared fragments display significant fronting (Figure 2.15 A2); (3) Citric acid and iso-citric acid: They have similar mass spectra and their EIC peaks of

common fragments are characterized by two distinct humps (Figure 2.14 B); (4) Creatinine and cysteine: Their individual CPFs show high similarity because they co-elute almost entirely (Figure 2.16 B). All of these four pairs were accurately identified and quantified based on the deconvolution results of ADAP-GC 2.0. The accuracy of deconvolution is largely dependent upon the steps to select good peaks and select/construct model CPFs. In the process of developing and testing ADAP-GC 2.0, we have systematically evaluated EIC peak qualities, and tested the multiple metrics (SNR, Gaussian similarity, sharpness, peak apex intensity, and mass) and the best way to combine them into the total score as calculated in Eqn. (2) for selecting/extracting model CPFs. With this optimal combination of parameters, ADAP-GC 2.0 identified a total of 425 components from Sample III, requiring that each component exist in more than four of the eight samples. Because Sample III comprises of mixtures of standard compounds and urine samples, a component could come from the standards mixture, urine, or both. For those components that come from both, including creatinine, citric acid, and mannitol, their R^2 values are on the low end (Table 2.5). After examining the raw data, we found that they exist in high concentration in the urine sample and are beyond the linear dynamic range of the mass analyzer. Among the 425 components identified from Sample III, 308 components were resolved from decomposing composite CPFs and 220 components co-elute with their neighbors within 2 seconds. This indicates that a high percentage of potential compounds co-elute in real biological samples. Therefore, accurate deconvolution is essential for accurate QUAL/QUAN.

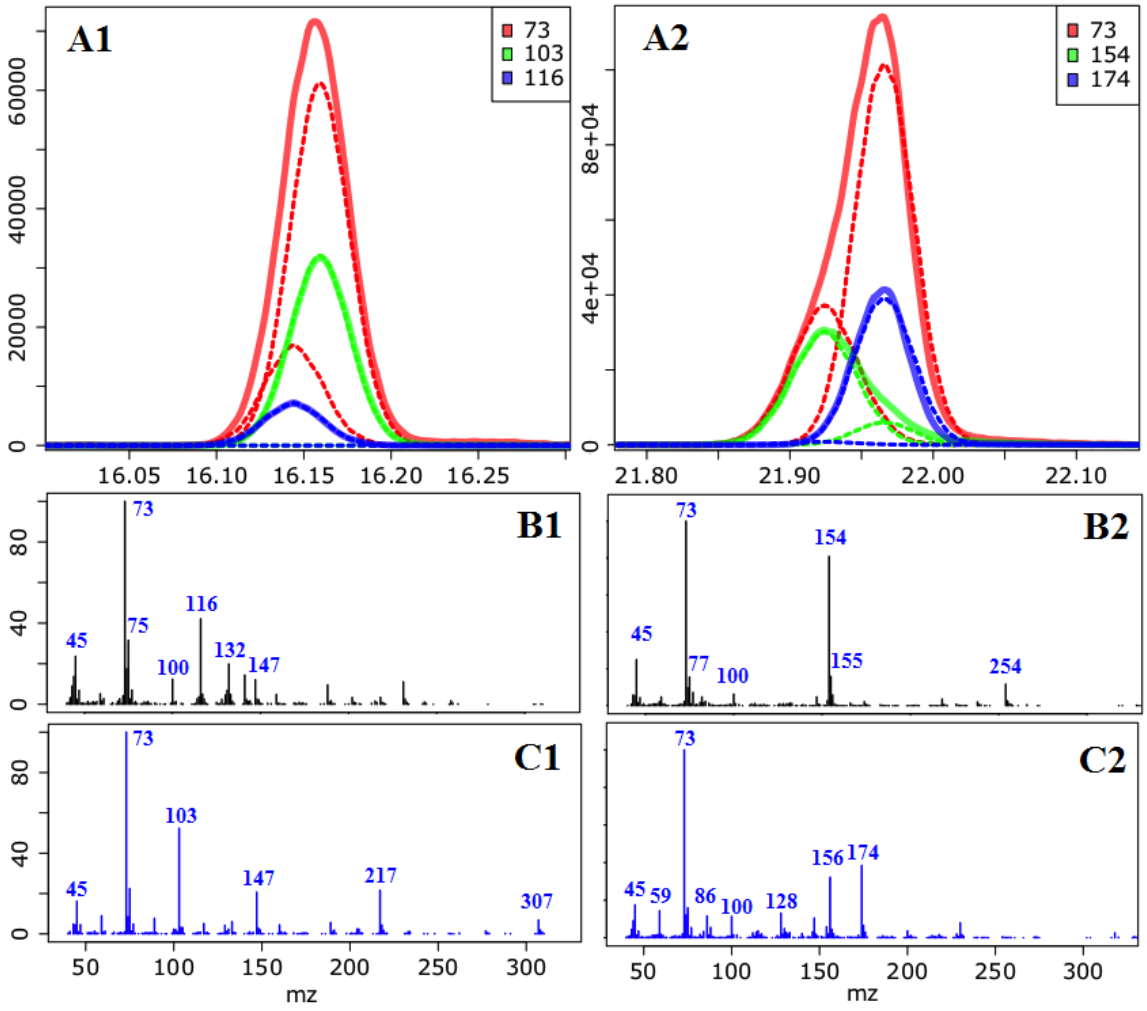

Figure 2.15. Deconvolution of two pairs of co-eluting components in two situations with different degrees of overlapping: almost complete overlap (Left) and partial overlap (Right) (A1) Raw EIC CPFs of masses 73, 103, and 116 (solid) and resulting simple CPFs (dashed) after deconvolution. The model CPFs are shown in green and blue corresponding to masses 103 and 116, respectively. (B1, C1) Mass spectra of the two components constructed from deconvolution results in (A1). They are identified as Asparagine and D-xylose, respectively, after library match. (A2) Raw composite EIC CPFs of masses 73, 154, and 174 (solid) and resulting simple CPFs (dashed) after deconvolution. The model CPFs are shown in green and blue corresponding to masses 154 and 174, respectively. (B2, C2) Mass spectra of the two components constructed from deconvolution results in (A2). They are identified as L-Histidine and L-Lysine.

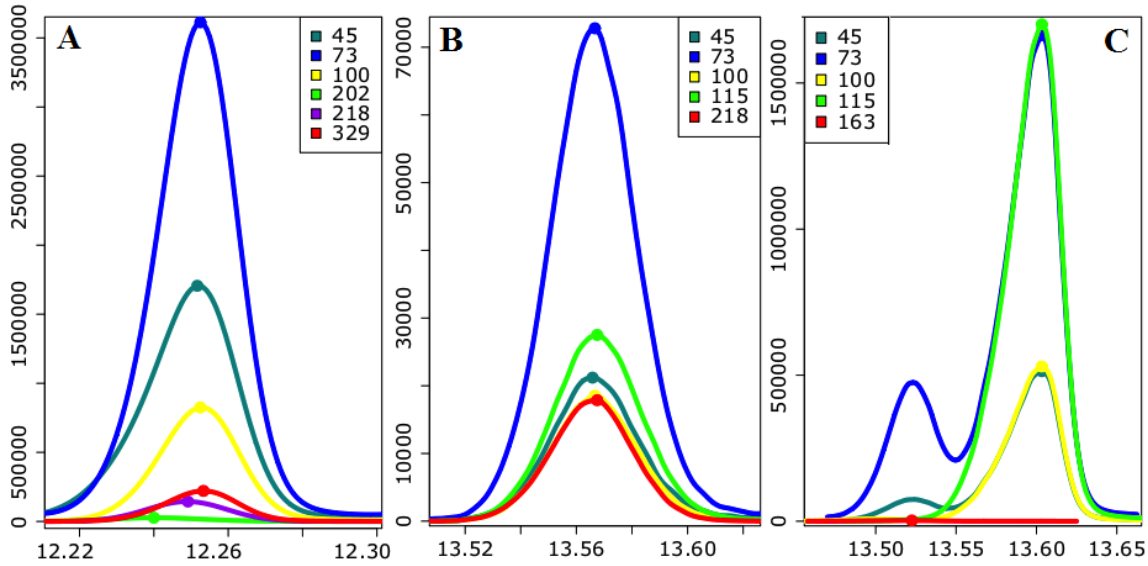

Figure 2.16. Comparison of cysteine co-eluting with neighboring compounds in Sample I, II, and III. (A) Cysteine co-elutes with cinnamic acid and creatinine in Sample I. ADAP-GC 2.0 successfully detected cysteine based on the slight difference in its elution profile from the co-eluting compounds. (B) The elution profile of cysteine overlaps with that of creatinine almost entirely in Sample II, which causes the failure of ADAP-GC 2.0 to detect cysteine from six of the seven samples. (C) Cysteine and creatinine were well resolved by chromatography in Sample III. ADAP-GC 2.0 successfully detected both of them.

2.4.2.2 Compound Splitting Issue

Compound splitting occurs when two or more model CPFs are constructed/selected for the same component within a deconvolution window. The last step of the deconvolution algorithm attempts to correct this issue and it is usually able to detect a majority of the cases. However, the solution is not targeted at the cause of the splitting and is therefore unable to resolve all the cases. The cause of the splitting issue lies in the hierarchical clustering, wherein a pre-specified distance cutoff is used to determine the number of components.

Prior to clustering, a combination of metrics is used to catch as many good candidates for model CPFs as possible. In particular, good candidates could be captured from low-intensity components that could otherwise be lost if without employing

multiple metrics. As a result, the probability of losing potential compounds is significantly reduced. On the other hand, too many CPFs participating in clustering dilutes the differences between the CPFs and can lead to more clusters than the actual number of components, i.e. the splitting issue. Therefore, there is a performance trade-off in the current deconvolution algorithm. The trade-off is between detecting as many compounds as possible and suffering from splitting issues. A better solution to the splitting issue demands a more robust method to determine the number of components in a deconvolution window.

2.4.2.3 Degree of Co-elution

The degree of co-elution of neighboring components influences the success rate of component detection. Take cysteine that exists in all of the three sample sets as an example (Figure 2.16). In Sample I, cinnamic acid co-elute with cysteine and creatinine sharing a number of intense fragment ions that include masses 45, 59, 73, 100, and 147 (Figure 2.16 A). However, the slight difference in apex elution time allows ADAP-GC 2.0 to extract model CPFs for the three compounds that correspond to masses 202, 329, and 218, respectively, and ultimately enables the successful identification of the three compounds. In Sample II, model CPFs from cysteine and creatinine (mass 218 and 115) are so similar (1.22° as calculated from the normalized dot product between two corresponding CPFs) that cysteine was detected only once from the seven constituent samples (Figure 2.16 B). In Sample III, cysteine and creatinine were well resolved by the chromatography and the resulting model CPFs are well separated (Figure 2.16 C). Masses of model CPFs for cysteine and creatinine are 163 and 115, respectively. With this distinct separation between the model CPFs, both compounds were identified with

matching scores above 800, even though cysteine appears to co-elute with another component based on CPFs of masses 45 and 73.

Clearly, the same set of neighboring components could have different degrees of co-elution in different samples. As long as a representative model CPF can be found for each co-eluting component, these components can usually be identified. However, when the difference between the elution profiles of co-eluting components is too small, we will have to resort to better chromatography systems to resolve them.

2.4.2.4 Robustness and Flexibility

ADAP-GC 2.0 has been developed using data generated primarily by a GC-TOF-MS platform that is configured to acquire spectra at a relatively high speed and produce integer mass measure. However, the pipeline can be applied to analyzing data from instruments with different scan acquisition rate, mass measurement resolution, and spectral bias (e.g, TOF is known to lose sensitivity at higher m/z) as well. This can be achieved by adjusting analysis parameters of ADAP, taking advantage of the built-in robustness of the deconvolution algorithm, and/or using existing capabilities of ADAP that are beyond the scope of this manuscript and thus not presented in the description of the data analysis method.

For instruments with lower scan acquisition rate, an EIC peak will consist of fewer sampling points. As a result, the moving window that is used to detect peak apexes and boundaries will cover fewer scans. If the width of the moving window is represented by the number of scans (alternatively by time), we will need to lower it to a smaller value. Additionally, since scan acquisition rate could affect the SNR of EIC peaks, the SNR threshold used to determine model CPFs usually needs to be adjusted accordingly. We

have tested ADAP-GC 2.0 on 14 standard mixture datasets generated from an Agilent 7890A gas chromatography system coupled with an Agilent 5975C inert XL EI/CI mass spectrometric detector (MSD) system (Agilent Technologies, Santa Clara, CA, USA). The scan acquisition rate was 2.57 scans/second. ADAP-GC 2.0 was able to successfully detect all the 17 standard compounds after parameter adjustment (data not shown).

For instruments with higher mass measurement resolution, three steps in the pipeline will need adjustment: extraction of EIC, detection of peaks, and selection of model CPFs. Extracting EICs for integer masses is achieved by simply grouping all of the observed intensity values based on the corresponding mass and then order each group by time or scan number. However, extracting EICs for high-resolution mass is more complicated. ADAP already has the module for this purpose. Peaks from the resulting EICs could have fewer sampling points than those from EICs of integer masses since masses that are within the same 1 m/z unit bin could be divided into multiple EICs when high-resolution mass measures are available. As a result, the same set of aforementioned parameters in the case of low scan acquisition rate need to be adjusted.

ADAP-GC 2.0 uses multiple factors in selecting/constructing model CPFs. The complementary nature of these factors makes the pipeline robust in processing data from different instruments. In the meantime, the possibility to adjust analysis parameters based on data makes the pipeline flexible. Currently, this adjustment has to be done manually. Ideally, the pipeline should be able to determine the optimal set of parameters based on the data to analyze, thus making the pipeline self-adjustable.

2.4.3 ADAP-GC 3.0

Based on ADAP-GC 1.0, we developed ADAP-GC 2.0 with improved deconvolution performance by implementing simple/composite peak feature detection, five metrics of peak qualities to select model peaks, and the constrained optimization to decompose shared peak features into a linear combination of simple ones [70]. Despite the significant progress in the accuracy of compound identification and quantitation, we continue exploring the limitations of ADAP-GC 2.0 especially in peak detection and model peak selection. During peak detection, the simple local extrema (maxima for peak apex and minimum for peak valley) method is very sensitive to the window width parameter. For model peak selection, peaks that are symmetric and resemble a Gaussian curve are favored. However, fronting and tailing do occur and cause asymmetric peak shapes to happen even when a compound elutes from the chromatography system alone. In addition, ADAP-GC 2.0 determines the number of co-eluting compounds by carrying out a hierarchical clustering of good candidates for model peaks. However, hierarchical clustering is very sensitive to the distance threshold and therefore it is very challenging to select an appropriate threshold and strike the balance between compound splitting that occurs when the distance threshold is too low and compound merging that occurs when the distance threshold is too high.

