An Automated Data Analysis Pipeline for GC–TOF–MS
Metabonomics Studies

Wenxin Jiang,†,# Yunping Qiu,†,# Yan Ni,† Mingming Su,§ Wei Jia,† and Xiuxia Du∗,†

Department of Bioinformatics & Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, Kannapolis, North Carolina 28081, Department of Nutrition, University of North Carolina at Greensboro, North Carolina Research Campus, Kannapolis, North Carolina 28081, and David H. Murdock Research Institute, North Carolina Research Campus, Kannapolis, North Carolina 28081

Received July 25, 2010

Recent technological advances have made it possible to carry out high-throughput metabonomics studies using gas chromatography coupled with time-of-flight mass spectrometry. Large volumes of data are produced from these studies and there is a pressing need for algorithms that can efficiently process and analyze data in a high-throughput fashion as well. We present an Automated Data Analysis Pipeline (ADAP) that has been developed for this purpose. ADAP consists of peak detection, deconvolution, peak alignment, and library search. It allows data to flow seamlessly through the analysis steps without any human intervention and features two novel algorithms in the analysis. Specifically, clustering is successfully applied in deconvolution to resolve coeluting compounds that are very common in complex samples and a two-phase alignment process has been implemented to enhance alignment accuracy. ADAP is written in standard C++ and R and uses parallel computing via Message Passing Interface for fast peak detection and deconvolution. ADAP has been applied to analyze both mixed standards samples and serum samples and identified and quantified metabolites successfully. ADAP is available at http://www.du-lab.org.

Keywords: gas chromatography–mass spectrometry • deconvolution • extracted ion chromatogram • clustering analysis • alignment

Introduction

Metabonomics/Metabolomics has emerged as an integral part of systems biology and functional genomics. It attempts to identify and quantify metabolites that are synthesized by a biological system1,2 and has found wide applications in drug discovery, disease diagnosis,3 plant biotechnology,4 nutrition science,5 and toxicology studies6 by profiling and fingerprinting metabolites in biofluids and tissues.

Common analytical platforms that are used in metabolomics research include NMR spectroscopy, gas chromatography–mass spectrometry (GC–MS), liquid chromatography–mass spectrometry (LC–MS), and capillary electrophoresis–mass spectrometry (CE–MS).7 The high sensitivity of modern mass spectrometers has made MS-based metabolomics the method of choice for studies that involve the identification and quantitation of low-concentration metabolites from complex samples.

For MS-based metabolomics, LC–MS and GC–MS are complementary technologies with the latter being especially well suited for metabolites that are volatile and do not ionize well by LC–MS techniques. In addition, GC–MS provides high chromatographic resolution and permits separation of structurally similar compounds that would be difficult to separate by LC. Primary mass analyzers that are coupled to GC separations are quadrupole and TOF instruments. TOF-MS allows fast spectra acquisition, which facilitates deconvolution of coeluting metabolites.8

Large volumes of many complex data sets can be produced from GC–TOF–MS platforms for systems biology or functional genomics studies. A number of software applications have been developed to analyze GC–MS data. Representative software include MetAlign,9 MathDAMP,10 MetaQuant,11 MET-IDEA,12 MCR (i.e., multivariate curve resolution),13–15 AMDIS,16 TagFinder,17 and ChromaTOF, among others. Despite their successful applications, these data analysis platforms have limited capacity to process data sets in a high-throughput fashion and cause data analysis to be very lengthy. Additionally, these existing software tools suffer from a range of other limitations. For example, AMDIS does not perform alignment. MCR is highly sensitive to the selection and number of coanalyzed chromatogram files.17 MetAlign appears not to perform deconvolution.9,17 ChromaTOF (LECO, St. Joseph, MI) was developed exclusively for the GC–EI-TOF–MS and GCxGC–EI-TOF–MS instruments of the vendor. Consequently, the source code is not available, which makes it very difficult for customized applications.
To enable large-scale, GC-TOF-MS-based metabonomics studies, a high-throughput data analysis platform is needed. In this paper, we report such a platform named Automated Data Analysis Pipeline (ADAP) that features peak detection, deconvolution, alignment, and library search and allows data to flow seamlessly through these analysis steps without any human intervention. Additionally, two novel algorithms have been designed for deconvolution and alignment. The first one applies clustering analyses to deconvolute coeluting compounds that are very common in complex samples. The second one minimizes the likelihood that different compounds are incorrectly aligned with each other by using a two-phase approach that is based on extracted ion chromatograms (EIC) rather than the total ion chromatogram (TIC).

ADAP is written in C++ and R and the computationally intensive peak picking and deconvolution take advantage of parallel computing via Message Passing Interface (MPI). As a result, the power of multiple computing cores that are common in modern computers can be fully utilized for fast data analysis. To our knowledge, none of the aforementioned software tools utilizes parallel computing. ADAP has been applied to analyze both mixed standards samples and serum samples and successfully identified and quantified metabolites. ADAP is available at http://www.du-lab.org.

### Experimental Procedures

**Sample Preparation.** Three different types of samples were analyzed on a GC-TOF-MS platform for the development, evaluation, and validation of ADAP.

1. **Mixed Standards (MS) Samples.** A total of 38 standard compounds (Supplementary Table S1) were carefully selected and mixed together with known ratios. Criteria for selecting those compounds are the following. (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-min time range.

2. **Calibration Curve (CC) Samples.** Ten calibration curve samples were prepared at different dilutions from the original mixture of 20 fatty acid standards (Table 1). These fatty acids vary in biochemical properties with different number of carbons and double bonds, or different double bond positions.

3. **Liver Injury (LI) Samples.** Serum samples were collected from male Sprague–Dawley rats. Ten rats had acute liver injury here to be consistent with the naming rule used in AMDIS. In brief, after TMS derivatization, each 1 µL aliquot of the derivatized solution was injected in splitless mode into an Agilent 6890N gas chromatography system (Santa Clara, CA) that was coupled with a Pegasus HT time-of-flight mass spectrometer (Leco Corporation, St. Joseph, MI). The spectral acquisition rate was 20 scans/s. In total, 15, 11, and 20 sample data sets were generated for experiments (1), (2), and (3), respectively. In the remaining part of the paper, we use “a sample” to refer to one raw data file that is produced from the mass spectrometer.

### Data Analysis Methods and Results

**Data Analysis Workflow.** The data analysis workflow consists of four sequential steps (Figure 1): peak detection, deconvolution, alignment, and library search. Peak detection finds local peaks in each EIC. Deconvolution separates coeluting components based on both the apex elution time and elution profiles of their fragment ions. Each component is potentially a compound if its identity can be determined. We use component here to be consistent with the naming rule used in AMDIS. After deconvolution, a mass spectrum is constructed for each EIC. The mass spectrum that is selected at the alignment stage as the optimal reference is used for compound identification. Lastly, the intensity of each component is determined for quantification. Next, we describe the details of each step.

**Peak Detection.** The first step of the workflow is to detect EIC peaks for each mass. 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. These algorithms usually perform a denoising step that includes chromatogram smoothing and/or baseline correction.
before peak detection. ADAP performs peak detection before denoising 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 denoising algorithm, and ultimately benefits identification of compounds in terms of both confidence and total number of identifications (Supplementary Figure S1). The rationale behind this is threefold. First, the observation of a larger number of fragments that belong to the same compound increases the likelihood that the identification is correct. Second, 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.\textsuperscript{16} 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. Because of these reasons, we chose to preserve as much information as possible at each stage of the data processing.

Each EIC peak is characterized by its apex elution time, left and right boundary, peak height, and shape. 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). For the experimental data described in this paper, this window spans nine scans which translates to 9/20 s. In the ADAP software tool, 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.

After peak detection has been performed for all the EICs, information in the raw data on the identity and concentration of metabolites has been extracted and represented as EIC peaks. The next step is to group these peaks so that fragment ions that correspond to the same component fall in the same group.

**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.\textsuperscript{22,23} 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 3A 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:

\[
\tilde{e}_x = (a_1, a_2, ..., a_n) \\
\tilde{e}_y = (b_1, b_2, ..., b_n)
\]

where \(a_i\) and \(b_i\) are the abundance values at retention time \(t_i\). The similarity between \(\tilde{e}_x\) and \(\tilde{e}_y\) can be measured by the normalized dot product:

\[
r = \frac{\tilde{e}_x \cdot \tilde{e}_y}{|\tilde{e}_x| \cdot |\tilde{e}_y|}
\]

where the center dot (\(\cdot\)) represents the dot product.
To separate coeluting 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 coeluting 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 3B depicts the clustering results for one time window within which two compounds coelute. The fragment ions of creatinine (shown in red and grouped in cluster 2) are well separated from those of trans-cinnamic acid (shown in blue and grouped in cluster 1). The two resulting spectra are matched with creatinine and trans-cinnamic acid in the NIST library with high matching scores, which demonstrates that ADAP can effectively detect differences in the shape of EIC peaks and successfully deconvolute coeluting components (Figure 3C–E). This deconvolution process is repeated for all deconvolution windows.

The k-medoids 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 is
research articles

used in this study to assess the clustering quality and determine the $k$ value.

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

where $S$ is the silhouette score of a cluster, $d_{\text{intra}}$ is the intracluster distance and is calculated as the average pairwise distance between objects within the cluster, $d_{\text{inter}}$ is the intercluster 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 intracluster distance threshold is specified. When $d_{\text{intra}}$ 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. Because of 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.

**Alignment.** A number of algorithms have been developed for chromatogram alignment. 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 component across samples, a measure of confidence needs to be defined that takes into account both the spectra similarity and retention time similarity between two components. In ADAP, this measure is:

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

where $s_i$ and $s_j$ denote two spectra. 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. Score$_{\text{RT}}$ is calculated by:

$$\text{score}_{\text{RT}} = 1 - |\Delta \text{RT}|/w$$

where $\Delta \text{RT}$ is the difference in their apex retention time and $w$ is the maximum retention time shift that is acceptable for the same component in a particular experiment. Components whose retention time difference exceeds $w$ should be considered as different components.

The incorporation of score$_{\text{RT}}$ into 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.