Thus, ADAP-GC 3.0 is developed in order to address these issues in peak detection, model peak selection, and hierarchical clustering. During peak detection, ADAP-GC 3.0 applies the wavelet transform to automatically identify peaks with different widths and peak shapes, and then uses local extrema to examine if there exists significant spikes to ensure the purity of unique/simple peak features. After that, unique

peak features with the highest sharpness values are considered as model peaks to represent the elution profiles of corresponding compounds. The criteria for unique peak features together with signal to noise ratio and sharpness values have replaced the previous five matrices of parameters to select model peaks, which is a linear combination of Gaussian curve fitting value, signal to noise ratio, mass, and intensity values at apex. Finally, two rounds of hierarchical clustering are applied to determine the total number of co-eluting compounds by measuring the closeness of retention time and similarity of peak shape of EIC peak features. With the sophisticated strategy to select model peaks and perform two rounds of clustering, and apply less number of parameters for tuning, ADAP-GC 3.0 has been developed as a robust and adaptive pipeline, and is able to identify metabolites at low concentration levels and/or co-eluting compounds overlapping very closely with neighbors. In this paper, we report the new computational approach and the associated algorithms of peak detection and deconvolution, as well as the improvement in terms of compound identification and quantitation as compared with the previous version. ADAP-GC 3.0 algorithms are prototyped in R and being incorporated into our developing stand-alone software tool which integrates data processing, statistical analysis, compound identification and visualization.

As compared to the previous version, ADAP-GC 3.0 has major improvements on peak feature detection, model peak selection and the clustering strategy to determine the total number of co-eluting components. Figure 2.17 illustrates the key steps of deconvolution in ADAP-GC 3.0 using a pair of co-eluting compounds from one standard mixture sample. Within the deconvolution window spanning from 8.62 to 9.91 minutes (Figure 2.17 A), the 1st hierarchical clustering produced three clusters, and the one with

only mass 89, whose absolute abundance is as low as 237, is considered as a noisy peak then removed (Figure 2.17 C). The 2nd clustering is performed on the selected unique peak features of each cluster, producing one component from each with the distance cutoff as 15 (Figure 2.17 D-E). Each component shows significantly similar unique peak features in terms of elution time and elution profiles (Figure 2.17 F-G). The unique peak features of mass 158 at 8.83 min and mass 147 at 8.78 min have the highest sharpness values thus are selected as model peaks (Figure 2.17 B). Finally, two mass spectra are resolved and identified as compound iso-leucine and proline with matching score 846 and 988, respectively (Figure 2.17 H-I).

By using the same testing datasets wherein 27 standard compounds were carefully selected and mixed at different concentrations, it is direct to compare the performance of compound identification and quantitation between ADAP-GC 3.0 and the previous version. Table 2.6 lists all of these 27 standard compounds identified from Sample II and III using ADAP-GC 3.0. It is clearly that compound identification and quantitation results have been improved in terms of the increased average matching scores against the user library and the linearity coefficient R^2 values. The result indicates that it is successful that ADAP-GC 3.0 applies the combination of wavelet transform and local extrema for peak detection and selects the unique peak features as model peak candidates. Also, the unique peak with the highest sharpness value is able to represent the elution profile of corresponding compound.

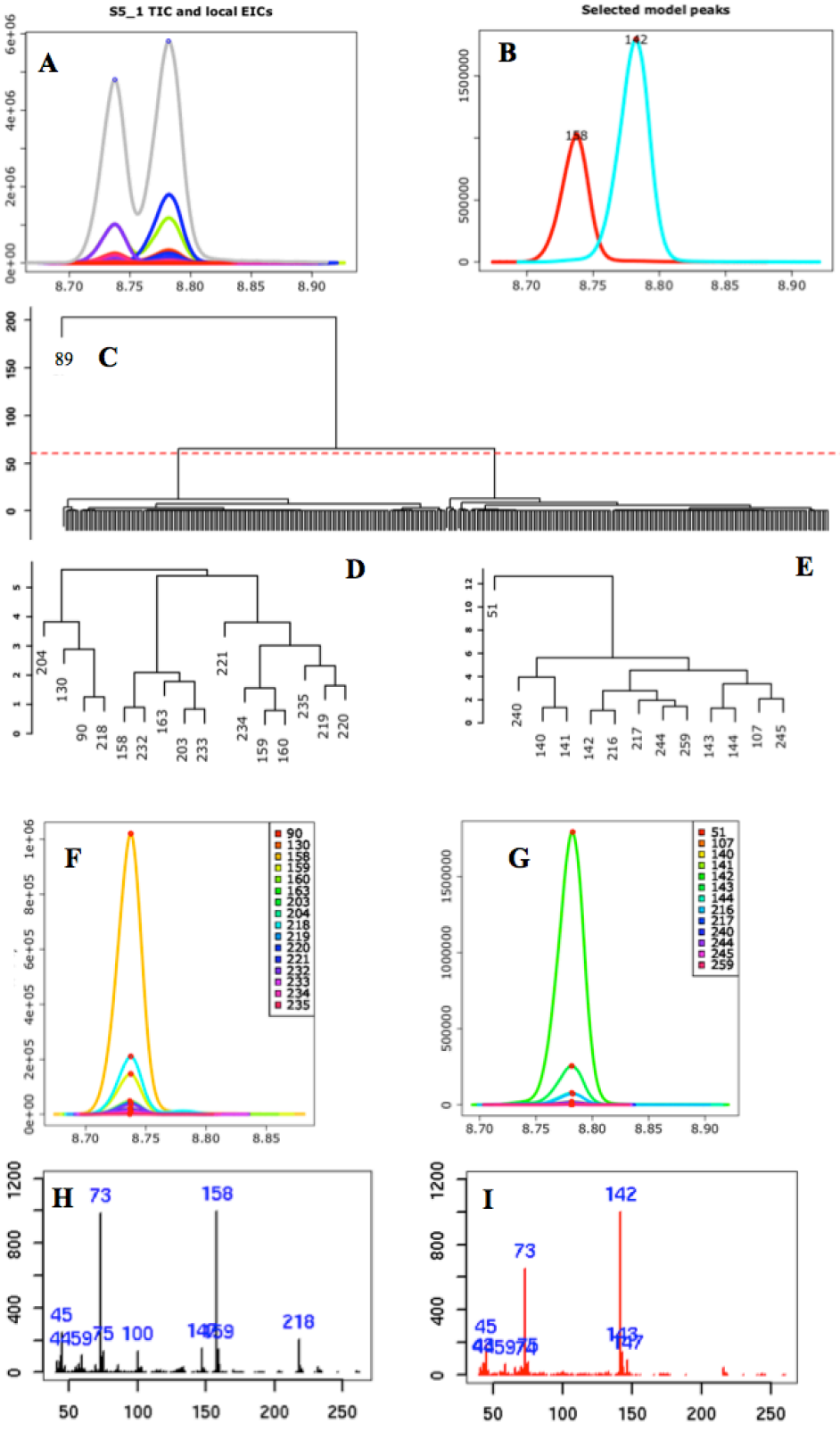
S5_1 TIC and local EICs
A
Selected model peaks
B
142
158
C
89
D
E
F
G
H
73
158
45
75
100
218
I
142
73
45
143

Figure 2.17. Illustration of the first and second round of hierarchical clustering in order to determine the number of co-eluting components and their corresponding model peaks.

For the previous ADAP-GC 2.0, most missing compounds were at lower concentration levels and their peak features generally exhibit noisy, zigzag, less close to standard Gaussian curves, so that none candidate peak features had met the criteria of five matrices of parameters for model peaks. For example, the iso-citric acid and histidine had not been identified previously from Sample II at the concentrations of 0.2 and 0.4 μg/mL. Iso-citric acid coelutes with citric acid, sharing most peak features together (Figure 2.18 A1, B1). Histidine exhibits much lower concentration as compared to the coeluting compound lysine (Figure 2.18 C1). So it is difficult to find all model peaks for both pairs of coeluting compounds using ADAP-GC 2.0. However, ADAP-GC 3.0 is able to identify these compounds in both Sample II and III (Table 2.6) since it is able to find the qualified unique peak features as model peaks or extract the mass spectra nearby through the careful examination of two rounds of clustering (Figure A2, B2 and C2). To note, among 27 standard compounds in Sample II, it is difficult to automatically deconvolute the co-eluting cysteine and creatinine due to that they elute within only two scans and most of their peak features overlap exhibiting as simple peak features. Except for this particular case, ADAP-GC 3.0 is able to correctly find back those standard compounds in Sample II undetected by the ADAP-GC 2.0, and has identified all the standard compounds from Sample III.

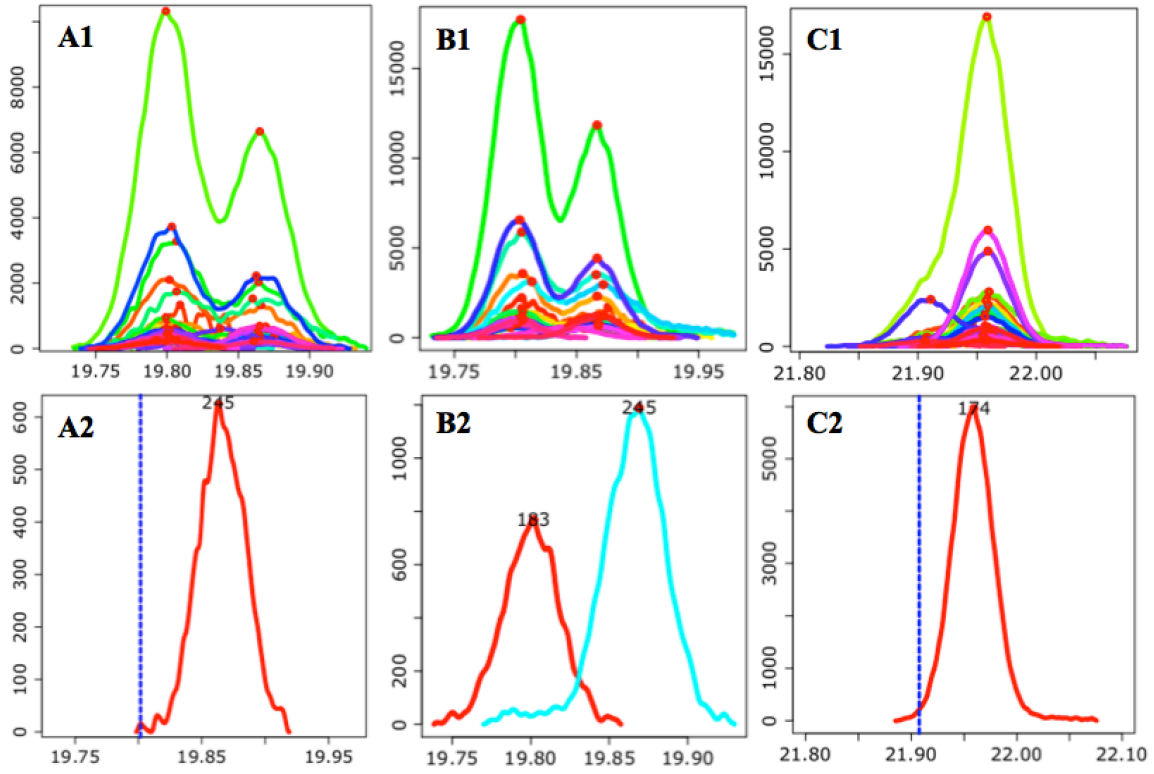

Figure 2.18. (A1-A2). At the lowest concentration of 0.2 µg/mL, citric acid and iso-citric acid share most peak features together and only the unique peak feature 245 is qualified as the model peak for iso-citric acid while the mass spectrum of citric acid can be extracted at the position at 19.81 min indicated by the 1st round of clustering. (B1-B2) At the higher concentration 0.4 µg/mL, both citric acid and iso-citric acid have identified their model peaks. (C1-C2). At the concentration 0.4 µg/mL, histidine exhibits much lower abundance than the co-eluting compound lysine, thus only the unique peak feature 174 is selected for compound identification while the mass spectrum of histidine can be extracted at the position 21.91 min.

Table 2.6. Identification and quantification results of 27 standard compounds from samples II and III using ADAP-GC 3.0

		Sample II (7 samples)					Sample III (8 samples)				
No.	Compound Name	RT (min)	Mass	R^2	Score	Count	RT (min)	Mass	R^2	Score	Count
1	Pyruvic acid	5.17	174	0.996	933	7	5.17	174	0.977	939	8
2	Propanoic acid	5.34	117	0.999	981	7	5.34	117	0.996	976	8
3	β-Amino isobutyric acid	7.47	102	0.999	938	7	7.47	102	0.885	897	8
4	L-leucine	8.4	158	0.996	915	7	8.4	158	0.998	852	8
5	isoleucine	8.73	158	0.994	856	7	8.74	158	0.998	847	8
6	Proline	8.78	142	0.996	982	7	8.78	142	0.998	938	8
7	Glyceric acid	9.34	73	0.994	974	7	9.34	189	0.996	968	8
8	Threonine	10.31	117	0.994	954	7	10.31	117	0.996	975	8
9	5-oxoproline	12.8	156	0.996	924	7	12.81	157	0.994	916	8
10	L-Cysteine#	13.57	73	/	842	2	13.54	307	0.373	715	8*
11	Creatinine#	13.57	73	0.999	867	5	13.59	115	0.347	968	8
12	Citrulline	14.85	73	0.992	947	7	14.85	142	0.994	925	8
13	D-Xylose	15.93	73	0.995	939	7	15.94	103	0.993	842	8
14	Asparagine#	16.15	116	0.993	756	7	16.16	116	0.992	785	8*
13(2)	D-Xylose#	16.16	103	0.989	958	7	16.17	103	0.998	965	8
15	1,4-Butanediamine	17.59	174	0.992	958	7	17.6	174	0.999	955	8
16	Glycerolphosphate	18.51	73	0.994	890	7	18.52	299	0.853	847	8
17	Chlorophenylalanine	18.95	218	/	954	7	18.96	218	/	932	8
18	Citric acid#	19.81	183	0.992	933	7	19.85	273	0.891	946	8
19	Isocitric acid#	19.87	245	0.992	901	7*	19.89	245	0.978	834	8
20	L-Histidine#	21.93	154	0.989	893	7*	21.95	154	0.958	899	8
21	L-Lysine#	21.96	174	0.989	950	7	21.97	174	0.992	908	8
22	Mannitol	22.61	73	0.992	945	7	22.63	103	0.859	942	8
23	Galic acid	22.87	73	0.994	970	7	22.88	281	0.961	912	8
24	N-Acetyl glucosamine methoxime	25.97	202	0.994	888	7*	25.96	129	0.996	848	8*

Table 2.6 (Continued)

25	L-tryptophan	27.94	73	0.993	965	7	27.94	202	0.995	964	8
26	Adenosine	31.38	73	0.991	894	7	31.38	230	0.995	927	8
27	Guanosine	32.31	73	0.988	821	7	32.31	324	0.991	865	8
	Average value			0.993	919*				0.926*	903*	

Note: the value of mass means the selected quantitation mass. Score is the average matching score by searching against the standard library. Count is the number of samples from which a compound is identified. The symbol * indicates the improvement as compared to the ADAP-GC 2.0. The symbol / means the R^2 value is not calculated.

2.5 Conclusion

ADAP-GC 1.0 has established the basic framework of an automated data analysis pipeline for GC-MS data analysis with features peak picking, deconvolution, alignment, and identification. In particular, we also developed the novel component-based alignment that has been validated the robustness and efficiency in multiple datasets. However, Deconvolution needs to be improved to deconvolute EICs of shared ions more accurately. So ADAP-GC 2.0 defined the concept of simple and composite peak feature during peak detection, and developed the model peak-based deconvolution method using five metrics of peak qualities to select model peaks. However, ADAP-GC 2.0 heavily relies on multiple parameters in peak detection and deconvolution, which made it difficult to select optimal parameter settings for datasets from different sources. Thus, ADAP-GC 3.0 has been developed as a flexible and robust pipeline with new algorithms implemented in peak detection, hierarchical clustering, and model peak selection. Furthermore, ADAP-GC 3.0 has been validated its superior performance in compound identification and quantitation compared to the previous ADAP-GC 1.0 and ADAP-GC 2.0.

CHAPTER 3: COMPARATIVE EVALUATION OF SOFTWARE FOR COMPOUND IDENTIFICATION AND QUANTITATION OF GC-TOF-MS DATA IN METABOLOMICS STUDIES

3.1 Introduction

Data processing plays a critical role in translating raw signals into mass spectra and peak abundance of biochemical compounds, thus has big impact on extent and quality at which metabolite identification and quantitation can be made as well as on the ultimate biological interpretation of results [71]. Among five critical steps for GC-MS data processing, i.e., de-noising, feature detection, deconvolution, alignment, compound identification and quantitation [71], deconvolution is particularly important during feature detection due to that a large number of compounds frequently co-elute after one-dimensional GC separation [44]. As summarized in Table 1.1, a variety of methods and software packages have been developed and applied in GC-MS based metabolomics studies for data processing in recent years. However, only five different software tools have their own capability of spectral deconvolution: AMDIS, ADAP-GC 3.0, AnalyzerPro, ChromaTOF, and MetaboliteDectctor. Among them, MetaboliteDetector frequently met crashes during the time of our testing, thus ADAP-GC 3.0, AMDIS, AnalyzerPro, and ChromaTOF are selected in this study for comparing their performance on compound identification and quantitation. Lu et al. compared the deconvolution performance of AMDIS, ChromaTOF, and AnalyzerPro [51]. As far as we know, this is the first time a comparison was carried out, which is invaluable. However, it was done from a user's point of view by simply calculating the false positive and negative rate of

compound identifications as the criterion to evaluate the deconvolution performance, and studied the effect of combinations of parameters used in the data processing stage. Further analysis of the challenges for data processing and causes of limitations for current software tools will be critical for future development and improvement.

In this study, four representative software tools, ADAP-GC 3.0, AMDIS, AnalyzerPro and ChromaTOF, are selected for comparison. Specifically, these four software tools perform peak feature detection and deconvolution on raw data of two different sample sets, which are seven standard mixture samples and eight pooled urine samples mixed with standards analyzed by GC-time of flight-MS platform. GC-TOF-MS provides higher mass resolution and mass accuracy compared to conventional GC-MS, and its faster scan rate improves Gaussian peak shape, which is very useful for accurate deconvolution [24]. Through identifying and quantifying a total of 27 standard compounds, we evaluate their advantages and limitations in peak feature detection and spectral deconvolution of GC-TOF-MS data for metabolomics studies. Our work is helpful for software users to better understand data processing results and select appropriate software tools for data analysis. Also, further discussion of common limitations and possible solutions will guide our software developers to develop novel computation algorithms and strategies efficient enough for processing metabolomics data, especially in peak detection and deconvolution.

3.2 Materials and Methods

3.2.1 Experimental Procedures and Testing Data

Two different sample sets are used and analyzed by four representative software tools for comparison, and they are sample II and III that have been introduced in details in chapter two.

3.2.2 Software Comparison

ADAP has been developed since 2009 for GC-MS-based metabolomics studies, with features of de-noising, peak detection, deconvolution, alignment, and compound identification and quantitation [43]. The second version ADAP-GC 2.0 applies "model peak" method to improve spectral deconvolution significantly [44]. ADAP-GC 3.0 is the newest version, which has improved peak detection and model peak selection, and has trayoptimized clustering strategy to determine the number of co-eluting compounds. National Institute of Standards and Technology (NIST) developed AMDIS in 1999. AMDIS (Version 2.71) has been a commonly used freeware for spectrum extraction and compound identification from GC-MS data [26]. A 15-day trial of AnalyzerPro (Version 3.0.0.0) is vendor-independent software, which is provided by Spectralworks Ltd, UK. AnalyzerPro enables batch processing of multiple datasets from metabolomics studies, and provides baseline correction, smoothing, peak finding, deconvolution, compound identification and quantitation for GC-MS data. ChromaTOF software (version 4.34) is developed exclusively for the use with LECO separation science instruments, e.g., GC-TOF-MS. It is powerful with comprehensive features including baseline correction, peak finding, deconvolution, alignment, compound identification and quantitation, as well as user-friendly interfaces for data overview, transfer and report.

Raw NetCDF format data produced from GC-TOF-MS instrument was processed at first to extract mass spectra and peak abundance information, which involves de-nosing, peak detection, and deconvolution. Parameters of data processing for each software tool have been set appropriately to keep their performance comparable. Their key parameters and specific report formats are listed in Table 3.1. To note, ADAP-GC 3.0 is able to output resolved mass spectra and peak abundance information automatically, while for others, extra programming work is necessary in order to interpret software specific reports into detailed spectra and abundance information.

After obtaining extracted mass spectra and peak abundance information of resolved components, we applied our user library to search for a total of 27 standard compounds in each sample of Sample II and III. The matching score threshold is set as 700 for identification, the greater matching score is (maximal value is 999), the more accurate mass spectra resolved from deconvolution step. Meanwhile, R^2 coefficient value (estimated quantity vs. true quantity) of each standard compound is calculated to measure its linearity across calibration samples. Since different model peaks or quantitation mass could be selected for a same compound from different samples, the most frequent one is considered as the common quantitation mass for R^2 coefficient calculation. Overall, performance of each software is evaluated based on: (1) compound identification: the total number of deconvoluted components from each dataset, matching scores of identified standard compounds and the true positive rate; (2) compound quantitation: linearity of estimated concentration of each standard compound across samples; (3) analysis of mass spectra accuracy from four software tools to examine their advantage and shortcomings in peak detection and deconvolution. The workflow to compare the

performance of compound qualification and quantitation of four software tools has been summarized in Figure 3.1.

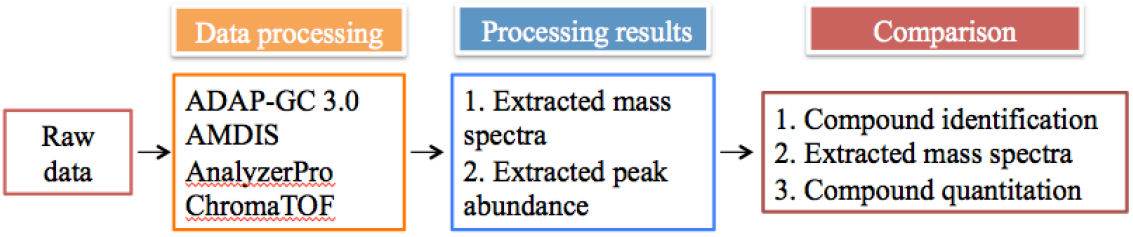

Figure 3.1. Workflow of comparative evaluation of software tools on the performance of compound qualification and quantitation

Table 3.1. List of key parameters and specific report formats of four software tools

Software	Key parameters	Report formats
ADAP-GC 3.0	Minimum S/N = 10; Hierarchical clustering distance cutoff = 15; Score cutoff for splitting correction = 750;	Extracted mass spectra list (.MSP); excel table saving the abundance information of all peaks (.CSV)
AMDIS	Minimum S/N = 10; Component width = 32; Adjacent peak subtraction = 1 Resolution = medium; Sensitivity = very low; Shape requirements = high.	Extracted mass spectra and peak abundance (.ELU and .FIN); excel table saving the quantitation mass abundance of each component (.CSV)
ChromaTOF	Minimum S/N = 10; Baseline offset = 1.0; Data points for averaging = 3; Peak width = 2.5.	Extracted mass spectra list (.MSP); excel table saving the quantitation mass abundance of each component (.CSV)
AnalyzerPro	Area threshold = 500 Height threshold = 1% Minimum mass = 6 S/N threshold = 10 Smoothing = 3 Width threshold = 0.01 min Resolution = low Scan window = 2	Excel table peak abundance and mass information (.XLS)

3.3 Results

3.3.1 Compound Identification

Table 3.2 and 3.3 summarize the identification and quantitation results of a total of 27 standard compounds from Sample II and III. ADAP-GC 3.0, AMDIS, AnalyzerPro,

and ChromaTOF could identify 25, 21, 20 and 24 standards, respectively, within seven datasets of Sample II, and 27, 15, 25 and 27, respectively, in eight datasets of Sample III. It seems that ADAP-GC 3.0 and ChromaTOF produced comparable results in terms of the number of identified compounds and their matching scores. Among five pairs of co-eluting compounds in Sample II, cysteine and creatinine co-elute so close within only one to two scans distance near 13.57 minutes and share most common peak features together (Figure 3.2 A), thus it is difficult to completely resolve them. But they elute much more independently in urine samples with 12 to 30 scans distance, so that it becomes easier to separate them during deconvolution. Besides cysteine and creatinine, ChromaTOF failed to identify histidine at the lowest concentration (0.2 μg/ml). It seems mass 154 is the only significant peak feature unique to histidine, however, its noisy profile and relatively low abundance compared to co-eluting lysine at 21.95 minutes prevents it to be detected easily (Figure 3.2 B). But ADAP-GC 3.0 had noticed that there existed at least one compound at 21.92 minutes based on the hierarchical clustering of eluting times of EIC peak features at the beginning of deconvolution.

Compared to ADAP-GC 3.0 and ChromaTOF, AMDIS and AnalyzerPro have missed more compounds in Sample II and III with different reasons. AMDIS tends to produce more than one mass spectrum for a compound, and co-eluting compounds with higher intensities could dominate and affect the extraction of mass spectra nearby. Take histidine for example which was only identified at 5 μg/ml in Sample II, a total of 12 mass spectra were identified as lysine from 21.92 to 21.97 minutes at 2 μg/ml (Figure 3.3), but none of them was qualified as histidine with confident matching score (the threshold is 700). AnalyzerPro lost even more compounds because more than 50% of

those missing compounds have obtained incomplete mass spectra and their common quantitation mass was failed to be extracted for quantitation. For example, there are two incomplete mass spectra resolved at 15.932 and 15.938 minutes matched against d-xylose with score 793 and 791, respectively (Figure 3.4). However, both of them seem to be a part of the standard mass spectrum of d-xylose. Further, the one with higher matching score does not have the quantitation mass 73, thus this compound had not been quantified successfully.

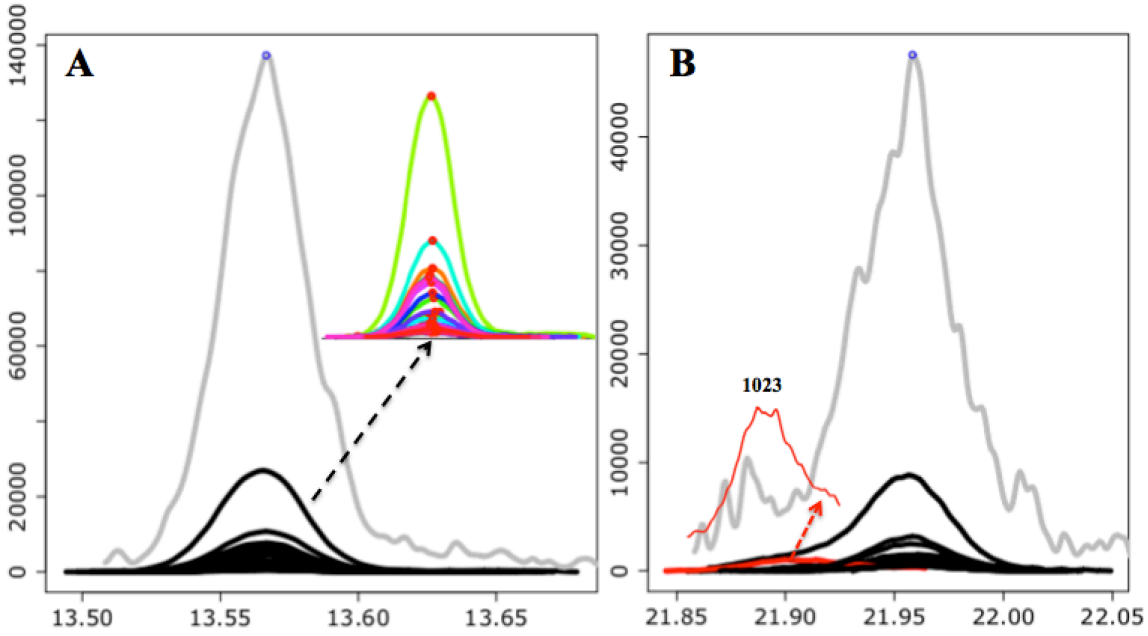

Figure 3.2. (A) An example of TIC (grey) and EIC (black and in color) peak features for cysteine and creatinine in Sample II. (B) TIC and EIC eluting profiles of histidine at the lowest concentration, peak feature of mass 154 is zoomed in and labeled in red.