The final 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 score$_{\text{total}}$ greater than a certain threshold (750 in this study). From these high-scoring components, select the one that produces the highest 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 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 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 (examples provided in Supplementary Figure S2). 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 1 demonstrates these improvements by analyzing the CC samples. Figure 4 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 4K. This deviation profile is consistent with the temperature change in the experimental process, that is, the maximum deviation occurs at the time when the temperature reaches its peak (approximately 20 min in this example).

**Compound Identification.** The identity of a component is determined by searching the corresponding mass spectrum against a library of spectra. ADAP is equipped with the capability to perform library search. Since the alignment references that have been found are the best representation of components, ADAP searches them against an in-house library that includes spectra for 500 human metabolites. In addition, ADAP provides the option to export spectra in the NIST
standard MSP format that can be read by NIST MS searching software or other third-party software to perform library search. The spectra files in the Supporting Information provide the mass spectra that were exported by ADAP in the standard NIST format. The high matching scores (searched against the NIST library) of the compounds listed in Table 1 demonstrate that ADAP has successfully identified these compounds from the CC samples.

**Compound Quantitation.** Among all the fragment ions that have been detected for a component, the most abundant, unique ion is invaluable for comparing the concentrations of the same compounds across samples. A unique ion is defined to be an ion that is unique to a component within a certain time range and multiple unique ions can exist for one component. 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. In the current version of ADAP, peak height is used. Supplementary Figure S3 depicts the abundance values of 19 compounds versus the true concentration of the compounds in the 11 CC samples. Both of the r-squared values are close to 1, which indicates the accuracy of ADAP in extracting quantitative information of compounds.

**Performance.** Table 2 lists performance measures of ADAP after it is applied to analyze the three sets of samples. These measures include the total number of peaks that are detected, total number of components after deconvolution, and the total number of components after alignment with the requirement that each component is observed in more than 20% of the total number of samples. Among those components that are aligned, most of them are correctly identified with the average score higher than 800.

ADAP is truly a high-throughput pipeline in that: (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 lists the number of samples and the corresponding
processing time for the three sets of data. We analyzed 15 MS samples and 20 liver injury samples with ADAP and HDA on the same workstation (2 × Intel Quad-core Xeon CPU X5570 2.93 GHz, 24 GB memory). ADAP used less than 3 min to analyze each entire set of samples, whereas HDA used 5 h for the same analysis job. To our knowledge, ADAP is much faster compared to other existing software tools as well including ChromaTOF. The average matching scores of common metabolites identified from ADAP and HDA are very similar, which indicates that the two pipelines have comparable performance in terms of identifying compounds.

Biological Application. ADAP has been applied to the metabolomics analysis in an animal experiment of liver injury. From 20 rat serum samples, a total of 277 components were produced after peak detection, deconvolution, and alignment. The concentration of each component in a sample was estimated as the total sum of the intensity values of all of the constituent fragment ions that have been identified. The resultant quantitation data for all of the components from the 20 samples were imported into the SIMCA-P 12.0 software package (Umetrics, Umeå, Sweden) for multivariate statistical analysis. Specifically, mean-centering and autoscaling were used for data pretreatment. Subsequently, PCA (Principal Component Analysis) was applied and a clear separation between the diseased and control groups (Figure 5A) 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 5B) to identify the differential 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. 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. These analysis results provide valuable pointers for further biological investigations about liver injury-induced metabolic disorder.

Conclusions and Discussions

A fully Automated Data Analysis Pipeline (ADAP) for GC–TOF–MS metabolomics studies has been presented. ADAP enables high-throughput data analyses by using parallel

---

**Figure 5.** 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.
computing and features peak picking, deconvolution, alignment, and identification without any manual intervention. In particular, two novel algorithms have been designed for efficient deconvolution and accurate alignment of components.

Currently, we are making further progress to deconvolute EICs of shared ions more accurately. Preliminary testing has demonstrated considerable improvements. Additionally, we plan to apply signal processing techniques to denoise EICs. Our goal is to filter out true noise interferences while preserving valuable information on low-concentration compounds.

ADAP is a module within a comprehensive data analysis framework that is equipped with the capability to perform comparative metabolomics studies and associated visualizations of large-scale data sets.

Acknowledgment. We thank Scott Woods and Steven Blanchard for their great help by making the high performance computing resources available to this project. These facilities are located on the campus of University of North Carolina at Charlotte and the North Carolina Research Campus.

Supporting Information Available: Figure S1, mass spectrum and EIC peak for l-histidine illustrating the importance of performing peak picking prior to denoising. Figure S2, example spectra demonstrating the improvement of alignment by using the two-phase over the one-phase approach. Figure S3, performance of ADAP in terms of extracting quantitative information. Figure S4, results from applying PCA and PLSDA to the quantitative data obtained by HDA after processing the liver injury samples. Table S1, list of 38 mixed standards compounds and the corresponding matching scores from searching against an in-house library using AMDIS. Table S2, quantitative data from ADAP after analysis of liver injury samples. Spectra S1, Spectra S2, Spectra S3, mass spectra files produced by ADAP in the standard NIST text format for MS, GC, and LI samples, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

References


PR1007703