Table 3.2. Standard compound identification and quantitation results from seven datasets of Sample II

			ADAP-GC 3.0				AMDIS				AnalyzerPro				ChromaTOF			
No.	Compound Name	RT	Mass	N	Score	R^2	Mass	N	Score	R^2	Mass	R^2	Score	N	Mass	N	Score	R^2
1	Pyruvic acid	5.17	174	7	933	0.996	73	7	891	0.995	73	0.996	896	7	174	7	932	0.996
2	Propanoic acid	5.34	117	7	981	0.999	73	7	971	0.998	73	0.999	962	7	117	7	978	0.999
3	β-Amino isobutyric acid	7.47	102	7	938	0.999	102	7	943	0.999	102	0.999	934	7	102	7	944	0.999
4	L-leucine	8.4	158	7	915	0.996	158	7	906	0.996	158	0.996	917	7	158	7	915	0.996
5	isoleucine	8.73	158	7	856	0.994	158	7	843	0.996	158	0.996	839	6	158	7	851	0.994
6	Proline	8.78	142	7	982	0.996	142	7	979	0.996	142	0.996	961	7	142	7	987	0.996
7	Glyceric acid	9.34	189	7	974	0.994	73	7	966	0.994	73	0.995	962	7	189	7	975	0.994
8	Threonine	10.31	117	7	954	0.994	73	7	954	0.993	73	0.995	947	7	73	7	975	0.995
9	5-oxoproline	12.8	156	7	924	0.996	156	7	924	0.993	156	0.996	899	7	156	7	928	0.996
10	L-Cysteine	13.57	73	2	842	/	73	5	827	0.985	73	0.997	783	3	115	2	822	/
11	Creatinine,	13.57	73	5	867	0.999	73	4	875	0.923	73	0.956	828	4	115	5	880	0.999
12	Citrulline	14.84	73	7	947	0.992	73	7	895	0.994	73	0.996	892	4	142	7	945	0.991
13	d-Xylose	15.93	73	7	939	0.995	73	6	894	0.995	73	0.999	872	4	103	7	936	0.996
14	Asparagine	16.15	116	7	756	0.993	116	5	783	0.998	116	0.996	743	5	116	7	795	0.993
13(2)	d-Xylose	16.16	103	7	959	0.989	103	7	927	0.981	73	0.997	833	5	103	7	968	0.989
15	1,4-Butanediamine	17.59	174	7	958	0.992	174	7	954	0.985	174	0.992	929	7	174	7	960	0.992
16	Glycerolphosphate	18.51	73	7	890	0.994	73	6	831	0.99	73	0.997	805	5	73	7	894	0.994
17	I.S.	18.95	218	7	954	/	73	7	932	/	73	/	900	6	218	7	954	/
18	Citric acid	19.81	183	7	933	0.992	73	7	929	0.995	73	0.996	855	6	183	7	957	0.992
19	Isocitric acid	19.87	245	7	901	0.992	73	7	867	0.994	73	1	819	4	245	7	909	0.992
20	L-Histidine	21.92	154	7	893	0.989	73	1	735	/	154	/	802	1	154	6	882	0.993
21	L-Lysine	21.96	174	7	950	0.989	73	7	943	0.985	73	0.988	887	7	174	7	950	0.989
22	Mannitol	22.61	73	7	945	0.992	73	7	928	0.99	73	0.992	888	7	73	7	943	0.992

Table 3.2 (continued)

23	Galic acid	22.87	73	7	970	0.994	73	7	948	0.996	73	0.994	890	7	281	7	984	0.993
24	N-Acetyl glucosamine methoxime	25.97	202	7	888	0.994	73	5	863	0.996	73	0.992	788	6	73	7	915	0.996
25	L-Tryptophan	27.94	73	7	965	0.993	73	7	959	0.99	73	0.995	954	6	202	7	966	0.991
26	Adenosine	31.38	73	7	894	0.991	73	7	887	0.992	73	0.991	858	7	73	7	913	0.991
27	Guanosine	32.31	73	7	821	0.988	73	7	822	0.991	73	0.991	789	6	73	7	826	0.987
	Average				919	0.994			899	0.990		0.994	873				924	0.994

Note: I.S. is internal standard.

Table 3.3. Standard compound identification and quantitation results from eight datasets of Sample II

			ADAP-GC 3.0			AMDIS				AnalyzerPro				ChromaTOF		
No.	Compound Name	RT	Mass	Score	R^2	Mass	N	Score	R^2	Mass	N	Score	R^2	Mass	Score	R^2
1	Pyruvic acid	5.17	174	939	0.977	73	8	932	0.973	73	6	910	0.977	174	941	0.977
2	Propanoic acid	5.34	117	976	0.996	73	8	974	0.987	73	8	966	0.994	117	978	0.996
3	β-Amino isobutyric acid	7.47	102	897	0.885	102	8	847	0.881	102	8	850	0.885	102	875	0.886
4	L-leucine	8.4	158	852	0.998	158	8	904	0.998	158	8	910	0.998	158	897	0.998
5	isoleucine	8.74	158	847	0.998	73	8	836	0.994	158	8	831	0.998	158	831	0.998
6	Proline	8.78	142	938	0.998	142	8	933	0.997	142	8	933	0.997	142	979	0.998
7	Glyceric acid	9.34	189	968	0.996	73	8	965	0.998	73	8	958	0.998	189	968	0.996
8	Threonine	10.31	117	975	0.996	73	8	962	0.991	73	8	956	0.994	219	972	0.996
9	5-oxoproline	12.81	157	916	0.994	73	8	938	0.983	156	8	913	0.995	156	948	0.994
10	L-Cysteine	13.54	307	715	0.373	220	8	783	0.405	220	6	729	0.9	218	764	0.847
11	Creatinine,	13.59	115	968	0.347	116	8	972	0.124	115	8	953	0.342	115	980	0.343
12	Citrulline	14.85	142	925	0.994	142	8	890	0.991	73	6	882	0.99	142	919	0.994

Table 3.3 (continued)

13	d-Xylose	15.94	103	842	0.993	73	8	929	0.848	73	6	849	0.952	103	921	0.993
14	Asparagine	16.16	116	785	0.992	75	7	837	0.954	116	8	759	0.993	132	788	0.997
13(2)	d-Xylose	16.17	103	965	0.998	73	8	966	0	73	8	946	0.998	307	963	0.997
15	1,4-Butanediamine	17.6	174	955	0.999	73	8	923	0.817	174	8	916	0.999	174	956	0.999
16	Glycerolphosphate	18.52	299	847	0.853	73	8	895	0.853	299	6	785	0.998	299	869	0.983
17	Chlorophenylalanine	18.96	218	932	/	73	8	912	/	73	6	855		218	931	/
18	Citric acid	19.85	273	946	0.891	73	8	978	0.149	73	8	940	0.84	273	970	0.702
19	Isocitric acid	19.89	245	834	0.978	245	8	766	0.831	245	7	741	0.976	245	775	0.976
20	L-Histidine	21.95	154	899	0.958	73	8	884	0.826	154	7	806	0.956	154	872	0.838
21	L-Lysine	21.97	174	908	0.992	73	8	922	0.978	156	8	864	0.994	174	934	0.993
22	Mannitol	22.63	103	942	0.859	73	8	949	0.152	73	8	917	0.967	319	945	0.937
23	Galic acid	22.88	281	912	0.961	281	7	919	0.957	281	6	856	0.975	281	926	0.966
24	N-Acetyl glucosamine methoxime	25.96	129	848	0.996	73	8	810	0.997	73	5	812	0.997	202	867	0.995
25	L-Tryptophan	27.94	202	964	0.995	202	8	961	0.992	202	8	950	0.994	202	973	0.994
26	Adenosine	31.38	230	927	0.995	73	8	921	0.997	73	7	890	0.997	236	933	0.995
27	Guanosine	32.31	324	865	0.991	73	8	847	0.998	73	7	820	0.995	324	877	0.99
	Average			903	0.926			906	0.803			875	0.952		913	0.940

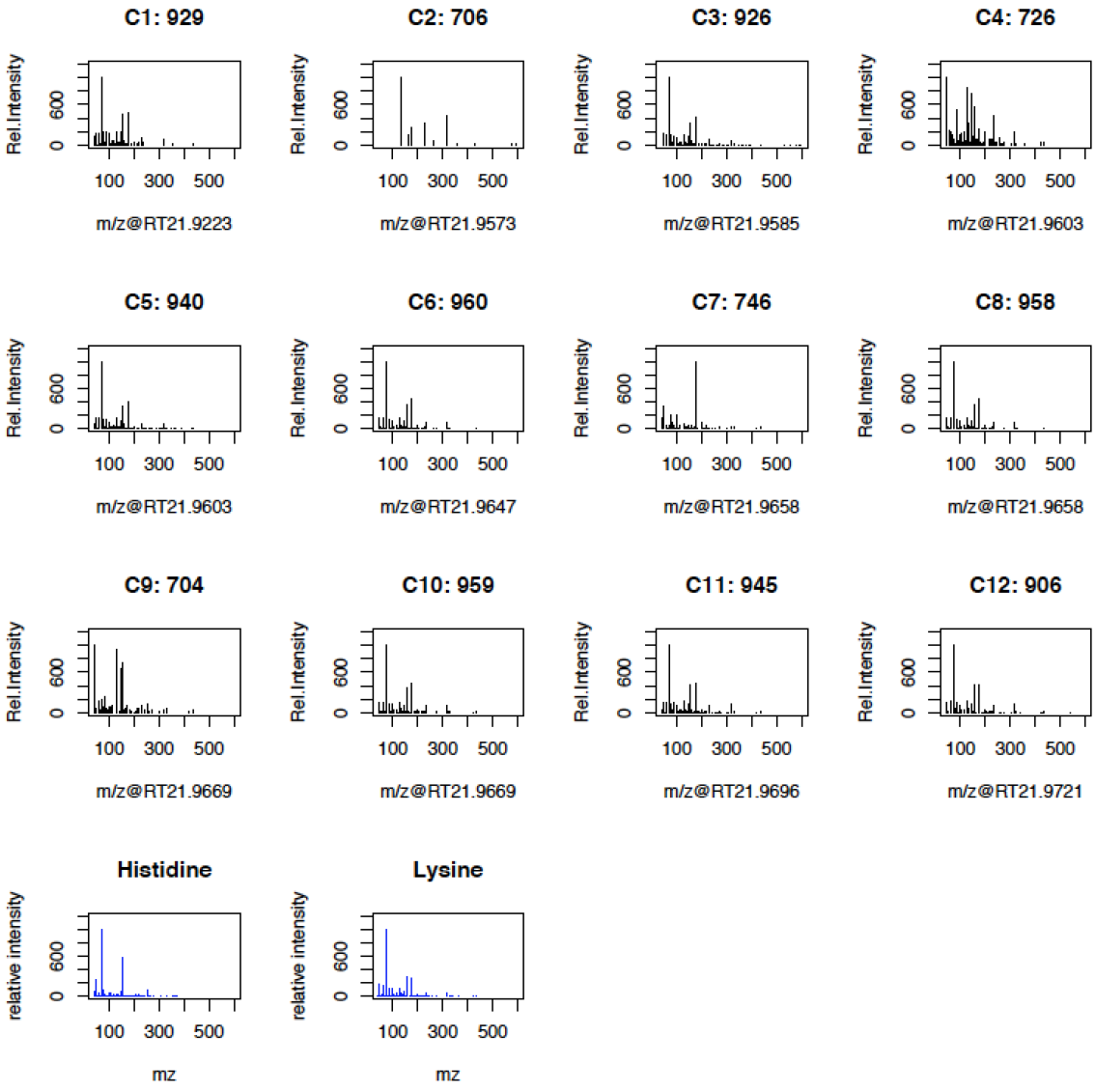

Figure 3.3. Comparison of extracted mass spectra from 21.92 min to 21.97 min from a sample of Sample II (2 µg/ml) that have been matched to lysine with scores greater than 700 and standard spectra of compound histidine and lysine.

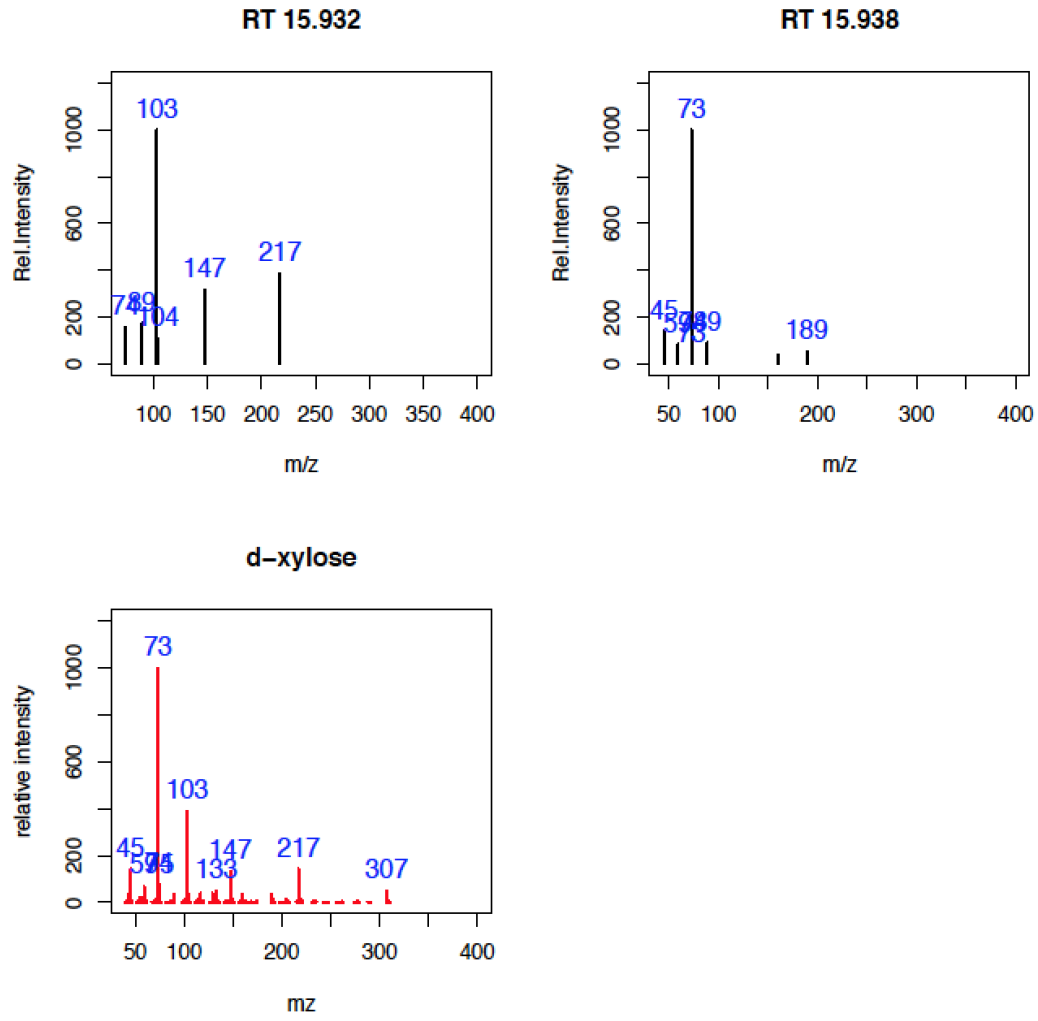

Figure 3.4. An example of two incomplete mass spectra resolved at 15.93 min by AnalyzerPro from a sample in Sample II (0.6 µg/ml), where theoretically exist d-xylose (bottom left).

Average matching scores of 27 standard compounds are 919, 899, 873, and 924 in Sample II and 903, 906, 875, and 913 in Sample III for ADAP-GC 3.0, AMDIS, AnalyzerPro, and ChromaTOF, respectively. Higher matching scores indicate more accurate extraction of mass spectra from original data. However, these software tools produced much more components than the theoretical number 28 in Sample II, and surprisingly, AMDIS produced around 5000 to 6000 thousands of components (Table

5.4). In the urine samples of Sample III, ADAP-GC 3.0, AnalyzerPro, and ChromaTOF produced relatively reasonable number of components, however, AMDIS still produced too many even a single component could employ different models for spectra construction. With the advantages of fast acquisition rate, GC-TOF-MS is able to produce chromatographic features with more scan points and improved Gaussian peak shapes. But for AMDIS, which is not originally developed for TOF data, if peak tops are broad, e.g. from GC-TOF-MS, with several local maxima present, more than one spectrum could be produced and identified for a compound. Similarly, noisy peaks are broad as well to AMDIS, thus extra false positive components have been produced in both datasets. Even though AMDIS has provided options of peak filtering, only signal to noise threshold was used and set as low as 10 to make parameter settings comparable with other two software packages. The parameter "sensitivity" during deconvolution was set "very low" to reduce the possibilities of noisy and broad peaks determined. However, these parameter adjustments did not change a lot.

Table 3.4. List of the number of components resolved by three software tools in each dataset of Sample II and III. Sample ID "S0.1" indicates the concentration of each standard compound is 0.1 ug/ml in this sample.

Sample ID	ADAP-GC 3.0		AMDIS		AnalyzerPro		ChromaTOF	
	II	III	II	III	II	III	II	III
S0.1		960		5563		743		938
S0.2	151	956	6015	5770	87	777	229	963
S0.4	134	978	5893	5746	90	757	221	986
S0.6	147	1044	5960	5740	110	837	237	1035
S0.8	141	1054	5942	5735	116	838	238	999
S1	144	1017	5975	5708	120	785	235	995
S2	178	982	5551	5776	145	788	260	999
S5	190	1031	5994	5742	217	813	302	1065

3.3.2 Compound Quantitation

ADAP-GC 3.0 selects model peak abundances to represent the relative concentrations of compounds, whose peak features are unique and have the highest sharpness values. AMDIS also selects model profiles through measuring sharpness characteristics of EIC peaks. But their measurements of sharpness are different: AMDIS considers noise factor while ADAP-GC 3.0 does not because it has de-noising as the first step of data processing. AnalyzerPro does not provide options how to select quantitation mass, thus the base peak with the highest intensity for each component is used for quantitation. Both ADAP-GC 3.0 and ChromaTOF provides multiple choices of quantitation mass for customized analysis, e.g. the most abundant unique peak or the summary of all EIC peak abundance. As a result, ADAP-GC 3.0 and ChromaTOF are found that they shared more than 70% of quantitation mass for 27 standard compounds.

Higher matching scores indicate more accurate mass spectra resolved for compound identification and quantitation. Together with high average matching scores for compound identification, all four software tools produced good quantitation results in Sample II with average R^2 values greater than 0.99. However, quantitation of standards in urine samples is complex because there exist hundreds of metabolites with diverse biochemical properties and a wide range of concentrations. As a result, a total of 17, 10, 17, and 17 compounds out of 27 have R^2 values greater than 0.99 in Sample III for ADAP-GC 3.0, AMDIS, AnalyzerPro, and ChromaTOF, respectively. The lower R^2 values of others indicate different degrees of impurity or inaccuracy of resolved mass spectra affected by noise or co-eluting compounds. Three out of 27 standard compounds (i.e., creatinine, citric acid and mannitol) have poor quantitation performance because

they themselves exist in the urine samples and their high concentrations have been beyond the linear dynamic range of TOF-MS analyzer.

3.3.3 Mass Spectra Comparison

It is interesting to compare resolved mass spectra of 27 standard compounds from four different software tools with standard spectra from our user library, as well as to compare their pairwise similarities to evaluate overall deconvolution performance. From Table 3.5, mass spectra from ADAP-GC 3.0, AMDIS and ChromaTOF have high similarities, even higher than their average matching scores against library. This indicates high consistency and accuracy of mass spectra resolved from these three software tools while the standard spectrum from our user library exists minor difference from them. For example, fragments 245 and 273 consistently appeared in the top candidate of iso-citric acid from ADAP-GC 3.0, ChromaTOF and AMDIS, and particularly, mass 273 has higher abundance level than the other. On the contrary, the standard spectrum has higher abundance level of mass 245 than that of mass 275 (Figure 3.6). We also noticed that ADAP-GC 3.0 and ChromaTOF showed the highest similarity score, indicating their deconvolution performance are very comparable. However, it is clearly that mass spectra extracted from AnalyzerPro have least similarities with others, which also explains the effects of incomplete mass spectra from deconvolution step.

3.4 Discussion and Conclusion

Four software tools with their own spectral deconvolution algorithms for GC-MS data are compared through identifying and quantifying a total of 27 standard compounds from standard mixtures and urine samples mixed with standards. All four software tools are able to identify most of standard compounds with matching scores greater than 700,

and quantify these compounds across calibration samples with R^2 coefficients greater than 0.99. Among them, ADAP-GC 3.0 and ChromaTOF performed the best and produced comparable results in terms of the percentage of true positives, the selected quantitation mass, average matching scores and R^2 coefficients. While AMDIS tend to produce multiple mass spectra for a compound from GC-TOF-MS data, which makes it difficult for automated compound identification. Also, abundant co-eluting compounds could easily affect the extraction of mass spectra of compounds nearby. AnalyzerPro produced much fewer false positives than AMDIS. However, incomplete mass spectra are found common from AnalyzerPro results, thus more compounds were failed identified.

Table 3.5. Average similarities of resolved mass spectra from four software tools against library of 27 standard compounds in each sample of Sample II (first four columns) and their pairwise similarities. “AD”, “AM”, “Chrom”, and “An” are short for ADAP-GC 3.0 AMDIS, AnalyerPro, and ChromaTOF, respectively.

Sample ID	AD	AM	C	An	AD-AM	AD-An	AD-C	C-Am	C-An	Am-An
S0.2	900	863	902	860	923	588	922	904	553	609
S0.4	911	885	924	870	948	549	950	927	506	411
S0.6	911	895	931	864	946	549	962	941	504	562
S0.8	923	907	931	860	949	549	959	934	547	550
S1	931	918	934	882	959	532	970	949	504	571
S2	930	929	935	900	960	507	971	966	497	522
S5	935	932	932	903	962	478	985	963	479	520
Average	920	904	927	877	950	536	960	941	513	535

Note: AD, AM, C, An represents software ADAP-GC 3.0, AMDIS, ChromaTOF, AnalzyerPro, respectively.

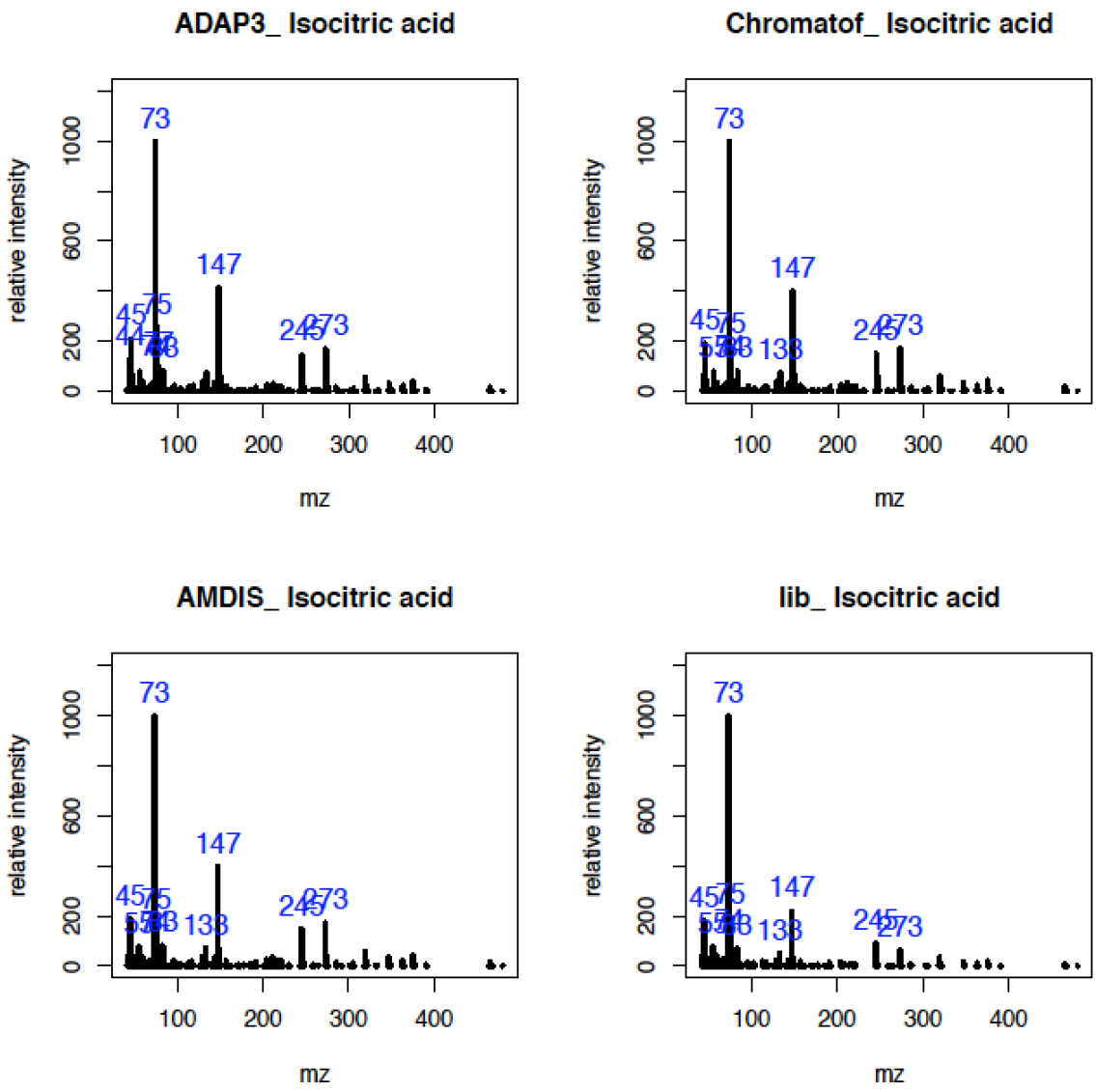

Figure 3.6. Comparison of mass spectra identified as iso-citric acid from ADAP-GC 3.0, ChromaTOF, and AMDIS with standard library.

Comparing different software in terms of compound identification and quantitation enables us to find their common issues in deconvolution: large amount of false positive components, and multiple mass spectra for a same compound (known as splitting issue), which affect the purity of extracted mass spectra for compound identification and quantitation. At least three steps of data processing are correlated with these problems: (1) de-nosing: noises that are wrongly considered as signals will increase

false positive components; (2) peak detection: it is difficult to keep the balance to find peaks with different widths because small peaks are easily undetected which results in incomplete mass spectra resolved and too sensitive methods, e.g. local maxima, could detect multiple local tops within a broad peak feature which result in multiple mass spectra resolved; (3) component perception which is the step to determine the number of co-eluting components, e.g., ADAP-GC 3.0 applies hierarchical clustering based on similarities of peak features and closeness of their retention times, and AMDIS evaluates whether there exists a sufficient number of ions maximizing together. The improper parameter setting related with component perception could directly result in compound missing or splitting issue.

In the future, development of software packages for GC-MS data processing with application in metabolomics should consider these factors to improve the performance of compound identification and quantitation. Based on our own experience, it is highly recommended to utilize samples from background runs and/or quality controls to reduce the interferences of background and random noises. In order to satisfy different chromatographic peak conditions, robust methods are highly required to comprehensively identify peaks with different feature characteristics, e.g., ADAP-GC 3.0 has combined transformed wavelets and local maxima together to improve peak detection. Lastly, it should be admitted that it is difficult to determine the number of components in an untargeted way for unknown biological samples, thus relevant parameter settings play a critical role in this step. The window or case specific parameter settings could be flexible and helpful to develop automated and robust methods during deconvolution.

In conclusion, both ADAP-GC 3.0 and ChromaTOF perform well in terms of compound identification and quantitation by processing two different sets of GC-TOF-MS data, while AMDIS and AnalyzerPro seems to be inappropriate to deconvolute GC-TOF-MS data in an untargeted way for compound identification and identification, and should require extensive correction, filtering and reorganization for metabolomics studies. ADAP-GC 3.0 is promising in the field of GC-MS based metabolomics studies because continuous efforts have been made to improve data processing performance since its' first version published in 2009. It aims to be developed as a freely available software tool with automated data processing and sophisticated statistical analysis capabilities. ChromaTOF is a commercial software tool, but it has been validated powerful in processing and analyzing GC-TOF-MS-based metabolomics data, and users benefit a lot from recent new features, e.g., statistical analysis methods and user-friendly interfaces.

CHAPTER 4: DEVELOPMENT OF VISUALIZATION SOFTWARE AND STATISTICAL ANALYSIS METHODS FOR GC-MS DATA ANALYSIS

4.1 Introduction

Modern analytical technologies afford comprehensive and quantitative investigation of a large number of metabolites. And running large-scale projects with hundreds to thousands of samples in metabolomics studies is on the verge of being routine. Thus, the resultant large and complex datasets require advanced bioinformatics tools for data processing, analysis and biological interpretation. Like other omics, sophisticated computational tools are vital for efficient and high-throughput analysis, to eliminate systematic bias and to explore biologically significant findings [72].

In metabolomics, data handling generally include three sequential steps: data processing, data pretreatment and statistical analysis. Data processing aim to extract identity and quantity of compounds from original data [71], however, only a few software packages (e.g., MetAlign [31] and MET-IDEA [38]) may own and develop novel algorithms to process GC-MS, particularly GC-TOF-MS data. Data pretreatment methods include normalization, centering, scaling and transformation, that have been applied in metabolomics with the goal to focus on biological information and to reduce the influence of disturbing factors such as measurement noise [67]. It has been pointed out that data pretreatment is a crucial step that can drastically change the

pertinence and the outcome of data analysis [73]. Finally, advanced multivariate statistical methods are often used together with univariate analysis, e.g., student *t*-test, to investigate relationships between different groups and to highlight differential metabolites that contribute to the relationship. Popular multivariate analysis methods used for metabolomics include principal component analysis (PCA) to examine natural clustering of samples and partial least squares discriminant analysis (PLS-DA), clustering analysis, and support vector machines (SVM) to supervise the group difference (e.g., case-control) [74, 75].

In this study, we develop ADAP-GC software based on the novel algorithms of data processing in ADAP-GC 2.0 and a statistical package for data pretreatment and analysis. Specifically, ADAP-GC software has four main features: (1) an integrated tool with seamless data processing, identification and quantitation (QUAL/QUAN) analysis, statistical analysis, visualization, and customized summary report. (2) handling and controlling each step internally so that we do not have to rely on other software that may change methods in the future. (3) Modular based pipeline: data processing and analysis happens in steps so that each step is saved to ensure rollback. It is very important to save time when running large scale data sets like in epidemiology studies: if certain parameters need to be adjusted at one step, the user could start from this step instead of starting over; (4) Quality checking and correction allow user interaction with ADAP-GC software. Both identification and qualification results could be checked manually by experienced analysts and allow semi-automatic correction of compound identification and missing values.

4.2 ADAP-GC Software

4.2.1 Workflow of ADAP-GC Software

The general workflow of ADAP-GC software can be divided into five modules: (1) analysis for parameter settings of automated data processing and analysis; (2) visualization of raw TIC/EIC chromatograms, extracted mass spectra, identified peaks and details about deconvolution, which is helpful to users to understand high-dimensional data, evaluate data processing performance and interpret analysis results; (3) Qual/Quan table listing the compound identification and quantitation result and allowing manual checking and correction; (4) statistics for basic data pretreatment and statistical analysis; and (5) a customized html report.

ADAP-GC software accepts netCDF format data produced from GC-MS platform. All the raw data and intermediate results from data processing will be organized and saved in a SQL database to facilitate easy and fast data retrieval for computation and visualization. To use ADAP-GC software, it usually starts with the raw netCDF files, and proceeds through data processing, Qual/Quan analysis, data pretreatment, statistical analysis and final report (Figure 4.1). Once a job finishes, a project folder is then created that stores raw netCDF files, intermediate data processing results and final data analysis report (Figure 4.2). Among them, the extracted mass spectra are saved in NIST format so that users could apply MSsearch software for further compound identification. And qual/quan tables are exported in CSV format allowing further statistical analysis and data exploration.

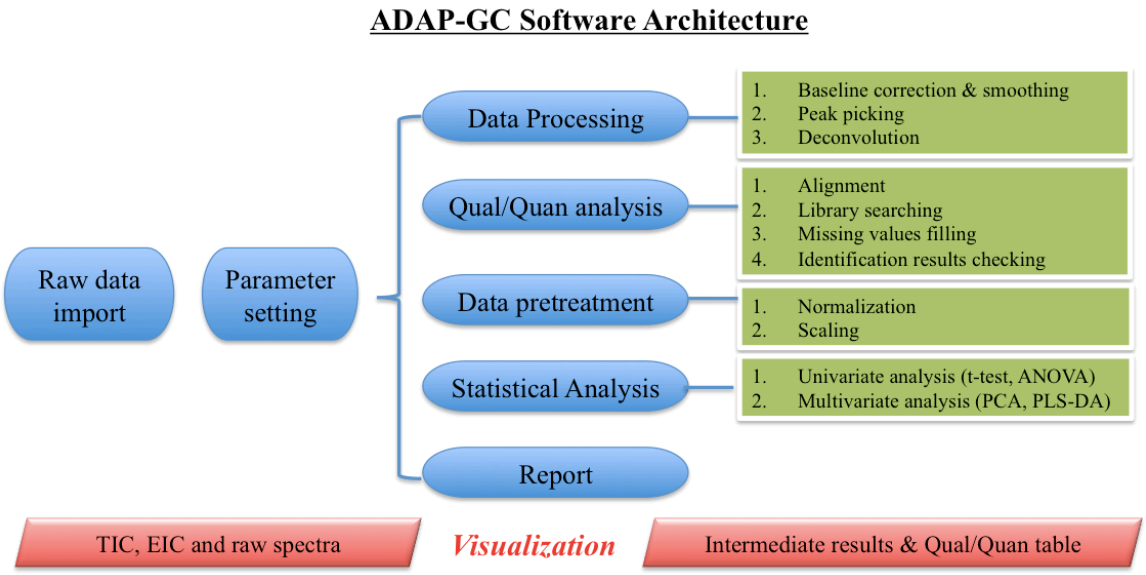

Figure 4.1. ADAP-GC software architecture

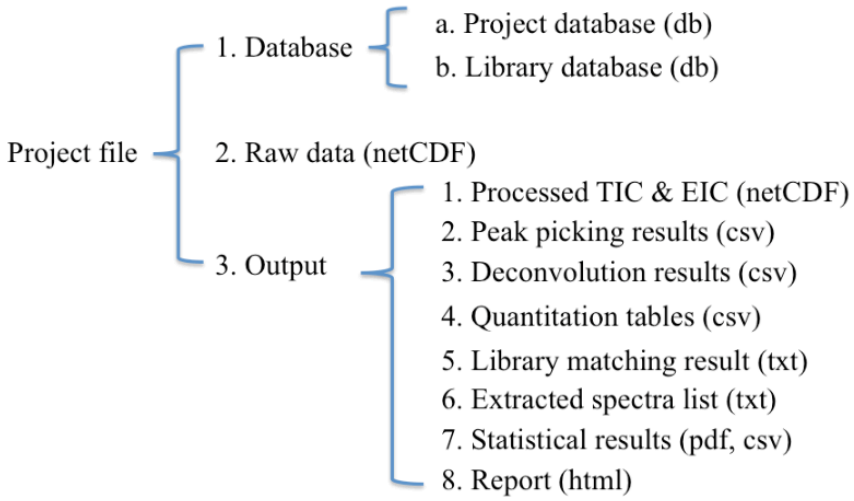

Figure 4.2. The structure of a project folder created by ADAP-GC software

4.2.2 Parameter Settings

ADAP-GC software provides full choices of parameter settings for each step of data processing and analysis (Figure 4.3): (1) smoothing and baseline correction have

window size to adjust, their default values are 20 and 240 scans respectively; (2) Peak picking covers the ratio of boundary intensities to peak apex intensity, window size for peak apex and boundary detection, the allowed maximal peak width and signal to noise ratio. (3) Deconvolution has a total of ten parameters, such as sharpness, signal to noise ratio (SNR), and Gaussian curve fitting score for model peak selection, the cutoff of pairwise spectra similarity to examine compound splitting issue. During deconvolution, one or more standard libraries in NIST format can be used for compound identification, where the default score cutoff for library matching is set as 750. (4) Alignment provides four parameters: instrumental acquisition rate, retention time tolerance, spectra similarity score, and the minimal number of samples having a same compound identified. (5) One or more standard libraries can be used for library searching after alignment, and users could decide the number of top candidates displayed in the Qual/Quan table. More details about parameters and data processing algorithms have been introduced Chapter three of ADAP-GC 2.0. Default parameters are optimal for GC-TOF-MS data with the acquisition rate set as 20 spectra per second; however, users are allowed to explore different parameter settings according to specific GC-MS instrument conditions and data analysis requirements.

4.2.3 Statistical Analysis

The goal to develop statistical package within ADAP-GC software is to analyze data directly from data processing steps without having to use third-party software, and to integrate commonly used statistical methods for metabolomics studies within a same pipeline. In the current version of ADAP-GC software, computational functions of statistical methods were written in R language by applying many available functions and

libraries from Cran (http://cran.r-project.org) and Bioconductor (http://www.bioconductor.org), including PCAMethods, mixOmics, pls, gplots, and limma. The statistical package can be divided into three parts: data pretreatment (normalization and scaling), data exploration (clustering and PCA) and significance analysis (univariate and multivariate analyses) (Figure 4.4). More statistical methods that are increasingly used for metabolomics studies will be added to our software in the future. For example, receiver operator characteristic (ROC) curves have been applied in recent translational biomarker discovery of clinical metabolomics [76].

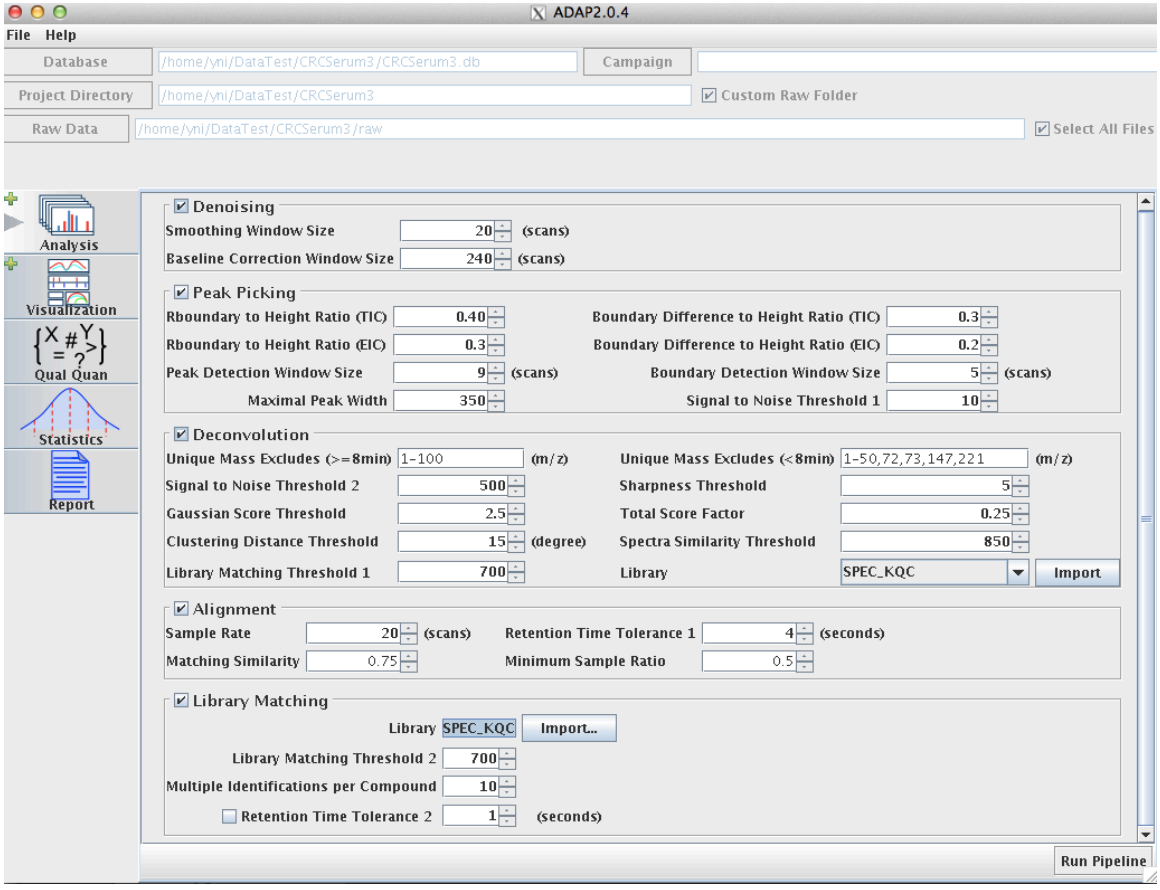

Figure 4.3. A screenshot of parameter settings within ADAP-GC software

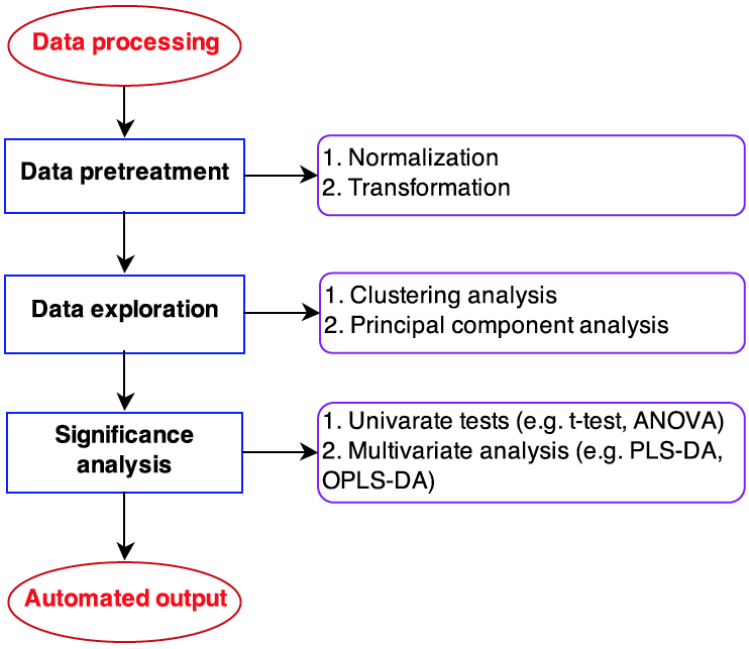

Figure 4.4. The ADAP-Stats architecture

4.3 Result Interpretation

A total of 20 data sets from Sample III in Chapter Two were selected and analyzed by ADAP-GC software, which were rat serum samples (10 animal models vs. 10 healthy controls) in an animal experiment of liver injury and analyzed by GC-TOF-MS platform. As introduced, ADAP-GC software is able to automatically analyze these raw netCDF data in a batch, proceeding through de-noising, peak picking, deconvolution, alignment, and compound identification and quantitation. Next, more details on data visualization, Qual/Quan analysis, and statistical analysis will be introduced.

4.3.1 Data Visualization

In the VISUALIZATION page, raw TIC and EIC chromatograms, extracted mass spectra, intermediate results from peak picking and deconvolution steps can be displayed. First, the identified peaks on TIC or each EIC can be labeled in black triangles, helping

users to decide whether all peaks of interest are successfully identified by ADAP-GC software (Figure 4.5 A). Within the 2nd panel, it is optional to display original mass spectrum alone, or 'head to tail' comparison between original/library and extracted mass spectra (Figure 4.5 B1-3). An extracted mass spectrum refers to a potential compound, thus the comparison of extracted mass spectrum and standard spectrum is helpful to examine the accuracy of peak identification and ion extraction during deconvolution. In addition, ADAP-GC software has provided comprehensive information regarding each identified component, including the retention time (Figure 4.5 C), compound information (Figure 4.5 D), model peak information (Figure 4.5 E), and the process of peak feature extraction (Figure 4.5 F). To summarize, our goal of developing VISUALIZATION module is to help users to examine characteristics of original GC-MS data and evaluate peak detection and deconvolution performance, which is crucial for quality control and troubleshooting and none software tools have provided such detailed information yet.

4.3.2 Qual/Quan Analysis

As introduced in the chapter of ADAP-GC 2.0, there are four options for users to select quantitation mass for compound quantitation: model peak mass, most intense mass, or the most intense unique mass, or summarization of all extracted masses. Correspondingly, four qualification and quantitation (qual/quan) tables could be displayed for visualization, manual check and semi-automatic modification. Each table includes both library searching and quantitation results (Figure 4.6 A): compound identification results provide unique ID, retention time, quantitation mass, molecular weight and formula, compound name, and matching score of each identified compound,

whereas quantitation results provide the estimated concentrations of all potential compounds in terms of extracted peak intensities of selected quantitation mass.

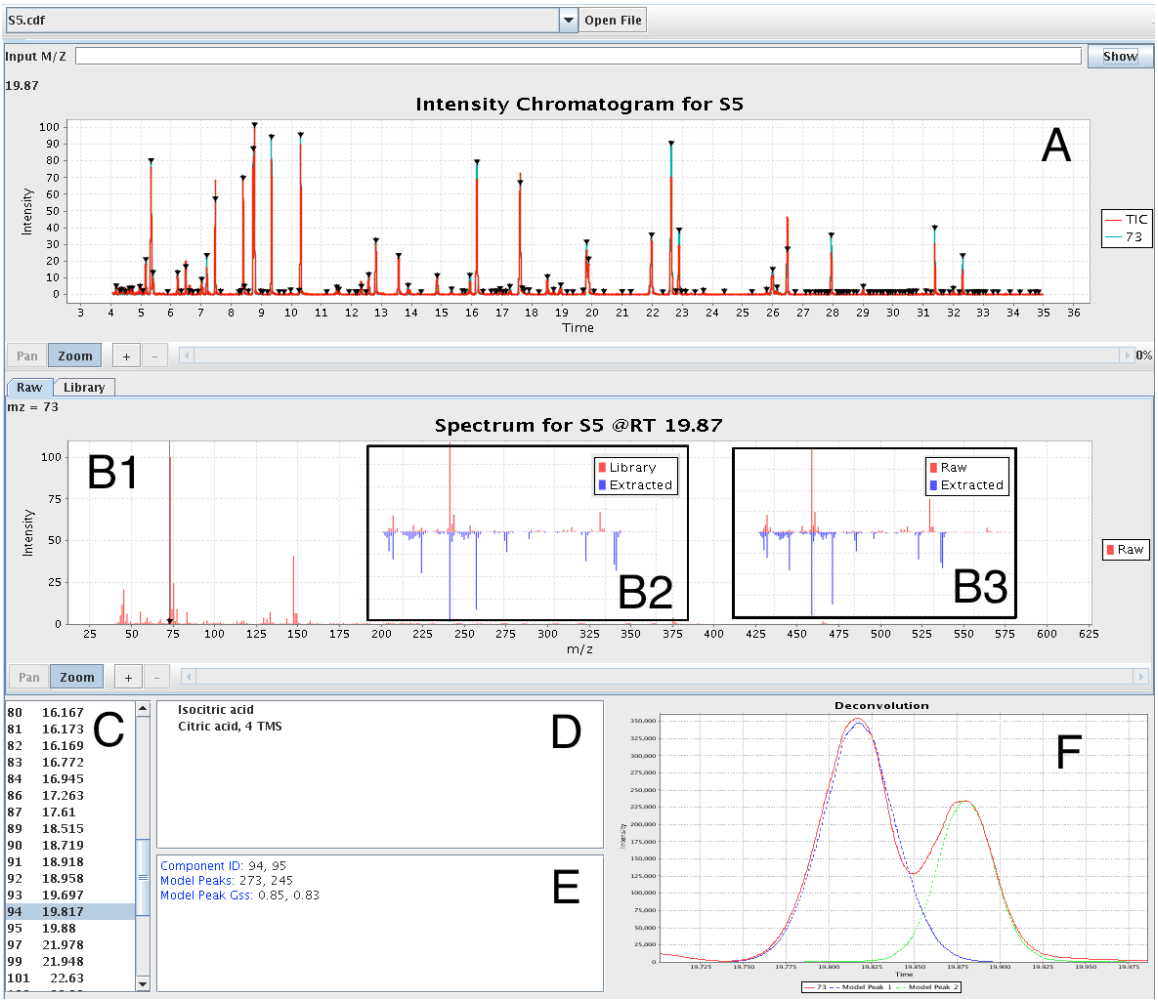

Figure 4.5. Visualization of chromatographic peak picking and component deconvolution within ADAP-GC software

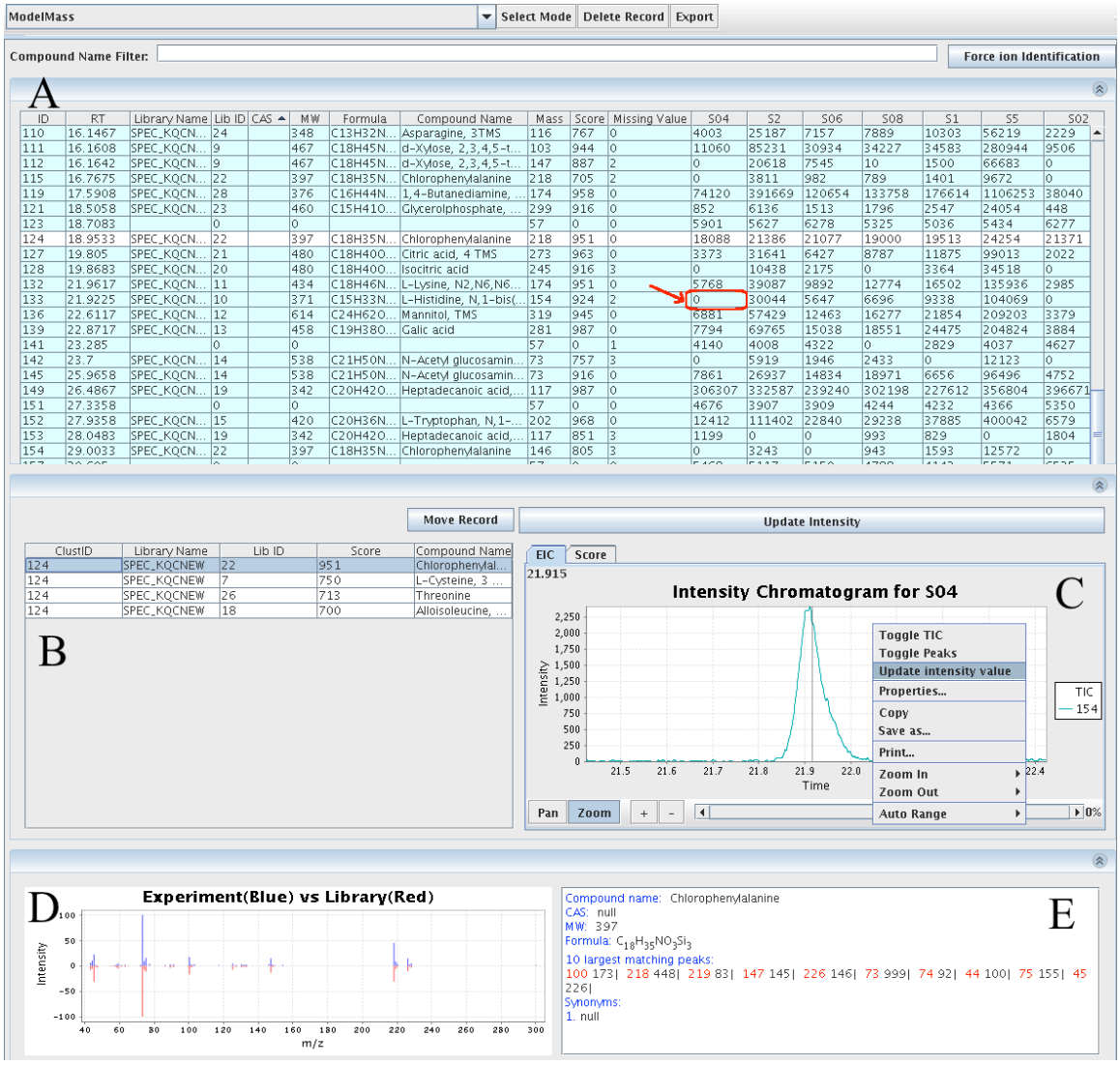

Figure 4.6. Module of qualification and quantitation analysis within ADAP-GC software

In practice, it is very useful to provide the option for users to check and modify the automatically produced results of compound quantification and qualification. (1) Compound identification check. Ten or less top candidates are listed in a table (Figure 4.6 B), and the one with the highest matching score is usually selected as the matched one. However, it is optional to select any candidate as the correct identified compound after examination of extracted mass spectra and quantitation results (Figure 4.6 D-E). (2)

Missing value correction. Missing value are very common in mass spectrometry that can result from several mechanisms [77]: (1) A compound can be present but with very low concentration that below the detection limit of mass spectrometry. (2) A compound is truly absent due to biological reasons thus could be not detected. (3) A compound fails to be detected due to technical issues related to data processing, such as inappropriate parameter settings in peak picking, deconvolution, and alignment. The missing value problem would directly affect the following statistical analysis and data interpretation, which usually take researchers tons of time for manual checking and correction. Thus, it is worthwhile to automate or semi-automate the checking and correction of missing values that are most likely coming from the inappropriate data processing. ADAP-GC software provides the capability for users to check raw data where a missing value exists (Figure 4.6 C) and correct automatically if these exists signals. After modifications, a new Qual/Quan table will be generated and updated in the backend database for subsequent statistical analysis.

4.3.3 Statistical Analysis

The qual/quan table with identification and quantitation results for all datasets is now moving forward for statistical analysis in ADAP-GC software. Normalization has two options, one is to select an internal standard and another is percentage normalization. Scaling has multiple choices including auto, pareto, centering, vast, range, level. Data analysis provides univariate, clustering, PCA and PLS-DA methods that are commonly applied in metabolomics. All the analysis can be performed automaticaly after users select appropriate parameters for data pretreatment and analysis, as a result, the corresponding results are immediately exported in two formats: (1) excel tables listing

significant metabolites with their corresponding p values, fold changes, and correlation coefficients. (2) high-resolution figures, e.g. volcano plot (Figure 4.7) , PCA and PLS-DA scores and locading plot, and summary of multivariate models (Figure 4.8).

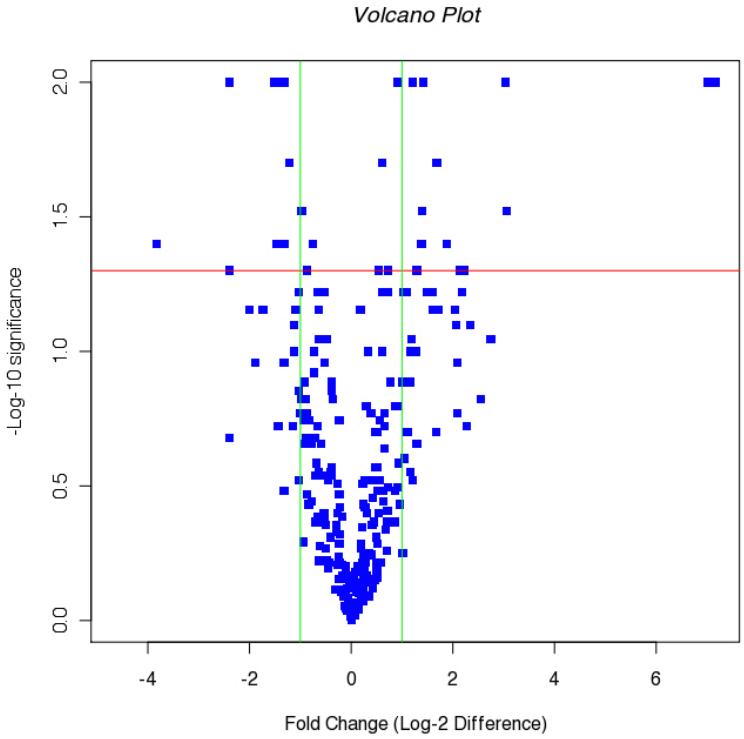

Figure 4.7. Volcano plot from ADAP-GC software.

4.4 Conclusion

ADAP-GC software is an integrated tool to process and analyze GC-MS data for metabolomics studies, which implement novel algorithms of data processing from ADAP-GC 2.0, advanced statistical analysis methods, and comprehensive capabilities of data visualization, compound identification and quantitation. Seamless data processing, automatic statistical analysis, and semi-automatic missing value imputation are very helpful to users, thus ADAP-GC 3.0 is promising in the field of metabolomics studies in the near future.

A stand-alone version of ADAP-GC software running in the LINUX system is available, however, user-friendly interfaces, structure optimization of SQL database, and accessibility across operating systems are expected and under development. Meanwhile, original algorithms of data processing are currently written in R, computational intensive parts, e.g. peak picking and deconvolution, are being recoded into Java using multithreading for fast data analysis.

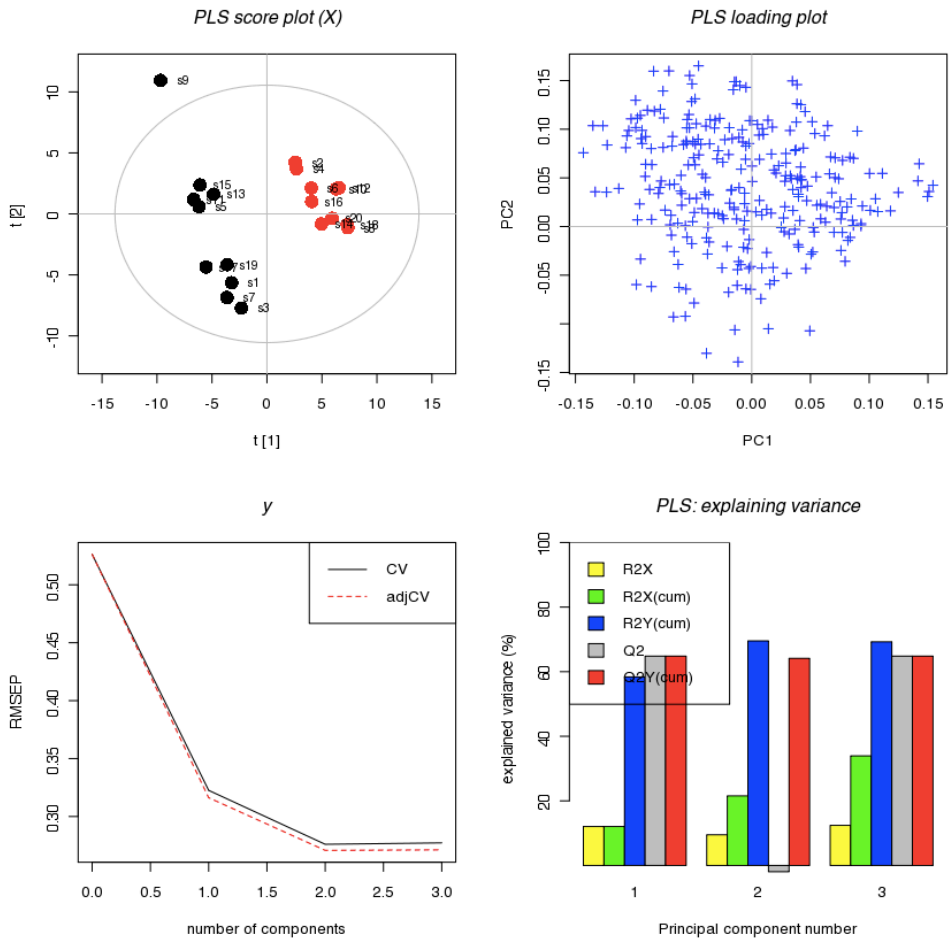

Figure 4.8. Example of automatic output from statistical analysis

REFERENCES

1. Bruggeman FJ, Westerhoff HV: The nature of systems biology. Trends Microbiol 2007, 15(1):45-50.

2. van der Greef J, Hankemeier T, McBurney RN: Metabolomics-based systems biology and personalized medicine: moving towards clinical trials? Pharmacogenomics 2006, 7(7):1087-1094.

3. Fiehn O: Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol 2002, 48(1-2):155-171.

4. Nicholson JK, Lindon JC, Holmes E: 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999, 29(11):1181-1189.

5. Fiehn O: Metabolomics - the link between genotypes and phenotypes. Plant MolBiol 2002, 48(1-2):155-171.

6. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN et al: Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature protocols 2011, 6(7):1060-1083.

7. Kim YS, Maruvada P, Milner JA: Metabolomics in biomarker discovery: future uses for cancer prevention. Future Oncol 2008, 4(1):93-102.

8. Nordstrom A, Lewensohn R: Metabolomics: moving to the clinic. Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology 2010, 5(1):4-17.

9. Gowda GA, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D: Metabolomics-based methods for early disease diagnostics. Expert review of molecular diagnostics 2008, 8(5):617-633.

10. Madsen R, Lundstedt T, Trygg J: Chemometrics in metabolomics--a review in human disease diagnosis. Anal Chim Acta 2010, 659(1-2):23-33.

11. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB: Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 2004, 22(5):245-252.

12. Wishart DS: Applications of metabolomics in drug discovery and development. Drugs R D 2008, 9(5):307-322.

13. Xu EY, Schaefer WH, Xu Q: Metabolomics in pharmaceutical research and development: metabolites, mechanisms and pathways. Current opinion in drug discovery & development 2009, 12(1):40-52.

14. Kell DB: Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 2006, 11(23-24):1085-1092.

15. Bando K, Kunimatsu T, Sakai J, Kimura J, Funabashi H, Seki T, Bamba T, Fukusaki E: GC-MS-based metabolomics reveals mechanism of action for hydrazine induced hepatotoxicity in rats. J Appl Toxicol 2010.

16. Beger RD, Sun J, Schnackenberg LK: Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity. Toxicology and applied pharmacology 2010, 243(2):154-166.

17. Wishart DS: Metabolomics: applications to food science and nutrition research. Trends Food Sci Technol 2008, 19(9):482-493.

18. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis S: Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics 2009, 5(4):435-458.

19. Dettmer K, Aronov PA, Hammock BD: Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007, 26(1):51-78.

20. Patti GJ, Yanes O, Siuzdak G: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Bio 2012, 13(4):263-269.

21. Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB: High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nat Biotechnol 2003, 21(6):692-696.

22. Dunn WB, Ellis DI: Metabolomics: Current analytical platforms and methodologies. Trac-Trends Anal Chem 2005, 24(4):285-294.

23. Williams MD, Reeves R, Resar LS, Hill HH, Jr.: Metabolomics of colorectal cancer: past and current analytical platforms. Analytical and bioanalytical chemistry 2013, 405(15):5013-5030.

24. Lei ZT, Huhman DV, Sumner LW: Mass Spectrometry Strategies in Metabolomics. J Biol Chem 2011, 286(29):25435-25442.

25. Bedair M, Sumner LW: Current and emerging mass-spectrometry technologies for metabolomics. Trac-Trends Anal Chem 2008, 27(3):238-250.

26. Stein SE: An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. Journal of the American Society for Mass Spectrometry 1999, 10(8):770-781.

27. Benton HP, Wong DM, Trauger SA, Siuzdak G: XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 2008, 80(16):6382-6389.

28. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G: XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 2006, 78(3):779-787.

29. Pluskal T, Castillo S, Villar-Briones A, Oresic M: MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC bioinformatics 2010, 11:395.

30. Katajamaa M, Miettinen J, Oresic M: MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics 2006, 22(5):634-636.

31. Lommen A: MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem 2009, 81(8):3079-3086.

32. Xia J, Psychogios N, Young N, Wishart DS: MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 2009, 37(Web Server issue):W652-660.

33. Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A: MeltDB: a software platform for the analysis and integration of metabolomics experiment data. Bioinformatics 2008, 24(23):2726-2732.

34. Bunk B, Kucklick M, Jonas R, Munch R, Schobert M, Jahn D, Hiller K: MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. Bioinformatics 2006, 22(23):2962-2965.

35. Reaves ML, Rabinowitz JD: Metabolomics in systems microbiology. Current Opinion in Biotechnology 2011, 22(1):17-25.

36. Hall RD: Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytologist 2006, 169(3):453-468.

37. Hiller K, Hangebrauk J, Jager C, Spura J, Schreiber K, Schomburg D: MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. Anal Chem 2009, 81(9):3429-3439.

38. Broeckling CD, Reddy IR, Duran AL, Zhao X, Sumner LW: MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. Anal Chem 2006, 78(13):4334-4341.

39. Schauer N, Fernie AR: Plant metabolomics: towards biological function and mechanism. Trends in Plant Science 2006, 11(10):508-516.

40. Cuadros-Inostroza A, Caldana C, Redestig H, Kusano M, Lisec J, Pena-Cortes H, Willmitzer L, Hannah MA: TargetSearch--a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data. BMC bioinformatics 2009, 10:428.

41. Luedemann A, Strassburg K, Erban A, Kopka J: TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments. Bioinformatics 2008, 24(5):732-737.

42. Luedemann A, von Malotky L, Erban A, Kopka J: TagFinder: Preprocessing Software for the Fingerprinting and the Profiling of Gas Chromatography-Mass Spectrometry Based Metabolome Analyses. Methods Mol Biol 2012, 860:255-286.

43. Jiang WX, Qiu YP, Ni Y, Su MM, Jia W, Du XX: An Automated Data Analysis Pipeline for GC-TOF-MS Metabonomics Studies. J Proteome Res 2010, 9(11):5974-5981.

44. Ni Y, Qiu Y, Jiang W, Suttlemyre K, Su M, Zhang W, Jia W, Du X: ADAP-GC 2.0: Deconvolution of Coeluting Metabolites from GC/TOF-MS Data for Metabolomics Studies. Anal Chem 2012, 84(15):6619-6629.

45. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS: MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. Nucleic Acids Res 2012, 40(Web Server issue):W127-133.

46. Styczynski MP, Moxley JF, Tong LV, Walther JL, Jensen KL, Stephanopoulos GN: Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. Anal Chem 2007, 79(3):966-973.

47. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G: XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 2012, 84(11):5035-5039.

48. Katajamaa M, Oresic M: Data processing for mass spectrometry-based metabolomics. J Chromatogr A 2007, 1158(1-2):318-328.

49. Mostacci E, Truntzer C, Cardot H, Ducoroy P: Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data. Proteomics 2010, 10(14):2564-2572.

50. Koh Y, Pasikanti KK, Yap CW, Chan EC: Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data. Journal of chromatography A 2010, 1217(52):8308-8316.

51. Lu HM, Dunn WB, Shen HL, Kell DB, Liang YZ: Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. Trac-Trends Anal Chem 2008, 27(3):215-227.

52. Qiu Y, Cai G, Su M, Chen T, Zheng X, Xu Y, Ni Y, Zhao A, Xu LX, Cai S et al: Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. J Proteome Res 2009, 8(10):4844-4850.

53. Li H, Xie Z, Lin J, Song H, Wang Q, Wang K, Su M, Qiu Y, Zhao T, Song K et al: Transcriptomic and metabonomic profiling of obesity-prone and obesity-resistant rats under high fat diet. J Proteome Res 2008, 7(11):4775-4783.

54. Yang C, He Z, Yu W: Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics 2009, 10:4+.

55. Colby BN: Spectral Deconvolution for Overlapping GC/MS Components. J Am Soc Mass Spectrom 1992, 3:558-562.

56. Stein SE: An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data. J Am Soc Mass Spectrom 1999, 10:770–781.

57. Rousseeuw PJ: Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 1987, 20:53–65.

58. G. DR, J. SM, C. RT, M. DA: Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry. Analystical Chemistry 1974, 48(9):1668-1375.

59. Du P, Kibbe WA, Lin SM: Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 2006, 22(17):2059-2065.

60. Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt A: High-accuracy peak picking of proteomics data using wavelet techniques. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing 2006:243-254.

61. Jiang W, Qiu Y, Ni Y, Su M, Jia W, Du X: An automated data analysis pipeline for GC-TOF-MS metabonomics studies. Journal of proteome research 2010, 9(11):5974-5981.

62. Byrd RH, Lu PH, Nocedal J, Zhu CY: A Limited Memory Algorithm for Bound Constrained Optimization. SIAM J Sci Comput 1995, 16(5):1190-1208.

63. Nederkassel AMv, Daszykowski M, Eilers PHC, Heyden YV: A comparison of three algorithms for chromatograms alignment. Journal of Chromatography A 2006, 1118(2):199–210.

64. Zimmer JS, Monroe ME, Qian WJ, Smith RD: Advances in proteomics data analysis and display using an accurate mass and time tag approach. Mass Spectrom Rev 2006, 25(3):450-482.

65. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA: PEPPeR, a platform for experimental proteomic pattern recognition. Mol Cell Proteomics 2006, 5(10):1927-1941.

66. Trygg J, Holmes E, Lundstedt T: Chemometrics in metabonomics. J Proteome Res 2007, 6(2):469-479.

67. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006, 7:142.

68. Yan N, Mingming S, Jinchao L, Xiaoyan W, Yunping Q, Aihua Z, Tianlu C, Wei J: Metabolic profiling reveals disorder of amino acid metabolism in four brain regions from a rat model of chronic unpredictable mild stress. FEBS letters 2008, 582(17):2627-2636.

69. Pan L, Qiu Y, Chen T, Lin J, Chi Y, Su M, Zhao A, Jia W: An optimized procedure for metabonomic analysis of rat liver tissue using gas chromatography/time-of-flight mass spectrometry. J Pharm Biomed Anal 2010, 52(4):589-596.

70. Ni Y, Qiu Y, Jiang W, Suttlemyre K, Su M, Zhang W, Jia W, Du X: ADAP-GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies. Analytical chemistry 2012, 84(15):6619-6629.

71. Katajamaa M, Oresic M: Data processing for mass spectrometry-based metabolomics. Journal of chromatography A 2007, 1158(1-2):318-328.

72. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M: Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Current bioinformatics 2012, 7(1):96-108.

73. Boccard J, Veuthey JL, Rudaz S: Knowledge discovery in metabolomics: an overview of MS data handling. J Sep Sci 2010, 33(3):290-304.

74. Trygg J, Holmes E, Lundstedt T: Chemometrics in metabonomics. J Proteome Res 2007, 6(2):469-479.

75. Madsen R, Lundstedt T, Trygg J: Chemometrics in metabolomics--a review in human disease diagnosis. Anal Chim Acta 2010, 659(1-2):23-33.

76. Xia JG, Broadhurst DI, Wilson M, Wishart DS: Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics 2013, 9(2):280-299.

77. Taylor SL, Leiserowitz GS, Kim K: Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. Statistical applications in genetics and molecular biology 2013, 12(6):703-722.